UNIVERSITY OF THESSALY

School of Engineering

Department of Electrical & Computer Engineering

AGGREGATION, ORGANIZATION AND DISSEMINATION OF ON-LINE DATA AND INFORMATION

Master Thesis

Of

Christou Nicola Theognosia

Thesis Advisor: Vavalis Emmanouil

Volos, 2013

UNIVERSITY OF THESSALY

School of Engineering
Department of Electrical & Computer Engineering

AGGREGATION, ORGANIZATION AND DISSEMINATION OF ON-LINE DATA AND INFORMATION

Master Thesis
Of
Christou Nicola Theognosia

Thesis Advisor: Vavalis Emmanouil
Professor UTH

(Signature)	(Signature)	(Signature)
……	……	……
Vavalis Emmanouil Professor D.E.C.E	Katsaros Dimitrios Lecturer D.E.C.E	Bozanis Panayiotis Chairman D.E.C.E

UNIVERSITY OF THESSALY

School of Engineering
Department of Electrical & Computer Engineering

(Signature)

……

Christou Nicola Theognosia

Acknowledgements

Upon the completion of this thesis, I would like to thank my advising professor Mr. Vavalis for his help and support during this period, but also for introducing me to the interesting field of Scientometrics. Moreover, I would like to thank the other two members of the committee, Mr. Katsaros and Mr. Bozanis for agreeing to participate in the evaluation process of my master thesis.

Finally, I would like to thank my parents and my dearest S.K. for their support and contribution over the years.

Abstract

The overall objective of this thesis is to study the theories, the common standards, the practices, and the effective tools currently available for aggregating and disseminating inhomogeneous data in a form that is as close as to information and knowledge. The specific aim of this thesis is to present the importance of knowledge dissemination and the technologies behind Research Information Systems (RISs). The objective is the research for best practices, and the research for appropriate technologies and standards needed for the implementation of such systems. Moreover, a model for the implementation of a RIS for the University of Thessaly is proposed. The model is based on the usage of Web technologies, and the usage of a standard (CERIF: Common European Research Information Format) proposed by the European Union for the development of this type of systems. Research related data are aggregated online from various sources, processed, organized, and stored, in order to be disseminated through the website of the University.

Keywords

Scientometrics, Bibliometrics, Research Information System (RIS), Common European Research Information Format (CERIF)

Contents

List of Figures

List of Images

List of Tables

1. Introduction

The communication of research results is a major enabling factor that makes science move forward. Scientists during Renascence may not have been willing to share their research output with everybody. They preferred to communicate them with a small group of colleagues. Perhaps this was due to the fact that the communication was solely based on letter exchanging and common meeting at the early years, or publications in scientific journals that were not easily available. This phenomenon does not occur anymore. Today, researchers share their findings not only with their colleagues, but also with the rest of the world. The communication of research helps them extend their knowledge, thus, creating a chain of knowledge that leads to new information and therefore to new research output.

At university, and Research and Development institutions level, it is nowadays crucial to disseminate this kind of information for two additional reasons. Firstly, because it gives them the ability to show their areas of excellence and expertise and through an evaluation process to identify strong and weak points. Secondly, it can be of help to other organizations, institutions, researchers and beyond in their search for research and associates with appropriate expertise.

1.1 Motivation and Objectives

Even though the concept of evaluating research output has been around for many decades, its use wasn't widely spread until recent years. At the beginning of the 21st century measures have been proposed, and the first services/systems for the dissemination of this information made their appearance. The latter flourished during the last five years, especially amongst universities; many of them have made commendable efforts implementing this type of systems.

The awareness of the need for quality Research Information Systems (RIS) keeps increasing due to the benefits that it offers. Researchers can easily access relevant information. Research managers and administrators can measure and analyze the research activity of an organization. Research councils can use a Current Research Information System (CRIS) in order to optimize the funding process. Moreover, entrepreneurs and technology transfer organizations can easily retrieve novel ideas and technology in a knowledge-assisted environment. This can be of help in identifying competitors and previously done similar research. Finally, a CRIS can be helpful for the public in order to access information. At institutional level CRISs can be used to monitor research activities and output, to evaluate research and to assist in project planning.

In this set of mind, our goal is to propose a model for a web observatory for the research activities of the researchers of University of Thessaly. This observatory will be providing comprehensive and easy to understand information, including research projects, publications, connections between researchers, patents, awards and honors, theses and dissertations. With the completion of the project, the web observatory of the University will be the first on national level and we hope that others will follow our example.

1.2 Background

Scientific measures: Definitions, measures and standards

The term **Bibliometrics** was coined by Alan Pritchard in a paper published in (Pritchard, 1969) as a better name for the term statistical bibliography. The latter was first used by E. Wyndham Hulme in (Hulme, 1923) in two lectures that he delivered as the Sandars Reader in Bibliography [1] at the University of Cambridge.

"Bibliometrics is the application of mathematics and statistical methods to books and other media of communication."

Research fields use bibliometric methods in order to explore the impact of the field, the impact of a set of researchers or the impact of a particular published paper.

Scientometrics is the science of measuring and analyzing science. In practice, scientometrics is often done using Bibliometrics, which is a measurement of the impact of scientific publications. Modern scientometrics is mostly based on the work of Derek J. de Solla Price and Eugene Garfield.

Price was a physicist, historian of science and information scientist. He is credited as the father of scientometrics and his major scientific contributions in the field include: studies of the exponential growth of science and the half-life of scientific literature, (Price, Little Science, Big Science, 1963), quantitative studies of the network of citations between scientific papers, (Price, Networks of Scientific Papers, 1965), and a mathematical theory of the growth of citations networks, (Price, A General Theory of Bibliometric and Other Cumulative Advantage Processes, 1976).

Eugene Garfield is an American scientist, one of the founders of Bibliometrics and Scientometrics, and the founder of the Institute of Scientific Information (ISI). ISI offered citation indexing and analyzing, a field in which Garfield pioneered. Garfield is also responsible for other innovative bibliographic products, including Current Contents [2], the Science Citation Index (SCI) [3], among other citation databases, the Journal Citation Reports [4], and the Index Chemicus [5].

The concept of **impact factor** was devised by Eugene Garfield in 1955 as a measure reflecting the average number of citations to recent articles published in a given journal. In a given year, the impact factor of a journal is the average number of citations received per paper published in that journal during the two preceding years.

Jorge. E. Hirsch, an Argentine American physicist, proposed the **h-index** in 2005, as a measure of "*the broad impact of an individual's work*", which gives an estimate of the total number of citations. Hirsch writes:

"A scientist has index h if h of his/her N_p papers have at least h citations each, and the other (N_p - h) have no more than h citations each."

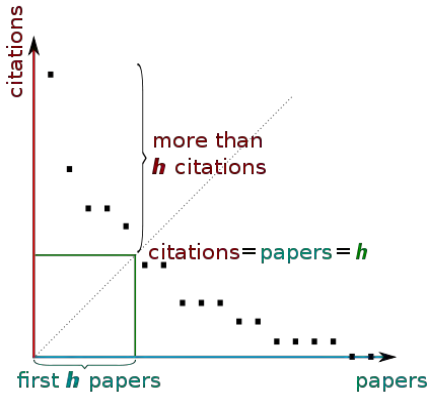


Figure 1 h-index from a plot of decreasing citations for numbered papers [http://en.wikipedia.org/wiki/File:H-index-en.svg]

The **g-index** is an index for quantifying productivity based on publication record. It was suggested by Leo Egghe, a Belgian mathematician and researcher in the areas of informetric theory and informetric laws amongst others, in 2006. It is calculated based on the distribution of citation received by a given researcher's publications. Specifically,

> *"Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received at least g^2 citations."*

$$g^2 \leq \sum_{i \leq g} c_i$$

A **Current Research Information System (CRIS)** is any information tool that is dedicated to provide access to and disseminate research information, such as People, Projects, Organizations, Results (publications, patents and products), Facilities, and Equipment. It consists of a data model describing objects of interest to Research and Development, and a tool or set of tools to manage the data. At institutional level CRIS is a tool for policy making, evaluation of research based on outputs, documenting research activities and output, and assistance in project planning. Individual end users of a CRIS can utilize it to evaluate opportunities for research funding, avoid duplication of research activity, analyze research trends, have links to full text of publications, locate new networks, and identify new markets for products of research [6].

The **Common European Research Information Format (CERIF)** is one type of CRIS standard implemented. It was proposed by the European Union and developed and maintained by euroCRIS, which is a non-profit organization dedicated to the development of Research Information Systems (RISs) and their interoperability. Besides CERIF, euroCRIS's areas of interest cover scientific datasets, institutional repositories, data access and exchange mechanisms, standards and guidelines and best practice for CRIS. CERIF emerged as a simple standard (CERIF91) not unlike a library catalogue card and was intended as a data exchange format. However, it was soon realized that in practice it was inadequate. This led to the generation of CERIF2000 that is a concept about research entities and their relationships (Specification), a description of research entities and their relationships

(Model) and a formalization of research entities and their relationships (Database Scripts). CERIF2000 has been revised over the years (CERIF2004, CERIF2006, CERIF2008, CERIF1.3, CERIF1.4, and CERIF1.5) and its latest version is CERIF 1.6 was released in the summer of 2013.

The data model, as described by euroCRIS:

- allows for a metadata representation of research entities, their activities and output, high flexibility with formal semantic relationships,
- enables maintenance, archiving, access and interchange of research information, and
- supports knowledge transfer to decision makers, for research evaluation, research managers, strategists, researchers, editors, and the public.

The four essential features of CERIF, listed by euroCRIS, are:

- It has a concept of objects or entities with attributes such as a project, person, and organizational unit.
- It supports n:m relationships (many-to-many relationship: is a type of cardinality referring to the relationship between two entities; e.g. An Author entity and a Book entity have n:m relationship because an Author can write many books, and a book can been written by several Authors) between them using "linking relations" thus providing rich semantics.
- It is fully internationalized in language and character set.
- It is extensible without prejudicing the core data model and character set.

Scientific collaboration

The internet and the greater affordability of international transportation have reduced distances between people and have made scientific collaborations between distant scholars very easy. This led to multi-university collaborations and the formation of larger research teams, which results to high impact publications, especially if they involve different countries.

In (Georghiou, 1998) two categories of motivations for research collaboration are presented:

Direct benefits to the Science and Technology concerned, allowing the research to be performed or applied at a higher quality, with a broader scope, more quickly or more economically than would be the case without cooperation.

Indirect benefits arising from the existence of the cooperation. These may accrue directly to the participants or more generally to the countries involved in terms of political economic or social benefits.

In (Pan, Kaski, & Fortunato, 2012) the results of a study of citation and collaborative interactions between different geographic locations are demonstrated. Their study showed that the average number of cites of a publication increases with the number of co-authors and with the number of affiliated countries and cities of its authors. By using the *k*-means clustering technique (partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest *mean*, serving as a prototype of the cluster) where k=2, they were able to classify countries in two groups (clusters); one with average spending less than about $100, 000 per researcher per year and other with average spending more than $100, 000 per researcher per year. Their separation revealed a striking feature:

"If the average spending is less than about $100, 000 per researcher per year there is an increase in the average number of citations with the spending. However, if the average spending exceeds this limit, it becomes scattered and independent of funding."

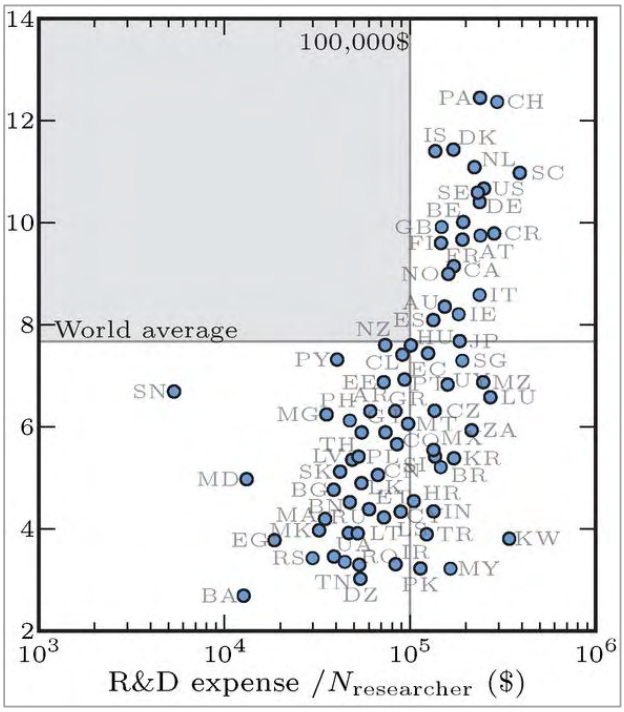


Figure 2 Relation between research outcome and funding [courtesy of (Pan, Kaski, & Fortunato, 2012)]

The latter phenomenon can be broken down in three cases; very rich countries like Luxembourg have high funding per researcher and yet their average number of cites below world average, countries like India and Brazil also show low number of cites, due to the fact that they invest more in infrastructure, whereas countries including Germany, Denmark and the United States have average number of cites above the world average. Moreover, it is impossible for a country to exceed the world's average, in terms of cites per paper, unless there is sufficient spending.

The importance of international collaboration

Apart from presenting the motivations, (Georghiou, 1998) also lists the direct and indirect benefits of international collaboration.

Direct benefits

Access to complimentary expertise, knowledge or skills to enhance scientific or technological excellence provides the principal motivation for cooperation between industrial countries and beyond. A program with wider geographical coverage has greater chance of finding the right partner. This has been very beneficial for smaller countries where national expertise may be absent in several areas.

Access to unique sites, facilities or population groups is a second source of motivation. Cooperation in this case derives from the need to perform research on a phenomenon present in one of the countries.

Sharing costs and risks is an important motive and in some cases may also be operational, where one country is the host to a large and expensive scientific instrument.

Addressing transactional or global problems forms another motivation and is exemplified by research on situations like fisheries and medical cooperation undertaken by several countries.

Establishing standards. Even though it has not emerged as a prominent activity, there are multilateral arrangements for laboratories working on measurement standards. The importance of this activity is constantly and rapidly increasing.

Indirect benefits

Indirect or strategic motivations describe the situation where the collaboration is driven by external goals of a political, economic or cultural nature. Learning benefits may also occur while working in another country or through the collaboration where a partner's strength is your own weakness.

1.3 Evaluation Bodies and Services

In 1960, Eugene Garfield's Institute for Scientific Information (ISI) introduced the first citation index for papers published in academic journals, the Science Citation Index (SCI). A citation index is an index of citations between publications, allowing the user to establish which later documents cite which earlier documents. Major citation indexing services include Web of Science, Scopus, CiteSeerX, Google Scholar and Microsoft Academic Search.

Thomson Reuters' **Web of Science** provides access to citation databases with content that covers over 12,000 of the highest impact journals worldwide, including Open Access journals and more than 150,000 conference proceedings. Web of Science can be found within **Web of Knowledge**, a comprehensive research platform.

Scopus is Elsevier's abstract and citation database that contains 49 million records, over 5.3 million conference proceedings and over 20,500 titles from 5,000 publishers worldwide. It is currently the world's largest database of its kind.

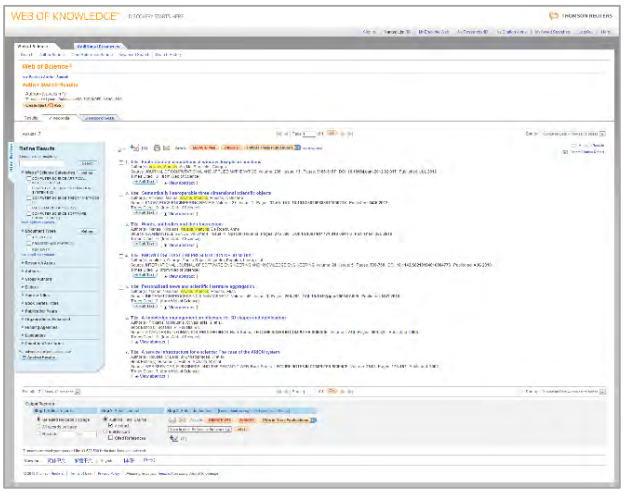


Image 1 Results on author search for Web of Science

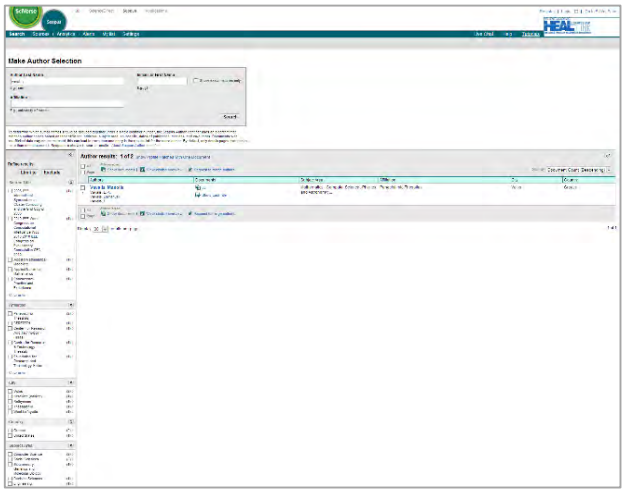


Image 2 Results of author search for Scopus

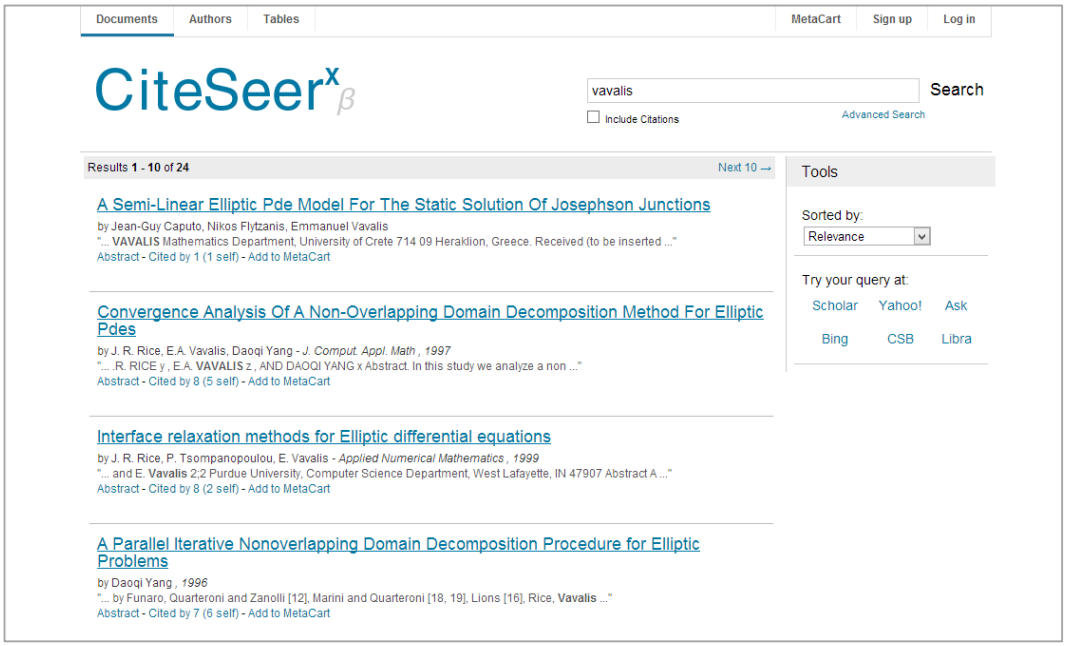
Web of Science and Scopus offer more or less the same functionality. A user can perform a query by defining multiple search fields, including author's name, title, keywords, DOI, affiliation, etc. On author search Web of Science will return records with authors matching the name in question, without offering author profiles. However, some information regarding the author, including address and email are given at the document's abstract page. Scopus on the other hand, is relatively more author-oriented returning author profiles that match the name, providing personal and research information; documents, citations, references, h-index, etc., regarding the author. The latter also offers visualizations on documents, h-index and citations. Example author search for Scopus and Web of Science are shown on Image 2 and Image 1, respectively.

CiteSeerX is the successor of CiteSeer that is considered to be the first automated citation indexing system and a predecessor of Google Scholar and Microsoft Academic Search. It was developed by researchers Dr. Isaac Councill and Dr. C. Lee Giles at the Pennsylvania State University. It is a public search engine and digital library and repository for scientific and academic papers. Its primary focus is on computer and information science. It currently has over 2 million documents with nearly 2 million unique authors and 40 million citations.

Google Scholar was released in 2004 and is a web search engine that indexes scholarly literature, including peer-reviewed and non-peer reviewed journals, scholarly books, theses, abstracts and citations.

Microsoft Academic Search was recently developed by Microsoft Research as an index of academic content, researchers, institutions and activities. It contains millions of publications and features the display of key relationships between and among subjects, content and authors.

CiteSeerX, Google Scholar and Microsoft Academic Search are developed as search engines. When a query takes place they return records and provide access to the author's profile. Unlike Scholar and Academic Search, CiteSeerX doesn't offer detailed information regarding the author and it was noticed that the records it contains are not up-to-date. On the other hand, Google Scholar and Microsoft Academic Search offer in great detail information regarding both the document and the author, including author's h-index and citations, and also provide links to each document's co-authors' profile pages. Academic Search presents information in a very well structured manner through a sidebar that lists co-authors, conferences, journals and keywords related to the author. Finally, it is the only one that offers co-author graphs, co-author paths and citation graphs. Example author search, with author's profile page, for CiteSeerX, Google Scholar and Microsoft Academic Search are shown on Image 3, Image 4 and Image 5, respectively.



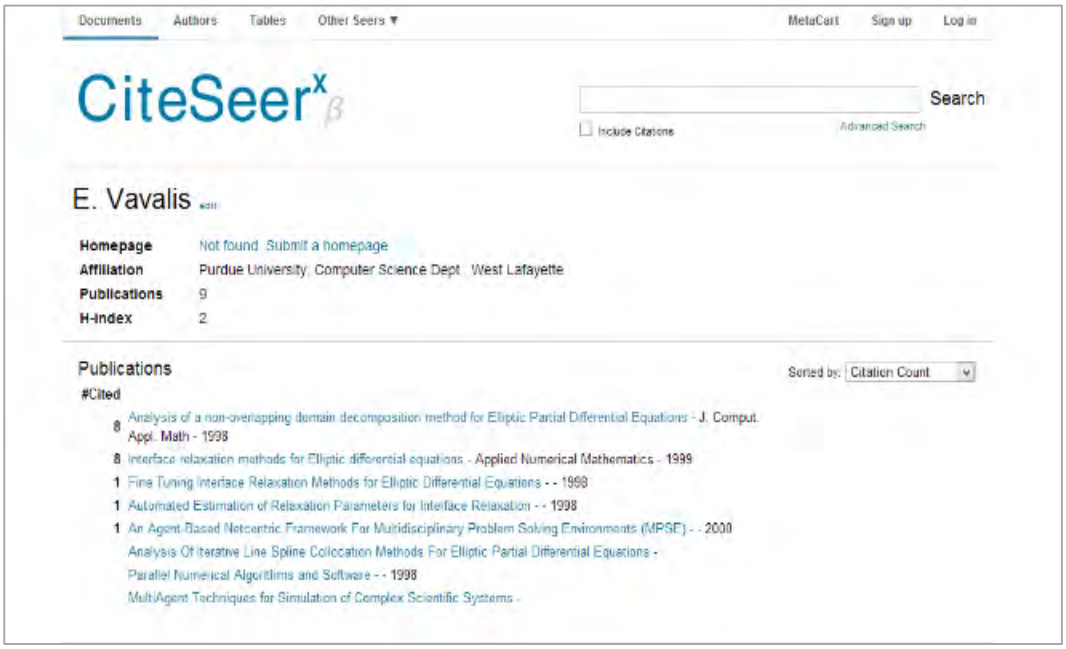


Image 3 Results on author search for CiteSeerX

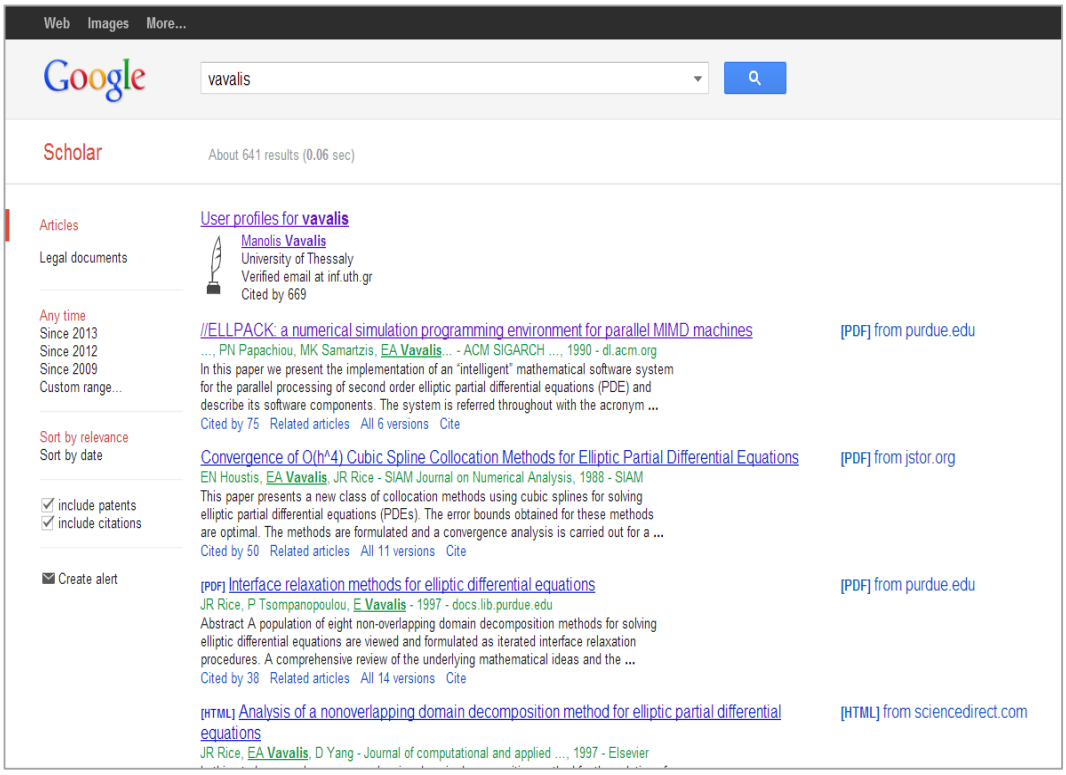
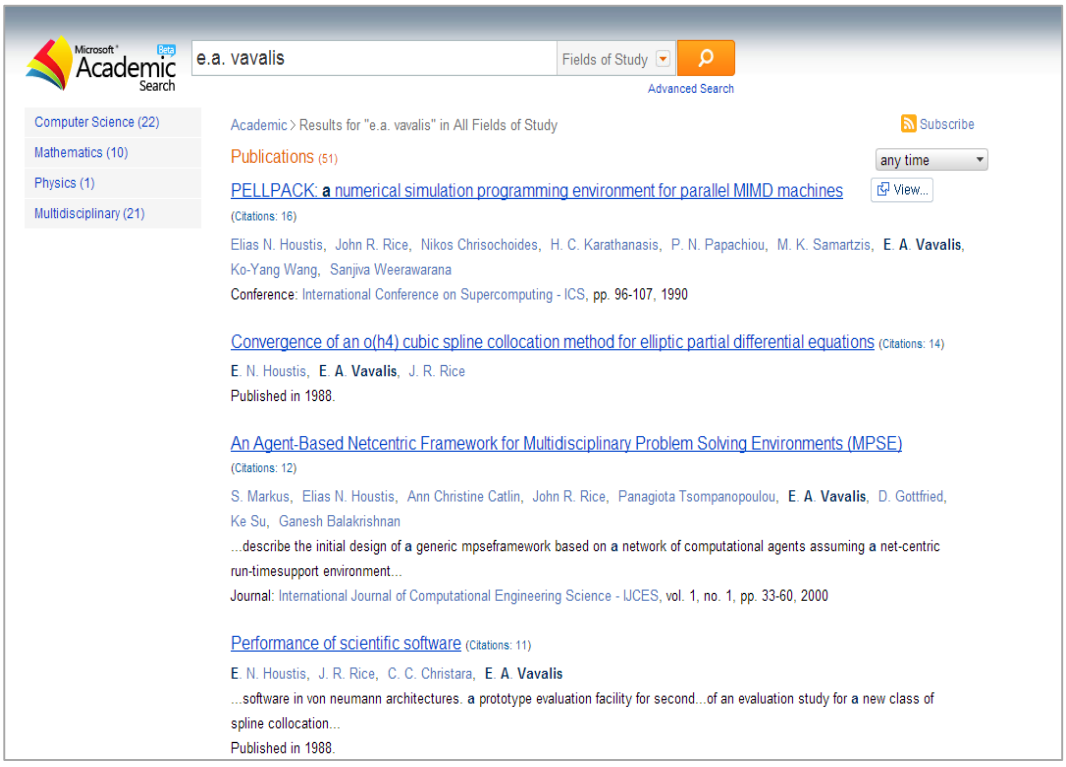


Image 4 Results on author search for Google Scholar



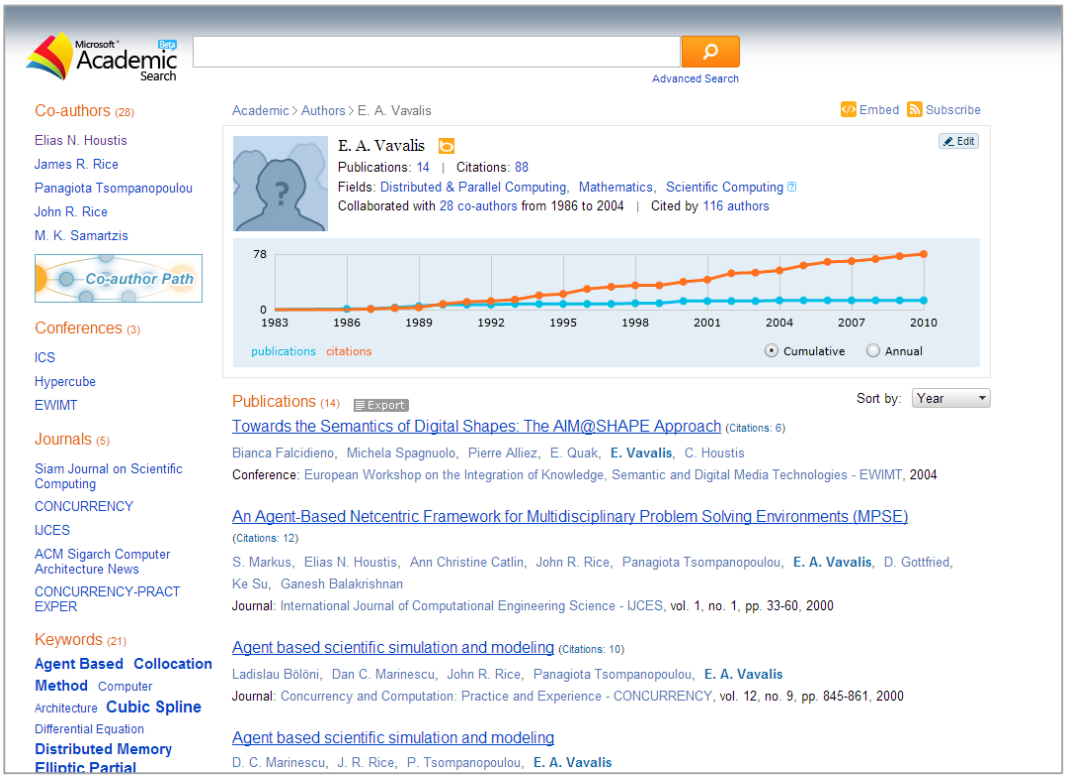


Image 5 Results on author search for Microsoft Academic Search

Thomson Reuters describes **InCites** as a customized, web-based research evaluation tool that allows organizations to analyze their productivity and benchmark their output against peers worldwide. InCites comes with three modules, the Research Performance Profiles module, the Global Comparisons module, and the Institutional Profiles module. The first two are built on citation metrics from the Web of Science, whereas the latter is using a combination of citation metrics from the Web of Science, profile information from institutions, and reputational data from the Global Institution Profiles Project [8]. Using the three modules institutions may extract quantitative data, as well as qualitative data.

The **Research Performance Profile** module allows institutions to monitor their contribution to scholarly literature and they can answer questions such as "*How many papers did my institution produce?*" and "*What are the strongest fields at my institution?*" Since a specific dataset was appointed to us for this module, the reports generated are regarding Greece. This module comes with seven submodules.

- Custom Report Generator: offers the ability to create custom reports of various types. Some of them are author, country, institution, and journal ranking for a specific time period. Moreover, the results can be limited to include only specific terms like, authors, subject areas, institutions, countries/territories, journals, titles, keywords, documents types, etc.
- Overview and Summary Metrics: offers the ability to view the executive summary, summary metrics, the citation frequency distribution, source articles per year, and funding agencies listings. Some of the information provided via this category includes top producing authors, most cited authors, most active subject areas, and other.
- Productivity and Researcher Output: this category includes rankings for journals, authors, and article types for documents included in a specific dataset.
- Collaboration and Research Networks: reports extracted by this category are based on a set of papers where the same institution or country, appear in all author addresses. Reports contain institutions or countries that are collaborating and the metrics regarding their produced documents. The number of Web of Science documents and metrics are listed for a country X, followed by the countries that are present in the address field of each paper. For example, Greece has 220.276 documents that were cited 2.082.093 times, from which 22.572 present the United States in the address field and were cited 485.078. This shows that the collaboration between Greece and the United States has made the greatest impact of any collaboration between Greece and any other country.
- Specialization and Subject Area Strengths: includes metrics regarding subject areas and keywords presented in documents, such as which subject area received the most citations, and under which subject area where the most documents produced.
- Trends and Time Series Analysis: this category displays bar graphs that show the citation and document production activity for the time period 1980 – 2012. The results are presented with a year increment each time, thus providing the overall activity for the country.
- Impact and Citation Rankings: this last category allows the creation of reports based on metric and bibliometric data for citing articles (number of citing article in the dataset) per year, and rankings for authors, institutions, countries, and fields citing one or more of the source articles that belong to the dataset. For example, in author rankings, values are ranked to times cited, which is the number of citations received by the author citing a document that belong to the dataset.

The **Global Comparisons** module lets institutions identify broader trends and benchmark their performance compared to others. Questions that can be answered include "*How many papers did a country produce?*" and "*Where is our work having the most impact?*" This module gives the ability to extract reports comparing up to 50 countries/territories and subject areas to see publication and citation information in a selected time period. Also, it

is possible to see this information regarding one country/territory, or subject area. The same functionality is available for institutions.

Last, the **Institutional Profiles** module provides profiles for nearly 665 of the world's leading institutions and presents details of faculty size, reputation, citation measures, etc. Using this module institutions can answer questions like "*How does our research output compare with the results it yields?*" and "*How does my institution's research perform against citation impact?*" Reports include a radar graph that demonstrates the overall research output of an institution by default, but results can be narrowed down to six major subject areas; Arts & Humanity, Clinical, Pre-Clinical & Health, Engineering & Technology, Life Sciences, Physical Sciences, and Social Sciences and by using a different indicator group; finances, research performance, etc. The report is accompanied by a table containing data reported by the institution and includes data regarding citations, papers, research income (funding), etc.

InCites isn't limited to only retrieve information about a specific institution. Due to the size of the Web of Science database, InCites offers access to an enormous volume of data and information. All information given by InCites is available for printing and downloading as excel or pdf files.

InCites can be used in order to assess the productivity and impact of an institution, but also to compare these results to other institutions. The information available from InCites, gives institutions the provisions needed to create a strategic financial and even promotional plan. This is very important, especially for private institutions.

Apart from being a useful tool for institutions, countries can benefit from its usage. Since InCites offers the ability to compare country metrics, the Global Comparisons module can be used to monitor publication and citation activity. Moreover, countries can detect in which subject areas their productivity or metrics were at a lower level and act accordingly, and national evaluation services can use this data to assess national institutions.

The **Turkish Scientific and Technological Research Council (TUBITAK)** [9] in a recent effort to trigger the development of innovation and entrepreneurship prepared the **Innovative Entrepreneur University Index** [10]. With the help of the index, they were able to rank universities according to the quality of education while collecting 400,000 sets of data regarding 209 institutions. The Index consists of 23 indicators that are divided into the 5 dimensions that are listed below.

1. Competence of Scientific & Technological Research (Weight ratio: 20%)
    a. Number of scientific publications.
    b. Number of citations.
    c. R&D and innovation, the project received support programs.
    d. Amount of funding received from R&D and innovation.
    e. Number of national and international scientific awards.
    f. Number of PhD graduates.

2. Intellectual Property Pool (Weight ratio: 15%)
    a. Number of patent applications.
    b. Number of patent documents.
    c. Utility model/industrial design, the number of documents.
    d. International patent application number.

3. Collaboration and Interaction (Weight ratio: 25%)
    a. University - industry collaboration in R&D and innovation in number of projects.
    b. University-industry collaboration in R&D and innovation projects in the amount of funds received.

c. International co-operation in R&D and innovation, the number of projects with.
d. International collaborations in R&D and innovation, the amount of funds available.
e. Circulation instructor/student numbers.

4. Entrepreneurship and Innovation Culture (Weight ratio: 15%)
    a. Undergraduate and graduate-level entrepreneurship, technology management and innovation management in the number of courses.
    b. Technology Transfer Office, technological parks, incubators, and the management of TDC number of people working full-time.
    c. Structuring the presence of the Technology Transfer Office.
    d. Held out of college for entrepreneurship, technology management and innovation management training/certification program number.

5. Economic Contribution and Commercialization (Weight ratio: 25%)
    a. Techno-parks academics, incubation centers, the number of firms active in TDC shared or owned.
    b. College students who graduated in the last five years or techno-parks, incubation centers, the number of firms active in TDC shared or owned.
    c. Techno-parks academics, incubation centers, the number of people employed in firms owned by public or TEKMER (Technology Development Center).
    d. Licensed patent/utility model/number of industrial designs.

TUBITAK intends to use the index once a year to evaluate the national universities and the results will be publicly available. Even if Turkish Universities are accessible only by the élite of the country, TUBITAK made this pilot effort, and they state that it has already attributed to the improvement of the quality of a number of institutions. Without mentioning which of them, they stated that seven out of the top ten showed a lot of change compared to the first trimester of 2013 [11].

The **Frascati Manual** is a standard for R&D surveys in OECD (Organization for Economic Co-operation and Development) member countries and worldwide. It is based on experience that was gained from collecting R&D statistics in OECD member countries. The Manual is a technical document and consists of two parts: (1) recommendations and guidelines on the collection and interpretation of established R&D data, (2) interpretation and expansion upon the basic principles outlined in the first part for the provision of additional guidelines for R&D surveys or topics relevant to such surveys. The Frascati Manual classifies research into three categories:

1. Basic research: experimental or theoretical work for the acquisition of new knowledge of the underlying foundation or phenomena and observable facts.
2. Applied research: original investigation undertaken for the acquisition of new knowledge and is directed towards a practical aim or objective
3. Experimental development: systematic work, drawing on existing knowledge that is directed to producing new materials, products or devices, etc. or to improving those already produced or installed.

The Manual also includes a Field of Science (FOS) classification, which consists of six high level groupings; Natural Sciences, Engineering and Technology, Medical Sciences, Agricultural Sciences, Social Sciences, and Humanities. The classification was revised in (OECD, Revised Field of Science and Technology (FOS) Classification in the Frascati Maunal, 2007) resulting to the alteration of Medical Sciences level to Medical and Health Sciences. The Frascati Manual is a part of a series of methodological manuals known as the “Frascati Family”, along with the Oslo Manual (innovation), the Canberra Manual (human resources), technological balance of payments and patents as science and technology indicators. (OECD, The Frascati Manual, 2002)

1.4 Thesis organization

The rest of this thesis is organized as follows. In the following section we present efforts that are closely related to our study. In section 3 the reader can find detailed information on the aggregation and organization of data, the procedure needed in order to collect data, any difficulties and how they were addressed and finally, how the aggregated information was organized. In section 4, we propose the structure of the web observatory, plus the visualization of information and data, that was developed by (Giakas, 2013) as a part of his dissertation. And finally, in section 5 we include information about Research in Greece, including which scientific fields are more popular, the country's citation, h-index and impact scores and a short and preliminary evaluation of 21 National Universities based on their research activities.

2. State of the art and related work

Some of the world's most pronounced scientists, many of them Nobel laureates, have made important discoveries or inventions while working at universities. For example, MIT is considered a pioneer in life changing discoveries in the field of The Physical Sciences, Engineering, and Information and Computer Sciences. On the other hand major Biological and Biomedical Discoveries took place at the University of California and at the University of Stanford.

The promotion of research output is of major importance since Universities rely on it for future funding. Moreover, Universities keep very competitive profiles; those ranked as best worldwide want to keep their status and others ranked in lower positions, want to improve their performance. Hence, their research activities are listed in great detail through their websites and their research news are updated constantly. A large percentage of international Universities offer RSS feed and/or use social media, publish monthly magazines and even hold science blogs to promote their research output.

2.1 Pure: A commercial CRIS

Pure has been developed by Atira (a Danish company, specializing in software engineering in the field of Research Information Management) in 2003 and is a commercial CRIS based on the CERIF standard. Atira was acquired by Elsevier in the summer of 2012 and Pure can be found now in SciVal. It can be integrated with local systems and its functionality covers the following content types:

- **Grant applications** – funder, program, amount applied, dates, etc.
- **Funding opportunities and awards** – funder, program, amount, contract documents, etc.
- **Projects** – participants, budget, external collaborators, students, outputs, etc.
- **Output and impact** – peer-reviewed journal articles, books, chapters, etc.
- **Equipment** – type, placement, etc.
- **Bibliometrics** – citations , impact factors
- **Activities** – conferences, peer reviewing, etc.
- **Press clippings** – national and international papers, etc.
- **Publishers and journals** – Sherpa RoMEO colors (RoMEO colors have been developed by Sherpa, which is a UK-based project team with expertise in open access and repositories, and are used to differentiate publishers' entries between four categories of archiving rights), issue, and page number, DOI, etc.
- **Organizations** – faculties, institutes, departments, external organizations, etc.
- **Persons** – researchers, postgraduate students, esteem, plus external persons.
- **Student theses** as a separate content type from other publications

Pure is now licensed to multiple institutions, including Universities, Hospitals, Government Research Institutions and Pharmaceutical companies in the UK, Denmark, Sweden, Finland, Germany and Belgium.

Aalborg University (Denmark), Roskilde University (Denmark), University of Helsinki (Finland) and Lancaster University (England) are some of the tens of Universities that use Pure. Since the aforementioned Universities have integrated Pure with their local systems the offered functionality is the same, so we will be presenting one of them as an example.

Aalborg University holds a publicly accessible research database called VBN. Through VBN, the University wishes to communicate its research activities, publications and research projects. As VBN is not a repository it does not offer immediate access to the full text of publications.

A user can explore VBN through a menu that gives him/her access to data about publications, researchers, research projects, activities, press clippings, research units and statistics. A publication page (Image 7) contains the list of authors, the research units involved, the abstract, some information regarding the document and the link to the University's library where the user can access the full text. The structure of a project's page is very similar to a publication's page, including the researchers and research units that are involved, miscellaneous information about the project and related projects. For every publication and project, VBN also offers graphical representations of the people and research units involved. A researcher's page (Image 6) contains information regarding his/her publications, research projects, most frequent journals, activities, press clippings and most frequent publishers, plus a map showing where the researcher's latest activities and conferences took place. Through VBN a user can also find information about a particular research unit. The information and structure are similar to those provided for a researcher and an example is shown on Image 9. A very interesting feature offered for a research unit is statistics (Image 8); one figure demonstrates the faculty's number of publications per year and a second figure that shows publications per type per year.

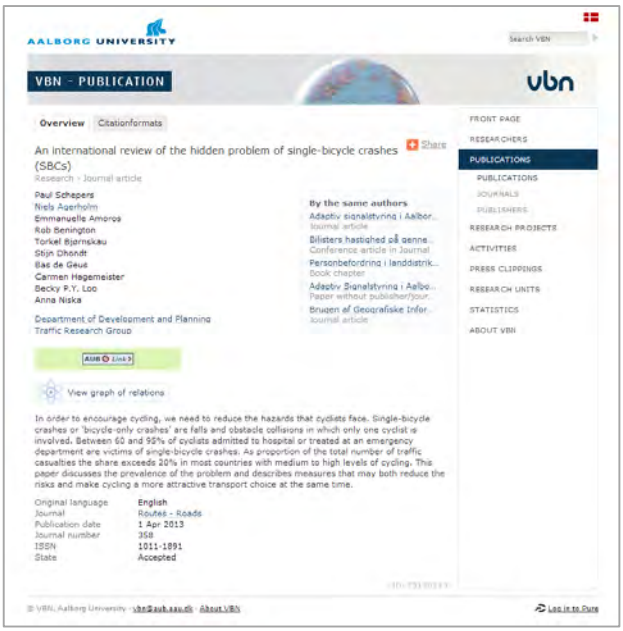


Image 7 VBN Publication Page

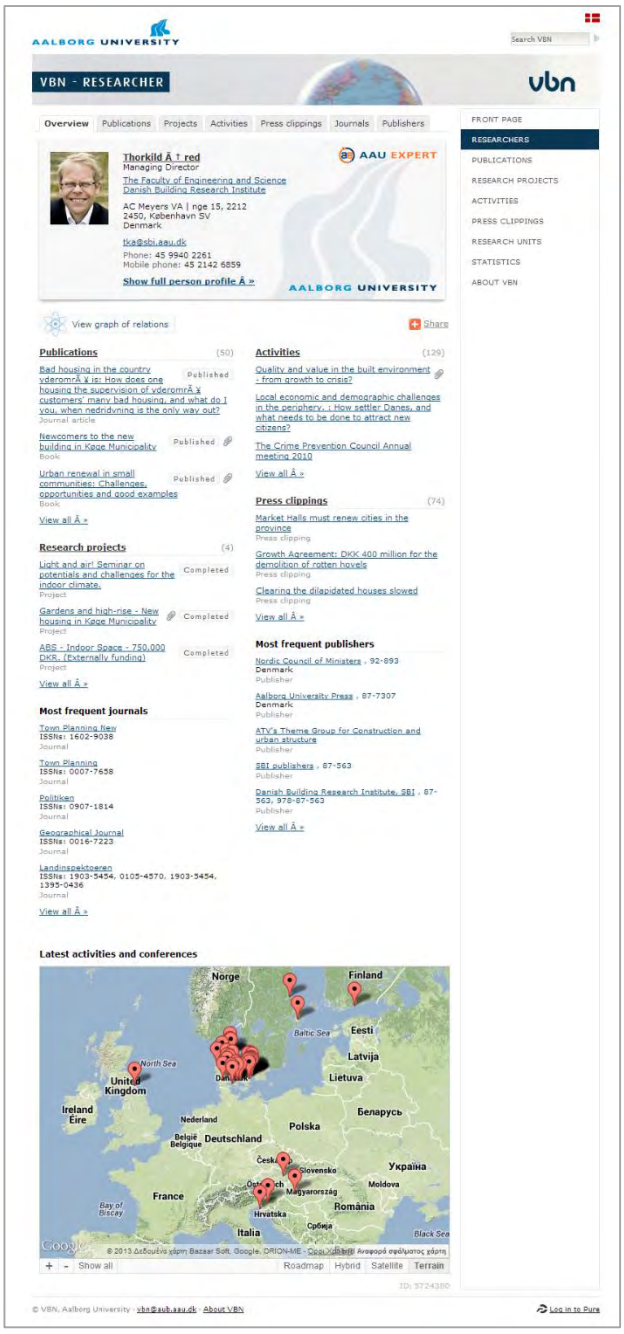


Image 6 VBN Researcher Page

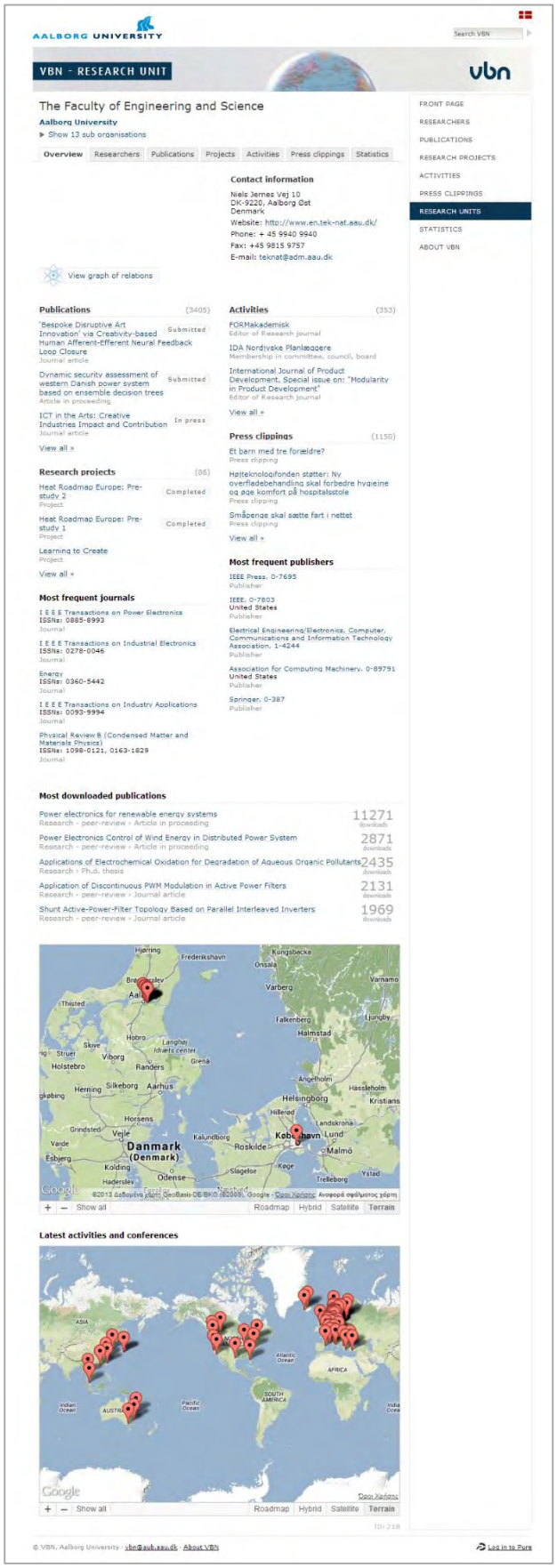


Image 9 VBN Research Unit Page

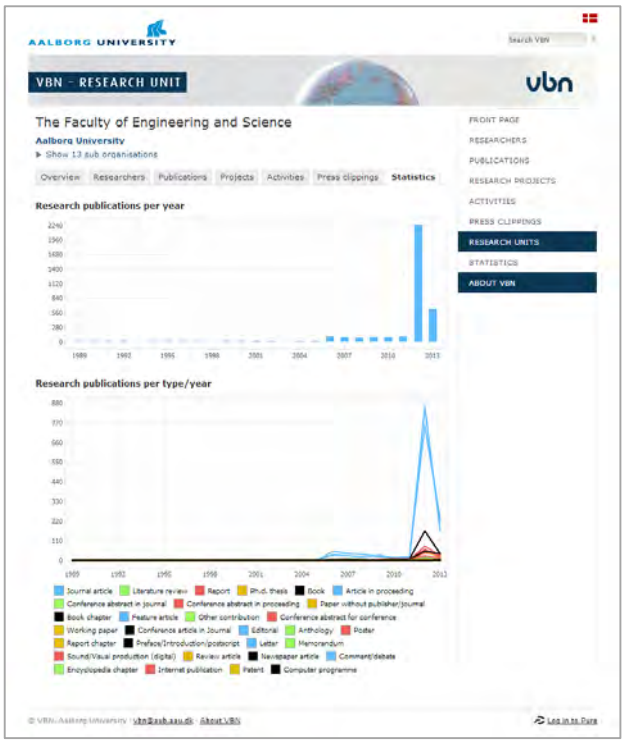


Image 8 VBN Faculty of Engineering and Science Statistics

Assessment

Pure is a well-developed product. Information is well structured, it is user-friendly, and the offered visualizations are uncluttered and easily read. Apart from the fact that we aim to provide as much graphically presented information as possible, it offers all the functionality that we are looking for. The reason for which Pure is excluded from our available options is the fact that the cost for its purchase is above our budget. For a research organization with up to 200 employed researchers the license costs 23, 700 EUR ex. VAT. Plus, 15% of the license price is charged each year for updates and patches.

2.2 VIVO: An open source tool for research discovery

VIVO is an open source semantic web application that was originally developed and implemented at Cornell University. At this point we should mention that the Semantic Web is a collaborative movement led by the World Wide Web Consortium (W3C). The term was coined by Sir Tim Berners-Lee for a web of data that can be processed by machines. W3C was founded in 1994 by Sir Lee and is the main international standards organization for the World Wide Web (WWW or W3). VIVO is a researcher and research discovery tool, aiming to connect scholars, communities, campuses and countries using Linked Open Data (LOD). VIVO started as a relational database for the Life Sciences department of Cornell University and during the period 2006 – 2008 it was converted to Semantic Web and expanded to all disciplines at the University. In 2009, VIVO begun partnerships with national institutions; Harvard University, Indiana University, Ponce School of Medicine, the Scripps Research Institute, University of Florida, Washington University in St. Louis School of Medicine and Well Cornell Medical College, and by 2012 it contained profiles for 16,231 people, 19,847 organizations, 15,660 courses, 28,839 academic articles and 6,162 grants. Significant partners include CASRAI (Consortium Advancing Standards in Research Administration Information), euroCRIS and the ORCID (Open Research and Contributor ID) initiative. Image 10 shows an example of how a researcher's profile is displayed via VIVO, whereas Image 11, Image 12 and Image 13 show how co-author and co-investigator networks are presented, respectively.

The VIVO project implements a number of W3C Semantic Web technology standards, including **RDF** (Resource Description Framework), **OWL** (Web Ontology Language) and **SPARQL** (SPARQL Protocol and RDF Query Language). The Integrated Semantic Framework ontology modules for VIVO **(ISF/VIVO ontology**) provide a set of classes and properties to represent researchers and the full context in which they work.

The **VIVO Harvester** is a collection of small Java tools designed to take data from external data sources and ingest it into VIVO. The *Fetch* is the first step of a typical harvest, which allows the collection of data from a target source. Since the fetched data are in its own format and needs to be converted into VIVO-compatible triples, it has to be translated. If the input is an XML format, the translation can be achieved using the *XSL Translator* tool and an .xsl file containing XSLT code specific to the data format being converted to RDF/XML triples. Next comes the comparison, via the *Score* tool, between VIVO and the input data. The *Score* tool assigns numbers to the comparisons, which are used by the *Match* tool to compare them to a threshold value. Input entities compared by *Score* and meet or exceed the threshold will have their identities changed to the URI of the person in VIVO in order to link existing data to the new data. The last step before importing the information into VIVO is to give the data proper URIs. This can be achieved via the *ChangeNamespace* tool. The Harvester runs multiple times, creating a "previous harvest model", which is used by the *Diff* tool in order to be compared with the new data and determine if any triples have been altered. The output of the *Diff* tool is an "Additions file" and a "Subtractions file", containing RDF/XML data that should be added and removed, respectively. The *Transfer* tool is then called four separate times to add the data from the "Additions file" to VIVO and the "previous harvest model", and to remove the data from the "Subtractions file" to VIVO and the "previous harvest model". The architecture of the VIVO Harvester can be seen in Figure 3.

In addition to localized installs of VIVO, users are able to search for people, publication, organizations, etc. through VIVO Search, which is a working example of the functionality of multi-institutional search. VIVO search allows the search across the seven partner institutions and across all disciplines. Image 12 shows an example of researcher search. Results include people with matching name compared to the person in question, publications and organizations. On the selection of a search result, the user is redirected to the corresponding page of the institution that is affiliated to the person or publication.

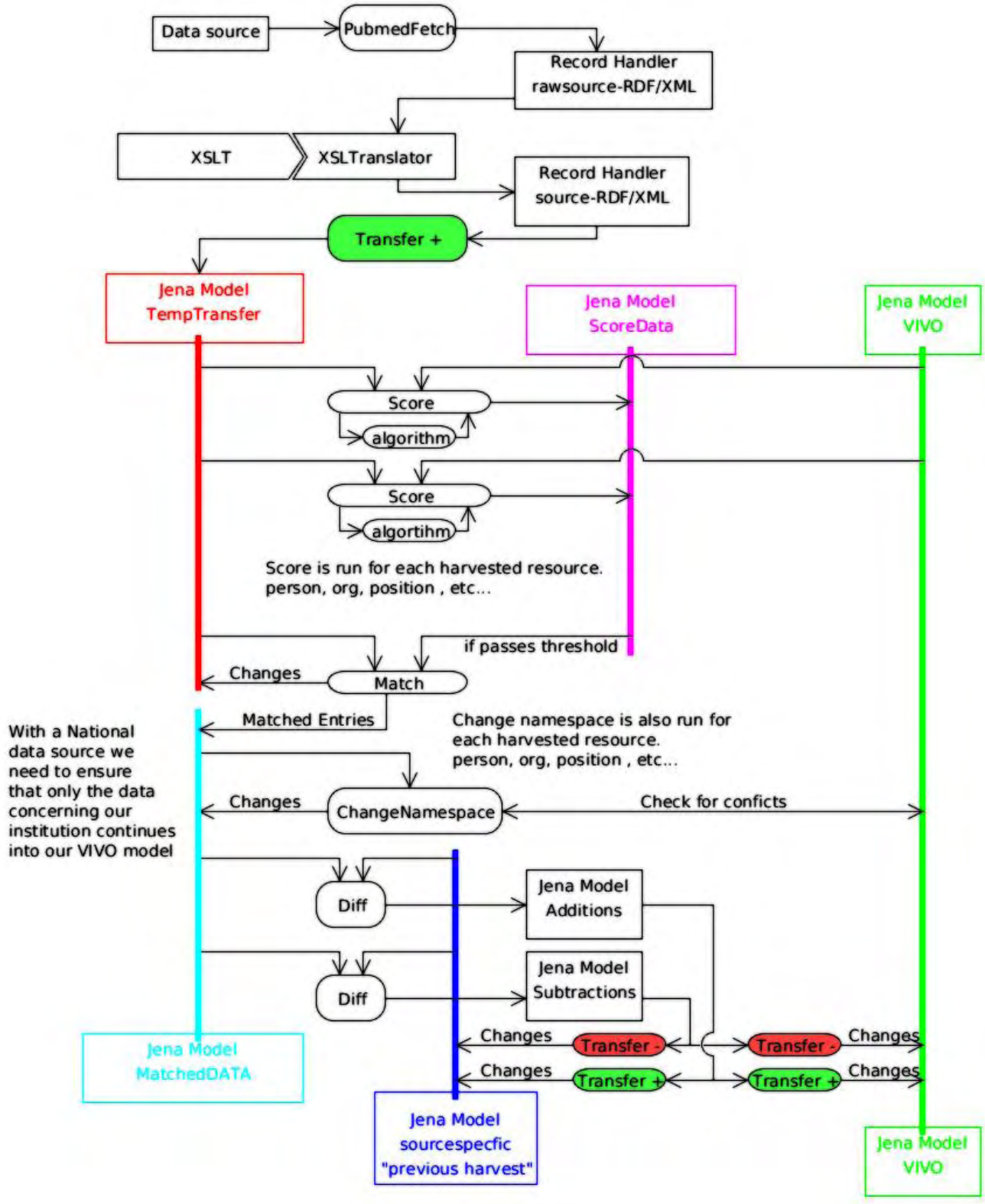


Figure 3 VIVO Harvester [source: https://wiki.duraspace.org/display/VIVO/VIVO+Harvester+User+Guide+1.0]

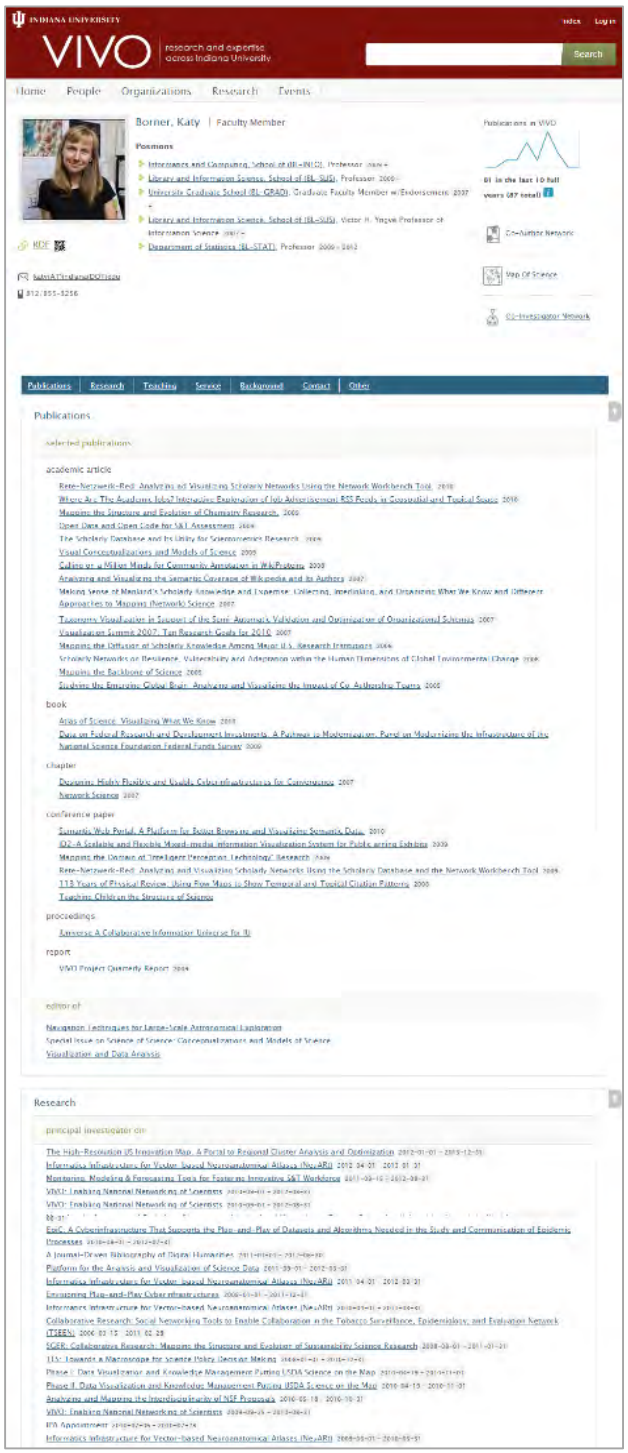


Image 10 VIVO Researcher Page

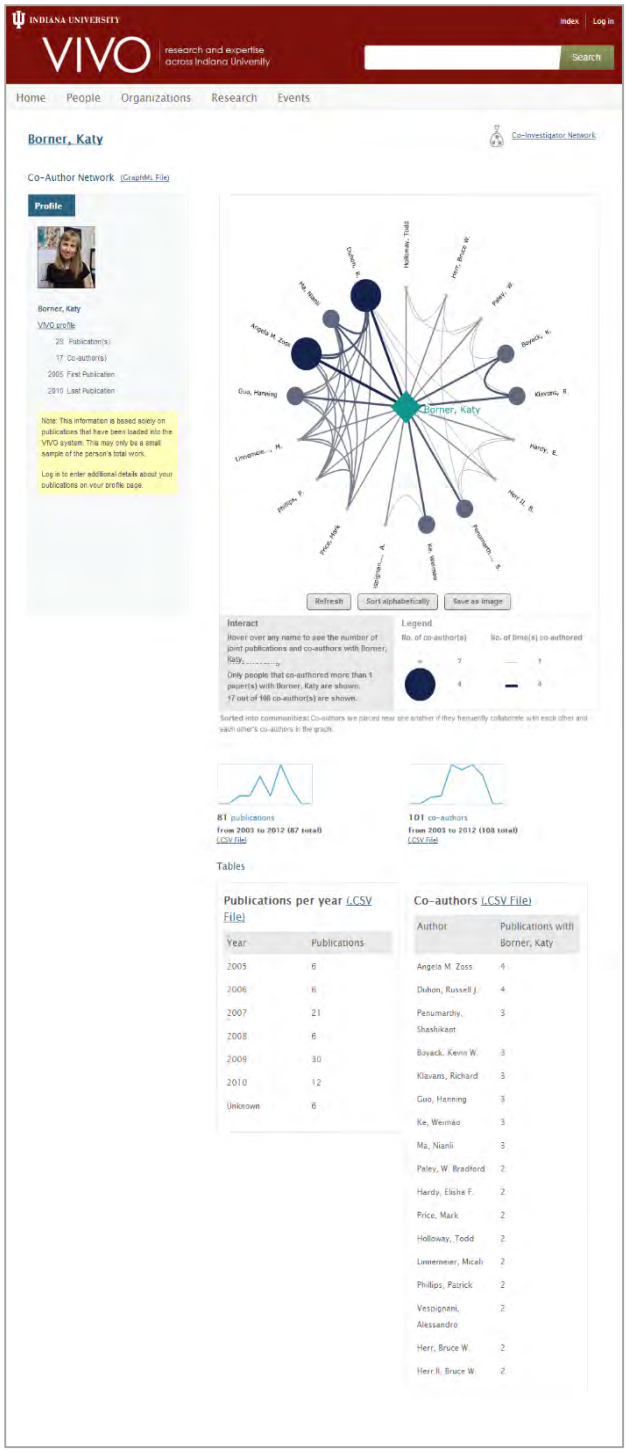


Image 11 VIVO Co-author Network

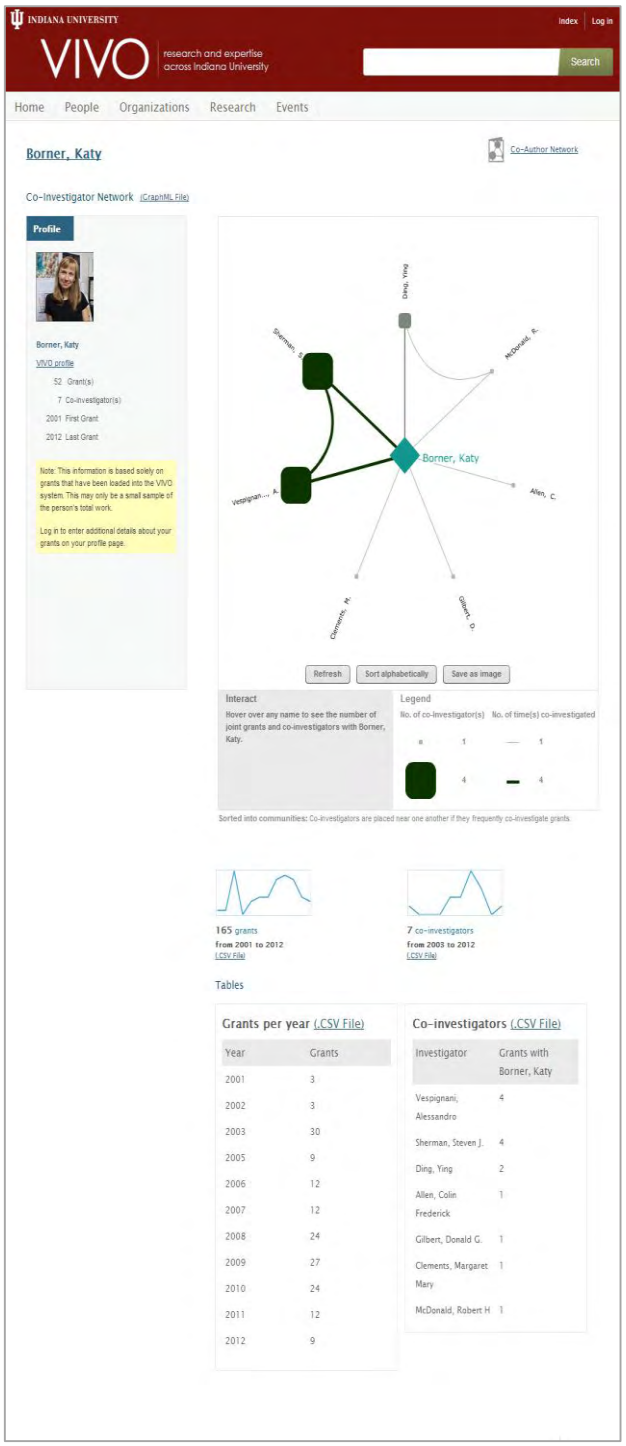


Image 12 VIVO Co-investigator Network

Image 13 VIVO Researcher Search

Assessment

VIVO has been evolving during the last decade and it has been adopted internationally by multiple European Universities and organizations, nine Oceania Universities and the Global Forum on Agriculture Research. The Chinese Academy of Sciences and the Tsinghua University are some of the 100 institutions that are part of the VIVO network. It is a well-developed application with great prospective, extensive documentation and a large community on an international level. It is being sponsored not only by the founding Universities, but from Brown University and Memorial University as well. Symplectic [12] (a company focused on the development of integrated research information management systems) and Elsevier have also invested on VIVO as founding corporate sponsors. With the support of its partners, VIVO seems to be heading towards being the largest researcher network on a global scale.

2.3 Best Practices

As mentioned above, the object of the thesis is the aggregation and dissemination of data, regarding the research activities of the faculty of the University of Thessaly. With this in mind the search for best practices on an international level was focused on finding Universities that provide this kind of information through their websites. To be more specific, the interest was in finding Universities that keep repositories containing information about publications, research projects and research activities.

The **Hong Kong University (HKU)** offers the HKU Scholars Hub [13] which is a repository that holds all information regarding the research activities of the faculty. The Hub uses the DSpace system software developed by MIT and Hewlett Packard. DSpace offers the Lucene search engine, which has been further enhanced by the Hub, in terms of search and retrieval capabilities by integrating content with Elsevier's full text retrieval software, and the SCOPUS citation analysis tracking system. The Hub can be easily accessed from the University's website, is very well structured and it offers to the user the capability to perform a quick search and/or to choose the field (publications, patents, etc.) that interests him/her from a menu. It offers search based on a researcher's name, publications, theses, grants and patents. When performing a search based on a scholar's name, the user is able to access all information concerning his/her research activities, such as networks of collaboration, publications, achievements, grants and bibliometric scores. On publication search, a user has the ability to choose between multiple fields in order to narrow down the results, including faculty, year and type. A publication page (Image 15) contains all information related to the document, including authors, issue date and affiliated department(s). As mentioned before, a researcher page contains contact information, publications, and visualizations of their network of collaboration, achievements, grants and bibliometrics. An example is shown on Image 14. HKU Scholars Hub is a great representation of what we are trying to accomplish. However, it would have been more interesting for the user if there was more visualized information.

Image 14 HKU Scholars Hub Researcher Page

Image 15 HKU Scholars Hub Publication Page

Tokyo Institute of Technology holds the Tokyo Tech Research Repository (T2R2) [14], a system developed to manage and disseminate academic data, such as papers and books published by the researchers of the University. Search options are divided into three categories: the Horizontal search for publications, patents and research highlights, the Researcher search and the Search by organization. T2R2 is very well structured; however, the information contained in publications and patents page is very limited (Image 16). A major disadvantage of the way T2R2 was developed is the fact that even though it was designed to disseminate academic data abstracts

and/or links to the full text of a publication are not included. The University also curates a repository that contains information related to its researchers called Tech STAR [15]. A user can perform a basic search using researcher's name, affiliation, title (professor, lecturer, visiting professor, etc.) or field of study. Moreover, an advance search can be performed by adding more search fields, such as researcher's telephone number, email, publication, course, etc. A researcher's page (Image 17) is presented in great detail, including contact information, top five most related researchers, professional experience, publications, books, awards and many more. However, it was observed that this was not the case for all researcher pages. Many of them were lacking data. Plus, even on the English version of the website, most of the content was written in Japanese, making the browsing difficult for non-speaking Japanese users. It should be mentioned here that multilingualism is certainly a crucial scientometrics issue for institutions whose mother language is not English. This issue is beyond the scope of this thesis.

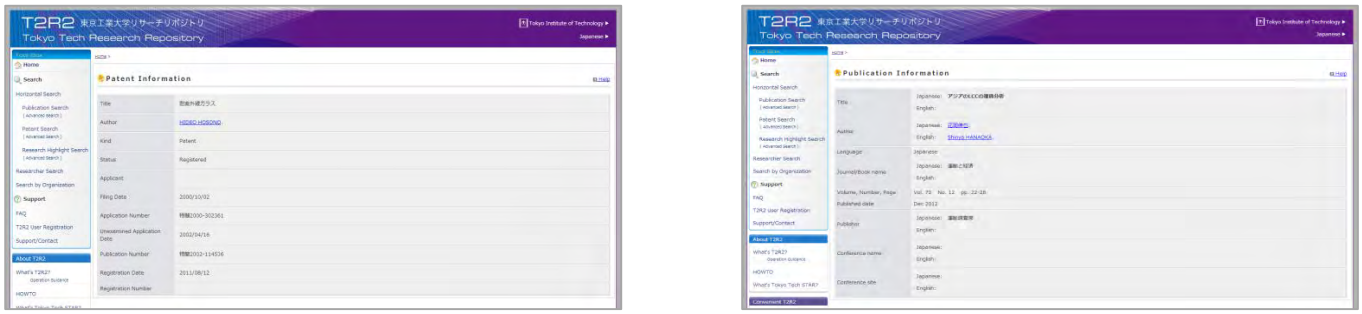


Image 16 T2R2 Publication and Patent Page

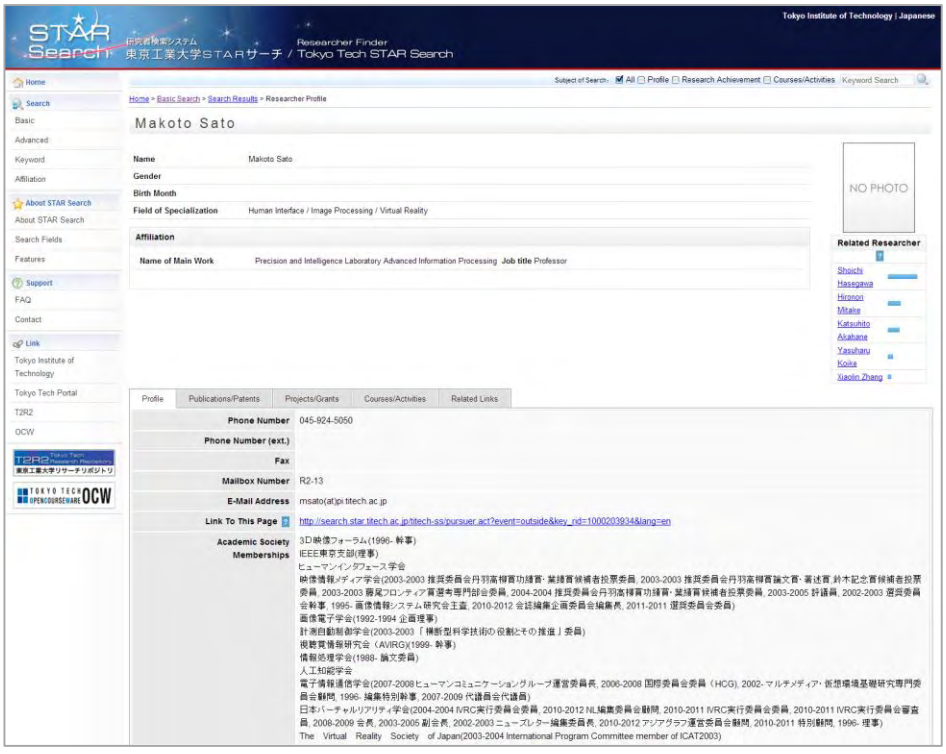


Image 17 TechSTAR Researcher Page

Wageningen University offers the ability to search for publications and researchers [16] using a single keyword. The query will return researchers whose expertise contains the selected keyword and publications that also contain the keyword. The user can refine the search criteria in both categories; for researchers: select organization and for publications: select publication type, year and organization. A publication page contains the summary, a link to download the document and miscellany information about the document, such as authors, type and year of publication. The user can also use social media to like, share or tweet (Facebook, Google Plus and Twitter) about the document. A researcher's page contains the researcher's contact information, publications, patents, books, dissertations, professional information, expertise and keywords. Image 18 and Image 19 show an example publication and researcher page, respectively. The University's repository is very well designed, user friendly and the fact that a user can perform a search using a single keyword makes the search very easy. The, initial but proper, integration to the web social mention is surely a not so common but very promising effort.

To sum up, the website of the Wageningen University was the easiest to browse, as information can be found by using just one keyword. However, it lacks visualizations and bibliometric data regarding the researchers. T2R2 was very well structured and easy to navigate. But, the lack of information about documents and links to full text was a major disadvantage, as well as the fact that the translation was not successful as far as the page content goes. HKU Scholars Hub contains all the information that we are interested in having in the University's observatory. It is also well structured and all information is given in great detail. The only drawback is the fact that is lacks visualizations as well.

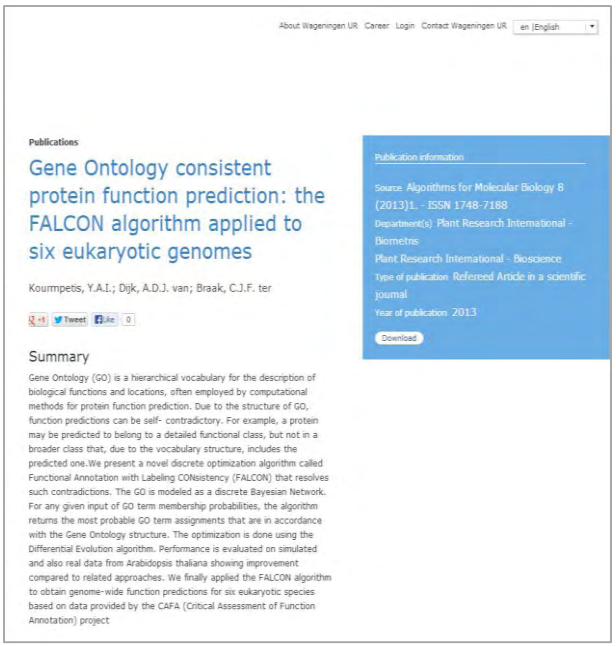


Image 18 Wageningen University Publication Page

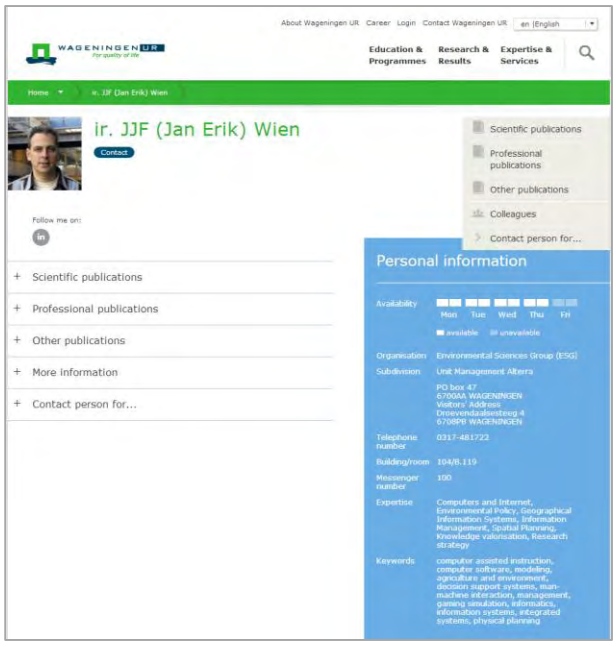


Image 19 Wageningen University Researcher Page

3. Aggregation and organization of data and information

3.1 Sources of information

In order to achieve the goals of this thesis, data and information needed to be collected from various sources and in miscellaneous forms. All the information accumulated so far is relevant to the research activities of the faculty of the University of Thessaly. It includes registered patents, publications, research projects, conferences, awards and honors, theses and dissertations.

Patents registered to faculty of the University were collected from the United States Patents and Trademark Office (USPTO) and the European Patent Office (EPO). Roughly 80% of registered patents belong to faculty from the Department of Electrical & Computer Engineering. Since many of the professors of the department used to live and work in the United States of America, the majority of these patents are registered at the USPTO. What is astonishing is that fact that among 12 professors there are 70 registered patents, from which 32 are registered to Dr. Ioannis Katsavounidis. Since the two offices do not offer APIs for automated data retrieval, data were collected and will be updated every six months manually. The process of collecting patents was time consuming since the search had to be performed name-by-name for each one of the faculty members.

Maria Markou, a PhD candidate in the Department of Electrical and Computer Engineering at the University of Thessaly, created a database containing information regarding **publications** with data that she collected from Scopus. In the database some additional tables were added, and all can be seen below:

- paper: the table contains publications and their metadata. Some of the fields include authors, publication year, cites and publisher.
- author_name: contains the miscellaneous name formats for each of the faculty members present in Scopus.
- author: the table contains author's name, publications, citations, co-authors, h-index, etc. For the name field a unique name format was chosen.
- patents: contains the title of the patent, the names of the inventors, the date the patent was filed, the abstract and the URI to the document.
- sarea_author and sarea: the tables contain the scientific fields that each faculty member is involved in.
- author_keyword and akeyword_paper: the tables consist of the author keywords and the id of the publication in which they are present.
- indexed_keyword and ikeyword_paper: correspondence with the above tables, only for index keywords.
- departments and depart_author: the former consists of the names of the departments of the university and the latter contains the IDS of both department and author for each faculty member.

The EER Diagram of the database is shown in Figure 4.

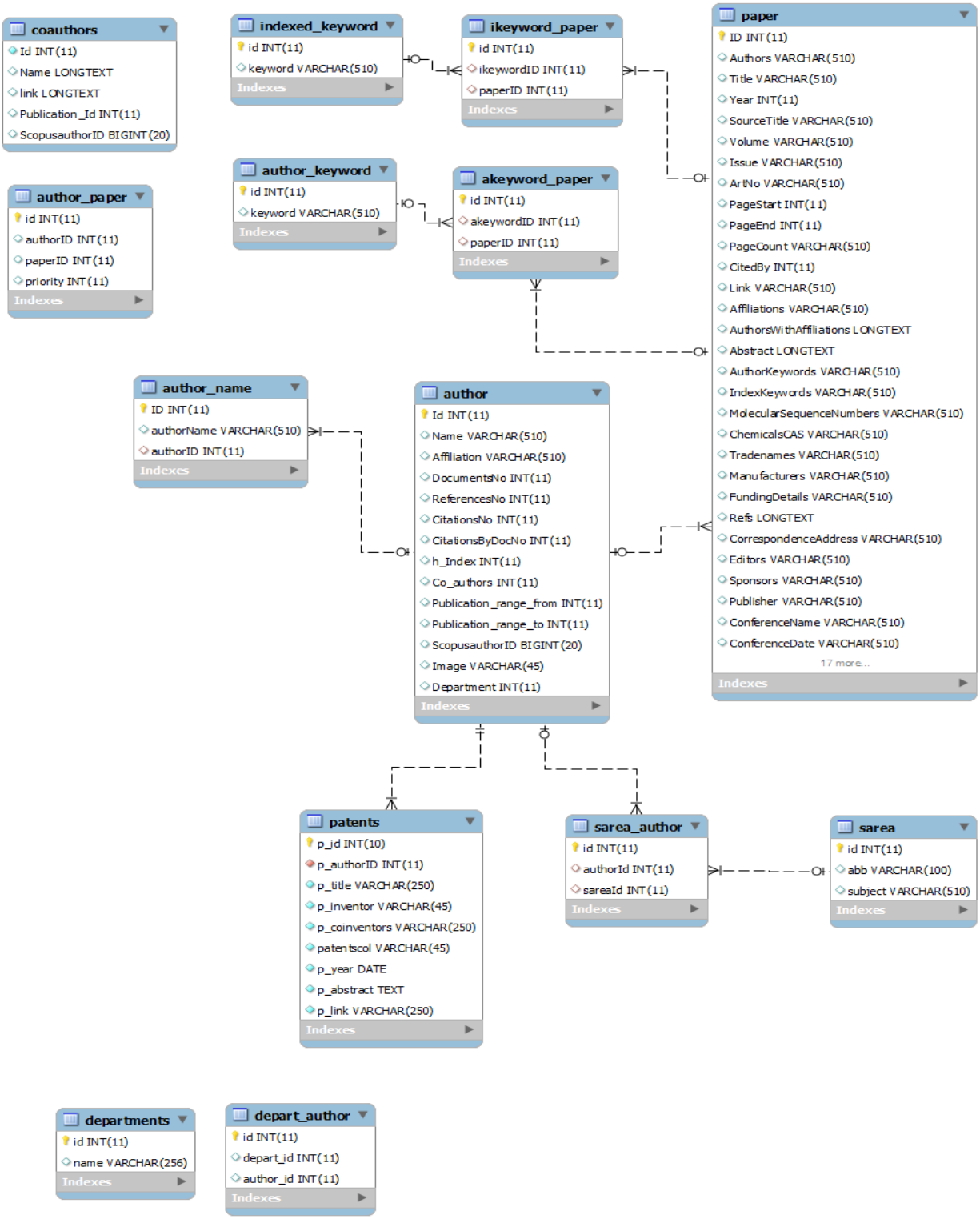


Figure 4 EER Diagram of the Database

Information about **funded research projects** is to be collected from different sources. The main source of information is the Research Committee of the University of Thessaly, of which part of its jurisdiction is the evaluation, selection and financing of research projects. The Research Committee offers a plethora of data on each project. We have chosen to collect the ones that we consider more important (in terms of our goals) and we list them below.

- Title, abstract and type of the project.
- Scientific coordinator and participants.
- Funding body and budget.
- Start and end dates.
- Type of research.
- Final report and,
- Year of completion.

The abstract of each project will later be parsed in order to extract keywords. This is needed for two main reasons: (1) to collect the keywords for each project in order to attribute them to the scientific coordinator, and (2) to offer the functionality of searching for research projects using keywords.

Unfortunately, the data provided by the Research Committee of the University of Thessaly are not complete. We believe that this is currently the case for many (perhaps all) Research Committees at the other Greek Universities. One reason for that is the fact that many faculty members are involved with projects in Research and Development Centers not directly related to their own University. There are several such centers that range from the prominent and most successful ones (like FORTH and CERTH) to newly establish and thematically restricted ones. It seems though that soon due to the on-going national research restructuring the above described situation will be resolved.

Moreover, information about European funded projects may be collected from the Community Research and Development Information Service (CORDIS). CORDIS allows the collection of metadata regarding a particular institution/organization. This was very helpful to us, as we were able to easily collect metadata in xml format explicitly for the University of Thessaly.

The Quality Assurance Unit (MoDiP) [20] in the context of the internal assessment of the departments of the University curates a database that contains information about the research activities of the faculty. In order to collect information regarding **conferences**, and **awards** and **honors** an API offered by MoDiP will be used. Each faculty member is responsible for importing his/her data in the information system of MoDiP, thus there is a possibility that not all data are up-to-date.

In addition to their research activities and teaching, faculty supervises students' **PhD** and **Master Theses**, and **dissertations**. Hence, their participation in such projects may reflect additional research activities, especially when referring to postgraduate theses. The metadata records of the National Archive of PhD Theses of the Hellenic National Documentation Centre (EKT) are available as open data, and information is provided for harvesting via an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol interface. However, all the records belong to one set, so we contacted their Help Desk team expressing our need to collect records for the University of Thessaly explicitly. Mr. Nikos Houssos (head of Software Development Unit at EKT) responded informing us that his team would develop a new set for our team. Moreover, he provided more information regarding euroCRIS and CERIF, and he offered to assist in case we were interested in using CERIF. As far as Master theses and dissertations go, metadata records will be collected from a special collection named Gray Literature (GL). GL is under development for several years now by the Library and Center of Information

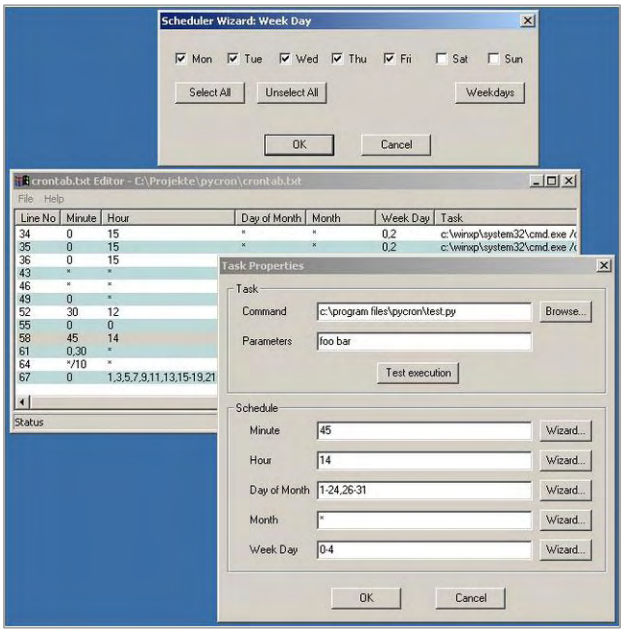
[21] of the University of Thessaly and it will presumably provide a typical API to access all data including all theses that have been defended at the University of Thessaly.

3.2 Procedure

For the needs of the project in (Giakas, 2013) a Windows Server 2012 was setup that has 100GB capacity, 8 processors and 8GB RAM, and has installed on the machine the necessary software, including Apache Tomcat and MySQL. Due to the interdependence (i.e. same database) software needed for the development of the web observatory will be installed on this server.

Aggregation

The first step will be to collect the metadata from the above sources. One way to achieve that is through the crawling agents like the ones developed in (Kallis, 2013). At any rate, since this harvesting of data has to be an automated procedure it is essential that a scheduling script should be implemented. For the scheduling, pyCron (cron service written in Python converted in an executable. Cron is a legacy time-based job scheduler in Unix-like computer operating systems) is a good solution for Windows. pyCron can be setup to run the appropriate script when needed. Since the data we wish to collect will not be altered very often, pyCron can be scheduled to run once a day. Image 20 shows an example usage of pyCron (left) and the setup of the time interval to run a script (right).




```
set_time_limit(60); // one minute, same as CRON ;)
ignore_user_abort(false); // you might wanna set this to true

while (true)
{
    $jobs = getPendingJobs();

    if ((is_array($jobs) === true) && (count($jobs) > 0))
    {
        foreach ($jobs as $job)
        {
            if (executeJob($job) === true)
            {
                markCompleted($job);
            }
        }
    }

    sleep(1); // avoid eating unnecessary CPU cycles
}
```

Image 20 pyCron [source: http://stackoverflow.com/questions/5758387/running-a-php-app-on-windows-daemon-or-cron]

A simple and efficient way to collect metadata from sources would be to store temporarily the data from the XML file offered by the source by targeting its URL. This can be achieved using a PHP script and the

```
int file_put_contents (string $filename, mixed $data [, int $flags = 0
[, resource $context]])
```

function. The new file will also be of XML format and will be named in the form of sourceNameDate.xml. An example of how this can be achieved follows.

As mentioned before Microsoft Academic Search offers RSS feed for publications related to Organizations. After searching for publications related to the University of Thessaly we can access the XML file containing the metadata for all the publications of the University. The URL to the XML file will be set as the inputFile and after reading it, its content is being written to the outputFile.

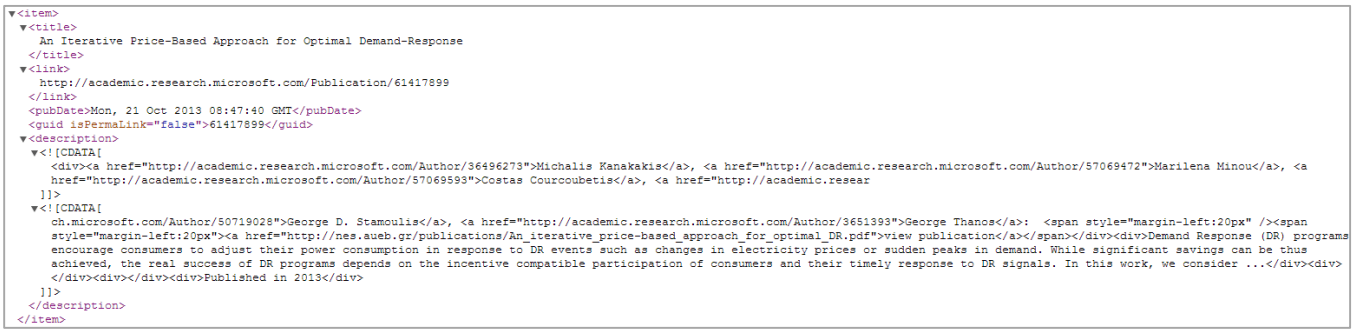


Image 21 Example of item in Microsoft Academic Search RSS

```
<?php

$inputFile = “http://academic.research.microsoft.com/Rss?query=org%3a%28University%20of%20thessaly%29&searchtype=0” //the input file

$currentDate = date(“d-m-Y”);

$outputFile = “/pathtofile/academic$currentDate.xml” //the output file

$fi = fopen($inputFile, ‘r’);

$source = ‘’;

while (!feof($fi)) {

	$source .= fgets($fi);

}

fclose(($fi));

file_put_contents($outputFile, $source);

?>
```

Organization

After the completion of the aggregation of data, the next step is to extract the desirable information and update the database. Since, the XML files that need to be parsed are not complicated; parsing can be achieved using PHP SimpleXML, an extension that provides a simple way of getting an XML element’s name and text. SimpleXML converts XML documents into objects, and it is suitable when performing tasks like reading/extracting data from

XML files. Continuing from the previous example, the following code snippets demonstrate how the inputFile will be parsed to extract the desirable data, which are then passed to the database.

```
<?php
//extracting data from outputFile
$xml = simplexml_load_file("outputFile.xml");
$paper_title = $xml->title;
$paper_link = $xml->link;
…
?>
```

At this point two issues arise:

1. The names of the authors are nested inside a `<description>` tag in a `<a>` tag and need to be extracted before being saved to the database. This can be achieved by using a regular expression to target the `<a>` and `</a>` tags and extract the enclosed text.
2. To avoid duplicated entries of documents in the database we have to check if the document in question is already stored. To do this, we must compare the `$paper_title` string with the appropriate title field of the paper table from the database. In case the document in question isn't already present, we continue to update the database.

```
<?php
…
$sql = 'INSERT INTO paper'.
       '(authors, title, year, …)'.
       'VALUES ("$paper_authors", "paper_title", "paper_year", …)';
…
?>
```

4. Development

The implementation of a CRIS can be achieved by using a subset of the superset of the full CERIF model and due to its neutral architecture and the fact that it is a recommendation to the EU Member States from the EC, we have also chosen to use it in order to accomplish the implementation of the web observatory.

4.1 CERIF in detail

CERIF is a conceptual model that describes the Research domain. It is maintained as an Entity Relationship Model (ERM) from which SQL scripts can be generated.

The Model Structure

CERIF describes entities in the Research domain including person, publication, patent, equipment, funding, etc. and the relationships between them. In (Jorg, Jeffery, Dvorak, Houssos, & al, 2012) the editors present the five entity types that CERIF consists of; **Base Entities**, **Result Entities, Infrastructure Entities**, **2nd Level Entities** and **Link Entities**. Each one of the entities maintains a system-internal identifier and some entity-specific attributes depending on the entity type (e.g. birthday and gender with person).

The **Base Entities** are Person, Project and OrganizationUnit (Figure 5). They allow for a representation of scientific actors and their various kinds of interactions.

The **Result Entities** are ResultPublication, ResultPatent and ResultProduct (Figure 6), and they represent the research output.

The **Infrastructure Entities** are Facility, Equipment and Service (Figure 7).

The **2nd Level Entities** allow the representation of the research context by linking to them form the base, result, and infrastructure entities.

A **Link Entity** always connects two entities, either Base, or Result, or Infrastructure or 2nd Level by id references. Additionally, each Link Entity carries semantics by reference to the so-called CERIF Semantic Layer.

Each one of the Base, Result and Infrastructure Entities recursively link to themselves and maintain relationships with other Base, Result and Infrastructure Entities, respectively.

The CERIF **Semantic Layer** supplies the means for maintaining the CERIF Semantics: types, roles, terminology, subject classifiers, or mappings.

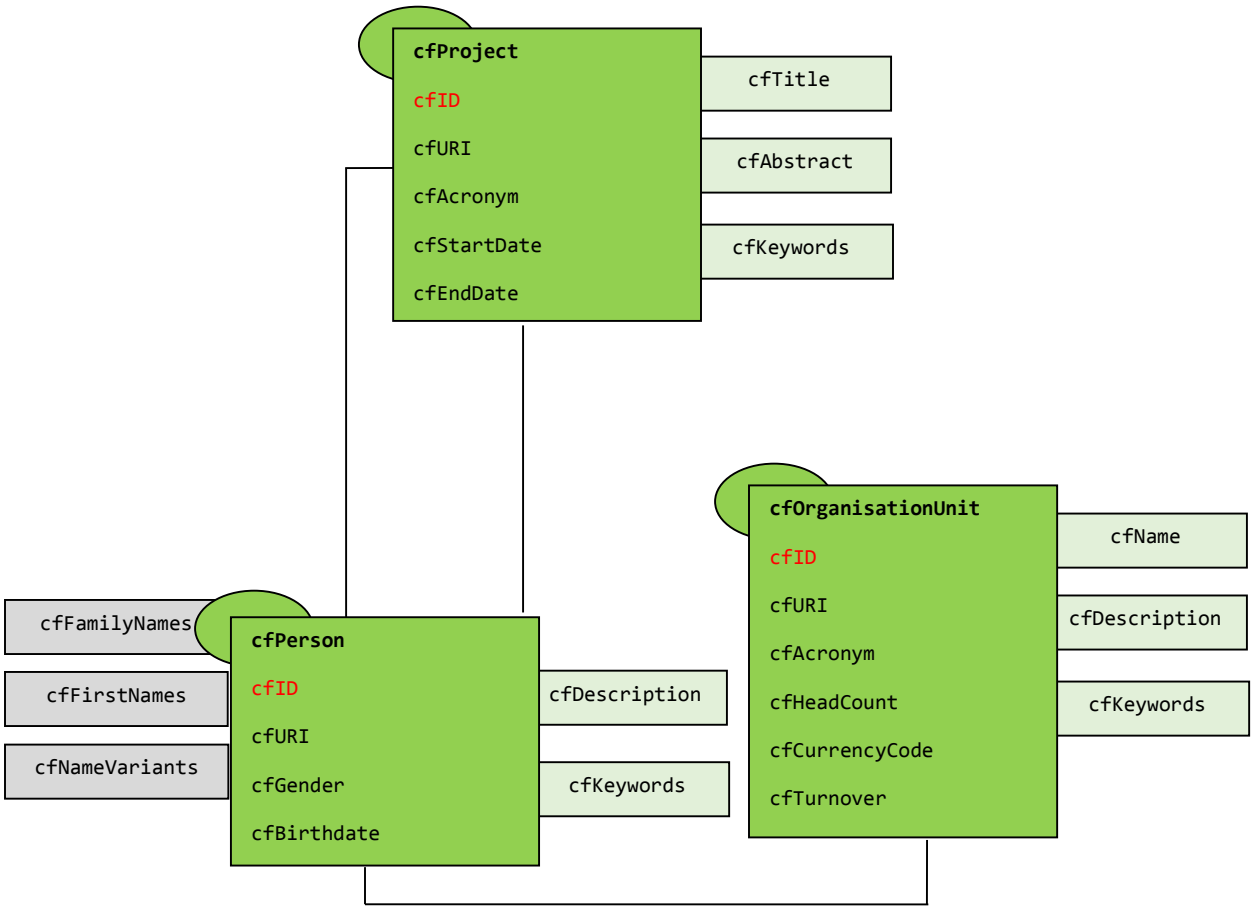


Figure 5 CERIF Base Entities (Jorg, CERIF Tutorial, 2012)

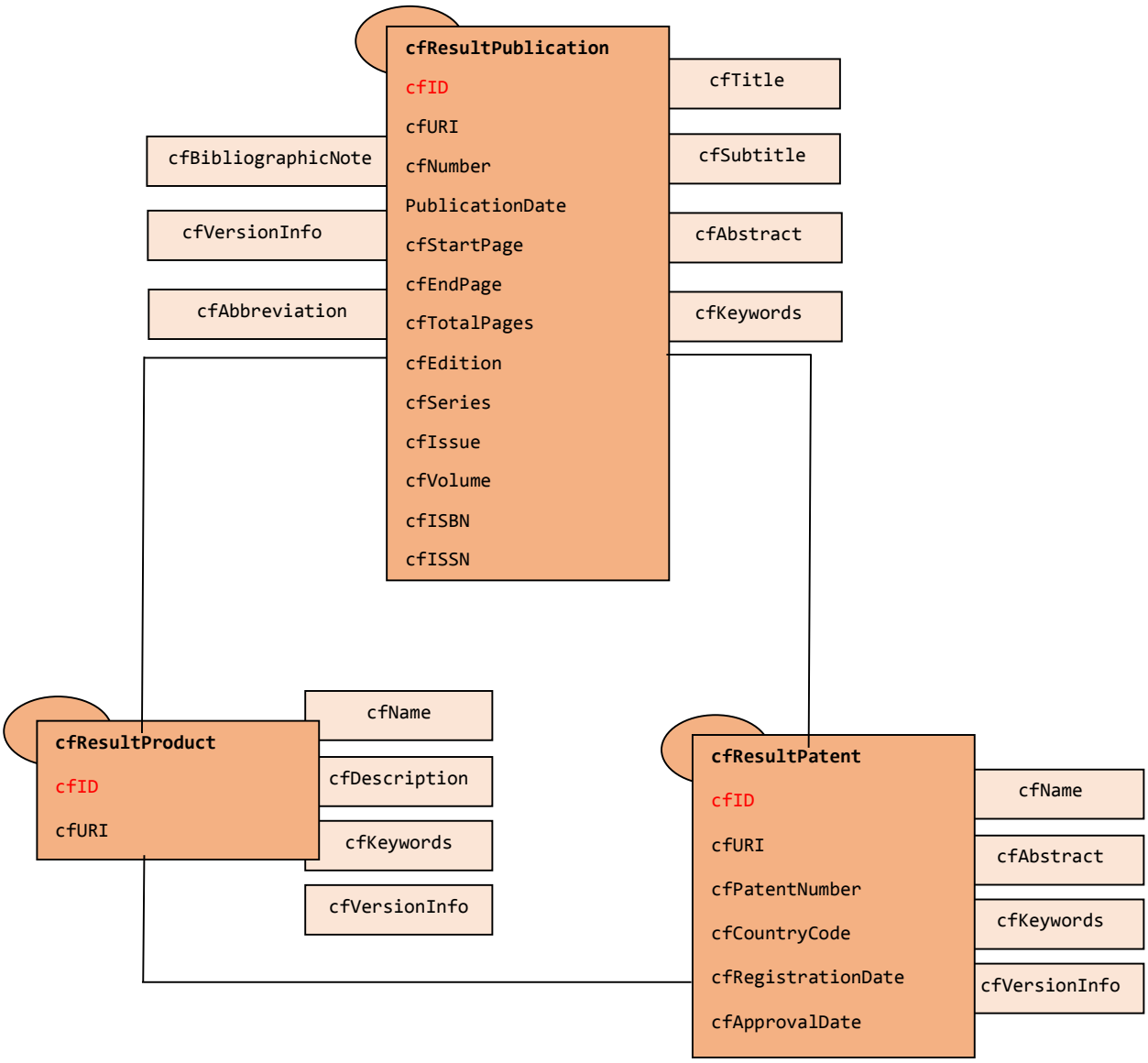


Figure 6 CERIF Result Entities (Jorg, CERIF Tutorial, 2012)

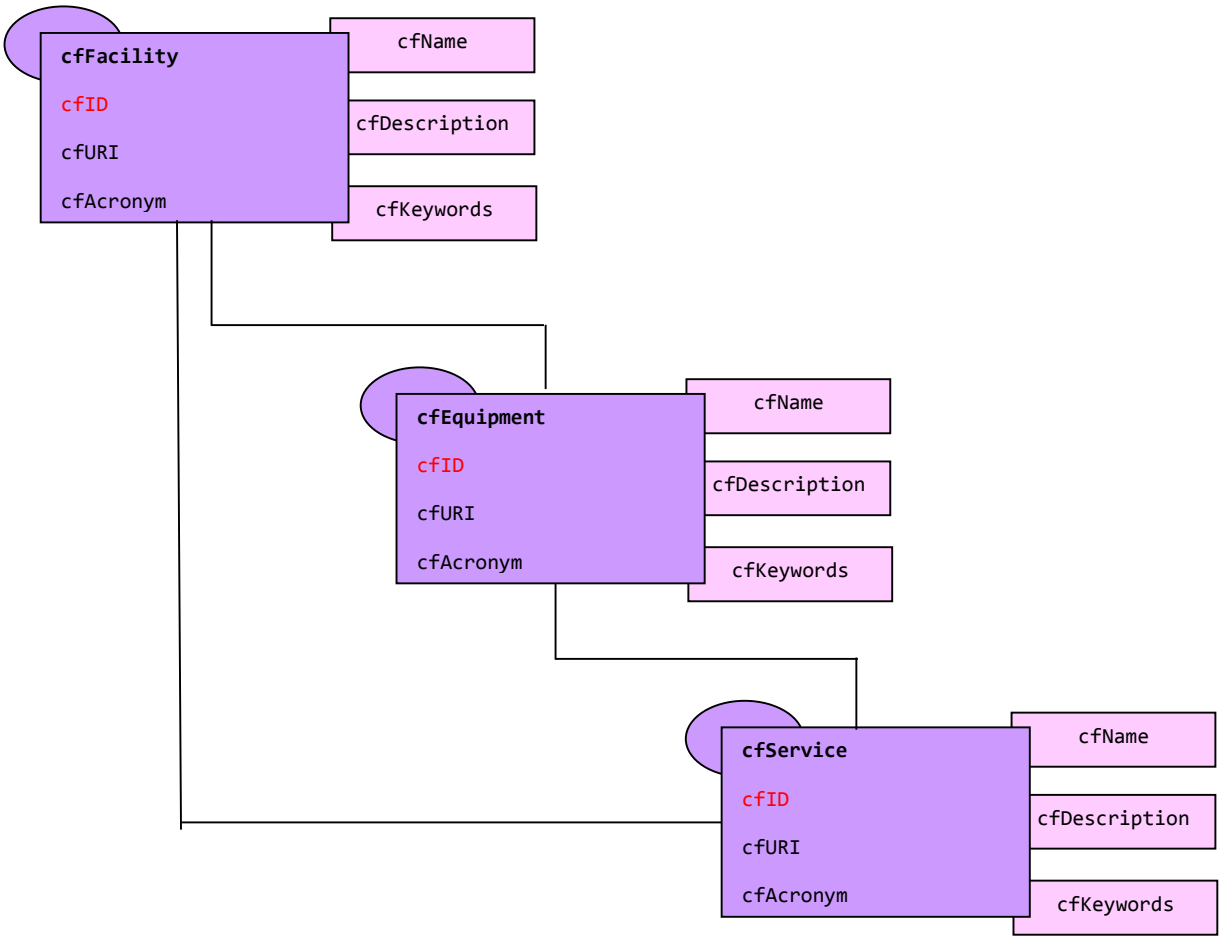


Figure 7 CERIF Infrastructure Entities (Jorg, CERIF Tutorial, 2012)

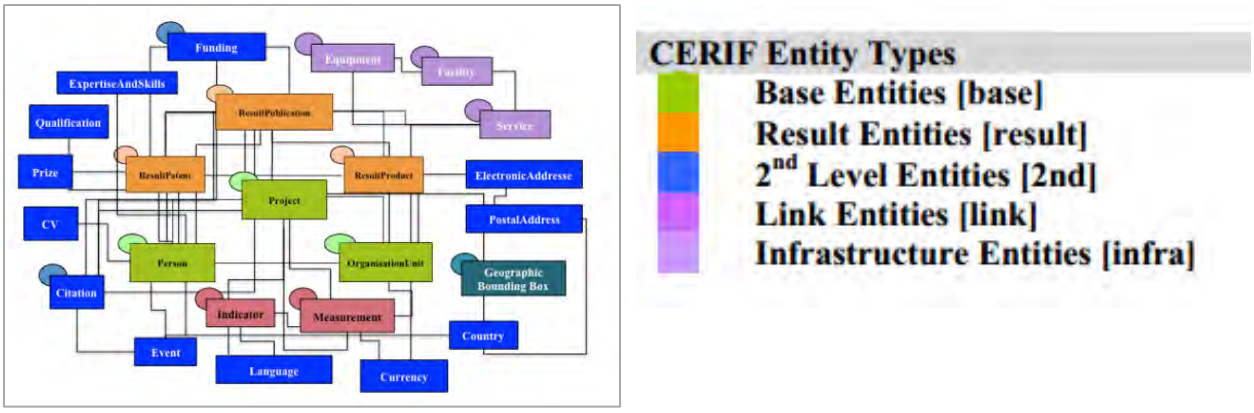


Image 22 Abstract View of CERIF Entities and their relationships (Jorg, Jeffery, Dvorak, Houssos, & al, 2012)

The Semantic Layer

It is a conceptual construct for the description of a sub model of or within CERIF that allows the management of controlled vocabularies, thus being a declared semantics that follows the formal syntax of the CERIF model. Vocabulary terms are applicable from within link entities (Binary and Unary). As mentioned before, a link entity hosts the two identifiers form the two linked entities, allowing a very detailed specification of each term by indicating its source.

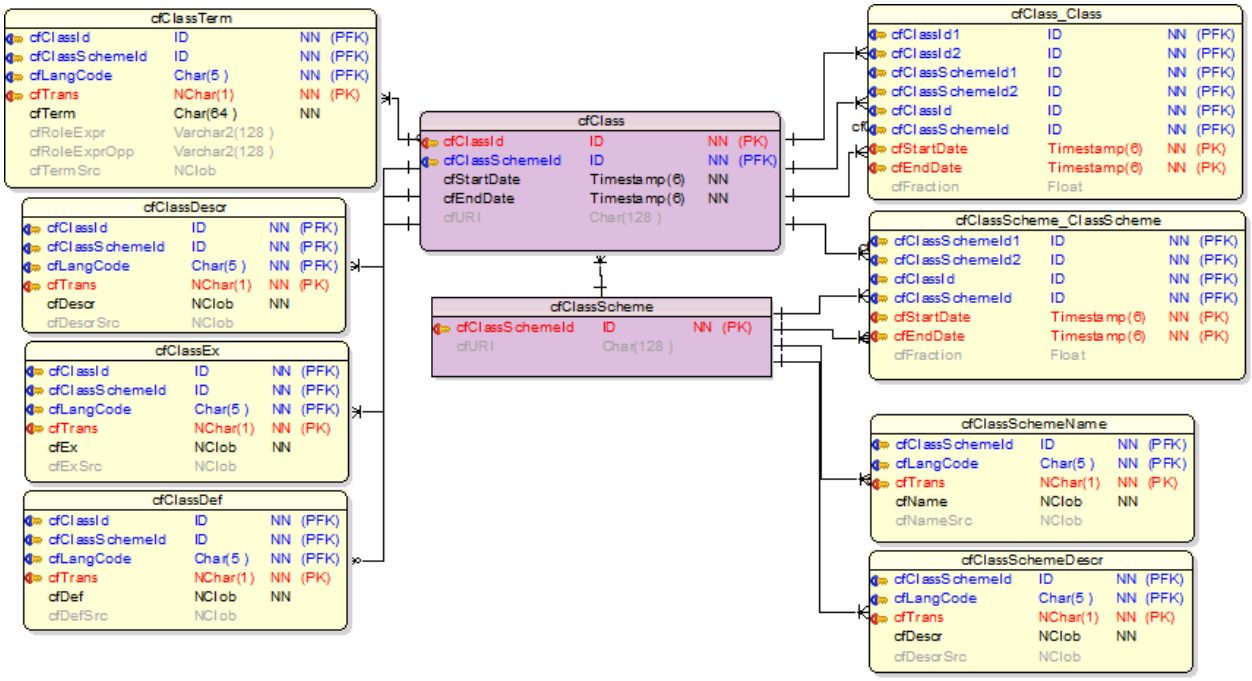


Figure 8 CERIF Semantic Layer [source: www.cerifsupport.org]

The XML

CERIF XML is a structured XML format that follows the CERIF model and aims at enabling consistent data interchange across systems in the RIS domain. It uses XML itself, XML Namespaces and XML Schema. To represent an instance of a CERIF entity, an enclosing element must be used, whose name is the physical name of the entity (e.g. the data of a cfCountry instance is enclosed in the <cfCountry>…</cfCountry> pair of tags). An example of how a CERIF XML person would look follows (courtesy of www.cerifsupport.org).

```
<?xml version="1.0" encoding="UTF-8"?>
<CERIF xmlns="urn:xmlns:org:eurocris:cerif-1.5-1" xsi:schemaLocation="urn:xmlns:org:eurocris:cerif-1.5-1
http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF_1.5_1.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" release="1.5" date="2013-01-07" sourceDatabase="More comprehensive Person Example Record">
<cfPers>
<cfPersId>pers-id0</cfPersId>
<cfGender>f</cfGender>
  <cfPersName_Pers>
    <cfPersNameId>persname-id1</cfPersNameId>
    <cfClassId>55f90543-d631-42eb-8d47-d8d9266cbb26</cfClassId> <!-- presented name -->
    <cfClassSchemeId>7375609d-cfa6-45ce-a803-75de69abe21f</cfClassSchemeId>
    <cfFirstnames>Jörg</cfFirstnames>
    <cfFamilyNames>Brigitte</cfFamilyNames>
```

```
  </cfPersName_Pers>
  <cfPersName_Pers>
    <cfPersNameId>persname-id2</cfPersNameId>
    <cfClassId>64f0eb00-462d-4737-8033-7efac82decf3</cfClassId> <!-- passport name -->
    <cfClassSchemeId>7375609d-cfa6-45ce-a803-75de69abe21f</cfClassSchemeId>
    <cfFirstNames>Joerg</cfFirstNames>
    <cfFamilyNames>Brigitte</cfFamilyNames>
  </cfPersName_Pers>
  <cfKeyw cfLangCode="EN" cfTrans="o">CERIF; CRIS; Information Systems Research Information Management</cfKeyw>
  <cfPers_EAddr>
    <cfEAddrId>b.joerg@ukoln.ac.uk</cfEAddrId>
    <cfClassId>35d43364-2160-4b6c-a487-5019458321e8</cfClassId> <!-- professional email -->
    <cfClassSchemeId>05cc5ff9-bc58-4743-ab59-46e5013e0039<cfClassSchemeId>
  </cfPers_EAddr>
  <cfPers_OrgUnit>
    <cfOrgUnitId>012345678</cfOrgUnitId>
    <cfClassId>ebd55ab0-1cfc-11e1-8bc2-0800200c9a66</cfClassId> <!-- researcher -->
    <cfClassSchemeId>e9616dbd-0d38-4b7d-a6cd-3c4df1e95462</cfClassSchemeId>
    <cfStartDate>2012-06-01T00:00:00</cfStartDate>
  </cfPers_OrgUnit>
  <cfPers_Pers>
    <cfPersId2>pers-id02</cfPersId2>
    <cfClassId>3ccd035b-bc79-477e-aa6c-0bd3606f85c8</cfClassId> <!-- supervisor -->
    <cfClassSchemeId>6b2b7d24-3491-11e1-b86c-0800200c9a66<cfClassSchemeId>
  </cfPers_Pers>
</cfPers>
</CERIF>
```

4.2 Visualization of data

For the front-end development of the web observatory [27] Athanasios Giakas (Giakas, 2013) used multiple technologies and/or software, including Quadrigram (commercial software that allows the creation of interactive visualizations), HTML5/CSS3, JavaScript, jQuery and Java Server Pages.

The web observatory offers the following features:

- **Search for publications**: the user has the ability to access the list of all publications, or he/she can search using a keyword. The keyword can be found either on the title, or abstract, or keywords of the document. The results can also be limited to a time period (Limited Search by Year), or by scientific field (Limited Search by Type), or limited to a minimum number of citations. Publications Search is shown on Image 23.
- **Search for researcher**: the user has the ability to place queries using the name of a researcher, the department, or a field of science, or he/she can chose from a list of names presented. The search for researchers is show on Image 24.

Researcher's Profile: each faculty member has its own profile page that includes information regarding their bibliometric scores (citations, h-index), keywords related to them, a list of their co-investigators, a list of their publications and conferences. An example researcher's profile can be seen on Image 25 Researcher's Profile.

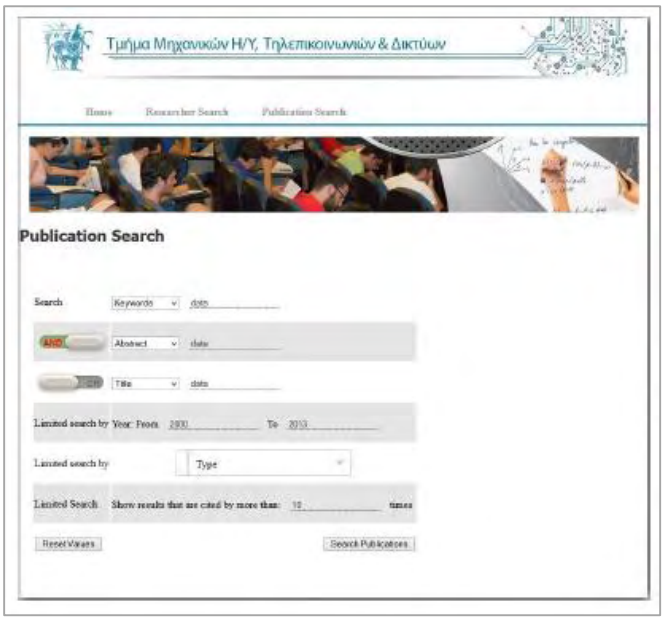


Image 23 Publication Search

Image 24 Researcher Search

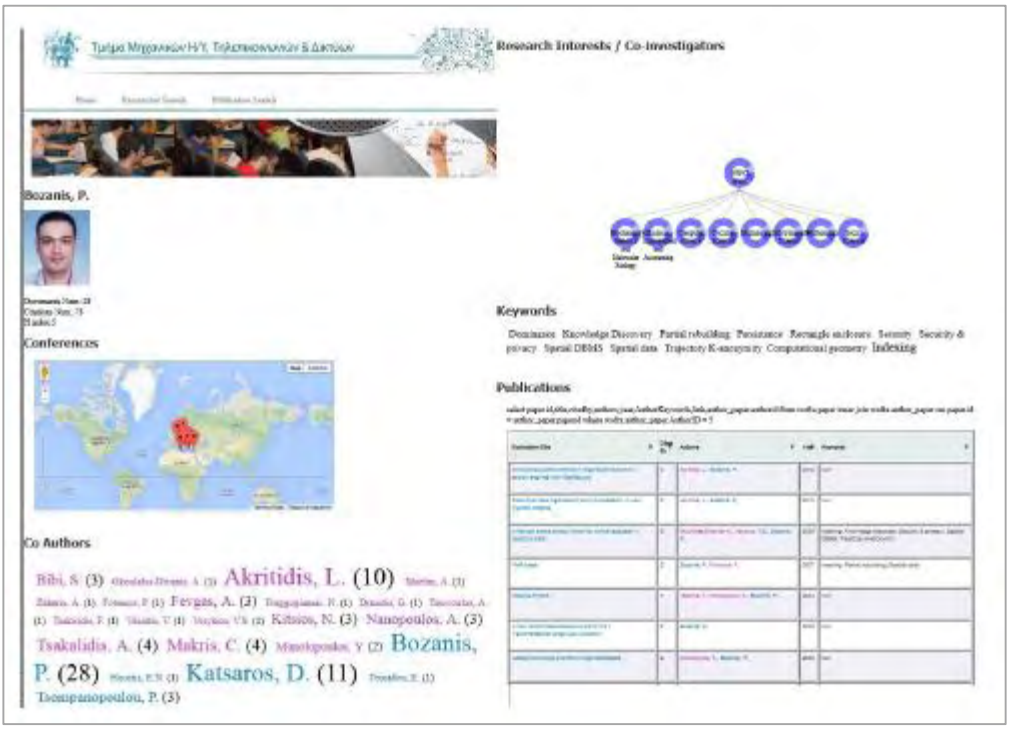


Image 25 Researcher's Profile

5. Test scenario: the Greek case

To elucidate few of the above mentioned issues and at the same time have an initial feeling of the effectiveness of certain practices and tools among the ones mentioned above, we have designed a case study concerning research in Greece. We should make clear here that our objective is not to evaluate the Greek R&D efforts but rather focus on scientometrics issues in general.

The most recent year for which figures are available, concerning the Research & Development (R&D) expenditure in Greece (exact shares of funding), is 2006. For the aforementioned year, gross expenditure accounted for €1.223 billion or 0.6% of the country's Gross Domestic Product (GDP); a percentage much lower than the European average of 1.9%. The two main sources of funding are government with 68% and the European Commission (EC) with 28%. The private sector funds only €3million worth of research. (Grant, Ling, Potoglou, & Culley, 2011)

One major drawback of this situation is the fact that there is poor collaboration between the universities and the industry and between research institutions. Compared to other countries, the Greek R&D system is inward looking and non-collaborative since roughly a third of all Greek research publications are affiliated to a single Greek institute. Having such a dependency on the funding by the EC, the Greek universities/institutes tend to compete each other for the funds instead of collaborating.

However little the R&D expenditure, the indicators of publications and citations reflect the high productivity of the Greek academic system. The country is ranked 26th worldwide, with 160 760 published documents and 28th with 1 589 963 citations for the period 1996 – 2011. During that period, Greece exhibits a remarkable growth rate in its publishing volumes something that ranked the country 8th among the OECD-34 (Organization for Economic Co-operation and Development) countries. In 1996 Greece published 3 729 documents, whereas in 2010 that number escalated at 10 219. This gives a rate of change equal to 2.74 which is above the average rate of change for the EU (1.54) and the OECD countries (1.41). (Sachini, Malliou, & Houssos, 2012)

Based on SCImago Journal & Country Rank, for the period 1996 – 2011, Greece had h-index of 247, 160 760 published documents of which 152 000 were citable, 1 589 963 citations, 289 460 self-citations and 9.89 citations per document.

Table 1 Top cited Greek journals [courtesy of SCImago Journal & Country Rank] shows the top ten most cited Greek journals for 2011. What is interesting here is the fact that even though publications in the field of Medical & Health Sciences represent the second highest share of Greek publications (the highest share of publications is in the field of Natural Sciences), they hold the top ten places in terms of cites.

S/N	Title	Cites (3 years)	H index
1	Anticancer Research	3 791	78
2	International Journal of Oncology	2 777	74
3	Oncology Reports	2 546	50
4	International Journal of Molecular	1 211	56
5	Hepato-Gastroenterology	1 034	55
6	In vivo	559	34
7	Journal of Musculoskeletal Neuronal Interactions	290	28
8	Hormones	276	17
9	Journal of B.U.O.N	256	10
10	Hellenic Journal of Cardiology	240	13

Table 1 Top cited Greek journals [courtesy of SCImago Journal & Country Rank]

5.1 Comparison with other countries

Latvia, Malta and Poland are three European countries that also spend around 0.6% of their GDP on R&D for the time period 2000 – 2010 based on the results from Eurostat. Figure 9 Number of published documents and cites for the four countries [courtesy of SCImago Journal & Country Rank] demonstrates the results of the four countries regarding number of published documents and cites. Similar results are found in the h-index scores. Poland holds the 1st place with an h-index of 275, Greece comes 2nd with 266 and, Latvia and Malta at 75 and 60, respectively.

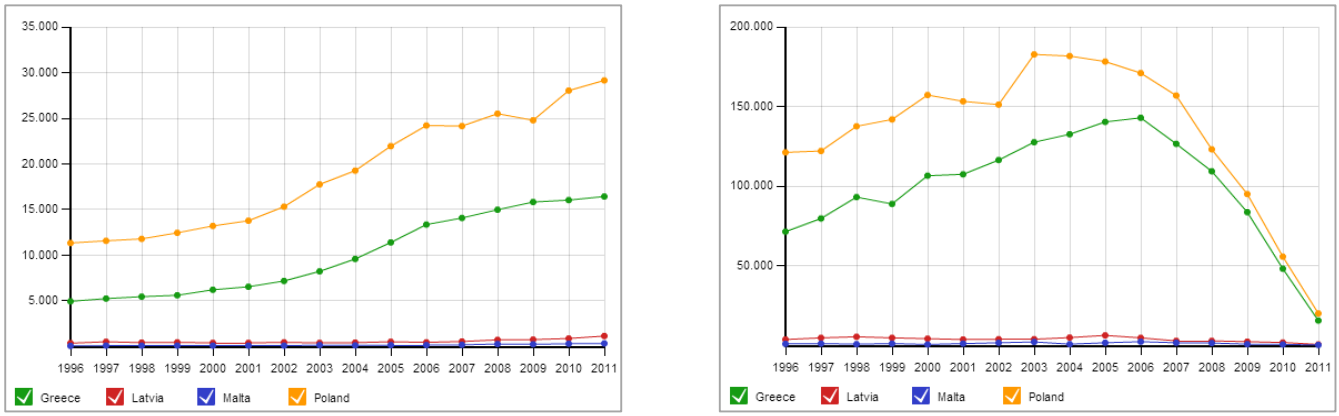
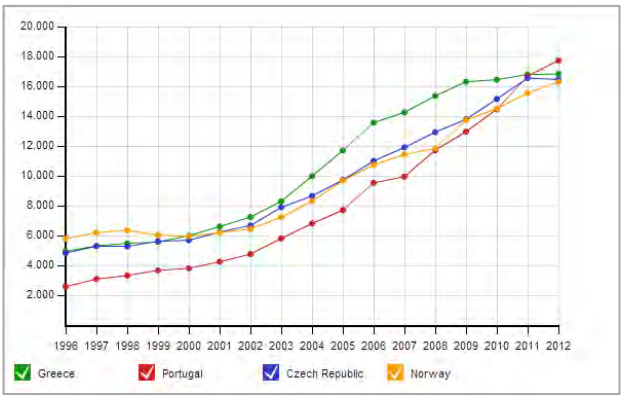


Figure 9 Number of published documents and cites for the four countries [courtesy of SCImago Journal & Country Rank]

On the other hand, Portugal, the Czech Republic, and Norway, are three European countries that spend a larger share of their GDP on R&D than Greece. To be more accurate, Portugal spends on average 1.03%, the Czech Republic spends on average 1.33%, and Norway spends on average 1.63%. Comparing these shares to Greece's expenditure, one would expect a big difference in terms of published documents, cites and h-index. However, that is not the case. In the time period 2000 – 2011 Greece shows the highest number of publications among them, while Norway shows the highest number of cites. Norway holds also the 1st place, in terms of h-index, with an h-index of 327, Greece holds the 2nd place with 266, followed by the Czech Republic with 239, and Portugal with 234. The number of published documents and cites for each country is shown in Figure 10.



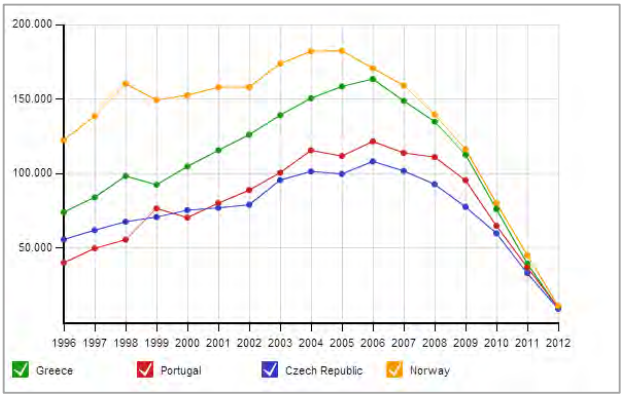


Figure 10 Number of published documents and cites for the four countries [courtesy of SCImago Journal & Country Rank]

5.2 Greek Universities

As mentioned before, Universities account for the majority of scientific output in Greece. In (Sachini, Malliou, & Houssos, 2012) the bibliometric indicators for the publications produced by national Universities were examined. Results concern 21 out of the 24 Greek Universities, due to the small amount of publications of the remaining 3. Table 2 shows the top ten Universities in terms of **publications** and **citations** for the periods 2004 – 2008 and 2006 – 2010.

S/N	University (full name)	2004 – 2008		2006 – 2010	
		Number of publications	Number of citations	Number of publications	Number of citations
1	National & Kapodistrian University of Athens	9 835	43 695	11 308	58 799
2	Aristotle University of Thessaloniki	7 602	25 504	8 577	32 469
3	University of Patras	4 536	15 642	4 989	22 093
4	National Technical University of Athens	4 058	9 715	4 261	13 036
5	University of Crete	3 458	18 006	3 854	23 755
6	University of Ioannina	3 262	16 642	3 481	20 705
7	University of Thessaly	1 738	5 591	2 136	8 443
8	Democritus University of Thrace	1 478	3 664	1 827	5 931
9	Agricultural University of Athens	995	3 001	1 140	4 182
10	University of the Aegean	839	2 024	959	2 340

Table 2 Number of publications and citations of Greek Universities [courtesy of EKT]

As one would expect, the National & Kapodistrian University and the Aristotle University are ranked 1st and 2nd, respectively, since they are two of the oldest and largest Institutions in Greece. However, the University of Crete is ranked 5th; a great accomplishment for a University that has been functioning for less than 40 years. According to QS World University Rankings the latter holds the highest score in terms of citations per faculty (51.0),

followed by the National & Kapodistrian University (42.6) and the Aristotle University (24.30). The University of Crete is also ranked 53 among the top 100 Universities worldwide that are under 50 years old.

In 2010 the greatest number of publications was attributed to the National and Kapodistrian University of Athens with 2,368 and the Aristotle University of Thessaloniki with 1,734. Their publications hold a share of 29.1% within National Universities. The share of each University is shown in Figure 11.

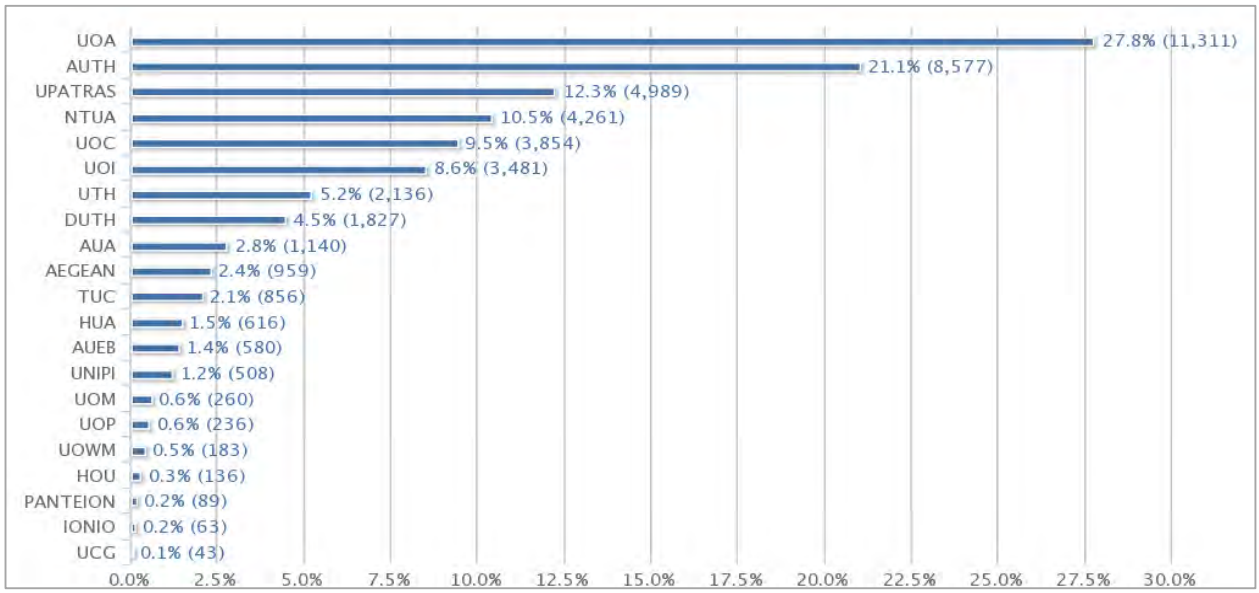


Figure 11 Share of publications [courtesy of EKT]

EKT uses six major **fields of science** to calculate the productivity of each university (publications and citation score) for each of the fields. These fields are "Natural Sciences", "Engineering and Technology", "Medical and Health Studies", "Agricultural Sciences", "Social Sciences" and Humanities.

In the field of "*Natural Sciences*" an increasing trend in publications among 18 of 21 universities was shown. The highest normalized citation score was achieved by the Harokopio University of Athens equal to 1.37, whereas, the University of Athens achieved the highest number of citations. The University of Crete, the Technical University of Crete and the University of Ioannina demonstrated citation scores that exceeded the world average.

The University of Crete, the Agricultural University of Athens, the Technical University of Crete, the University of Piraeus, the University of Western Macedonia and the University of Ioannina had publications with citation scores higher than the world average, in the field of "*Engineering and Technology*". What is interesting here, if the fact that even though the National Technical University of Athens achieved the highest number of publications, its field normalized citation score was below the world average.

In the field of "*Medical and Health Sciences*", 10 universities hold the majority of publications, with highest citation scores those of the Agricultural University of Athens, the University of Ioannina, the Harokopio University of Athens and the University of Crete. The University of Athens achieved the highest number of publications and citations, 7,271 and 40,377, respectively.

In "*Agricultural Sciences*", publications emerged mainly from 9 universities. The highest citation scores were reached by a small number of publications, which were achieved by the National Technical University of Athens, the University of Patras and the University of Ioannina. The citation score of the Agricultural University of

Athens was just above the world baseline with the second highest number of citations (934), following the Aristotle University of Thessaloniki (1,752).

Even though in the field of "*Social Sciences*" there were a larger number of active Universities (15), the publication output was much lower in comparison with other scientific fields. However, there was an increment in the citation scores and the University of Piraeus, the Demokritos University of Thrace, the Technical University of Crete, the University of Ioannina and the National Technical University of Athens, reached above world average regarding their publications.

Last, in "*Humanities*" only 3 universities were found to produce publications, of which the highest citation score (0.82) was achieved by the Aristotle University of Thessaloniki that also achieved the highest number of citations (219).

Even though the two pillars of the Greek Higher Education System continue to increase their research output, some of the smaller Universities show higher growth rates in terms of number of publications and citations. What is interesting is the fact that the Aristotle University of Thessaloniki, the University of Patras and the National Technical University of Athens (among other Universities) demonstrate quite an inward looking character in terms of share of publications with no collaboration; national or international. 44% of the National Technical University's publications were produced without partners, followed by the Aristotle University with 36.6% and the University of Patras with 35.6%. The University of Crete accounted for the highest share of publications produced with international collaboration (51%), whereas, the University of Central Greece accounted for the highest share of publication produced with national collaboration (83.7%).

5.3 Research at the University of Thessaly

Table 2 (Paragraph 5.2) shows that the University of Thessaly is ranked 7th amongst the National Universities. The University shows a growth rate of 22.89% in publications and a growth rate of 51% in citations. The growth rates of the University are the second largest among the top ten Greek Universities; first is the Demokritos University of Thrace (23.61% in publications and 61.87% in citations).

Figure 12 demonstrates the number of publications and rate of change in the number of publications, and the number of citations for the University of Thessaly for the period 1996 – 2010. The growth trend in the number of publications reached a peak in 2009 with 466, to be reduced to 452 in 2010 probably due to the economic recession. At this point it is worth mentioning that for the period 2006 – 2010 the University produced 15 publications which were ranked among the 1% most highly cited publications worldwide.

John Ioannidis, Evangelia Ntzani, Ioanna Tzoulaki, Athina Tatsioni, Evangelos Evangelou, Georgios and Konstantinos Siontis and Orestis Panagiotou, curate two databases [28] of highly cited Greek scientists. Their aim was to capture scientists with Greek names who have high impact in scientific literature. The first database is limited to biosciences, whereas the second one covers all scientific fields. Dr. Leandros Tassiulas from the Department of Electrical and Computer Engineering, the most cited professor at the University, is also in the top 50 most cited Greek scientists. He is placed 39th, having 12,000 citations as of January 2012.

The results from (Sachini, Malliou, & Houssos, 2012; Borner, Conlon, Corson-Rikert, & Ding, 2012) showed that the highest number of publications produced from the University was in the field of Medical and Health Sciences (1,071), followed by publications in the field of Natural Sciences (776). However, proportionally, publications in the latter field were cited more times than in the field of Medical and Health Sciences. Figure 12 demonstrates the distribution of publications and citations across the aforementioned fields of science (Paragraph 5.2) for the period 2006 – 2010.

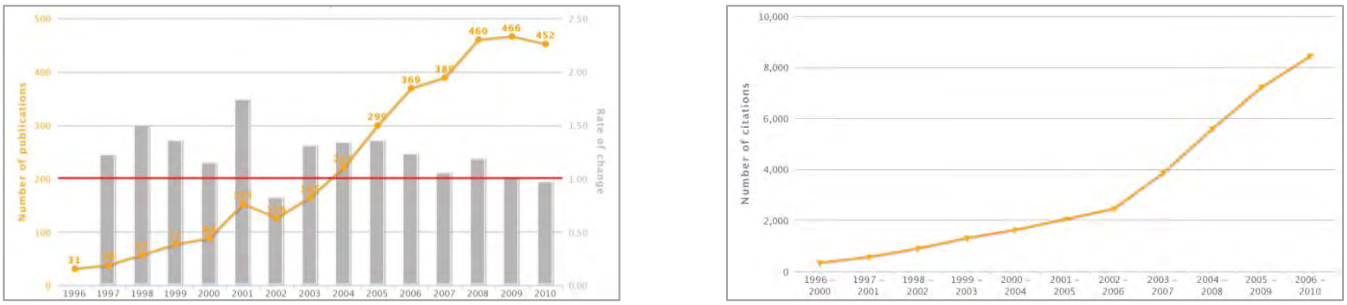


Figure 12 Number of publications and rate of change in their number, and number of citations [courtesy of EKT]

The University was active in five out of six major fields of science presented in Paragraph 5.2. The highest number of publications and citations was achieved in the field of "*Medical and Health Sciences"* where the University produced 1,071 documents with 4,949 citations. However, the highest field normalized citation score (0.93) was achieved in the field of "*Agricultural Sciences"*. In the aforementioned field the University produced 238 publications with 690 citations, ranking the University 3rd after the Aristotle University of Thessaloniki and the Agricultural University of Athens. In the field of "*Natural Sciences*", there were 3,400 citations among 776 publications and a citation score of 0.79 and in the field of "*Engineering and Technology*" the production of 411 publications accrued 1,382 citations and the second highest citation score for the University of 0.87. Finally, in the field of "*Social Sciences*" the University had the smallest number of publications and citations, 135 and 255 respectively, with a citation score of 0.80.

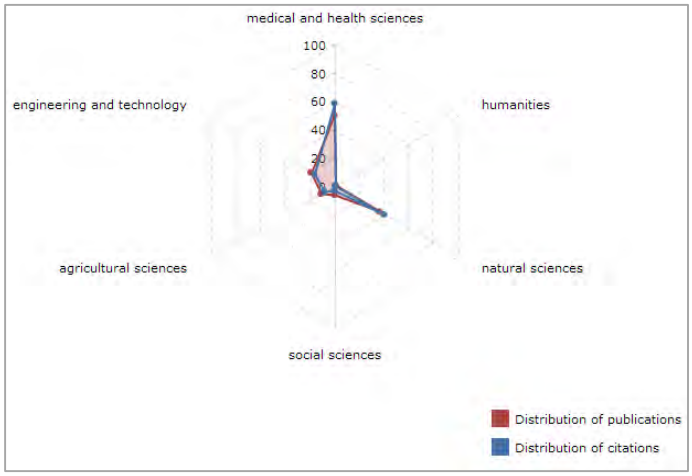


Figure 13 Distribution of publications and citations for the University of Thessaly [courtesy of EKT]

6. Conclusions and future work

In this thesis we present how the need for communicating knowledge and the importance of international collaboration have inspired people to develop systems that are able to do so. We presented in detail a commercial solution, Pure, and an open source application, VIVO, as well as some independent efforts made by various Universities across the globe. Moreover, we proposed a model for the development of a web observatory for research activities of the faculty of the University of Thessaly. The model includes technologies to be used for the aggregation, organization and dissemination of data, and also how the data and the associated information and knowledge could be visualized. Finally, we present some preliminary but indicative information regarding the research output of Greece and the research output of the University of Thessaly.

The proposed model accompanied by the data visualizations developed in (Giakas, 2013) will be used for the implementation of the web observatory for the University of Thessaly. This project is a part of the University's effort to promote the research activities of the faculty and to show their academic course over the years. It is interesting to point out here that there are several other efforts that fall in the general thematic area of the present thesis. We could mention here a preliminary study carried out by our University's administration and recently announced to our community. Although no details are so far available to us it seems that is has been carried out in a rather outdated and hardwired manner that does not follow the common modern practices. It has been heavily criticized by several members of our University's academic community and beyond. On the other hand, we should also mention another recent effort carried out at low system programmer level. Its objective was not really to collect scientometric data, but rather to expose the author (Kallis, 2013) to a very interesting, from the system and web programming view problem, application that is also easy to understand and even easier to justify its importance.

It is worth to mention here that the economic crisis in Greece has led to the reduction of funding of National Universities. Taking into account that approximately 80% of the national research and innovation comes from Universities one can easily imagine how this percentage is affected from such reductions. However, if Universities collect and share data as proof of their productivity and significance of their research output, they will be able to push political decision-makers to end further reductions. Even if such scenario fails, these data can be used as a mean for many Universities to explore alternative funding sources, such us the private sector.

The observatory, as briefly described above, may be considered as our team's first effort and may need additional modifications in order to reach the desirable level of success. It can be integrated using W3C Web Technologies standards to become over the years a network that covers all the national universities. However, this endeavor is beyond our means and time limitations, and would of course require the participation of public bodies or other institutions.

References

Borner, K., Conlon, M., Corson-Rikert, J., & Ding, Y. (2012). VIVO: A semantic approach to scholarly networking and discovery.

Garfield, E. (1955). Citation indexes for science.

Garfield, E. (1994). Introducing the impact factor.

Georghiou, L. (1998). Global cooperation in research. *Res. Policy, 27*, 611–626.

Giakas, A. (2013). Web based interfaces for inhomogeneous datasets in real time.

Grant, J., Ling, T., Potoglou, D., & Culley, D. M. (2011). *A rapid review of the Greek research and development system.* RAND Corporation.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output.

Hulme, E. W. (1923). *Statistical bibliography in relation to the growth of modern civilization.* London.

Jorg, B. (2012, June 8). CERIF Tutorial. Prague.

Jorg, B., Jeffery, K., Dvorak, J., Houssos, N., & al, e. (2012, January 28). CERIF 1.3 Full Data Model (FDM) Introduction and Specification. euroCRIS.

Kallis, P. (2013). Academic Publications Tracker.

OECD. (2002). *The Frascati Manual.*

OECD. (2007, February 26). Revised Field of Science and Technology (FOS) Classification in the Frascati Maunal.

Pan, R. K., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: uncovering the role of geography in science. *Nature*.

Price, D. J. (1963). *Little Science, Big Science.* Columbia University Press.

Price, D. J. (1965). Networks of Scientific Papers. *SCIENCE*, 510-515.

Price, D. J. (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, 292-306.

Pritchard, A. (1969). Statistical Bibliography of Bibliometrics?

Sachini, E., Malliou, N., & Houssos, N. (2012). *Greek Scientific Publications 1996-2010: A Bibliometric Analysis of Greek Publications in International Scientific Journals.* National Documentation Centre.

Web References

1. http://www.lib.cam.ac.uk/sandars/duty.html
2. http://thomsonreuters.com/current-contents-connect/
3. http://thomsonreuters.com/web-of-science/
4. http://thomsonreuters.com/journal-citation-reports/
5. http://thomsonreuters.com/index-chemicus/
6. http://www.eurocris.org/Index.php?page=concepts_benefits&t=1
7. http://incites.isiknowledge.com/
8. http://ip-science.thomsonreuters.com/globalprofilesproject/
9. http://www.tubitak.gov.tr/
10. http://www.tubitak.gov.tr/sites/default/files/gyue_2013_gosterge_seti.pdf
11. http://www.tubitak.gov.tr/sites/default/files/gyue_2013_bilgi_notu.pdf
12. http://www.symplectic.co.uk/
13. http://hub.hku.hk/
14. http://t2r2.star.titech.ac.jp/index_en.html
15. http://search.star.titech.ac.jp/titech-ss/search.act
16. http://www.wageningenur.nl/en/
17. http://reports.metrics.ekt.gr/en/report02/
18. http://www.uspto.gov/
19. http://www.epo.org/
20. http://inf.modip.uth.gr/
21. http://www.lib.uth.gr/LWS/el/el_hp.asp
22. http://phdtheses.ekt.gr/eadd/opendata
23. http://cordis.europa.eu/newsearch/index.cfm?page=resultListGET&formid=form_proj&exactphrase=University%20of%20Thessaly&typeEvent=PROJECTS_SEARCH&useraction=advanced_search&controlsession=false&js=1
24. http://ee.uth.gr/index.php?option=com_browse&view=browse&Itemid=8
25. SCImago. (2007). SJR — SCImago Journal & Country Rank. Retrieved May 14, 2013, from http://www.scimagojr.com
26. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php?title=File:Gross_domestic_expenditure_on_R%26D,_2000-2010_(%25_share_of_GDP).png&filetimestamp=20121016060906
27. http://83.212.108.14/research/allauthorsControl
28. http://www.highlycitedgreekscientists.org/
29. http://en.wikipedia.org/wiki/Current_Research_Information_System
30. http://en.wikipedia.org/wiki/EuroCRIS
31. http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1
32. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/MImage.html
33. http://en.wikipedia.org/wiki/CiteSeer#CiteSeerX
34. http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=S27CLKc@kAMim1K1KF7&preferencesSaved=&highlighted_tab=WOS
35. http://www.scopus.com/home.url
36. http://vbn.aau.dk/en/
37. http://www.atira.dk/en/pure/
38. http://researchanalytics.thomsonreuters.com/incites/
39. http://www.cerifsupport.org/cerif-in-brief/
40. http://vivoweb.org/

41. https://wiki.duraspace.org/display/VIVO/VIVO+Main+Page