



**University of Thessaly**  
School of Engineering  
Department of Electrical and  
Computer Engineering

**“Evaluating scientific research”**

**Diploma Thesis Project**

**Elvis Papa**

Supervisors:

Manolis Vavalis  
Professor  
University of Thessaly

Spyros Lalis  
Assistant Professor  
University of Thessaly

Volos, Greece  
October 2014

Blank page



**Πανεπιστήμιο Θεσσαλίας**  
**Πολυτεχνική Σχολή**  
**Τμήμα Ηλεκτρολόγων Μηχανικών**  
**και Μηχανικών Υπολογιστών**

**“Αξιολόγηση ερευνητικού έργου”**

**Διπλωματική Εργασία**

**Έλβις Πάπα**

Επιβλέποντες Καθηγητές:

Εμμανουήλ Βάβαλης  
Καθηγητής  
Πανεπιστήμιο Θεσσαλίας

Σπύρος Λάλης  
Αναπληρωτής Καθηγητής  
Πανεπιστήμιο Θεσσαλίας

Εγκρίθηκε από τη διμελή εξεταστική επιτροπή την .....

.....  
Εμμανουήλ Βάβαλης

.....  
Σπύρος Λάλης

Διπλωματική εργασία για την απόκτηση του Διπλώματος του Μηχανικού Ηλεκτρονικών Υπολογιστών, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας, στα πλαίσια του Προγράμματος Προπτυχιακών Σπουδών του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Πανεπιστημίου Θεσσαλίας.

.....

Έλβις Πάπα

Διπλωματούχος Μηχανικός Ηλεκτρονικών Υπολογιστών, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας

Copyright © Elvis Papa, 2014

Με επιφύλαξη παντός δικαιώματος, All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό.

Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα.

Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς το συγγραφέα

# Acknowledgements

*I would like to thank my supervisors professors Manolis Vavalis and Spyros Lalis. Especially, I would like to thank professor Manolis Vavalis, who believed in me and in my abilities. Our perfect collaboration, his advice during all this period and his assistance on each step of the implementation, have contributed to the accomplishment of this thesis project. Also, I would like to thank the former student Paulos Kallis for his help at some stages in the development of the project. Finally, I would like to thank my family for supporting me all these years and all my friends for all the happy moments we had together.*

*Elvis Papa  
Volos  
October 2014*

## Abstract

The objective of this diploma thesis project is the development and design of a scientific information system for the management and monitoring of the research activities of the University of Thessaly. Our system purpose is to provide capabilities of searching and retrieving qualitative and quantitative information of research activities, which information is produced and collected automatically through web crawling techniques in real time. Each department after installing the software package will be able to categorize and manage its continuous flow of raw data in order to:

- produce information that leads to specific qualitative and quantitative evaluation metrics
- present both the raw data and the added valuable data and information in an intelligent and effective way, adapted to the background and the interests of the viewer using data visualization techniques.

## Περίληψη

Ο στόχος αυτής της διπλωματικής εργασίας είναι η σχεδίαση και υλοποίηση ενός επιστημονικού συστήματος πληροφοριών για τη διαχείριση και παρακολούθηση των ερευνητικών δραστηριοτήτων του Πανεπιστημίου Θεσσαλίας. Σκοπός του συστήματός μας είναι να παρέχει δυνατότητες αναζήτησης και ανάκτησης ποιοτικών και ποσοτικών πληροφοριών ερευνητικής δραστηριότητας, οι οποίες πληροφορίες παράγονται και συλλέγονται αυτοματοποιημένα με τεχνικές web crawling σε πραγματικό χρόνο. Κάθε τμήμα μετά την εγκατάσταση του λογισμικού θα είναι σε θέση να ταξινομήσει και να διαχειριστεί τη συνεχή ροή ακατέργαστων δεδομένων του με στόχο να :

- παράγει πληροφορίες οι οποίες θα οδηγήσουν σε συγκεκριμένες ποιοτικές και ποσοτικές μετρήσεις αξιολόγησης της ερευνητικής δραστηριότητάς του
- παρουσιάσει τα ανεπεξέργαστα δεδομένα και τα προστιθέμενα με αξία δεδομένα και τις πληροφορίες τους, με ένα έξυπνο και αποτελεσματικό τρόπο προσαρμοσμένο στο παρασκήνιο και τα ενδιαφερόμενα του παρατηρητή, χρησιμοποιώντας τεχνικές οπτικοποίησης δεδομένων.

# Table of Contents

|       |  |    |
|-------|--|----|
| 1     | Introduction .....   | 12 |
| 1.1   | Introduction to Information Systems for scientific research .....      | 12 |
| 1.2   | About this diploma thesis project .....                                | 12 |
| 1.3   | Content's organization .....   | 13 |
| 2     | Data Visualization .....   | 14 |
| 2.1   | The World Wide Web and its big data .....                              | 14 |
| 2.2   | What is data visualization .....                                       | 15 |
| 2.3   | Data visualization patterns .....                                      | 19 |
| 2.3.1 | Independent quantities .....   | 20 |
| 2.3.2 | Continuous quantities .....  | 21 |
| 2.3.3 | Proportions .....  | 21 |
| 2.3.4 | Correlations .....   | 22 |
| 2.3.5 | Networks .....   | 23 |
| 2.3.6 | Hierarchies .....  | 24 |
| 2.3.7 | Cartographic .....   | 24 |
| 2.4   | Data visualization tools .....   | 24 |
| 2.4.1 | SVG, HTML 5 Canvas .....   | 24 |
| 2.4.2 | JavaScript libraries .....   | 25 |
| 2.4.3 | Programming languages .....  | 27 |
| 2.4.4 | Software platforms .....   | 27 |
| 3     | Web Crawling .....   | 28 |
| 3.1   | Introduction .....   | 28 |
| 3.2   | What is a web crawler and how it works .....                           | 30 |
| 3.3   | Implementing an efficient web crawler and best strategies .....        | 31 |
| 4     | Scientific research .....  | 35 |
| 4.1   | Introduction .....   | 35 |
| 4.2   | Big repositories and information systems for scientific research ..... | 35 |
| 4.3   | Bibliometric Analysis of research activity .....                       | 36 |
| 4.4   | ORCID .....  | 38 |
| 4.4.1 | What is ORCID .....  | 38 |



|       |  |    |
|-------|--|----|
| 4.4.2 | What are the benefits of ORCID.....                                | 39 |
| 4.4.3 | ORCID & PlumX .....  | 40 |
| 4.5   | Altmetric .....  | 42 |
| 4.6   | HUBzero .....  | 44 |
| 5     | Implementation .....   | 47 |
| 5.1   | Scientific information systems for research around the world ..... | 47 |
| 5.2   | The objective of the project.....                                  | 51 |
| 5.3   | The architecture of the system .....                               | 52 |
| 5.3.1 | Start Crawling.....  | 54 |
| 5.3.2 | Crawler Manager .....  | 55 |
| 5.3.3 | Our Crawlers .....   | 55 |
| 5.3.4 | Database Processing .....  | 57 |
| 5.3.5 | Request Handler.....   | 57 |
| 5.3.6 | Visualization Handler .....  | 59 |
| 5.4   | The databases of the system .....                                  | 59 |
| 5.4.1 | The database of the University's central page.....                 | 59 |
| 5.4.2 | The database of a department .....                                 | 61 |
| 5.5   | Functionalities.....   | 63 |
| 5.6   | User Interface .....   | 63 |
| 5.7   | Admin interface .....  | 78 |
| 6     | Conclusion and future work.....                                    | 80 |
| 6.1   | Conclusion.....  | 80 |
| 6.2   | Future work.....   | 80 |
| 7     | Tools and technologies we used .....                               | 81 |
| 8     | References .....   | 84 |

# Table of figures

|  |    |
|--|----|
| Figure 1 – An infographic about the four v’s of Big Data (source: <a href="http://www.ibmbigdatahub.com/infographic/four-vs-big-data">http://www.ibmbigdatahub.com/infographic/four-vs-big-data</a> ) .....  | 15 |
| Figure 2 - The process from data to wisdom in the theory of data visualization (source: <a href="http://harlotofthearts.org/blog/2011/01/08/data-information-knowledge-wisdom/">http://harlotofthearts.org/blog/2011/01/08/data-information-knowledge-wisdom/</a> ) .....  | 16 |
| Figure 3 - An example of knowledge acquisition (source : <a href="https://visualisingadvocacy.org/blog/disinformation-visualization-how-lie-datavis">https://visualisingadvocacy.org/blog/disinformation-visualization-how-lie-datavis</a> ).....  | 19 |
| Figure 4 - How a website looks like from the user’s perspective .....  | 29 |
| Figure 5 - How a website looks like from the crawler’s perspective .....   | 30 |
| Figure 6 - How a web crawler works (source: <a href="http://blog.datafiniti.net/?p=280">http://blog.datafiniti.net/?p=280</a> ) .....  | 31 |
| Figure 7 - A pseudo code of a web crawler (source : <a href="http://www.devbistro.com/articles/Misc/Implementing-Effective-Web-Crawler">http://www.devbistro.com/articles/Misc/Implementing-Effective-Web-Crawler</a> ) .....  | 34 |
| Figure 8 - The ORCID as an identifier hub (source: <a href="http://informatics.mit.edu/blog/2014/07/integrating-researcher-identifiers-funding-identifiers-and-institutional">http://informatics.mit.edu/blog/2014/07/integrating-researcher-identifiers-funding-identifiers-and-institutional</a> ) .....                         | 39 |
| Figure 9 - The categories of metrics and types of research output that of PlumX’s research tracking (source: <a href="http://blog.plumanalytics.com/post/97218263655/orcid-and-plum-analytics-work-together-to-easily">http://blog.plumanalytics.com/post/97218263655/orcid-and-plum-analytics-work-together-to-easily</a> ) ..... | 41 |
| Figure 10 - All the services that PlumX provides (source <a href="http://www.plumanalytics.com/downloads/ORCID_Poster_Final.pdf">http://www.plumanalytics.com/downloads/ORCID_Poster_Final.pdf</a> ).....  | 41 |
| Figure 11 - The Altmetric Explorer panel (source: <a href="http://www.diglib.org/archives/5132/">http://www.diglib.org/archives/5132/</a> ) .....  | 42 |
| Figure 12 - The page details of a badge (source: <a href="http://www.altmetric.com/details.php?citation_id=571540">http://www.altmetric.com/details.php?citation_id=571540</a> ) .....   | 44 |
| Figure 13 - A community center for developing and sharing knowledge and tools for environmental systems analysis, build with HUBzero (source: <a href="http://iemhub.org/">http://iemhub.org/</a> ) .....  | 46 |
| Figure 14 - How the University of Pittsburgh monitor its research activites (source: <a href="https://plu.mx/pitt/g/?artifact_tab=all_artifacts&amp;sort=year&amp;order=DESC">https://plu.mx/pitt/g/?artifact_tab=all_artifacts&amp;sort=year&amp;order=DESC</a> ) .....   | 48 |
| Figure 15 - The architecture of the system.....  | 54 |
| Figure 16 - A description about all of our system’s views .....  | 58 |
| Figure 17 - Showing how a client’s request is handled.....   | 58 |
| Figure 18 - A representation of the system’s database.....   | 61 |
| Figure 19 - The home page .....  | 65 |
| Figure 20 - All the publications that cite the department’s publications .....   | 67 |
| Figure 21 - The collaborations page. We can see on the map all the affiliations what collaborate with the department.....  | 68 |
| Figure 22 - The content of a maker.....  | 68 |
| Figure 23 - All the counties of the affiliations collaborated with the department.....   | 69 |
| Figure 24 - The research areas of the department.....  | 70 |
| Figure 25 - The publications search page .....   | 71 |
| Figure 26 - The authors page which shows all the authors of the department .....   | 72 |
| Figure 27 - The author main page.....  | 74 |
| Figure 28 - How the author’s publications looks like.....  | 75 |

|   |                                     |
|---|-------------------------------------|
| Figure 29 - What we see if we click on the number of citations..... | 75                                  |
| Figure 30 - The Altmetric metrics of a publication .....            | <b>Error! Bookmark not defined.</b> |
| Figure 31 - The way we are visualizing the author's keywords .....  | 76                                  |
| Figure 32 - The result if we click on a keyword.....                | 77                                  |
| Figure 33 - All the publications that cite the author.....          | 77                                  |
| Figure 34 - All the collaborators of an author.....                 | 78                                  |
| Figure 35 - The admin interface.....                                | 78                                  |
| Figure 36 - What we see if we want to add an author .....           | 79                                  |

# Chapter 1

## 1 Introduction

### 1.1 Introduction to Information Systems for scientific research

The scientific research is a very important sector that aims to continue development and progress of any society. The recording, the conclusions and the study of these conclusions which are carried out by different organizations around the world, are undoubtedly important and valuable information. This information can give us the opportunity to be able to observe which research communities are making progress, how these communities' research production amount is evolving through time and how these communities work together worldwide.

This produced research activity information is certainly extremely useful for the researchers themselves. Through this information they can broaden their knowledge and can form a network of knowledge that will lead to new research results.

Undoubtedly, the study of research activities and its results are also very important for the research that different universities and institutions are operating around the world. Through this, they will have the opportunity to show and highlight the scientific fields in which they have expertise, but also observe their weakness fields of research and so be able to strengthen them in the near future.

Also, it gives the opportunity to promote their work in universities and other institutions through scientific collaboration between researchers. Finally, it allows them to promote their general results and existing knowledge and expertise to the public.

The big question that arises is how and where we can get all this information that would lead us to the above mentioned conclusions. Therefore, it is necessary to develop and use appropriate **information systems for scientific research** which will be able to record this whole research activity and produce in a more easy and efficient way this information which is valuable for any organization, institution and researcher.

### 1.2 About this diploma thesis project

In this thesis project we are developing a scientific information system. Especially, we are trying to visualize the research activities of the University of Thessaly, using **data visualization** techniques. Through this system, every department will be able to access valuable information about its research activity, information that is collected in real time, from big web repositories, using **crawling** techniques.

Our goal is to use this information to visually represent the scientific profile of each department, its scientific productivity, its evolution and its collaborations with other universities, institutions and researchers around the world.

Finally, we developed our system in a way that can visualize the above information with an intelligent and effective way adapted to the background and the interests of every viewer.

### 1.3 Content organization

This thesis is structured in the following way:

In **chapter 2**, we describe the theory and fundamentals of data visualization. At first, we make a small introduction about the huge amounts of data and information of the web, what we call Big Data. Then, we describe in deep the fundamentals of the theory of data visualization and also the reason we use data visualization in our lives. Finally, we present some of the best tools and software that exist today, and are used to visualize data.

In **chapter 3**, we describe the term of “Web Crawling”. We start by describing what a web crawler is and how it is related with the web. Then, we show how a modern web crawler works and finally we describe the structure of a web crawler and how to select the appropriate technique to crawl, in order to have efficient and intelligent web crawling.

In **chapter 4**, we first describe the bibliometric analysis and then we deal with the scientific information systems for research that exist today and the big repositories and search engines about research activities.

In **chapter 5**, we describe in deep the system we have developed by analyzing its functionalities and all the visualizations we have used.

Finally, in **chapter 6**, we describe the conclusions that we came to and discuss possible future work.

# Chapter 2

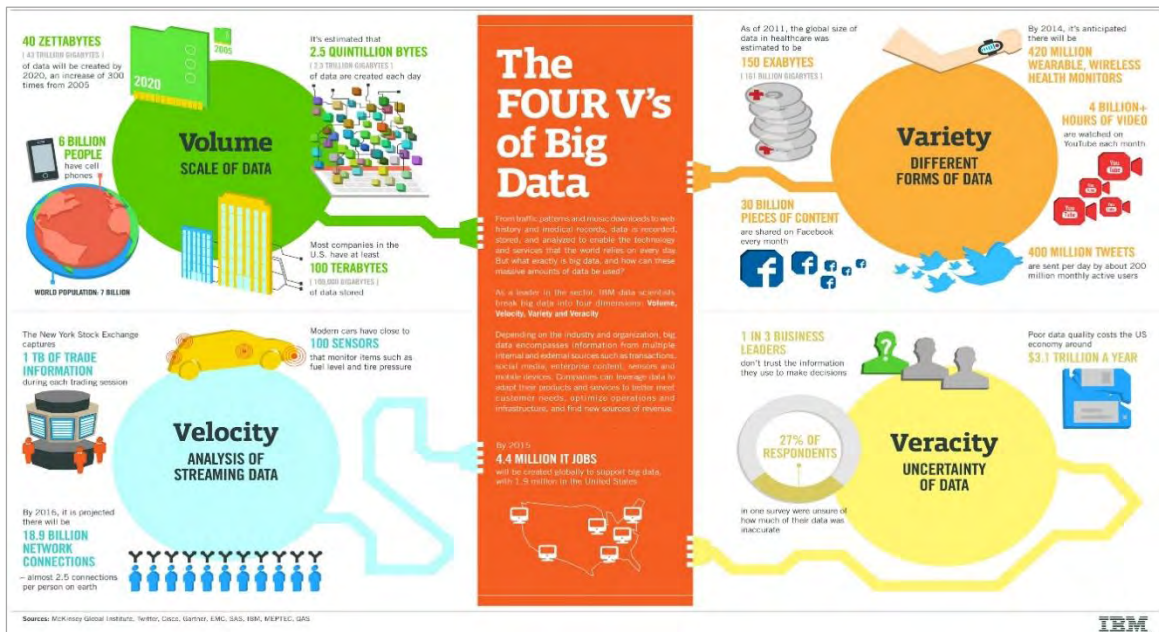
## 2 Data Visualization

### 2.1 The World Wide Web and its big data

Since the formation of the World Wide Web our lives and the way we live and think have changed. Every day we come across huge amounts of data. Data never gets old and it is going to stay there forever. Moreover, we can see the Internet has grown exponentially with a half billion of websites and this has caused a revolutionized way about how we access information. Every day, the web is becoming a growing universe of interlinked web pages and web apps teaming with huge amounts of valuable information and data at our fingertips. As the Internet every day is gaining more and more attraction, this information is growing and is shared instantaneously and so data has begun to surge and a new medium has been born: Big Data. Big Data is an umbrella term. It encompasses everything from digital data to health data, to the data collected through years and years. Specifically, Big Data can be described by following main characteristics:

- **High volume:** refers to the quantity of data that is generated which is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. Also, the name 'Big Data' itself contains a term which is related to size and hence the characteristic.
- **High velocity:** refers to the speed of generation of data (through multiply sources such as online systems, sensors, social media, smartphone etc...) or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.
- **High variety:** refers to the category in which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.
- **High veracity:** refers to uncertainness of the data. With the many forms of Big Data, the quality and accuracy are less controllable. So, we need trusted, certain and precise data capable of working with these.

As we can understand, the big question is how we will use all this data to propel forward. Big Data needs to be linked and correlated in order that we be able to grasp the appropriate information which is necessary for us to use with an effective and representative way in order to enhance our knowledge, solve problems and make decisions. So, all the complexity of Big Data we mentioned above needs to be visualized in an appropriate way, and so the need for visualizing information was born.

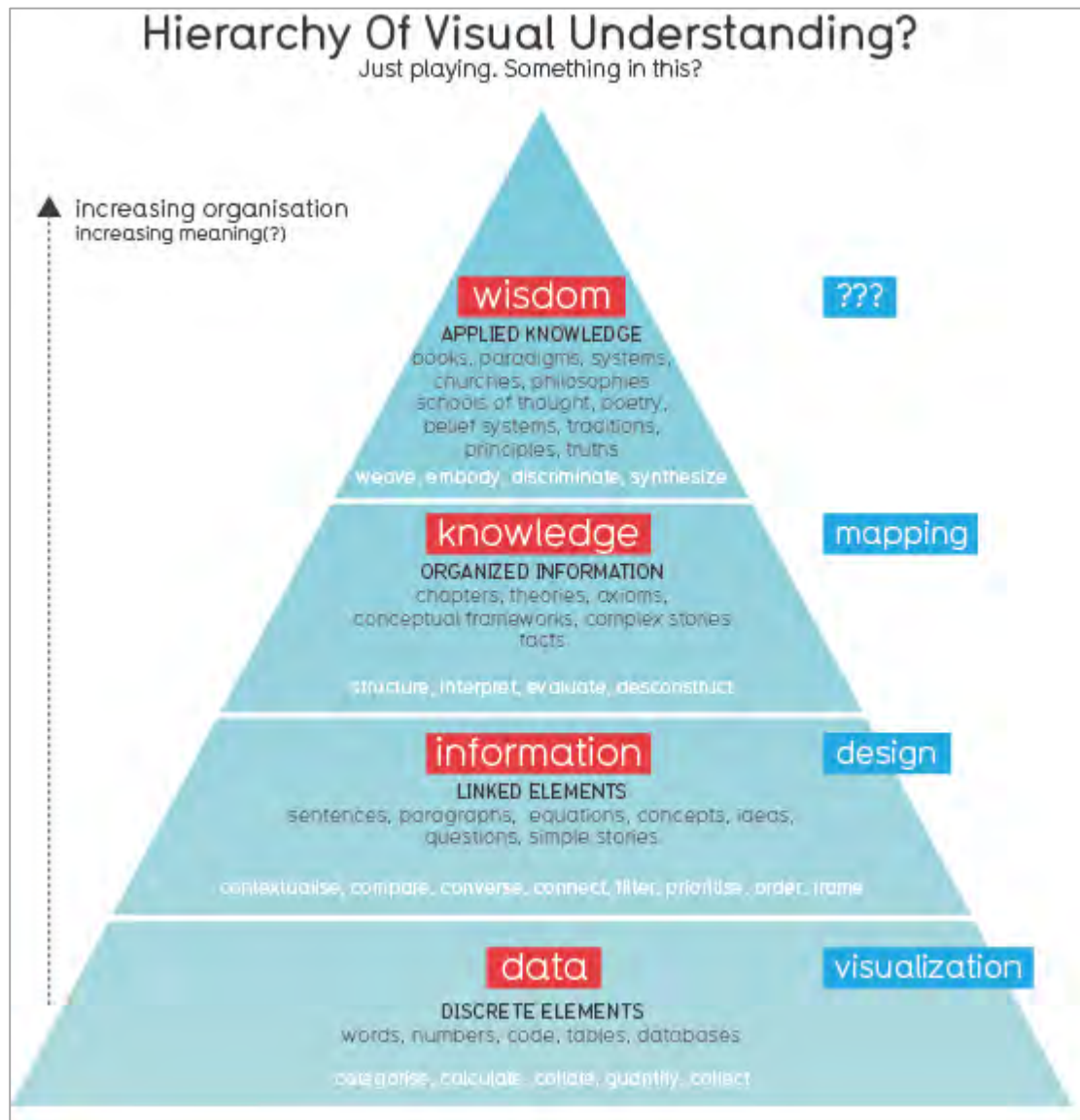


**Figure 1** – An infographic about the four v's of Big Data (source: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>)

## 2.2 What is data visualization

As we mentioned in par 2.1 every day we come across huge amounts of abstract data. We want all this abstract data to be visualized in order that we be able to enhance our knowledge and make decisions which are based on specific information. This data is abstract from the aspect that it describes things that are not physical and is very difficult for the human brain to understand in order to get valuable information. So we need a way to translate this abstract data into physical attributes of vision such as length, position, size, shape, color and more. This process of translation is made through data visualization. So, data visualization is the presentation of abstract data in a pictorial or graphical

format, a presentation which must follow specific design principles that are derived from an understanding of human perception.



**Figure 2** - The process from data to wisdom in the theory of data visualization (source: <http://harlotofthearts.org/blog/2011/01/08/data-information-knowledge-wisdom/>)

Data visualization is derived from three main concepts. Visualization, interactivity and knowledge & wisdom. All these concepts make up what we call data visualization. Let's describe every concept in detail.

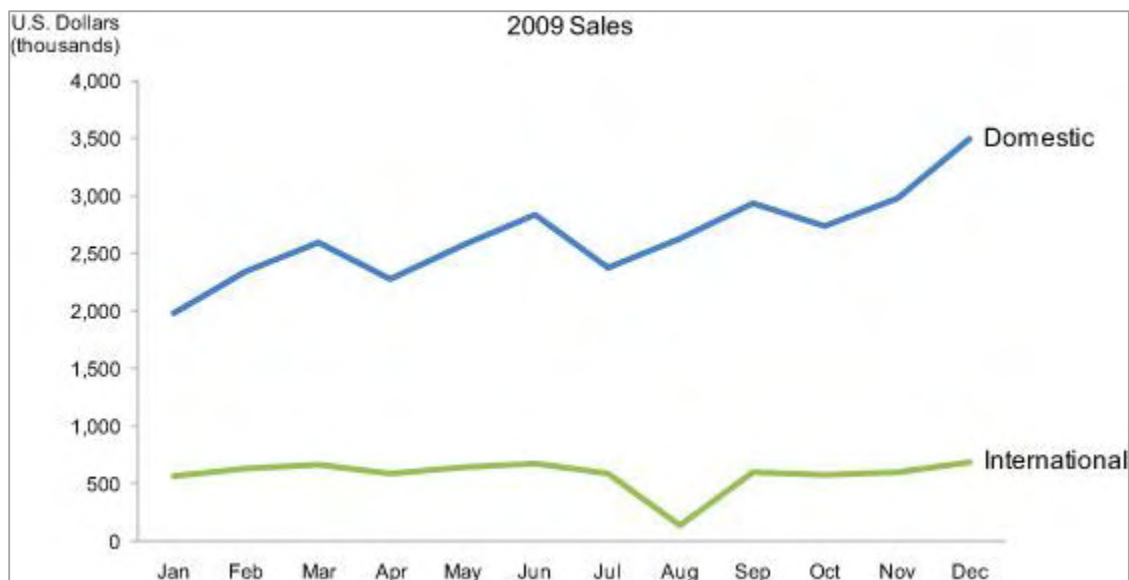


## Visualization

Visualization is the use of computer software programs and tools for the translating of abstract data into visual form. Its purpose is to improve the fundamental understanding of large amounts of data through graphical representations. In order to do this correctly, visualization makes effective use of human visual perception. Therefore, one of the main reasons why the visualization is a valuable tool for inspection, analysis and transmission of information is the fact that benefits from the considerable abilities of human vision. Human vision is powerful, like a channel of information transmission in the brain with greater range, compared to the other senses. As the saying goes, "a picture is worth a thousand words", but only when the data (here, the data of the picture) is best presented graphically rather than verbally and the picture is well designed. Let's look at a simple example to understand better how the representation process, we described above, works from the aspect of data visualization. Let's suppose we have the following table of numbers, which is about some domestic and international sales of the U.S.

| 2009 Sales (thousands of U.S. \$) |       |       |       |       |       |       |       |       |       |       |       |       |        |
|-----------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| Region                            | Jan   | Feb   | Mar   | Apr   | May   | Jun   | Jul   | Aug   | Sep   | Oct   | Nov   | Dec   | Total  |
| Domestic                          | 1,983 | 2,343 | 2,593 | 2,283 | 2,574 | 2,838 | 2,382 | 2,634 | 2,938 | 2,739 | 2,983 | 3,493 | 31,783 |
| International                     | 574   | 636   | 673   | 593   | 644   | 679   | 593   | 139   | 599   | 583   | 602   | 690   | 7,005  |
| Total                             | 2,557 | 2,979 | 3,266 | 2,876 | 3,218 | 3,517 | 2,975 | 2,773 | 3,537 | 3,322 | 3,585 | 4,183 | 38,788 |

This table does two things extremely well: it expresses these sales values precisely and it provides an efficient means to look up values for a particular region and month. But if we're looking for patterns, trends, or exceptions among these values, if we want a quick sense of the story contained in these numbers, or we need to compare whole sets of numbers rather than just two at a time, this table fails. Now, look at the following visualization of the above table in the form of a line chart:



From the above chart several facts now leap into view:

- Domestic sales were considerably and consistently higher than international.
- Domestic sales trended upward over the year as a whole.
- International sales, in contrast, remained relatively flat, with one glaring exception: they decreased sharply in August.
- Domestic sales exhibited a cyclical pattern - up, up, down - that repeated itself on a quarterly basis, always reaching the peak in the last month of the quarter and then declining dramatically in the first month of the next quarter.

What these numbers could not communicate when presented as text in a table, which our brains interpret through the use of verbal processing, becomes visible and understandable when communicated visually. This is **the power of data visualization**.

### Interactivity

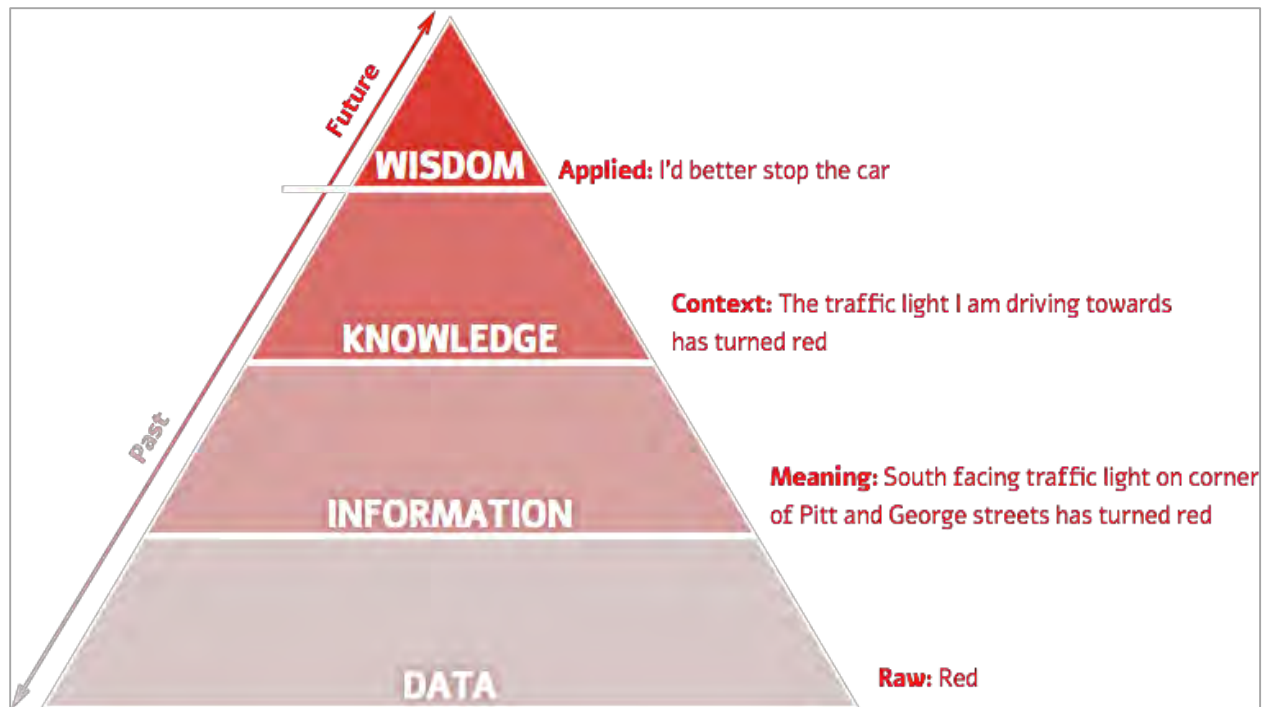
Static visualizations can offer only pre-composed “views” of data, so multiple static views are often needed to present a variety of perspectives on the same information. The number of dimensions of data is limited, too, when all visual elements must be present on the same surface at the same time. Representing multidimensional datasets fairly in static images is notoriously difficult. A fixed image is ideal when alternate views are neither needed nor desired, and required when publishing to a static medium, such as print.

Dynamic, interactive visualizations can empower people to explore the data for themselves. The basic functions of most interactive visualization tools have changed little since 1996, when Ben Shneiderman of the University of Maryland first proposed a “[Visual Information-Seeking Mantra](#)”: overview first, zoom and filter, then details-on-demand.

This design pattern is found in most interactive visualizations today. The combination of functions is successful, because it makes the data accessible to different audiences, from those who are merely browsing or exploring the dataset to those who approach the visualization with a specific question in search of an answer. An interactive visualization that offers an overview of the data alongside tools for “drilling down” into the details may successfully fulfill many roles at once, addressing the different concerns of different audiences, from those new to the subject matter to those already deeply familiar with the data.

### Knowledge & wisdom

As we mentioned above, our main purpose is to exploit the potential of visualization and interactivity enhance the process of knowledge acquisition including reasoning, judgment and decision making. Below, we see an example of how we acquire knowledge and wisdom.



**Figure 3** - An example of knowledge acquisition (source: <https://visualisingadvocacy.org/blog/disinformation-visualization-how-lie-datavis>)

## 2.3 Data visualization patterns

With the huge amount of abstract data, we must choose an appropriate visualization which will highlight with the best way the valuable information that is hiding inside the data. Selecting the appropriate visualization it is not an easy task since several factors such as the nature of data, the type of information we want to produce and the intended users, are involved.

Specifically, there are many factors to take into consideration in order to design a visualization and each of them indicates the use of a different visualization technique. The main factors that should be considered is the number of dimensions of our problem we want to solve and the type of data structure we want to visualize. Let's look in more details each one of them.

### Number of dimensions

The number of dimensions of data is the number of features that will be represented. Each dimension is a visualization of a distinct type of information encoded in the visual graph. Moreover, the number of dimensions of a graph can be described as the level of complexity of the visualization which complexity is increasing as the number of dimensions is increasing. Generally, we have the following variable dimensions:

- One dimension: when we have only one feature to represent
- Two dimensions: when we have two features to represent
- Three dimensions: when we have three features to represent
- Multi dimensions: when we have more than three features to represent

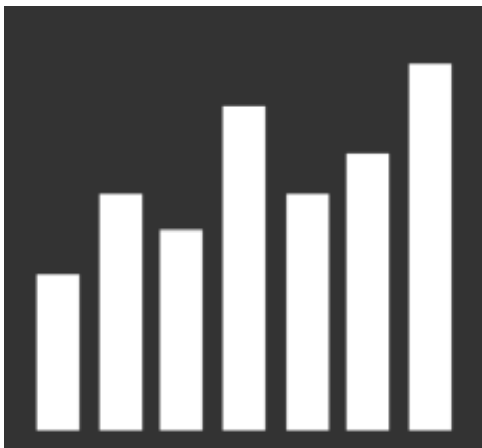
Let's now look an example to understand better the above concept of dimensions. Let's suppose we want to visualize the number of sales (such as the example of par 2.2) for a company in time. For this visualization we have two dimension, because we have to represent only two features, the time and the number of sales in a particular time. But if want to add another feature in our visualization, for example, the type of products the company has sold in time, we have three dimension, since we now have three features, the time, the number of sales and the type of products.

## Type of Structure Data

All data visualizations patterns we use to visualize our data are categorized in following way:

- **Independent quantities:** are used when we have to visualize independent quantities and want to compare the values of these independent variables.
- **Continuous quantities:** are used when our data is continuous, for example when visualizing data over a period of time.
- **Proportions:** are used when we want our data to represents parts of a whole.
- **Correlations:** are used when each piece of our data has measurable variables which can be plotted on a grid.
- **Networks:** are used when the most important feature of the visualization is to show which data is connected to each other.
- **Hierarchies:** are used when our data has a strict hierarchy that needs to be communicated.
- **Cartographic:** are used when our data we want to visualize is relevant to specific locations or regions which can be plotted on a map.

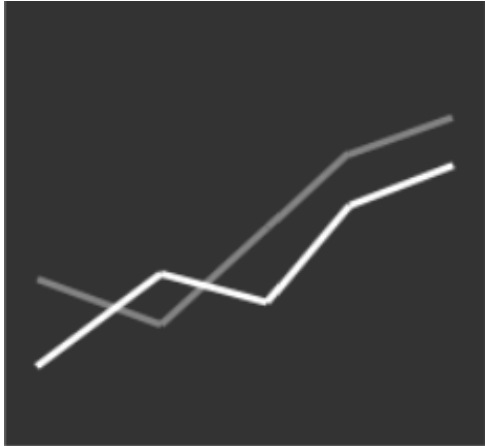
### 2.3.1 Independent quantities



## Bar charts

Simple bar charts are the most common form of data visualizations. Typically they only display different quantities of single variable data or more. However other variations, such as stacked bar charts or multi-set bar charts can be used to compare multiple variables using bars.

### 2.3.2 Continuous quantities



#### Line graphs

Line charts are created by plotting points on a Cartesian grid, usually with the horizontal axis representing time. They are very powerful because without looking at the specific data, they show how a variable develops over time (from left to right)



#### Stacked area charts

Similar to line charts but with the added value of filled areas. The data is stacked adds up to a total of all variables combined. For example a business might use stacked area charts to visualize their total income, with each stacked area a different income channel.

### 2.3.3 Proportions



## Pie charts

Simple pie charts are the most common visual used to compare proportional data. They give viewers a very quick understand of the distribution of the data. Pie charts are not useful when comparing many pieces of data with relatively close values.



## Ring charts

Similar to pie charts, ring charts are used to visualize the distribution of a data set. The advantage is they compare similar data sets. Also, the alternative would be to place multiply pie charts next to each other, this can also be viewed as a space-saver.

### 2.3.4 Correlations



## Scatterplots

They are created by plotting independent points on a Cartesian relationship between data or to reveal information such as trends within the data which are not easily visible when in a table. A disadvantage is that they only works with two dimensional data.



### Bubble charts

Similar to scatterplots, but bubble charts can display more dimensions of data by varying size, color or texture of the bubbles. It therefore can display multiply dimensions of data in a two dimensional display.

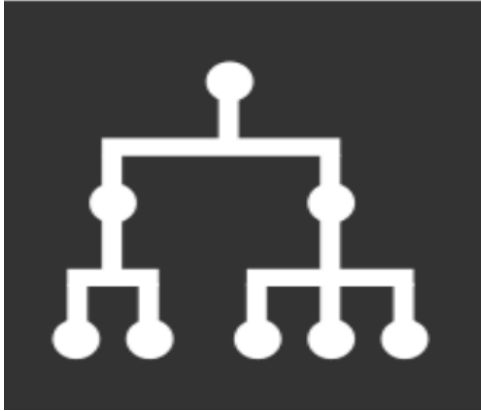
#### 2.3.5 Networks



### Diagrams maps

These visualizations are used to primarily represent the connections between different nodes or points. Their purpose is to show which nodes or points are connected to each other. Common examples of diagrams maps are metro maps and also social network visualizations.

### 2.3.6 Hierarchies



#### Tree diagrams

Tree diagrams are often used when wanting to represent the strict hierarchy of data. They are most often used to represent strict hierarchies such as family trees or how data is stored in a computer system.

### 2.3.7 Cartographic



#### Maps

Maps are used when the data is related to a specific location or region on the map. The advantage is that their spatial representation directly relates to a real world situation. However at times can be difficult to read.

## 2.4 Data visualization tools

### 2.4.1 SVG, HTML 5 Canvas

All modern web browsers supports two major technologies for creating rich graphics: SVG and HTML5. Both technologies are used for drawing inside a web browser, but they are fundamentally different.

#### SVG

Short for Scalable Vector Graphics, SVG is a vector graphics file format that enables two-dimensional images. Any program such as a Web browser that recognizes XML (Extensible Markup Language) can



display the image using the information provided in the SVG format. In contrast to JPEG and GIF images on the Web, which are bitmapped and always remain a specified size, SVG images are scalable to the size of the viewing window and will adjust in size and resolution according to the window in which it is displayed. Also, it supports text label and descriptions on its graphics that can be searched by search engines.

## HTML5 Canvas

HTML5 canvas is a new tag of HTML which is part of the HTML5. The real power of the canvas element is that it takes advantage of the HTML5 Canvas API. This API is used by writing JavaScript that can access the canvas area through a full set of drawing functions, thus allowing for dynamically generated animated graphics. To use it we must put the `<canvas>` tag in our webpage. Also, the HTML5 Canvas is supported by all modern web browsers and can be accessed on a wide range of devices including desktops, tablets, and smartphones.

### 2.4.2 JavaScript libraries

The majority of all modern visualization tools make use of the famous web technology JavaScript. Let's explore the best JavaScript visualization tools that exist today.

#### General libraries

##### D3.js

D3.js is probably the best free JavaScript library that exists today providing hundreds of visualizations to use. Specifically, it is used for manipulating documents based on simple data using HTML, SVG and CSS. Also, D3 emphasizes on web standards giving full capabilities of modern browsers and combining powerful visualization components and a data-driven approach to DOM manipulation. In addition, it supports a really extensive documentation with all the source code hosted on GitHub. Finally, in our project we have used the D3.js library in some of our data visualizations.

##### Highcharts

Highcharts is a powerful charting library written in pure JavaScript. It offers a variety of interactive visualizations such as: line, spline, area, area spline, column, bar, pie, scatter, angular gauges, area range, area spline range, column range, bubble, box plot, error bars, funnel, waterfall, polar chart types some combinations of them. Also, setting the Highcharts configuration options requires no special programming skills. The options are given in a JavaScript object notation structure, which is basically a set of keys and values connected by colons, separated by commas and grouped by curly brackets. Moreover, it lets you modify the appearance of your chart by removing specific parts of it and also provides download of the chart with various formats, such as image (PNG, JPEG), pdf and SVG vector image. Finally, in our project we have used Highcharts in some of our chart visualizations.

## Google Charts

Google Charts is a powerful, free and simple to use, and it has everything from simple line charts to complex hierarchical tree maps, the chart galley provides a large number of well-designed chart types. It's an especially useful tool for specialist visualizations such as geocharts and gauges, and it also includes built-in animation and user interaction controls. Also, it provides a good community-forum.

## JavaScript InfoVis Toolkit

JavaScript InfoVis Toolkit is another JavaScript library which provides tools for creating beautiful and interactive visualizations mainly based on hierarchical data. It implements advanced features of information visualization like TreeMaps, an adapted visualization of trees based on the SpaceTree, a focus+context technique to plot Hyperbolic Trees, a radial layout of trees with advanced animations effects and other visualizations. The feature that sets it apart is the combination of aesthetics and its simplicity of code and also is free to use.

## KoolChart

KoolChart is a powerful library and is based on pure HTML5 and JavaScript offering 90+ types of charts, which have animation, 3D and gradient effects. Beyond the basic offerings, KoolChart provides also advanced functions and chart types such as Real-Time Monitoring Charts, Target vs Actual Charts, Slide Charts and Candlestick Charts in a single product package. Also, the library offer a XML-based and ease-of-use template design feature and pre-packaged with over 250 ready-to-use sample components, which cuts the initial time, cost and complexity of embedding charts for any applications. KoolChart offers a 30-day trial fee edition but then you have to buy a license.

## Graphs visualization libraries

### Arbor.js

Arbor.js is a graph visualization library built with web workers and jQuery. It provides an efficient, force-directed layout algorithm, abstractions for graph organization and screen refresh handling. The library doesn't force a specific method for screen-drawing and you can use it with canvas, SVG, or even positioned HTML elements. Also, Arbor.js simply helps you focus on the graph data and its style rather than spending time on the physics math that makes the layouts possible

### Sigma.js

Sigma.js is an open-source lightweight JavaScript library to make network visualizations using the HTML canvas element.

### Cytoscape

Cytoscape is not a JavaScript library, but we must refer it because it is a fantastic tool. Especially, Cytoscape is an open source software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general

platform for complex network analysis and visualization. Cytoscape core distribution provides a basic set of features for data integration, analysis, and visualization. Additional features are available as Apps (formerly called Plugins). Apps are available for network and molecular profiling analyses, new layouts, additional file format support, scripting, and connection with databases. They may be developed by anyone using the Cytoscape open API based on Java technology and App community development is encouraged. Most of the Apps are freely available from [Cytoscape App Store](#).

## Time series visualization libraries

### Cubism

Cubims is a D3 plugin for visualizing time series. Its scalability allow us to construct better realtime dashboards, pulling data from sources such as [Graphite](#) and [Cube](#).

### RichShow

RickShow is a pure JavaScript toolkit for creating interactive time series graphs. It is based all on the D3.js library we mentioned above, so graphs are drawn with standard SVG and can styled with CSS. RickShow, provides many elements such as renderers, legends, hovers, range selectors etc.

## 2.4.3 Programming languages

### R

R is a language and environment for statistical computing and graphics supporting a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques. Except from its highly extensibility, another R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

Also, R is available as free software runs on a wide variety of platforms such as Windows, Linux and MacOS.

### Processing

Processing is a programming language, development environment, and an online community which makes it a wonderful environment for writing generative, interactive and animated applications. An offshoot of Processing is Processing.js, a JavaScript library that creates data visualizations using web standards and minus the need for any plug-ins. Also, Processing is very easy to learn and you don't need to know JavaScript. So, you simply write code using the Processing language, include it on your web page, and Processing.js take care of the rest. Also, it can be simply downloaded and installed on any platform.

## 2.4.4 Software platforms

## Tableau

Tableau is groundbreaking data visualization software for simple and also complex visualizations. It connects easily to nearly any data source, such as Microsoft Excel or web-based data. You can drag and drop fields onto the work area and ask the software to suggest a visualization type, then customize everything from labels and tool tips to size, interactive filters and legend display. Also, Tableau offers a variety of ways to display interactive data. You can combine multiple connected visualizations onto a single dashboard, where one search filter can act on numerous charts, graphs and maps. And once you get the hang of how the software works, its drag-and-drop interface is considerably quicker than manually coding in a programming language such as JavaScript or R, making it more likely that you'll try additional scenarios with your data set. In addition, you can easily perform calculations on data within the software.

## Gephi

Gephi is an interactive visualization and exploration software. It supports visualizations for all kinds of networks and complex systems, dynamic and hierarchical graphs. With Gephi we can represent large data sets in the form of graphs or hierarchical network and provides easy analysis and exploration. Also, it includes several features such as algorithms for optimizing the layout of nodes, dynamic filtering, and grouping.

# Chapter 3

## 3 Web Crawling

As we mentioned in chapter 2 of this thesis, the Word Wide Web consists of great amounts of information. This large amount of information is changing dynamically and semantically unstructured, making us difficult finding the related and valuable information we need. We need a way for an automatic discovering of this information. So this need created what we nowadays called “Web Crawling”. So, Web Crawling is the methodical and automated process of collecting data from websites and finally return us this data.

### 3.1 Introduction

Before examining what a web crawler is and how it works, we must define what a web page is. A lot of people think of a web page as what they see in their browser window, which is right, but that's not what a web page is when a web crawler sees it. So let's look at a web page as a web crawler with an example.

When we visit <http://www.theguardian.com/us>, we see something like this:



Figure 4 - How a website looks like from the user's perspective

In fact, what we see is the combination of many different “resources”, which our web browser combines together to show us the page we see. Here’s an abridged version of what happens:

- 1) We type in “<http://www.theguardian.com/us>”.
- 2) Our browser says ok, let me GET “<http://www.theguardian.com/us>”.

The Guardian’s server says, hey browser, here’s the content for that page. At this point, the browser is only returning the HTML source code of “<http://www.theguardian.com/us>”, which looks something like the following figure:

```

<!DOCTYPE html>

<html>

<head>

    <title> Latest news, world news, sport and comment from the Guardian | theguardian.com | The Guardian </title>

    <meta property="fb:app_id" content="100444840287"/>
    <meta charset="utf-8" />

    <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1"/>

    <link rel="canonical" href="http://www.theguardian.com/us" />

    <meta name="description" content="Latest news, world news, sports, business, comment, analysis and reviews from the Guardian, the world's leading liberal voice" />
    <meta name="DC.date.issued" content="2011-08-24">

    <meta name="llt" content="4Ujcp60Q" />

    <meta name="keywords" content="US Network front" />
    <meta name="news_keywords" content="US Network front" />

    <link rel="shortcut icon" href="http://static.guim.co.uk/favicon.ico" type="image/x-icon" />

    <meta name="application-name" content="The Guardian"/>
    <meta name="msapplication-filecolor" content="#000080"/>
    <meta name="msapplication-tileimage" content="http://static.guim.co.uk/static/74214e9be05c32f5e5796e00883734f031c51027/common/images/favicons/windows_tile_144_b.png"/>

    <link rel="shortcut" href="http://gu.com/p/3sduh" />

    <meta name="content-id" content="/us"/>

    <link rel="publisher" href="https://plus.google.com/111000007143113820257a"/>

    <meta name="p:domain.verify" content="4fa576e6ac27d86fd926c4579b07f23"/>
    <meta name="p:site-verification" content="be6d97be1f6961ce6348e7ced4f14" />
    <link href="http://feeds.theguardian.com/theguardian/us/rss" rel="alternate" type="application/rss+xml" title="rss" />
    <link rel="stylesheet" type="text/css" href="http://static.guim.co.uk/static/74214e9be05c32f5e5796e00883734f031c51027/common/styles/network-front-grid.css" media="all" />
    <link rel="stylesheet" type="text/css" href="http://static.guim.co.uk/static/74214e9be05c32f5e5796e00883734f031c51027/zones/news/styles/zone-accents.css" media="screen" class="contrast" />
    <link rel="stylesheet" type="text/css" href="http://static.guim.co.uk/static/74214e9be05c32f5e5796e00883734f031c51027/zones/news/styles/grid-zone-accents.css" media="screen" class="contrast" />
</-- [if lte IE 6] >

```

**Figure 5** - How a website looks like from the crawler's perspective

- 3) Our browser looks through this code and notices a few things. It notices there are a few style resources needed. It also notices there are several image resources needed.
- 4) The browser now says, I need to GET all of these resources as well.
- 5) Once all the resources for the page are received, it combines them all and displays the page we see.

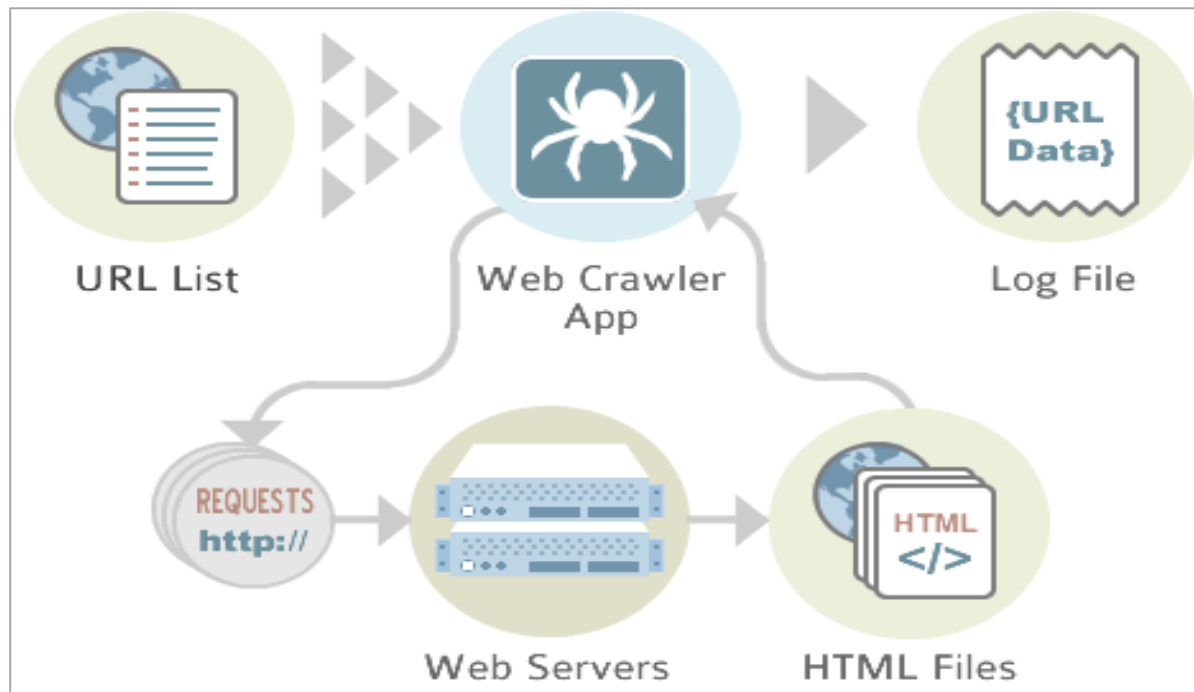
This is what our browser does. A web crawler can get all the same resources, but if we tell it to GET “http://www.theguardian.com/us”, it will only fetch the HTML source code we wrote above. That’s all it knows about it until we tell it do something else (possibly with the information in the HTML). Now, that we understand the structure of a website let’s deep into about what a web crawler is and how it works.

## 3.2 What is a web crawler and how it works

Generally, a web crawler is an automated program (or script) that, given one or more seed URLs list (such the one we mentioned in par 3.1), downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks. Moreover, a web crawler simply sends HTTP requests for documents to a web server on the Internet, just as a web browser does when the user clicks on links. Then, it handles the result



(HTML files) the server sends by filtering its content (depending on its policy) and stores it in a file (Log File). All this process is shown on the picture below:



**Figure 6** - How a web crawler works (source: <http://blog.datafiniti.net/?p=280>)

### 3.3 Implementing an efficient web crawler and best strategies

We saw the basic concept behind the structure of a web crawler, but implementing this concept is not merely a bunch of programming. Let's examine the difficulties involved in implementing an efficient web crawler and the best strategies.

#### Difficulties in implementing efficient web crawler

There are two important characteristics of the web that generate a scenario in which web crawling is very difficult to be implemented.

- **The large volume of webpages.**

This implies that web crawler can only download a fraction of the webpages and hence it is very essential that web crawler should be intelligent enough to prioritize download.

- **The rate of change on webpages.**

This implies the today's problem of Web that its content is changed dynamically and very frequently. As a result, by the time a web crawler is downloading the last page from a website, the webpage may change or a new content has been placed or updated to the website.

### Best strategies and selecting the right algorithm

Every web crawler has its own policy and therefore the implementer of the web crawler must define the crawler's behavior that corresponds to its policy. Defining the behavior of a web crawler is the outcome of a combination of the following strategies.

- **Selecting the best algorithm to decide which web page to download.**

Due to the huge size of the web, it is essential that every crawler should crawl on a fraction of the web. The crawler should obtain that the fraction of pages crawled must be most relevant pages and not random pages. Moreover, the crawler's implementer must keep in mind that algorithm must make sure that webpages are chosen depending upon their importance, which lies in their popularity in terms of links or visits. Let's see two main algorithms types:

**Path-ascending crawling:**

With this algorithm we want the crawler to download as many resources as possible from a particular website. That way the crawler would ascend to every path in each URL that it intends to crawl. Let's see an example. If we want to crawl the URL `http://website.com/a/b/page.html`, then our crawler will attempt to crawl every directory of the URL, `/a/b`, `/a/` and `/`. The advantage of this technique is that the crawler is very effective in finding isolated resources, or resources for which no inbound link would have been found in regular crawling.

**Focusing crawling:**

The importance of a page for a crawler can be expressed as a function of the similarity of a page to a given query. So, with this algorithm we intend our crawler to download pages that are similar to each other. The only problem of this technique is that we would like the crawler to be able to predict the similarity of the text of a given page to the query before actually downloading the whole page.

- **How to re-visit web pages**

By the time a web crawler has finished its crawl, many events could have been happened, including creations, updates and deletions. From a web crawler's point of view, there is a cost associated with not detecting an event, and thus having an outdated copy of a resource. The most-used cost functions are freshness and age.

**Freshness** is a measure that indicates whether a local copy is accurate or not. The freshness of a page  $p$  in the repository at time  $t$  is defined as:



$$Fp(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

**Age** is a measure that indicates how outdated the local copy is. The age of a page in the repository at time  $t$  is defined as:

$$Fp(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

Moreover, the optimum method to re-visit a page and maintain its average freshness high, is to ignore the pages that change too often. This approach could be if re-visiting all pages regardless of their rates of changes or re-visiting more often the pages that change more frequently. Both techniques regards all pages as homogenous in terms of quality, something that is not a realistic scenario on the Web, since pages are not worth the same.

#### ▪ How to avoid overloading

A web crawler in order to accomplish its purpose can perform multiple requests per second to a web server or download large amount of files. But, this has a big cost. First of all, we have a cost on network resources as crawlers require considerable bandwidth and operate with a high degree of parallelism during a period of time, are overloading. Also, we have cost on the web server, due to the multiple requests, as servers crash or routers especially if the crawler is poorly written. To resolve all these problems each web server uses a robots exclusion protocol, which is a text file (**robots.txt**) including some specific directories of the website. This protocol is a contract to prevent accessing specific part of a website or whole website from a web crawler or from a cooperation of web crawlers. To be more specific, let's look an example. Suppose we have the following robots.txt file:

```
User-agent: *
Disallow: /local/
Disallow: /images/
Disallow: /system/
```

This example tells all crawlers not to enter the above directories, `/local/`, `/images/` and `/system/`. In addition if the file was the following:

```
User-agent: *
Disallow: /
```

it tells all crawlers to stay out of the website.

Finally, a web crawler's architecture (algorithm) is shown in the following pseudo code:

```
Ask user to specify the starting URL on web and file type that crawler should crawl.

Add the URL to the empty list of URLs to search.

While not empty ( the list of URLs to search )
{
    Take the first URL in from the list of URLs
    Mark this URL as already searched URL.

    If the URL protocol is not HTTP then
        break;
        go back to while

    If robots.txt file exist on site then
        If file includes .Disallow. statement then
            break;
            go back to while

    Open the URL

    If the opened URL is not HTML file then
        Break;
        Go back to while

    Iterate the HTML file

    While the html text contains another link {

        If robots.txt file exist on URL/site then
            If file includes .Disallow. statement then
                break;
                go back to while

        If the opened URL is HTML file then
            If the URL isn't marked as searched then
                Mark this URL as already searched URL.

        Else if type of file is user requested
            Add to list of files found.

    }

}
```

**Figure 7** - A pseudo code of a web crawler (source:  
<http://www.devbistro.com/articles/Misc/Implementing-Effective-Web-Crawler>)

# CHAPTER 4

## 4 Scientific research

### 4.1 Introduction

As we mentioned at the beginning of this thesis, scientific research is a very important sector of any society which undoubtedly leads to a lot of benefits. On this chapter we describe how modern scientific research works. We start, by referring to some of the most famous repositories and search engines that exist today providing us with huge amount of research activity information. But, all this amount of information needs to be “filtered” through a mechanism-analysis that can produce significant metrics associated with this information. This analysis is called “bibliometric analysis” which we will describe in the next paragraphs. Finally, since the modern scientific research needs to line with the latest technology, we describe some new tools and technologies that aim to provide the whole community of scientific research with a lot of benefits.

### 4.2 Big repositories and information systems for scientific research

#### Scopus

Scopus was launched in November 2004. It is the largest web repository for research that exists today containing 53 million records, 21.915 titles and 5.000 publishers from several fields of science such us technology, medicine, social sciences, arts and humanities. Scopus provides tools for intelligent tracking, analysis and visualization of research activity.

#### Microsoft Academic Search

Although Microsoft Academic Search is still in a beta edition is can be easily classified to the big web repositories and search engines that exists today. It provides a good search engine with a rich and interactive visualizations, in which we can search based on a researcher name, a publication name, a keyword or a field of study. We can also search for a whole organization or institution and we can even compare organizations or institutions and see the differences between them associated with their research activity. Search results are displayed in a rich graphical interface, which enables user to display statistics such as the number of publications, number of citations, an h-index value and collaborated authors which is visualized through an interactive graph. In addition, every researcher can create a profile that represents his research interests and also he can informed with email for new publications relating to a specific author name, a field of study or affiliation.

## The Web of Science

The Web of Science of the company Thomson Reuters is a system which has the oldest database of scientific publications providing access to databases covering scientific journals with high impact worldwide. The functionality is similar to that of Scopus. A user can make a query specifying any search fields such as the name of the researcher, title of publication, keywords and so on. The major difference is that the Scopus results returned by Scopus is turned more to the profile of the researcher.

## Incites

The Incites is another service of Thomson Reuters, which allows the analysis of productivity of institutions and organizations, through data visualization and reporting tools. For this analysis the Incites system uses data and metrics derived from the Web of Science and aimed at taking decisions and conclusions in a general field.

## Google Scholar

Google Scholar is a comprehensive web repository for monitoring academic publications, providing an integrated search engine. Its records are collecting from several sources such as [Scopus](#), [ACM library](#), [Science Direct](#) and others. Moreover, every researcher can create a profile by adding all the necessary elements that represent his research profile such as his personal elements, his research interests and the affiliation he belongs. Also, a good functionality of Google Scholar is that every researcher can monitor the publications which refer his name by taking updates or watching results in the application.

## Research Gate

Research Gate was founded in 2008 by a virologist and computer scientist, Ijad Madisch. It is it's not just a search engine for scientific research but a social network of researchers and scientists. Research Gate offers tools and applications for researchers to interact, exchange knowledge and collaborate with researchers of different fields. Finally, its search engine it is based on an intelligent way based on semantic and contextual correlations.

## CiteULike

CiteULike is another service for managing and discovering scholarly references. It is based on principles of social sharing bookmarks (social bookmarking), aiming to promote the development and sharing of scientific references among researchers. A researcher may enter in CiteULike an indication of a publication which is an online database. CiteULike automatically searches for metadata relating to the publication and will enter in its own database. Then it will search for references to the publication.

## 4.3 Bibliometric Analysis of research activity

All the repositories and search engines we saw above in par 4.1 provide us with huge amount of scientific research data. But, having all these data we are not able to create any measurable image associated with any entity, such as a researcher or institution. So we need a process that receiving all these data as input to able to provide us a measurable image that represent best these data. This


process we are talking about in the area of scientific research is called Bibliometric analysis which we describe next.

Bibliometric analysis of research activity is the process of recording and processing data associated with scientific publications and the exportation of the appropriate "bibliometric indicators" such as the number of publications, the number of citations from other publications etc.

These bibliometric indicators contribute to the creation of a measurable image for any research, technological development and innovation system. Having these bibliometric indicators we are able to evaluate any organization, group or researchers that produce research activity. More specifically, we can identify the characteristics and trends of the research production of any institution or country. Also we can assess the impact of research activity in our society and identify or create national and multinational networks among researchers aim to achieve common research goals.

Some of the main bibliometrics indicators and which we also used in our system are the following:

- **Number of publications:** This indicator shows the amount of publication production and can be refer to every entity, from a researcher to a university or a specific scientific field of research.
- **Number of citations:** This indicator shows the number of publications which refer to other(s) publication(s). Also, this indicator indicates the recognition and the influence of a scientific publication.
- **Citation impact:** This indicator indicates the average number of citations per publication and is calculated as the ratio of the number of references which are listed in a certain period of time to a total number of publications of the same period.
- **H-index:** This indicator measures both the productivity and the impact of the published work of a researcher. The index is based on the set of the scientist's most cited publications and the number of citations that they have received in other publications. A researcher has h-index value  $n$  when  $n$  of his  $N$  publications have at least  $n$  citations each, while the rest ( $N-n$ ) publications have fewer than  $n$  citations each. Let's look an example to understand how it is calculated. Suppose a researcher has the below eight publications. The first step is to order all publications in descent order based on their number of citations and then we cross all the elements with a way we describe next:



| <u>Publication</u> | <u>Number of citations</u> |
|--------------------|----------------------------|
| 1                  | 33                         |
| 2                  | 30                         |
| 3                  | 20                         |
| 4                  | 15                         |
| 5                  | 7                          |
| <b>6</b>           | <b>6 = h-index</b>         |
| 7                  | 5                          |
| 8                  | 4                          |

The first publication gives us a 1 – there is one publication that has been cited at least once, the second gives a 2, there are two publications that have been cited at least twice, the third publication, 3 and all the way up to 6 with the sixth highest paper –the final two publications (7 and 8) have no effect in this case as they have been cited less than six times.

In addition, there also some others bibliometric indicators that can arise from the above indicators:

- **Share of publications:** This indicator is used to indicate the participation of an entity (researcher, university, country etc) regarding to the production of publications inside a “group” of the same entities.
- **Field normalized citation score:** This indicator arises from the normalization of citation impact indicator under different scientific thematic areas. It is used to compare the impact of a publication relative to the impact of other publications worldwide of the same scientific thematic area.
- **P Top X%:** It is used to indicate the number of publications that every year are ranking high in the percentage classification of worldwide publications of the corresponding scientific field. The classification is based on the citations number. This indicator is calculated for intervals five years and concerns the percentiles the number of publications that are classified awards at 1%, 5%, 10%, 25% and 50% of the publications with the higher impact.
- **I10-index:** H10-index is a simple bibliometric indicator introduced by Google in July 2001 as part of their search engine, Google Scholar. This indicator indicates the number of publications of an author what have at least ten citations from other.

## 4.4 ORCID

### 4.4.1 What is ORCID

ORCID (Open Researcher and Contributor ID) which was firstly introduced in October 2012, is an open, non-profit and community-driven sponsored and funded by Elsevier. ORCID effort to create and maintain a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. ORCID is unique in its ability to reach across disciplines, research sectors and national boundaries. This registry is available free of charge to individuals, who may obtain an ORCID identifier, which is 16-digit number, manage their record of activities, and search for others in the registry.

Moreover, ORCID is a hub that connects researchers and research through the embedding of ORCID identifiers in key workflows, such as research profile maintenance, manuscript submissions, grant applications, and patent applications. It is worth to point out that the ORCID community is rapidly growing having so far issued 900.799 ORCID ids. In the next paragraphs we will see how researchers, universities and research funders can be benefited from ORCID.



**Figure 8** - The ORCID as an identifier hub (source: <http://informatics.mit.edu/blog/2014/07/integrating-researcher-identifiers-funding-identifiers-and-institutional>)

#### 4.4.2 What are the benefits of ORCID

##### For researchers

With the number of publications and authors being rapidly increasingly nowadays all the researchers and scholars face the ongoing challenge of distinguishing their research activities from those of others with similar names. As we know, many researchers share the same name, while others have different names during their career, or different variations of the same one. For example, A. Lazare, Aaron Lazare and Aaron A. Lazare could all refer to the same person. With a unique identifier, as provided by ORCID, which researchers can associate with their name variations and their research works, is a way to ensure that these links can be made accurately and reliably. This will help A. Lazare to get credit for her publications by uniquely identifying him as the author of his work across all systems integrated with the ORCID registry.

It is therefore obvious that with ORCID researchers can easily and uniquely attach their identity to research objects such as datasets, equipment, articles, media stories, citations, experiments, patents, and notebooks. As they collaborate across disciplines, institutions and borders, they must interact with

an increasing number and diversity of research information systems. So, with ORCID they will avoid entering biographical and bibliographic data in multiple systems, which can be time-consuming, and often frustrating. Finally, the registration for an ORCID id is very simple, researchers should visit [ORCID website](#), where they can create a complete online record of their research and publications. This is made open and freely available via a web page and data feeds. More importantly, once created a researcher's unique ORCID can be used as a linking identifier throughout the entire chain of the scholarly communication process to allow reliable attribution of research.

### **For universities & institutions**

ORCID is also potential for universities and institutions. With ORCID they can easily identify and bring in all information associated with their staff, to their scientific research systems, having a full record of their staff's research activities and achievements which could be facilitated by ORCID.

With all these records they can accurately benchmark their research strengths and impact and also keep their repository up to date. An advantage of that is that researchers only need to provide information about their publications and research outputs once. This goes in ORCID's repository and is then re-used from the universities and institutions research systems, in their staff profiles and to facilitate transfer of information about researchers' publications when they move organization. Finally, all this process can easily done using the ORCID [API](#).

### **For funders**

An ongoing challenge for funding organizations is tracking the outputs and outcomes of research they have funded. Understanding the impact of funding is a vital input into funding strategy, programming design, and mission alignment. ORCID provides the foundation to address this challenge, by providing a registry of persistent unique identifiers for researchers and methods for linking to digital research objects.

Funders can integrate ORCID identifiers into their research workflows, such as grant application processes and grant progress reporting protocols. Combined with efforts by research organizations and publishers, embedding ORCID identifiers in critical funding workflows will make it possible to link a researcher's contributions across their career. Finally, a greater precision and transparency of funded research and associated outputs can improve information on global research and development resource flows, vital for funding agency gap analysis and strategy.

#### **4.4.3 ORCID & PlumX**

In the previous paragraph we saw what ORCID is and how we can benefit from it. Let now see a new promising service named "PlumX" which works together with ORCID to track metrics associated with the research activity. PlumX is a research dashboard that analyzes newly available impact metric data to provide information about how research is being utilized and talked about around the world. More specifically, it works by using the ORCID API to retrieve a list of the researcher's public works and then

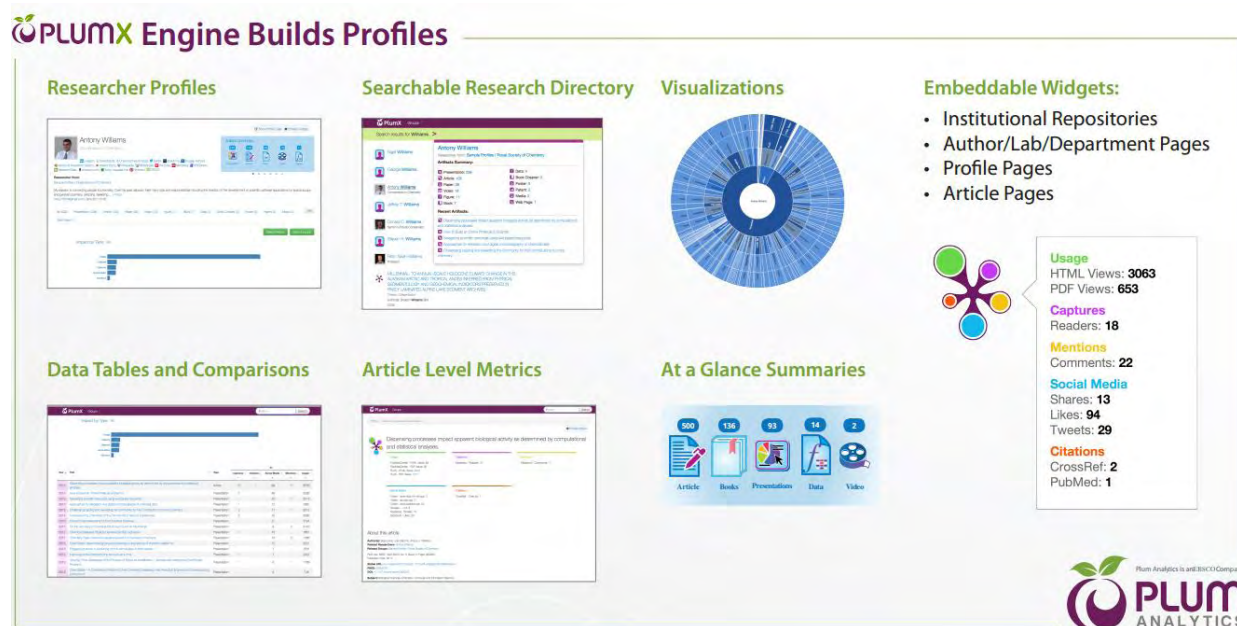


uses the information to establish a PlumX researcher profile and gather metrics about his research's impact.



**Figure 9** - The categories of metrics and types of research output that of PlumX's research tracking (source: <http://blog.plumanalytics.com/post/97218263655/orcid-and-plum-analytics-work-together-to-easily>)

As we can see in the above image PlumX metrics are based on five categories of metrics, usage, captures, mentions, social media and citations. Also, as we can see, PlumX tracks twenty and more types of research output which undoubtedly shows the power and the significance of this service. Moreover, the metrics about the research is viewable via the PlumX dashboard which include a lot of functionalities and beautiful visualizations or through one of the Plum embeddable widgets which can be used for a whole institutional repository, an author profile or an article. All this we can see in the image below:



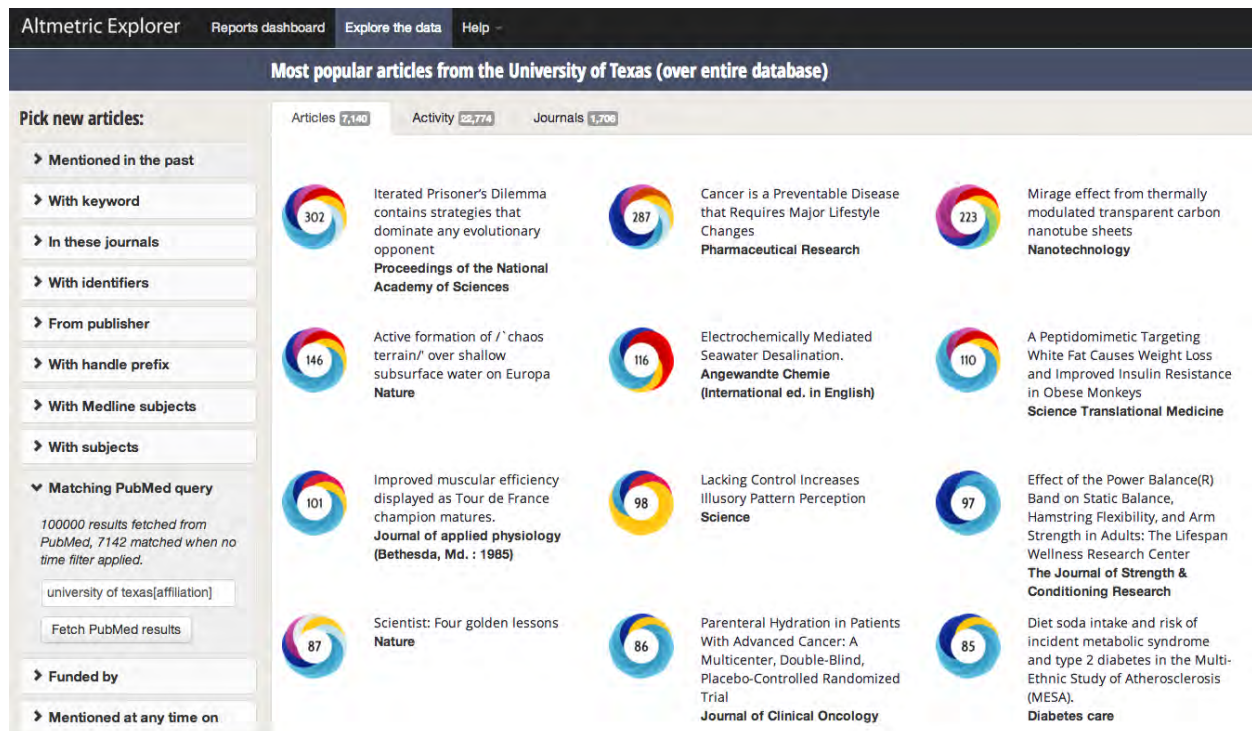
**Figure 10** - All the services that PlumX provides (source: [http://www.plumanalytics.com/downloads/ORCID\\_Poster\\_Final.pdf](http://www.plumanalytics.com/downloads/ORCID_Poster_Final.pdf))

## 4.5 Altmetric

Altmetric is another one service similar to the PlumX we saw above. Specifically, Altmetric is a London-based start-up focused on making article level metrics easy. Their mission is to track and analyze the online activity around scholarly literature. They have created and maintain a cluster of servers that watch social media sites (Facebook, Twitter, YouTube), newspapers, government policy documents and other sources for mentions of scholarly articles. Their belief is that researchers care about what people are saying about their work and need to show the impact of their papers beyond just the number of citations. Also, it is worth to mention that they track approximately five thousand papers a day, from hundreds of different publishers, preprint databases and institutional repositories. The company main services are:

### The Explorer

This service offers a panel with a lot of functionalities for the researchers. Specifically, they can monitor, search and measure conversations about their publications. In the below image we can see some of the functionalities it offers. For more information click [here](http://www.altmetric.com).



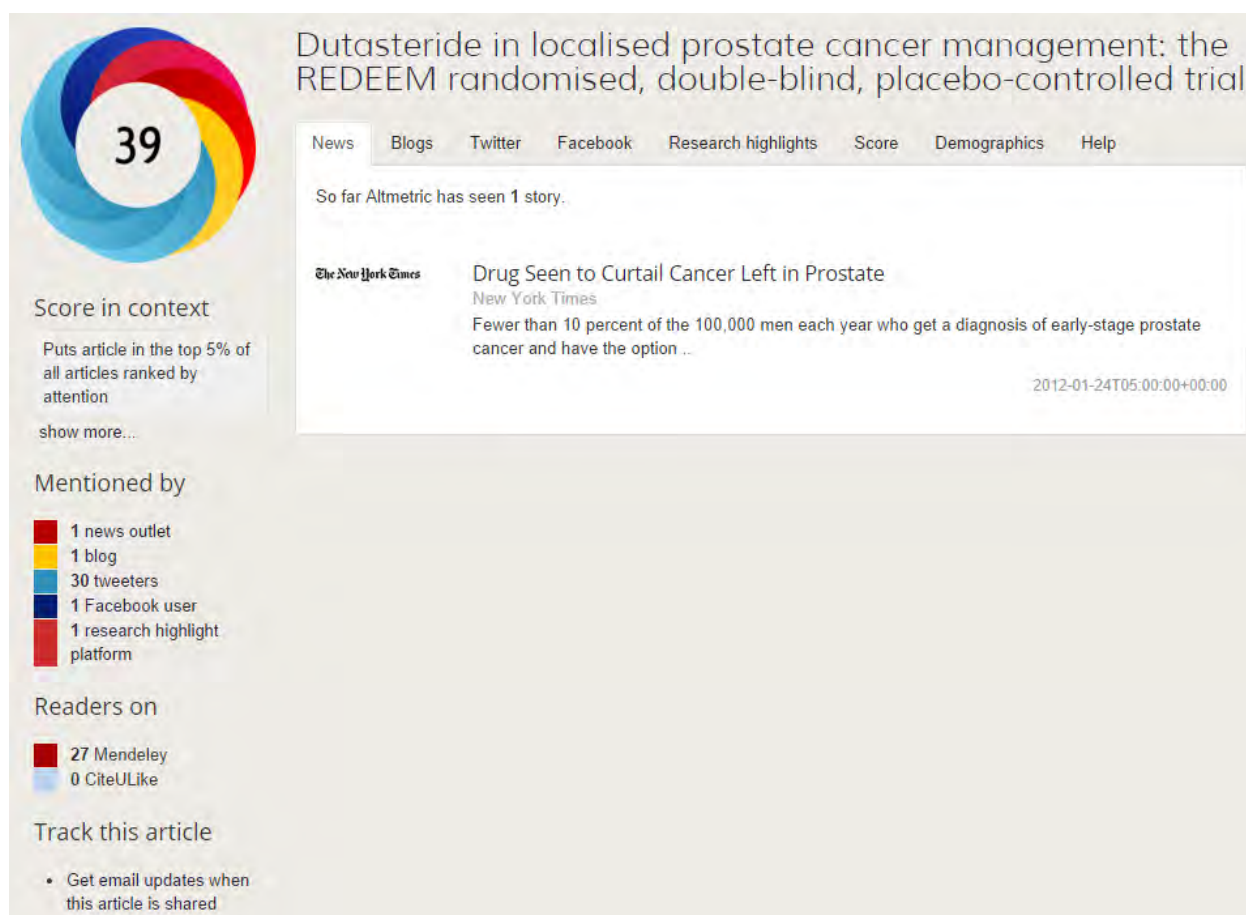
**Figure 11** - The Altmetric Explorer panel (source: <http://www.diglib.org/archives/5132/>)

## The embeddable badges

These are small badges that show the social impact of an article in a fast and effective way and can be added at any website. Each badge has a score which is a measure of the attention that article has received. This score is derived from three main factors. The number of people who have mentioned an article, the category of mentioning and how often the author of each mention talks about this article. The process of adding a badge into a website is very simple. We have just to add a `<div>` element wherever we want the badge to appear, specifying the article DOI. Also, we can customize the appearance of the badge by settings attributes on the `<div>` element. Let's look at an example. Suppose our article DOI is '10.1016/j.yuro.2012.07.006', so we add the following line of code in our website: `<div class='altmetric-embed' data-badge-type='donut' data-doi='10.1016/j.yuro.2012.07.006'></div>`. This will render the following badge:



Now, if we click on the badge we will see all the details about this badge:



**Figure 12** - The page details of a badge (source: [http://www.altmetric.com/details.php?citation\\_id=571540](http://www.altmetric.com/details.php?citation_id=571540))

Moreover, the company also offers an [API](#) (application programming interface) for the developers with a full programmatic access to data associated with articles and datasets collected by Altmetric. Finally, they offer a very interesting tool for web browsers, named [Bookmarklet](#). With this tool we can get article level metrics for any paper we read in a website. The process, is very easy. At first, we have to add the [bookmarklet button](#) to our browser's bookmark toolbar and then every time we see a paper in a website, we have just to click to the bookmarklet button we put in our browser's bookmark toolbar, and we will see all the level metrics associated with this paper. For more information about how it works watch [this](#) video.

#### 4.6 HUBzero

HUBzero is a powerful open source software platform for scientific collaboration. With HUBzero we can build scientific websites that allow researchers and educators share materials and collaborate online. HUBzero is used to create hubs for more than twenty research areas such as nanotechnology,

pharmaceutical engineering, cancer caring, biofuels, earthquake engineering, environmental modelling and more.

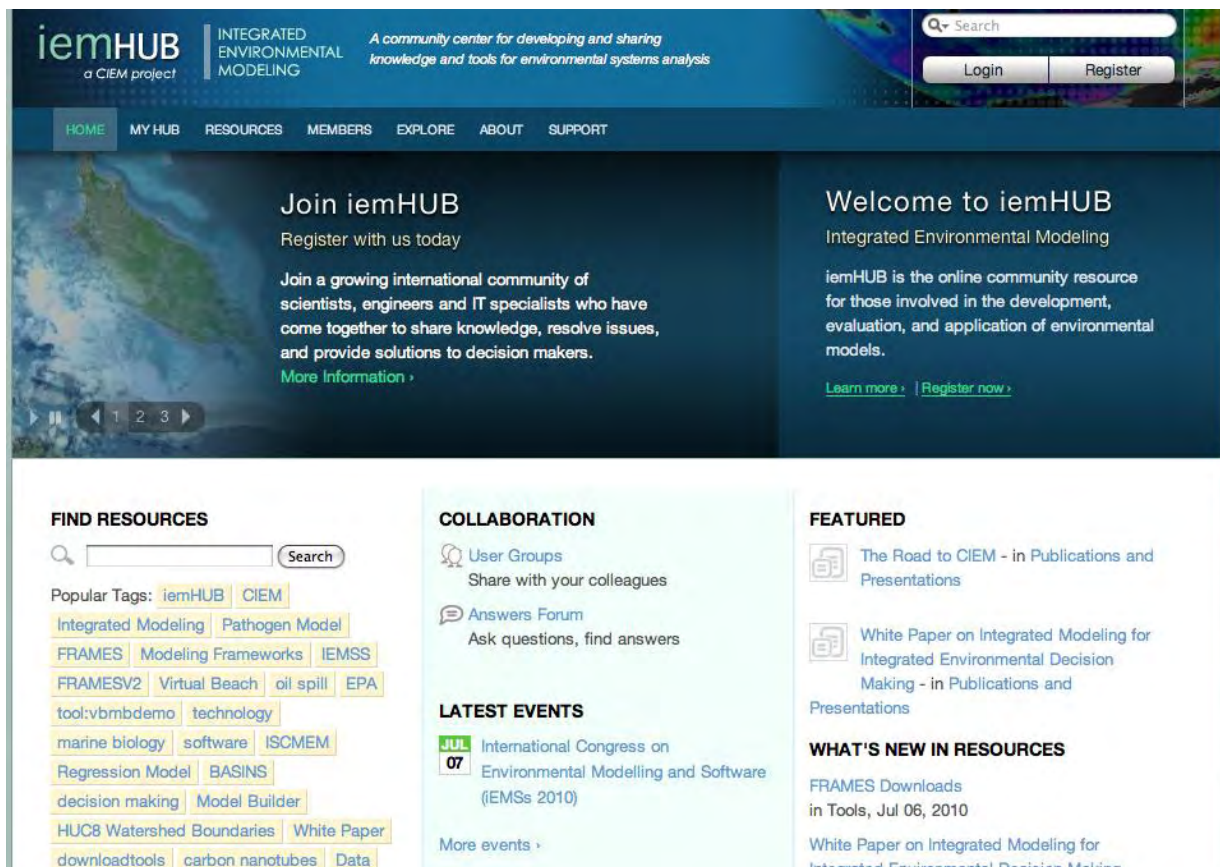
These hubs work with a way that their content is generated from users. This content can be from a complex simulation tool and animations to a seminars or even homework assignments. This interactive content helps researchers to share their scientific ideas and build their reputation by publishing online with a powerful way that hubs offer.

Hubs, are more than simple repositories. They help researchers to collaborate and build new scientific models. Hub's power is that the researcher don't have to download and install extra software since HUBzero provides fingertip access to several simulations tools that can be launched with only the click of a button. Tools that are powerful research codes that run a national grid resources. Tools with intuitive controls that are well-documented with results they're readily visualized. Moreover, using a hub we can post questions to a community and get answers from others researchers and experts from all over the world.

We can explore new areas of science by watching seminars from leading experts or we can download homework assignments and other resources. Also, we can share everything we want to any social website such as LinkedIn, Twitter or Facebook. The power of HUBzero is that as we contribute to a hub by uploading our own resources, by posting comments and reviews or answering answers to others, we are been rewarded with points. Points help us gain trust within a community and increase our reputation. Moreover, we can use our points to use premium services that HUBzero offers, or we can merchandize a new hub. We can also 'trade' our points with other users as payment for helping answering in a question or helping in a specific feature of a tool. Also, everyone can post their own ideas and other users can vote up or down. With that way, the community capture new innovative ideas and work together to create even more "bigger" ideas.

This collaboration among researchers, often leads to seminars, tutorials and other educational materials that benefit the entire community. All these materials and all this knowledge, can even help educators to share new methods and advancements with their students in order to be in touch with the latest research from a variety of fields. Final, in order to start building a hub we have just to download their virtual machine image which is available for Windows, Mac and Linux. For more information about how to build a hub just click [here](#) to read the documentation.





**Figure 13** - A community center for developing and sharing knowledge and tools for environmental systems analysis, build with HUBzero (source: <http://iemhub.org/>)

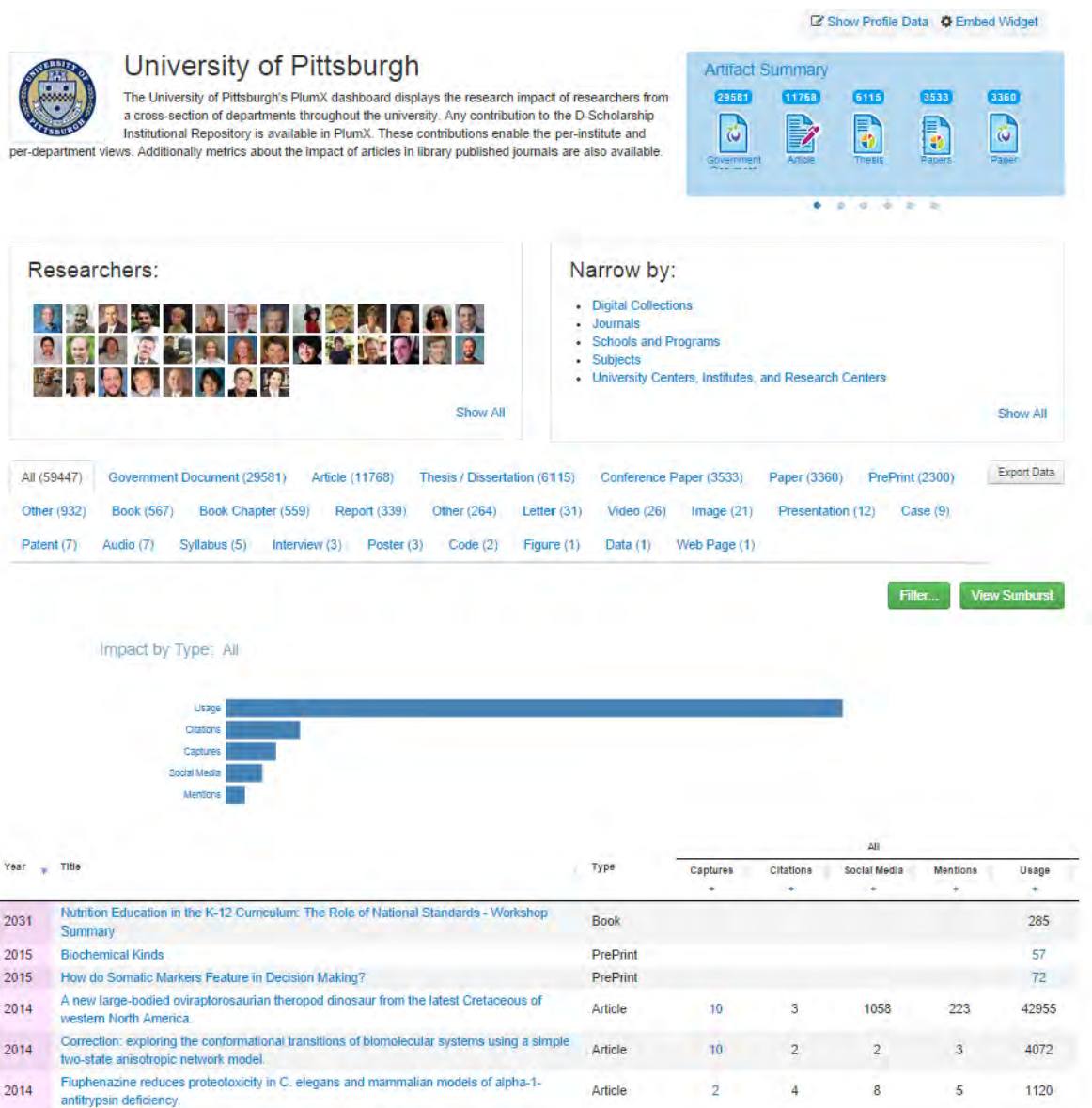
# CHAPTER 5

## 5 Implementation

### 5.1 Scientific information systems for research around the world

Before we start designing and developing our system, we took a look to find out how other universities and institutions manage and monitor their research activities. From Greece, we observed that no university has any research system for monitoring and analyzing its research activities. As for the foreign universities and institutions we liked the following three universities.

[The University of Pittsburgh](#) has built an integrated dashboard based on the PlumX and the ORCID we described above. We liked it because it is the first University which has make full use of the PlumX and the ORCID functionalities and undoubtedly is something innovative in the area of research. Through the dashboard as we can see below, they can monitor their research activity which consist of several sources such as papers, government documents, books, letters, reports and etc, providing valuable information metrics about their department research output. For each type of their documents they make full use of the visualizations and functionalities of Plumx we described in par. 4.4.3. Finally, we liked also the way they visualize each of their researcher profiles from which we see all the information that corresponds exactly to the researcher productivity.



**Figure 14** - How the University of Pittsburgh monitor its research activites (source: [https://plu.mx/pitt/g/?artifact\\_tab=all\\_artifacts&sort=year&order=DESC](https://plu.mx/pitt/g/?artifact_tab=all_artifacts&sort=year&order=DESC))

The Aalborg University has a very interesting scientific research system called VBN which registers the university's publications, research projects, research activities and press cuttings. We can have access to data about researchers, publications, research projects, activities, press clippings, research units and statistics (Figure 15). The statistics page demonstrates the university's number of publications per year and publications per type per year. As for the researcher's page, it contains a lot of information such as his publications, research projects, most frequent journals, press clippings and most frequent publishers, plus a map showing where the researcher's latest activities and conferences took place. A publication page 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. Finally, we noticed that for every publication and project, VBN offers beautiful visualizations of the people and research units involved.



**Figure 15 - VBN faculty of Engineering and Science Statistics Page**

The Hong Kong University (HKU) monitor its research activities with the HKU Scholars Hub. The Hub is very well structured offering capabilities to perform quick searches based on miscellaneous choices such

as researcher name, title of publication, type, date etc. We liked very much the way they visualize each researcher's page (Figure 16) in which we have access to the researcher's networks of collaboration, publications, achievements, grants and bibliometric scores. More specifically, in the bibliometric scores the hub shows some of the researcher's bibliometric indicators such as number of publications, citations and h-index from Scopus and bibliometric indicators from Microsoft Academic Search. Another functionality we liked, it is that we can follow a specific researcher and the system will notify us with email associated with the researcher's research activity. Finally, a publication page contains all information related to the document, including authors, published date and affiliated department(s).

The screenshot displays the HKU Scholars Hub website. At the top, the HKU logo and name are visible, along with navigation links for Home, Researchers, Publications, Theses, Grants, Patents, and Community Service. The main header reads 'The HKU Scholars Hub 香港大學學術庫'. Below this, the page title is 'HKU ResearcherPage: Yung, BHW'. The left sidebar contains a 'Profile' section with links to Contact Information, Media Contact Directory, and Professional Societies. It also lists 'Publications' (Articles: 18, Conference Papers: 45, Books: 7, Book Chapters: 15) and 'Bibliometrics' (External Metrics: Monthly Increases, Internal Metrics: Monthly Increases). The main content area, titled 'Contact Information', features a portrait of Dr. Yung, Benny Hin Wai, and his details: Title (Associate Professor), Dept. (Faculty of Education), Faculty (Faculty of Education), and Research Interests (Teaching and learning of ideas about science, Teacher belief, pedagogical content knowledge, Comparative studies in science education, Teachers and students conceptions of good science teaching, Science classroom discourse and student learning of science, School-based assessment, Students' alternative conceptions in science and the relevant teaching strategies, Teacher professional development especially uses of video for this purpose). It also lists 'My URLs' (Personal Page, ORCID) and 'Also Cited As' (Yung, Hin-wai, Benny). At the bottom, there are buttons to 'Follow by Email Alert' and 'Follow by RSS Feed'.

**Figure 16** - HKU Scholars Hub researcher page

Also, relating to works from the following former students, Athanasios Giakas on his thesis, tried to develop a very interesting scientific research system for the Department of Electrical and Computer Engineering with a lot of functionalities visualizing all the information in a beautiful and easily accessible way. Also, Paulos Kallis for his thesis project, developed an integrated publication tracker with capabilities of real time notifications. Finally, Magdalene Ntirogiannh on her second part of her thesis visualized some data associated with the department's publications and the authors' research

production. In the next paragraphs, we'll analyze the system we have developed and the functionalities it provides.

## 5.2 The objective of the project

The majority of systems we have mentioned in par 4.2 provide many opportunities and are addressing to any researchers and also to simple users who just want to monitor the progress of scientific research activities. But, for the purposes of a research institution, these systems are not appropriate and do not provide a satisfactory solution, because of the general non-customizable functionality they provide. Specifically, the information they provide is only a set of raw data which is a little difficult to be found and also to become perceptible and difficult to be understood notably from a simple viewer with no background. Also, this information, for example, for a researcher, doesn't represent all his/her research activity, since all the researcher's publications and generally his/her research work is recorded in different sources and not in only one source.

So, after having taken into account all the above factors, we decide to develop a scientific information system which will focus on the needs of research of the University of Thessaly and its members. More specifically, our goal is to develop a system that will visualize all the information which represents the scientific research profile of each department of the University of Thessaly, with an intelligent and efficient way adapted to the background and the interests of every viewer. We want our system to produce through an efficient bibliometric analysis, valuable information which aims to lead to specific qualitative and quantitative evaluation metrics which will be immediately and easily understood by any viewer. Moreover, we developed our system in a way that all the information that is visualized to be able to be generated in an automated way without any intervention.

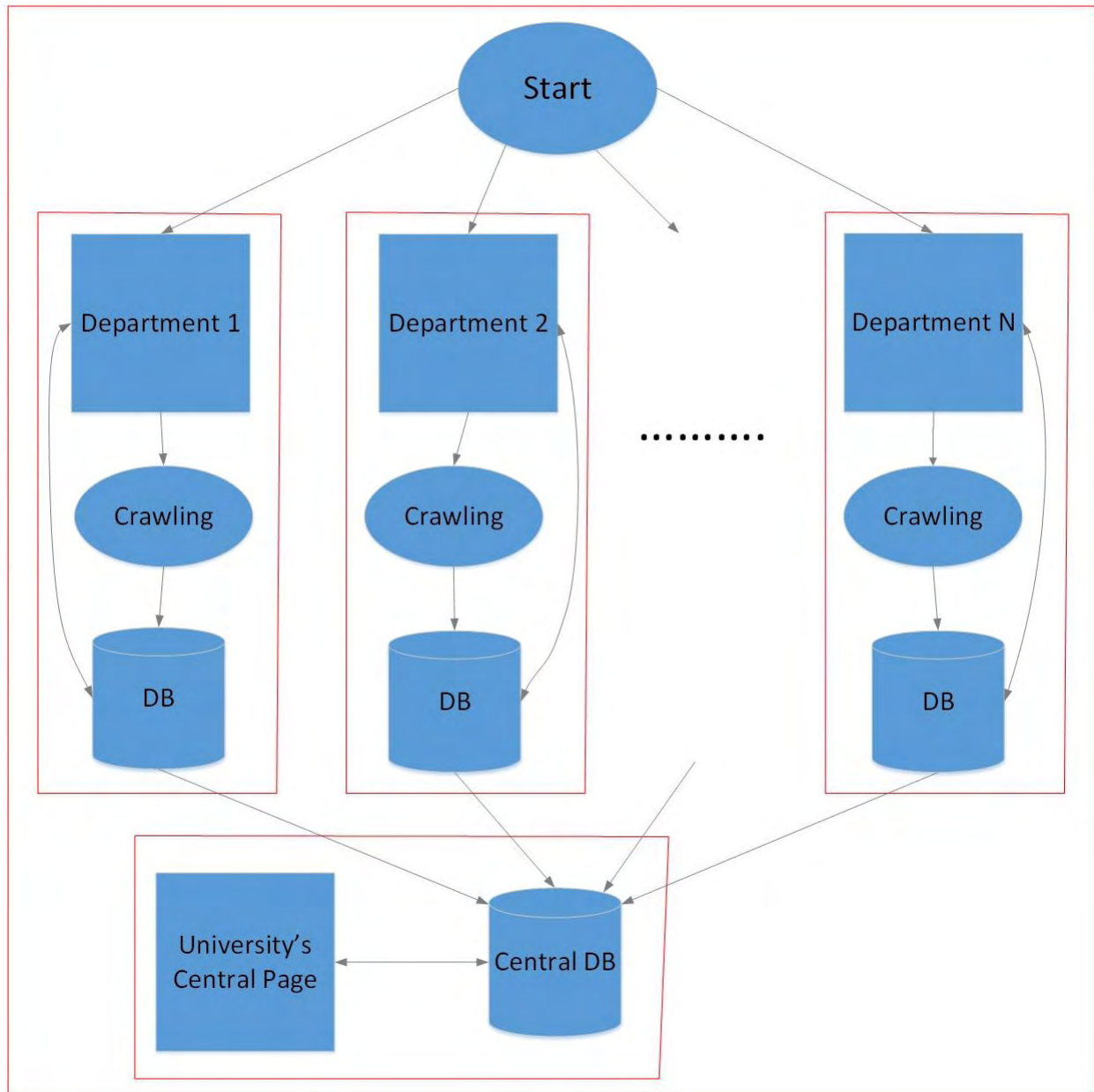
Using our system the University of Thessaly and each one of its departments will have the opportunity to have a system capable of generating all the above valuable information which undoubtedly will lead to the following benefits:

- The collection and analysis of the scientific production of the whole University of Thessaly will allow reliable evaluative comparisons with other universities and institutions. This will result in an increase of the University's reputation to the rest of the world.
- Each department will be able to monitor its research production activity and how it evolves in time. They will have the opportunity to show and highlight the scientific fields in which they have expertise, but also observe their weakness fields of research and so to be able to strengthen them in the near future.
- Each researcher of the department will have the opportunity to promote his/her knowledge and work which can lead to new network formation of knowledge that will lead to new research results.
- The research organizations that are responsible for the evaluation of research will have access to valuable information which can help them to observe and evaluate easily the progress of each department's research activity.
- Finally, it will give the opportunity to local and external companies or organizations to search for scientists with the expertise and seek for new or emerging technologies, practices and theories.

### 5.3 The architecture of the system

Our main goal was to build a full dynamic system whose content changes without any external intervention. In addition, our system was built in a way that any error, especially the part responsible for crawling, will not affect the functionality of the rest of the system. Moreover, we wanted any task that is responsible for a specific functionality and is part of the whole system, to be done from a specific isolated component. So, in order to have an easy management system and also a system that is easily maintained we develop our system with the Django Framework which uses the MTV (Model-Template-View) pattern. This pattern fitted perfectly our needs since it gave us the opportunity to build a system that could separate the logic of how our data is been displayed (Template) from the logic of the communication between the client and the server (View) and the logic of the communication with the database (Model).

As we have mentioned several times in this thesis, our primary goal was to develop the system in a way that every department of the University of Thessaly can use it by managing and monitoring its research activities. But, this goal also includes the visualization of the whole University of Thessaly's research activities which arise from the aggregation of each department's research output. In that way, we will be able to have valuable information for the whole University's research output. Let's now see an abstract representation of the system's architecture we are talking about.

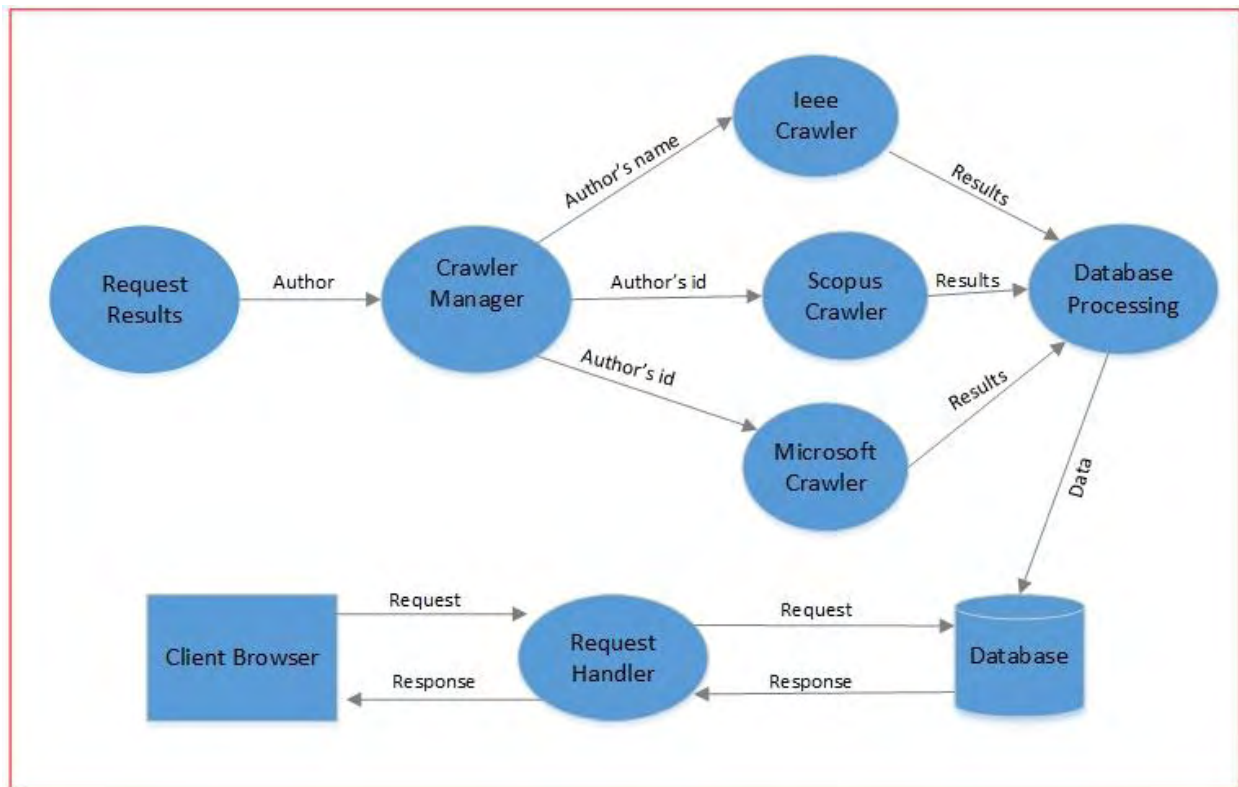


**Figure 17** - The whole architecture of the system. This architecture represents the whole system of the university

In the above figure, each component inside the vertical-rectangular red frame represents a department of the University. All these components constitute the whole system which is represented by the big external rectangular red frame. The whole process of the whole system, as we can see, starts by crawling for each department and saving the results in its own database. This process is the same for each department, since we crawl to the same three sources, Scopus, Ieee Xplore and Microsoft Academic Research. Then, after the process of crawling ends, we store in a central database some of the results. Again, this is done in the same way from each department. These results, include all the necessary information so we will be able to have a total image which represents the research output of

the whole University by combining the output of each department, as we have again mentioned above. We will see how it looks like in par 5.5 where we describe the user interface.

Let's now describe the architecture of each department's system inside the vertical-rectangular red frame in the above figure we saw and then describe it.



**Figure 18** - The system architecture of a department

As we can observe in the above architecture, our system consists of six components which we describe below.

### 5.3.1 Start Crawling

This component is responsible for starting the functionality of crawling. This functionality is need to be done automated without our interference. So in order to accomplish that we follow the following steps:

- 1) First, we created a custom management command in Django. A custom management command is a command that lets us register our own actions to our Django application beyond the existing commands that Django offers. In our case, we register our command to start crawling for each



author of the department, from our three sources, Scopus, Ieee Xplore and Microsoft Academic Search.

- 2) Then, in order that the above command can be executed in an automated way in a specific time we installed and used Cron. Cron is a time-based job scheduler used to execute commands or scripts (group of commands) automatically at a specified time and date. So, with it, we scheduled our custom management command to run every week. This scheduled task can change any time we want, just by changing the command inside the Cron file.

### 5.3.2 Crawler Manager

This component is responsible for the whole functionality of the crawling. Firstly, for each input received from the Start crawling component, it calls every crawler. For the Scopus crawler module, we give as input only the Scopus id which we find by its name. Similarly, for the Microsoft crawler the input is the author's Microsoft id. For the Ieee crawler the input is only the author's name since on the Ieee each author is not represented by a unique id. Finally, after calling each crawler for every author the component work ends.

### 5.3.3 Our Crawlers

Each of our three crawlers was implemented with python. First, since not Scopus, nor Microsoft, nor Ieee provide a good API for retrieving our data, we had to make requests and after retrieving the HTML page to parse it. But, before this, we had to study the structure of each of our sources (Scopus, Ieee & Microsoft) and then to write each of the crawler's policy. In that way we also have the flexibility to access everything we want. For the parsing we used the [Beautiful Soup](#) library, which is one of the best Python HTML parsers. With Beautiful Soup we can get particular content from a webpage, remove the HTML markup and save the information we need. So, each of our three crawlers after being called returns a list of publications that it found. This list looks like:

```
Publications = [publication_1, publication_2, publication_3.....]
```

and each publication has the following structure:

```
publication = {'title': 'pub_title', 'url': 'pub_url', 'venue': 'pub_venue', 'date': 'pub_date', 'type': 'pub_type',  
'citations': 'pub_citations', 'pub_cited': '[(publication title_1),(publication url_1), (publication  
title_2),(publication url_2)]', 'authors_affiliations':  
:[(author_name_1,author_affilaition_1,author_profile_url_1),(  
author_name_2,author_affilaition_2,author_profile_url_2)....], 'keywords': 'pub_keywords', 'doi': 'pub_doi' }
```

We must mention that, we faced the problem of publications duplicates. So in order to avoid it, from the above fields of the publication's structure, we first find the publication's title and the number of

citations for this publication and then check the database if any publication exists with this title. If not, we must ignore this publication and continue to the next one, otherwise (if it exists) we must check if the number of citations for this publication has changed, and if yes, we must update the database. We have to do this process inside the crawler module and not in the Database Processing component because if we do it on the Database Processing, we have to find the rest of fields (date, venue, keywords, authors....) which we don't need since the publication exists and will not be written in the database and so we would unnecessarily waste time which would negatively affect the performance of our system.

Another problem we faced was that due to the many requests we had to make on each server (Scopus, leee, Microsoft) the connection was lost due to limited timeouts. To overcome this issue, we had to use a [retry](#) decorator on each function in our crawler code we make request to server. So, each time a timeout error occurs, the decorator of the function catches the error (exception), which can be a socket error or a HTTP error, and try again using an exponential back-off algorithm. So, each time the decorated function finds an exception, the decorator will wait for a period of time, which time we set to grow by 10 sec with the increase of tries, and retry calling the function up to a maximum number of tries, which we set to 4. Thus, we achieved to establish a continuous connection without interruptions.

Let's now take a look at how our crawlers' structure is like.

The crawler firstly, makes an HTTP request. After retrieving the result, which is a HTML page, we parse the page with BeautifulSoup. Because the results may be many, and so they may be on different pages, we had to find the number of results and the number of pages. So, after iterating through each page we collect the information we need and then store it. Moreover, in order to find the DOI (digital object identifier) for each publication we had to make a request to <http://search.crossref.org/> in the following way: `http://search.crossref.org/?q=%s&page=1&rows=1" % title`, where title, is the title of the publication, and then parse the results and save the DOI id. So each crawler has approximately the following structure:

```
Publications = []

publication[subject_areas] = find author's subject areas

for each page do:

    for each publication on this page do:

        pub_title = find_publication_title
        pub_citations = find_publication_citations
        if exists(pub_title, pub_citations) == True:
            continue
        else:
            publication['title'] = pub_title
            publication['citations'] = pub_citations
            publication['type'] = find type
            publication['cited_by'] = find the cited publications
            publication['url'] = find url
            publication['venue'] = find venue
            publication['date'] = find date of published
            publication['doi'] = find doi
```



```

        publication['keywords'] = find keywords

        publication['authors_affiliations'] = find authors and their
affiliations

        Add publication to Publications

return Publications

```

#### 5.3.4 Database Processing

The role of this component is to collect all the publications which every crawler returns and save these to our database after communicating with the corresponding models. Moreover, for each table we want to insert data, we first check if the data exists so as to avoid duplicates. As we mentioned in par 5.2.3 these publications look like:

```
Publications = [publication_1, publication_2, publication_3.....]
```

So, for each list of publications we read all its fields and write it on our database.

#### 5.3.5 Request Handler

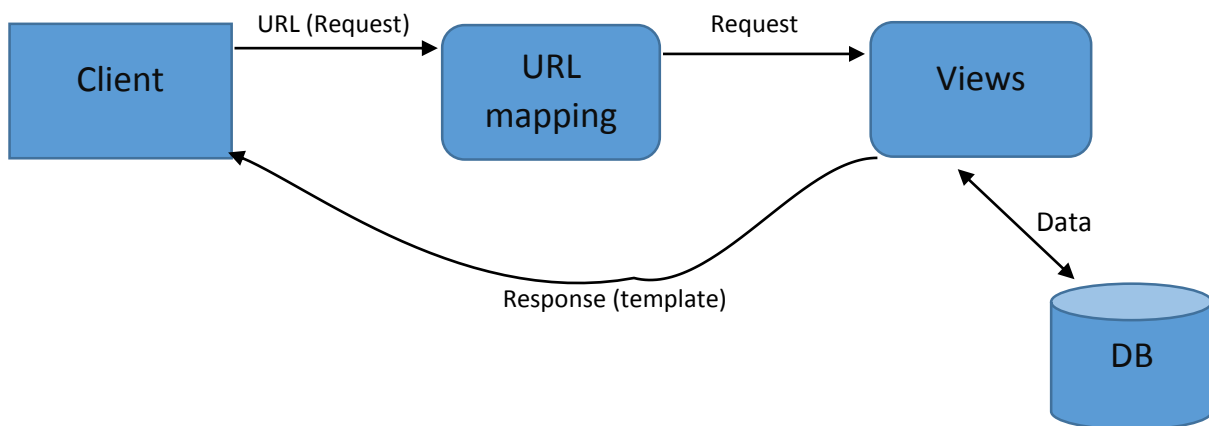
This component is responsible for the logic about the communication between the client and the server. For each client's request, the Request Handler processes this request by executing specific tasks which the request demands or querying the database if necessary, after communicating with the corresponding models. Then, the Request Handler, after collecting all the data, create the response and then load this response into the appropriate template which contains all the logic of data presentation of the client's request. Moreover, with the power of Django MTV pattern, our system was built in a way, that the above logic to be represented with a unique view, which has respectively its unique URL. Let's see all of our views and their URL with a description about what the view does.

| URL               | View             | Description                                 |
|-------------------|------------------|---|
| /home             | home             | Show the home page                          |
| /publications     | publications     | Show the publications search page           |
| /all_publications | all_publications | Show all the publications of the University |
| /all_citations    | all_citations    | Show all the citations of the University    |

|   |                             |  |
|---|-----------------------------|--|
| /publications/search                            | publications_search         | Show the results of a searching  |
| /publication/keywords/{keyword_value}           | publications_keyword        | Show the publications which contains the keyword clicked {keyword_value}       |
| /search_by_keyword                              | home_search_by_keyword      | For searching of a publication which contains a keyword (home page)            |
| /collaborations                                 | collaborations              | Show the collaborations page   |
| /authors  | authors                     | Show all authors   |
| /authors/{author_name}                          | author_profile              | Show the profile of an author  |
| /authors/{author_name}/keywords/{keyword_value} | author_publications_keyword | Show the publications of the author which contains the keyword {keyword_value} |
| /citations                                      | Citations                   | Show all the publications that cite the department's authors                   |
| /about  | about                       | Show the about page  |
| /access_db                                      | access_db                   | For Ajax requests from JavaScript  |
| /admin  | admin                       | For accessing to admin page  |

**Figure 19** - A description about all of our system's views

All the above client's requests (URLs), as we mentioned, are handled from the Request Handler which is represented by the below figure:



**Figure 20** - Showing how a client's request is handled – Request Handler

### 5.3.6 Visualization Handler

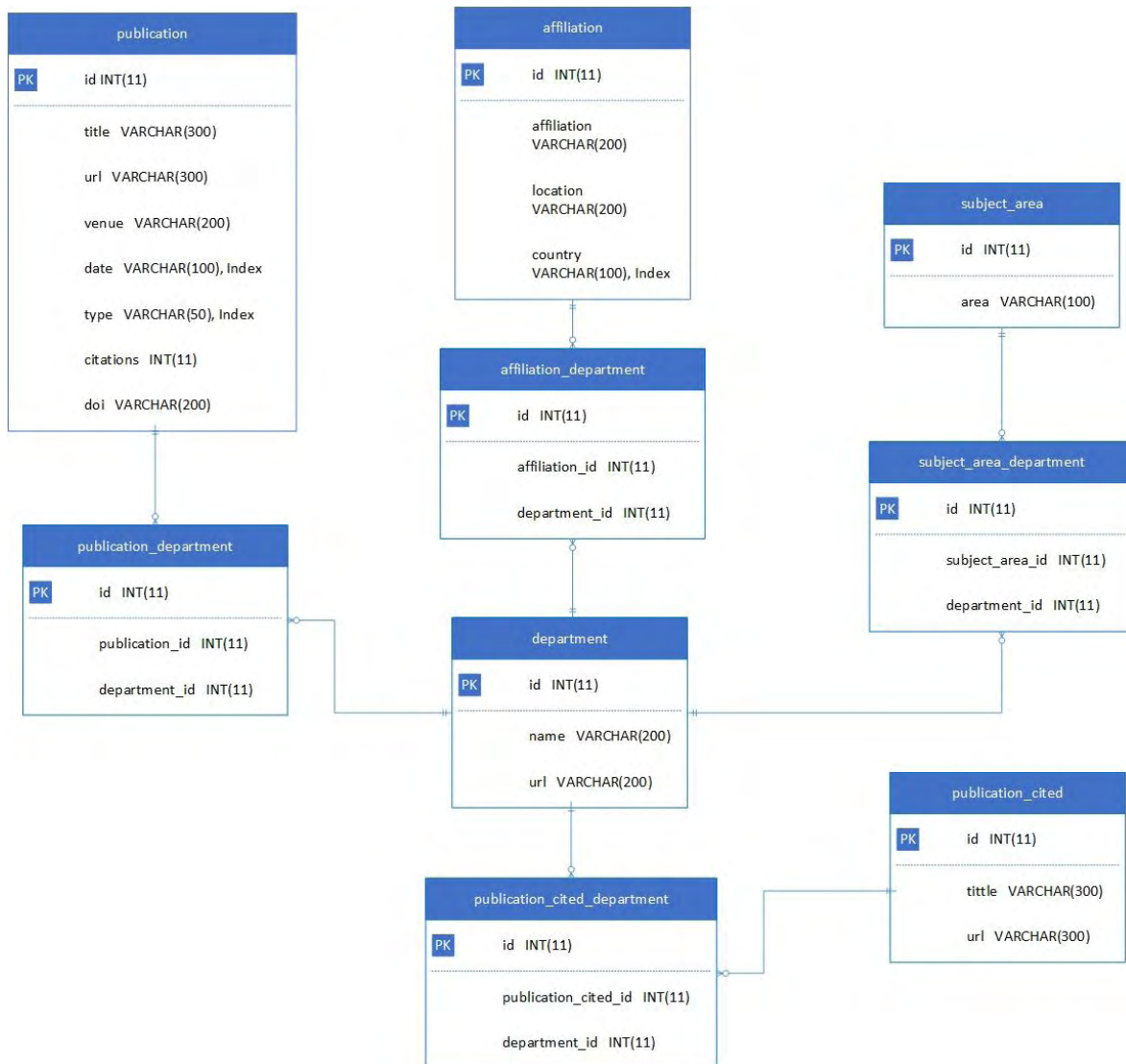
This component is responsible for the majority of the visualizations we use in our system. More specifically, for each visualization we make a request to our database through Ajax from our JavaScript files to retrieve all the information which is necessary for the visualization. Next, having that information we can draw the visualization. With this technique, all the process of visualization is done in the background and the page does not have to be refreshed.

## 5.4 The databases of the system

Our system mainly depends on the data that is stored on our databases. So, in order to have a system that its data is easily accessible and a system that has a fast response on client's queries, we design our databases considering this need.

As we mentioned in par 4.3 our primary goal was to have a total image that represents best the whole research activity of the University. So, our whole system consists of two databases, a central database for the central page of the University and the database of each department. As for the central page in which only a part of the research output of each department is stored, we decided to store only the necessary information which is sufficient to represent the above image. This information, is the total number of citations that all departments' publications have received in time, the type of publications produced by all departments in time, the affiliations around the world and the research areas on which the University's researchers are working. Next, we describe our databases.

### 5.4.1 The database of the University's central page



**Figure 21** - A representation of the central page's database

Let's now describe each table of the above database architecture:

**Publication:** This table represents each publication. Each publication is represented by its title, its url, its venue, its date of publish, the number of citations it has, its DOI (document object identifier) and from its type, which may be conference, journal, book or other.

**Affiliation:** This table represents an affiliation that the University is collaborate with. Its fields are affiliation which is the name of the affiliation, location, which is the coordinates (longitude and latitude) on the map and a country, which is the country where the affiliation comes.

**Subject area:** This table represents a research area that the University's researchers are working. Its field is area, which is the value of the subject area.

**Publication\_cited:** This table represents each publication which cites a publication of the University.

**Department:** This table is used to represent the each department of the University. Its fields are name, which is the name of the department and its url.

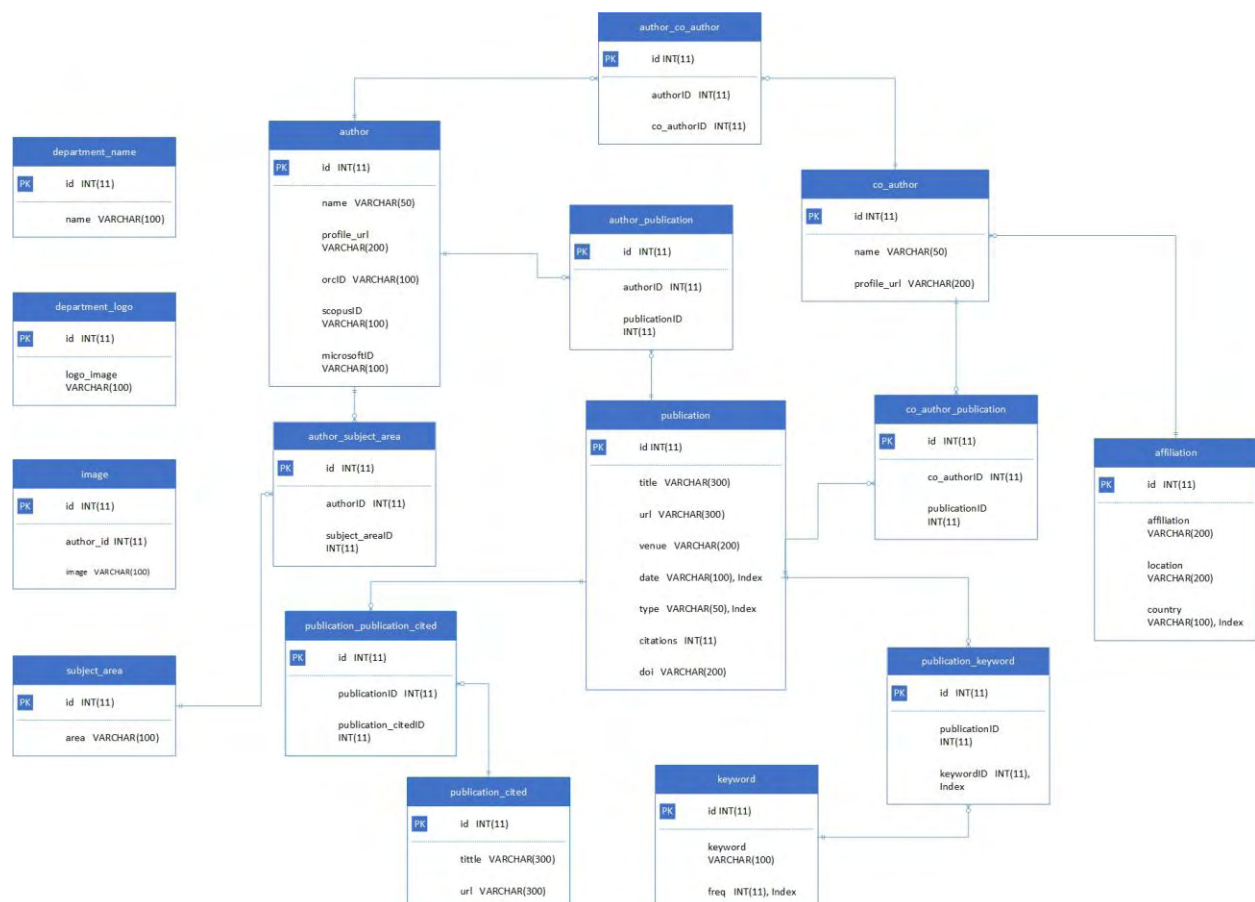
**Publication\_Department:** This table is used to describe the many to many relationship between a publication and a department. Having this table, we are able to know the number of publications that each department has.

**Publication\_Cited\_Department:** This table is used to describe the many to many relationship between a cited publication and a University's publication of a department. Each cited publication may cite many departments' publications and a cited publication may cites many publications.

**Subject\_Area\_Department:** This table is used to describe the many to many relationship between a research area and a department. Having this table we are able to know the research areas of each department.

**Affiliation\_Department:** This table is used to describe the many to many relationship between an affiliation and a department. Again, having this table we are able to know the affiliations which each department collaborate.

#### 5.4.2 The database of a department



**Figure 22** - A representation of the department's database

Let's now describe each table of the above database architecture:

(In the above database we don't show the Django's tables such as the table of users, groups etc..)

**Author:** This table represents each author of the department. Each author of the department is represented by a full name, a profile url, an orcid id, a scopus id, and a microsoft id.

**Co Author:** This table represents each collaborated author that the department is collaborated with. Each collaborated author is represented by a full name and a profile url, which it can be from scopus or microsoft.

**Publication:** This table represents each publication that is written from the department's authors. Each publication is represented by its title, its url, which is from which source (ieee or scopus or microsoft) it comes, its venue, its date of publish, the number of citations it has, its DOI (document object identifier) and from its type, which may be conference, journal, book or other. The fields date and type are indexes in order our system to have a better performance for queries about dates and types.

**Affiliation:** This table represents the affiliation of each collaborated author. Its fields are: affiliation which is the name of the affiliation, location, which is the coordinates (longitude and latitude) on the map and a country, which is the country where the affiliation comes. The field country, is an index since in order to handle better queries about countries.

**Keyword:** This table represents each keyword which is used in a publication. Its fields are: keyword, which is the keyword value, and freq, which represent how many times this keyword is used. Although, we have the table publication\_keyword, we use the freq field in order to have a fast response about a query about how many times a specific keyword is used. Also, the field freq is an index for a better performance.

**Subject area:** This table represents each subject area for an author of the department. Its field is area, which is the value of the subject area.

**Publication\_cited:** This table represents each publication which cites a publication of the department.

**Author\_Subject\_area:** This table is used to describe the many to many relationship between an author and a subject area. Each author may has many subject areas and a subject area can may be used from many authors.

**Author\_co\_Author:** This table is used to describe the many to many relationship between an author of the department and a collaborated author. Each author can collaborate with many authors and a collaborated author can collaborate with many authors.

**Author\_Publication:** This table is used to describe the many to many relationship between an author and a publication. Each author may has written many publications and a publication may has be written from many authors.

**Co\_Author\_Publication:** This table is used to describe the many to many relationship between a collaborated author and a publication. Each collaborated author may has written many publications and a publication may has be written from many collaborated authors.

**Publication\_Keyword:** This table is used to describe the many to many relationship between a publication and a keyword. Each publication may has many keywords and a keyword can may be used from many publications.

**Publication\_Publication\_Cited:** This table is used to describe the many to many relationship between a publication and a cited publication. Each publication may cited from many publications and a cited publication may cites many publications.

**Co\_Author\_Affiliation:** This table is used to describe the one to many relationship between an affiliation and a collaborated author. A collaborated author can belong only to one affiliation and an affiliation may be used from many collaborated authors.

**Department\_name:** This table is used to represent the department's name. We need to store this value because when our crawlers are storing data to the central database we must associate this data with the corresponding department.

**Department\_logo:** This table is used to represent the department's logo. Again, we need to store this value since each department has its own logo image.

**Image:** This table it used to store the photo profile of each author.

## 5.5 Functionalities

Let's now refer in brief to all functionalities our system provides which a viewer can perform when visiting our system:

- View the whole university's research profile from which a viewer can get valuable information which is visualized with full interactive visualizations representing the whole research activity of the University and how it evolves in time.
- View a department's research profile from which a viewer can explore valuable information about the department's research activity and how it evolves in time.
- View a researcher profile. Through the researcher's profile the viewer can get information that represents in deep the researcher.
- Using an integrated search engine for searching publications of the department. The viewer can select various criteria depending on his/her interests or background.

## 5.6 User Interface

Our goal was to design a friendly user interface from which any viewer regardless of his background, can be able to get valuable information and knowledge about what he sees. All visualizations that we use were designed and developed considering the above purpose. Also, in our effort to design this interface we liked many features of Scopus, Microsoft Academic Search and [CitEc](#) (a system for citations in economics). Especially, we liked the way they visualize the profile of a researcher and so we were

influenced. Finally, it is worth noting, as we have mentioned again on this thesis, that every visualization and bibliometric indicators we present are generated dynamically on the server side, and are not static information.

Next, we analyze in deep all the functionalities of our system and what a viewer can explore through the visualizations we use.

## 1. University's central page



**Figure 23** - The home page

In the University's home page we are visualizing the whole University's research activities, which arises from the collection of each department's research output. More specifically, we can see the total number of publications, citations, collaborations, countries and research areas.

Also, the two bar charts exactly below from the red arrows, visualize the number of citations that the University's publications have received in time, the first one, and the second one, the type of publications produced by the University in time. We will examine what happens if we click on this visualizations and numbers in the next paragraph we describe the department's home page, since each department's home page looks the same with the University's home page, except the division which contains all the University's departments, as the last red arrow indicates at the bottom in the figure above. On that division, we see all the departments and if we hover one of them we will see the number of publications and citations of the department and we can click to the button that appears to visit the department's page. Next, we describe what we will see if we click to visit a department.

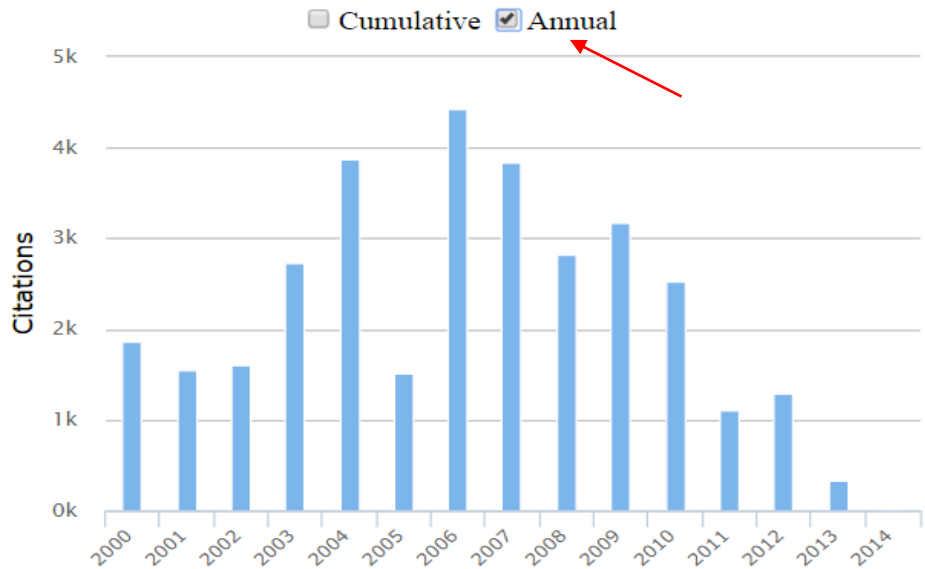


## 2. Department's home page

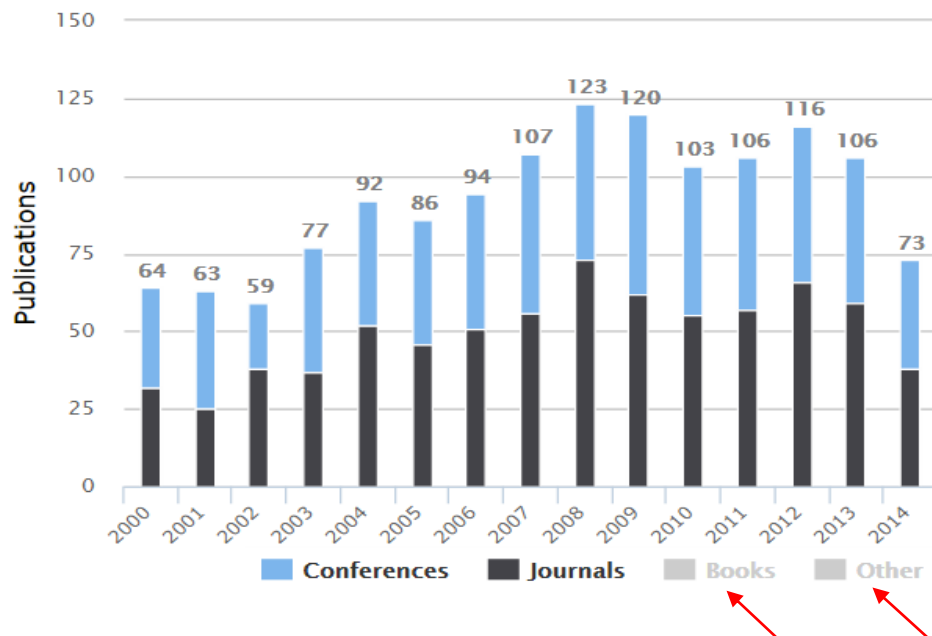


**Figure 24** – The department's home page

As we can see from the above figure, we must focus again on the points that the red arrows indicate. Let's start from the two main charts we see, from which was implemented with the Highchart library. The first one, on the left, visualize the number of citations that the department's publications have received in time cumulatively using a bar chart pattern. If we go the mouse over a column, a box will appear showing the number of citations for the specific year which corresponds with the column the mouse hovers. As we can see, on the top of the chart, we have two choices to select. The first one, "Cumulative", is the default value, but if we click on "Annual" we will see that the chart is changed, showing this time the number of citations the publications have received, annually and not cumulative:

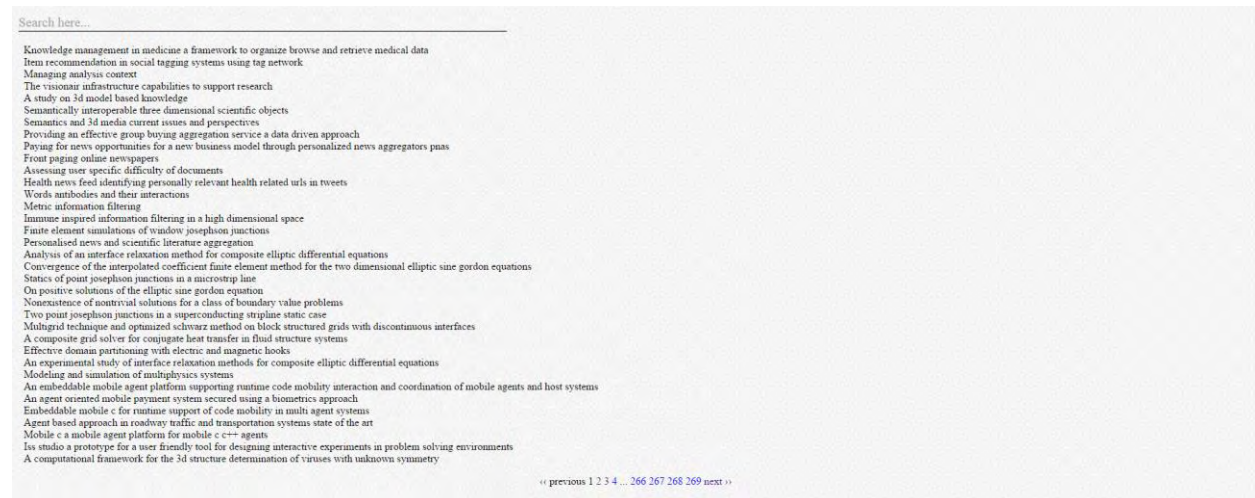


The second one, on the right, visualize the type of publications produced by the department in time. The type as we can see can be conference, journal, book or other. In addition, the visualization let us click on any type of the above for better comparing. For example, if we click on “Books” and “Other” we will see the following:



Finally, from the department’s home page, we can search for publications based on a keyword. When we type a character a division appears containing all the stored keywords which starts with this character and if we click to search will see all the publications which contains the keyword we typed.

Let's see now the other options a viewer has on the home page. If we click on the total number of publications we will navigate to the publications' search page which we describe in the next paragraphs of this charter. Now, if we click on the total number of citations, which is the total number of citations of all publications, we will navigate to the following page:



**Figure 25** - All the publications that cite the department's publications

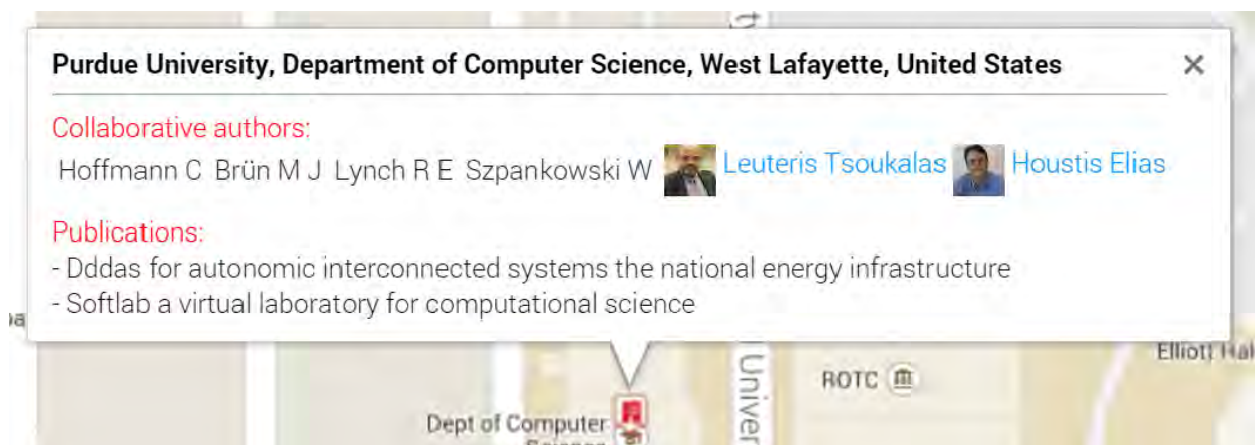
On the above page, we see a list of all publications that cite all the department's publications. If we click on any of these we will be navigated to the source that this publication belongs to. In addition, at the top of the page we can search for a specific publication and also we have added the pagination functionality that Django offers for faster loading of the page's content. Finally, we used the pagination that Django offers for a better structure of the page.

Let's see now the other options we have. If we click on the number of collaborations, which is the total number of affiliations around the world that the department is collaborated with, we will see the following:



**Figure 26** - The collaborations page. We can see on the map all the affiliations what collaborate with the department

As we can see above, on the right side, we see on the map all the affiliations. Also, we make use of google map clustering, which grouped all the affiliations that are close together on the map based on their distance. Now, if we zoom in and click on a marker we will see the following:



**Figure 27** - The content of a maker

As we see above, the marker's content consist of the title of the affiliation, the collaborative authors and a list of publications. On the collaborative authors, with black color, we refer to the collaborated authors names who belong to this affiliation. On the other hand, with blue color and their photos we refer to the

departments' authors who are collaborated with this affiliation and more specifically, they have written the list of publications which are shown. Moreover, we can click on the authors name to see their profile and also to the publications. It is worth to say, that every time a marker is clicked we load its content with AJAX. With that way we achieved to have better performance when the map and its markers load. Otherwise, due to the big number of markers, we would expected a big delay. Finally, on the left side of the page, we see all the affiliations, with each one to consist of its name and a number, which refer to the times that this affiliation is collaborated with the department. Also, we can click on an affiliation and with navigate to the map where this affiliation is.

Another visualization we have on the home page is the number of countries, which refer to the total number of countries over the world from which the affiliations the department collaborated with, came from. For this visualization we use the Google charts library. If we click on it we will see the following map:

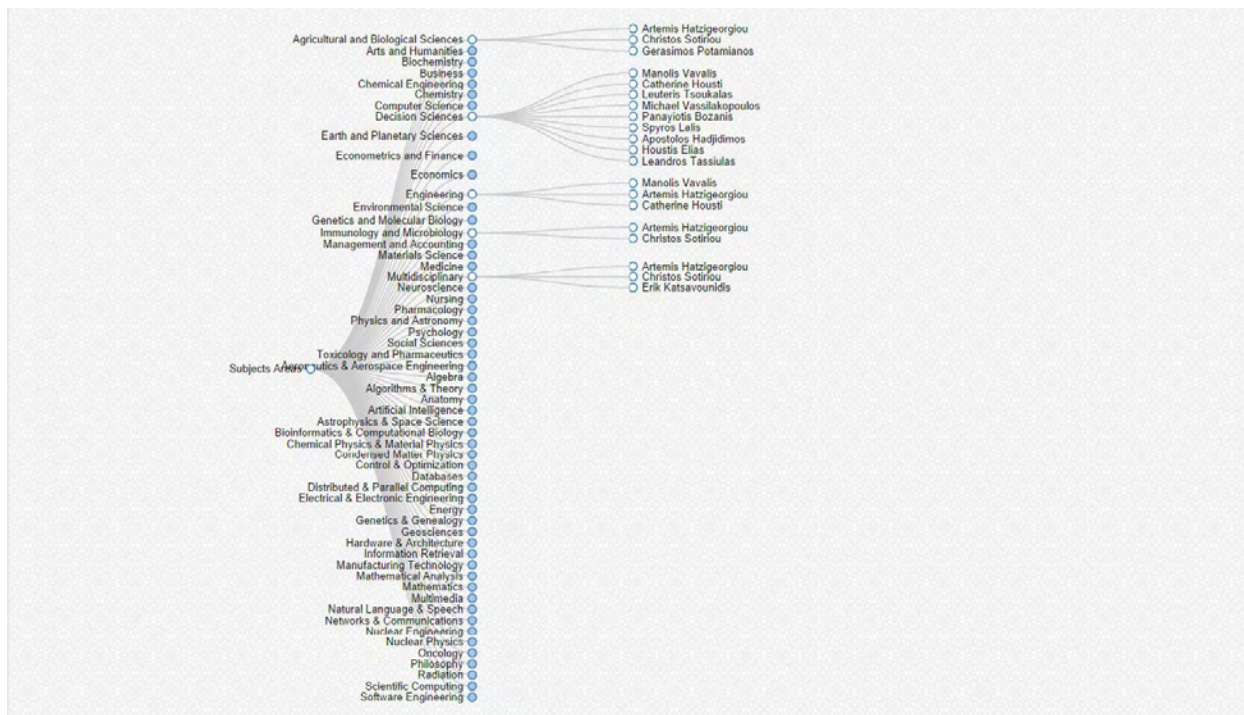


**Figure 28** - All the counties of the affiliations collaborated with the department

As shown on the above map, the color intense of a country, which range from 1 to 365, indicates the number of affiliations for this country. In addition, if we hover a country with the mouse we will see a box showing the number of affiliations (collaborations). Moreover, we must say that on the map are visualized only the countries which contains affiliations that collaborate with the department. So, if we hover a country that doesn't contain any affiliation, nothing will happen.

Finally, another interactive visualization a viewer can use on the home page, is the number of subject areas. For this visualization we use the D3.j library. This number indicates the total number of research areas of the department's authors. If we click on it we will see the following visualization:





**Figure 29** - The research areas of the department

In the above visualization, if we click in a subject area will see the department's authors who are working in this area. Also, if we click on an author name we will navigated to his profile.

## 2. Publications search page

Document title:

Document's title...

Document's keywords:

associated tool x

Document's authors:

Manolis Vavalis x

Document type:

Number of citations:

(between 0 and 1068):

Published:

From: To:

Search Reset

Top publications :

- Breast cancer classification and prognosis based on gene expression profiles from a population based study with 1068 citations
- Gene expression profiling in breast cancer understanding the molecular basis of histologic grade to improve prognosis with 872 citations
- Maximum lifetime routing in wireless sensor networks with 690 citations
- Energy conserving routing in wireless ad hoc networks with 689 citations
- Validation and clinical utility of a 70 gene prognostic signature for women with node negative breast cancer with 603 citations

Top keywords :

Used 115 times

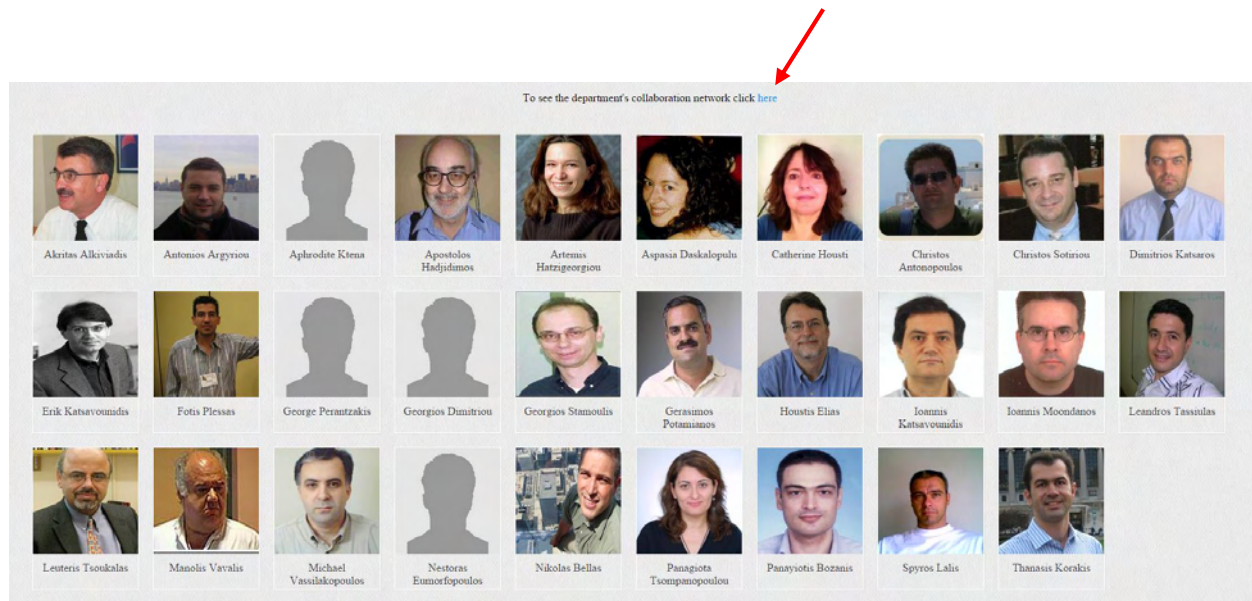
memetic algorithms epidermal growth factor receptor 2 estrogen receptor mathematical models breast cancer wireless sensor networks computer simulation predictive location tracking in cellular and in ad hoc wireless networks optimization gravitational waves

**Figure 30** - The publications search page

This page provides a full search functionality for retrieving publications which are stored on our database. More specifically, the user can search base on: publication title, publication keywords, publication authors, type of publication, number of citations and the published year of the publication. When the user is going to fill a keyword for a publication and is typing a character, a box containing all the keywords that start with this character, will appear. The same functionality exists if he fill a publication's author name.

In addition, on the right side of the publication search page, we can see the top five publications and just below the top ten keywords. If we hover a keyword with mouse, we will see the total number that this keyword is used. Also, we can click on any of these keyword, and will see all the publications that use this keyword.

### 3. Authors page

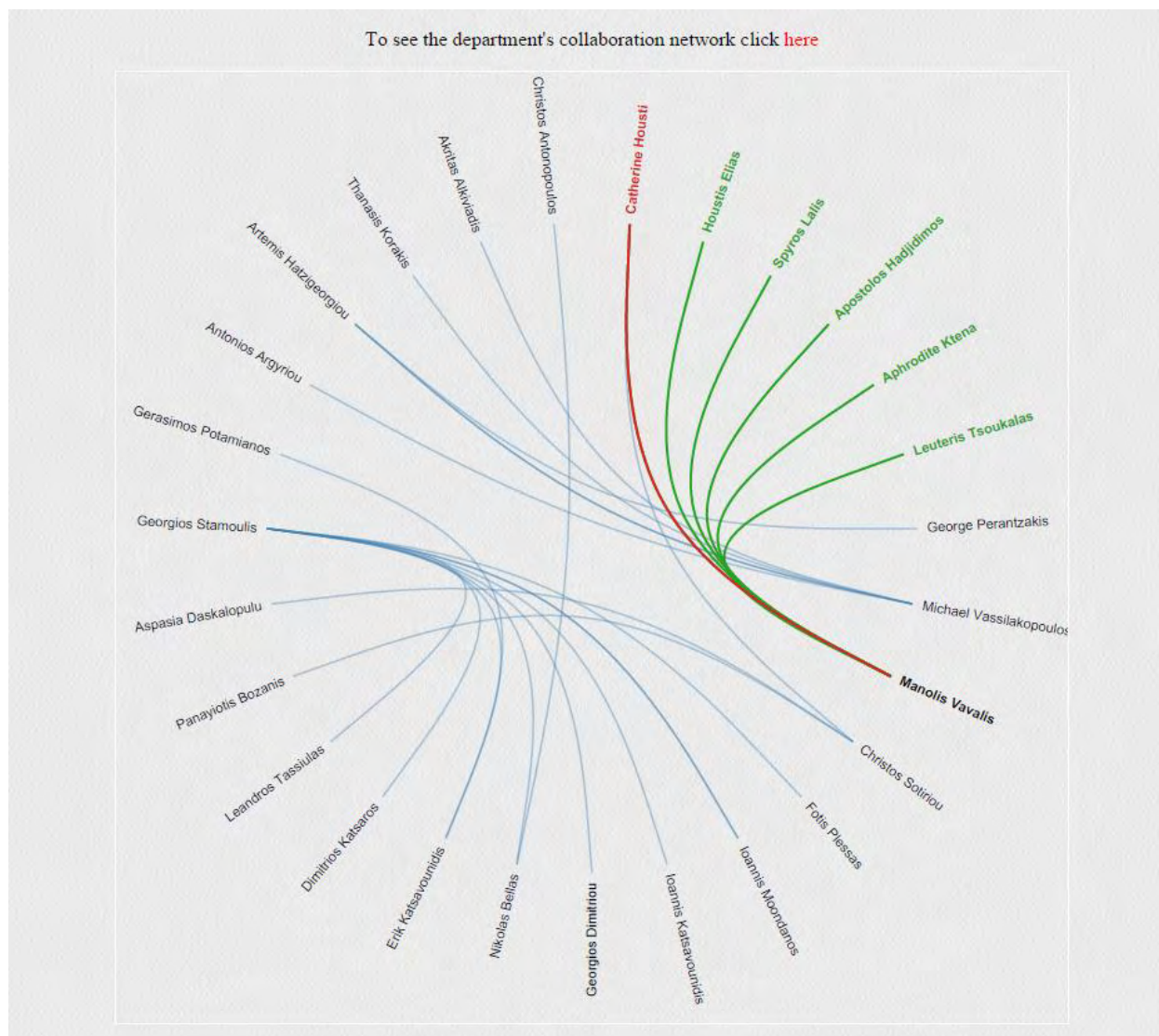


**Figure 31** - The authors' page which shows all the authors of the department

On the authors page we can see all the department's authors ordering by their names. If we hover with our mouse an author's photo profile we will see an animated window sliding down containing the number of publications, the number of citations for this author and a button which if we clicked navigate us to his profile which describe next. We decided to put the number of publications and the number of citations because these are the most significant bibliometric metrics that describes best a researcher.

Now, if we click on the point that the red arrow indicates in the above figure, we will see an animated division to appear containing the department's collaborations network (figure 31). On this visualization, which we develop with the D3.js library, we try to visualize the departments' collaborations among its researchers. As shown in the below figure, if we hover with our mouse a researcher name, we will see with bold red and green lines the researchers he or she is collaborating.





**Figure 32** - The collaboration network of the department's authors

#### 4. Author page

When we visit an author profile, we will see the following window:



**Figure 33** - The author main page

As we can observe from the above image, at the point the blue arrow indicates, each author is described by the following metrics:

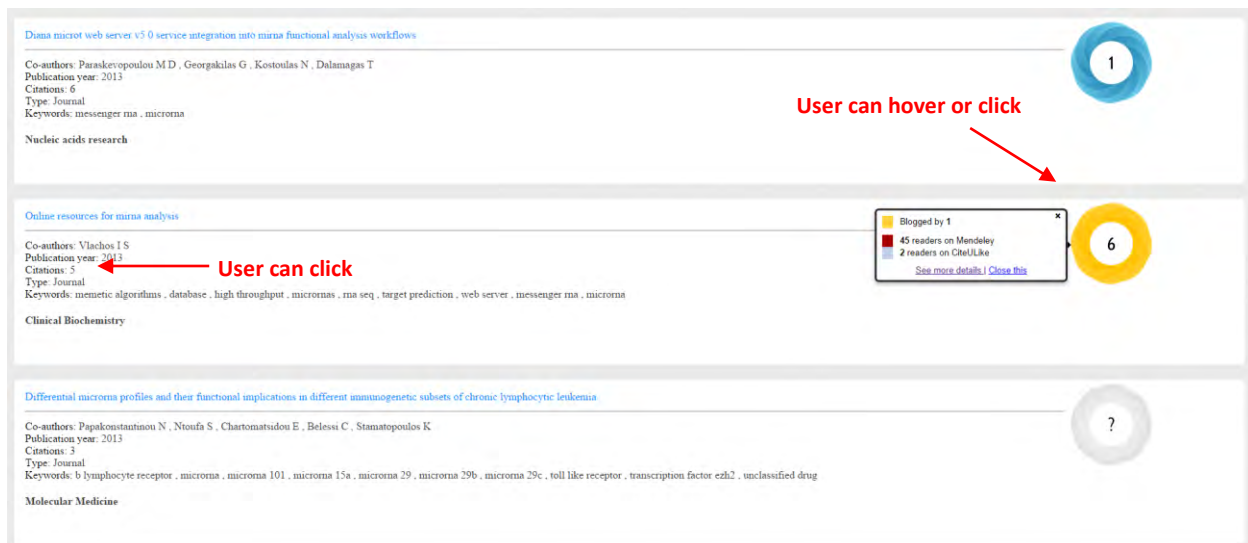
- Publications: the total number of publications of the author.
- Citations: the total number of citations the author has received so far.
- H-index: the h-index value of the author.
- Citation impact: the citation impact value of the author. As we described on par 4.1.1 this indicator indicates the average number of citations per publication and is calculated as the ratio of the number of references which are listed in a certain period of time to a total number of publications of the same period. In our case, we calculate it for the past ten (10) years.
- Collaborative authors: the total number of collaborative authors the author is collaborated with so far.

Just below of all the above metrics we can see also all the author's subject areas he works. Let's now describe the all the options we have, on the points the red arrows indicate. The first one, the "Main Data graphs", which is the default when we visit the author's profile, when clicked we can see just below a window which shows four interactive visualizations, which we designed with the Highchart library:

- Cumulative publications published: This bar chart shows the author's publication production in time.
- Citations received: This line chart shows how the author's number of citations is evolved in time. Also, if we hover a specific point-year we can see the number of citations what corresponds with this year.

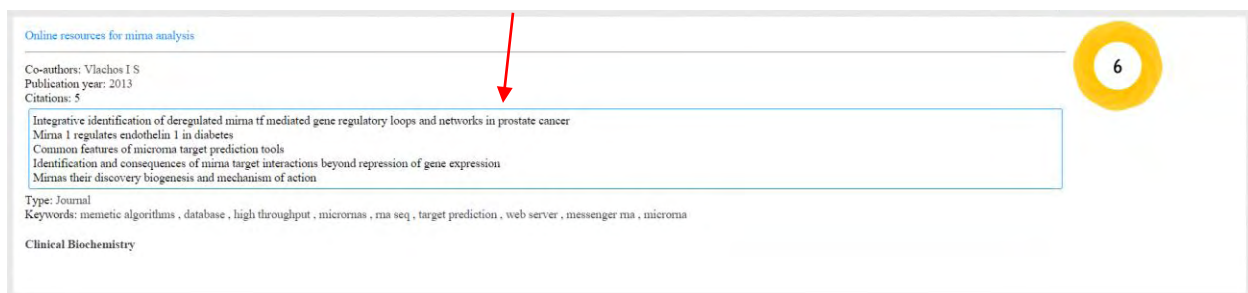
- **Publications type:** This pie chart is used to show what type of publications the author published. The type can be conference, journal, book or other. Also, we can click on any type and will see that it is separated from the others.
- **Production by publication type:** This stacked column chart shows us how the author's production type is evolved in time. From this chart, an author can get valuable information about what type of publication he is published in time.

Now, if we click on the “Publications” button we will see all the author’s publications which we have ordered by date and are visualizing with the following way:



**Figure 34** - How the author’s publications looks like

From the above image we can see that each publication consists of its title, its authors, its number of citations, its type, its venue, its keywords and finally its Altmetric badge on the right side. If we click on the title of the publication we will navigate to the source from we crawled it. Similarly, if we click on an author name we will navigate to his profile page. It is worth to mention that if we click on the number of citations of a publication we can see which publications cite this publications. So, if we click on the where the red arrow indicates on the above image, we will see the following box appearing:



**Figure 35** - What we see if we click on the number of citations



2 publications which contain the keyword "dna sequences"

**Knowledge based tdn architectures for features recognition in dna sequences**

Co-authors: Potamian G, Papamkelov E  
Publication year: 2001  
Citations: 0  
Type: Conference

**Proceedings of the International Joint Conference on Neural Networks**

Keywords: genetic algorithms , backpropagation , computational methods , database systems , **dna sequences** , feature extraction , target genes , knowledge based systems

?

**Feature recognition on expressed sequence tags of human dna**

Co-authors: Reczko M  
Publication year: 1999  
Citations: 0  
Type: Conference

**Proceedings of the International Joint Conference on Neural Networks**

Keywords: cloning , **dna sequences** , fault tolerant computer systems , automated pattern recognition , target proteins , statistical methods

?

**Figure 37** - The result if we click on a keyword

Finally, let's see the last two visualization from the last two buttons, the "cited by" and "Collaborators".

### Cited by:

If we click on the "cited" button, we will see a scrolled window which contains all the publications what cite the author:

Here we can search

Search here...

- Let represent dna elements via human proteins interaction
- Integration of microRNA databases to study microRNAs associated with multiple sclerosis
- High content imaging of presynaptic assembly
- Microna 137 regulates a glucocorticoid receptor dependent signalling network implications for the etiology of schizophrenia
- Mir 21 is an up regulated microRNA that supports ngf signaling and regulates neuronal degeneration in pc12 cells
- MicroRNAs in neuronal communication
- Rna binding proteins a common denominator of neuronal function and dysfunction
- MicroRNAs and the regulation of neuronal plasticity under stress conditions
- New advances of microRNAs in the pathogenesis of rheumatoid arthritis with a focus on the crosstalk between dna methylation and the microRNA machinery
- MicroRNAs and intellectual disability sd in down 2 syndrome x linked sd and fragile x syndrome
- Wnt signaling pathway in rheumatoid arthritis with special emphasis on the different roles in synovial inflammation and bone remodeling
- MicroRNAs and autoimmunity
- Fant and microRNA
- Microna 323 3p with classical potential in rheumatoid arthritis alzheimer's disease and ectopic pregnancy
- Classical nf- $\kappa$ B activation impairs skeletal muscle oxidative phenotype by reducing akt  $\alpha$  expression
- Microna 129 3p a new biomarker and potential therapeutic target for rheumatoid arthritis
- Increased mir 223 expression in t cells from patients with rheumatoid arthritis leads to decreased insulin like growth factor 1 mediated interleukin 10 production
- Sexual dimorphism of miRNA expression a new perspective in understanding the sex bias of autoimmune diseases
- Unmet needs in the treatment of autoimmunity from aspirin to stem cells
- Epigenetic and disease targets by polyphenols
- Epigenetics in rheumatoid arthritis a primer for rheumatologists
- Online resources for miRNA analysis
- The role of mir 155 in regulatory t cells and rheumatoid arthritis
- Microna mediated regulation of innate immune response in rheumatic diseases
- Impact of microRNAs on the understanding and treatment of rheumatoid arthritis
- Maternal stress induces epigenetic signatures of psychiatric and neurological diseases in the offspring
- Bioreactive potential of microRNAs in rheumatic diseases
- Mir microarray a microarray based microRNA target prediction method
- Mirna profiles in plasma from patients with sleep disorders reveal dysregulation of miRNAs in narcolepsy and other central hypersomnias
- Omni dynamic semi automated ontology development for the microRNA domain
- Embryonic miRNA profiles of normal and ectopic pregnancies
- Regulation of acv1 and sd2 by cell secreted exosomes during follicle maturation in the mare
- Involvement of miRNAs in the early phase of halothane induced liver injury
- Mir 192 induces p21 in growth arrest in aristocholic acid nephropathy
- Assessing the impact of copy number variants on miRNA genes in autism by micro array simulation
- Characterization and comparison of lactate dehydrogenase and lactate dehydrogenase related microRNAs

**Figure 38** - All the publications that cite the author

Also, at the top of the window, we can search for a publication, with the results to be shown immediately and also we can click on any publication title, and will navigate to its main source.

### Collaborators:

When we click on the "Collaborators" button the following window will be appear, which contains all the collaborated authors the author collaborate:



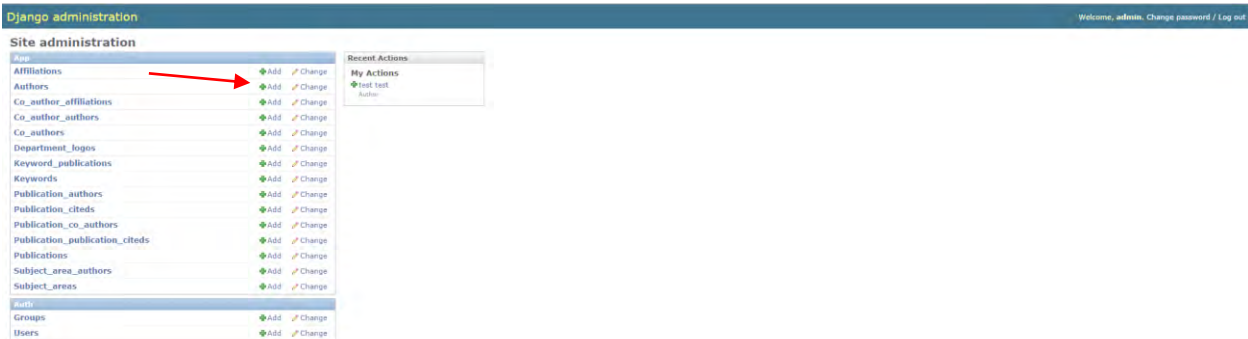
Camps C (1), Saini H K (1), Mole D R (1), Enright A J (1), Ragoussis J (1), Sacharidis D (1), Sartzetakis S (1), Sellis T (3), Paraskevopoulou M D (3), Georgakilas G (2), Kostoulas N (3), Dalamagas T (7), Vlachos I S (4), Papakonstantinou N (1), Ntoufa S (1), Chartomatsidou E (1), Belessi C (1), Stamatoopoulos K (1), Reczko M (14), Maragkakis M (11), Alexiou P (10), Bikalas N (1), Paschou M (1), Dorvalis E (1), Pandis I (2), Ospelt C (2), Karagianni N (2), Gay S (2), Kalajdzic P (1), Oehler S (1), Grosse I (2), Theodore D (1), Marotta D (1), Karar J (1), Jenkins W T (1), Koch C (1), Koumenis C (1), Arend D (1), Weinholdt C (1), Iizasa H (2), Wulff B E (2), Allin N R (2), Lieberman P (2), Nishikura K (4), Megraw M (11), Sethupathy P (10), Gumireddy K (1), Huang Q (2), Riback J (1), Gleditsch M (1), Tsanakas P (2), Simossis V A (2), Pereira F (2), Kreider E (1), Volinia S (1), Bonome T (1), Croce C M (1), Coutos G (3), Gupta A (2), Gartner J J (3), Fraser N W (3), Gagnebin M (1), Antonarakis S E (1), Bernal A (2), Crammer K (2), Zushreya B (1), Corda B (2), Mukerjee R (1), Sandri Goldin R M (1), Baev V (1), Rusinov V (1), Weber B L (1), Barrasa M I (1), Kirilaiden M (1), Nelson P T (2), Konranov A (1), Mourelatos Z (2), Bajic V B (1), Fiziev P (2), Hansenhalli S S (1), Fickett J W (3), Mache Niels (3), Bucher P (1), Botz J (1), Zerfass Thome K (1), Spitkovsky D (1), Jansen Dürr P (1), Sulai S (1), [Wageningen University & Research Centre](#), K Takada (1), Karol Szafranski (1), Martini A (4), Hanno Hinsch (2), LaDeana W Hillier (1), Miller L D (1), Ewan Birney (1), Wesley Warren (1), Ross C Hardison (1), Chris P Pouting (1), Peer Bork (1), David W Burt (1), [Martien A M Groenen](#) (1), Mary E Delany (1), Jerry B Dodgson (1), Christian Derst (1), Paul Levi (1)

**Figure 39** - All the collaborators of an author

We use the tag-cloud visualization to represent this case, which we implemented with pure Django. As we can observe, the font size of a collaborative author corresponds to its weight. The more the author is collaborate with the collaborative author the bigger the font size will be. Finally, if we hover a collaborative author with the mouse, a small box will appear showing his affiliation he belongs.

## 5.7 Admin interface

As mentioned on the start of this thesis, our goal is each department of the University of Thessaly to be able to use our system's services. The first time we will install and deploy our web application we will notice that the content is empty since nor the department's logo nor its author names with their ids have initialized. More specifically, when our crawlers will start to crawl, will note that the author names list is empty and so they can't crawl since their only input to start crawling are an author name and his id. So to overcome this problem, we make use of the Django admin interface. With this powerful interface each department will be able to manage all its content by adding, deleting or updating content. The only thing that needs to be done to use the admin interface, is to be created an administrator account. Then if we log in will see the following page:



**Figure 40** - The admin interface

As we can see, we have adjust the admin interface to show all the tables that our web application uses. From here, we can remove or update any table we want. For example, if we want to add another author

which has just become to the department we have to click to the “Add” button next to the “Authors” table, where the red arrow indicates on the above image, and will see the following window:

The screenshot shows the Django administration interface for adding a new author. The page is titled "Add author" and contains several input fields for profile information. Red arrows point to specific elements: one to the "Name" field labeled "Add content to profile", and another to the "Image" section labeled "Add profile photo". A "Save" button is visible on the right side of the form.

**Figure 41** - What we see if we want to add an author

Then we have just to add his profile elements such as his name, his ids, his profile url of the department and a profile photo if he has. Finally, we click to the “Save” button and will see a message that says that the author was successfully added. We follow this simple process every time we want to add an author or we want to edit his profile elements. Finally, a good functionality that Django admin interface offers is that we can create groups of users with specific permissions. For example, we can create a group that consists of all the department’s authors so that, each author can edit only the elements which are associated with him.

# Chapter 6

## 6 Conclusion and future work

### 6.1 Conclusion

The objective of this thesis was the development of a scientific information system for the management and monitoring for the research activities of the University of Thessaly. A system of which information is generated in an automated way through web crawling techniques and that makes use of all the best data visualizations techniques to present this information.

Our main goal was for each department of the University Of Thessaly to use our system which aims to provide it with qualitative and quantitative evaluation metrics. With these metrics the department and the whole University's research community, will be able to monitor its scientific productivity, its evolution in time and its collaborations with other institutions. Also, any research organization that is responsible for the evaluation of research will have valuable information that will help them for a more efficient evaluation.

Finally, with our system the collection and analysis of the scientific production of the whole University of Thessaly will allow reliable evaluative comparisons with other universities and institutions. This will result in an increase of the University's reputation to the rest of the world.

### 6.2 Future work

As for future work, a possible extension we are planning, is to modify our crawlers in order to collect our data from the ORCID registry using their API. In that way, we will solve the problems of duplicates about researchers or publications and generally we will be benefited by using all the advantages that ORCID provides which we described in par. 4.4.4 on this thesis.

Another possible extension for our system is to add a functionality of notifications. More specifically, a user who has registered to our system can be able to receive notifications on his email or on the system's dashboard, for a new publication that has just been added to the system or for any other events that interest the user, such as updates from specific authors or departments.

In addition, an important extension that must be done is the categorization of the research areas of the home page. So we would categorize them into some main categories with each one to include its subcategories. In that way, we will have a better and more efficient overview of all the areas where the department's authors work.



# CHAPTER 7

## 7 Tools and technologies we used

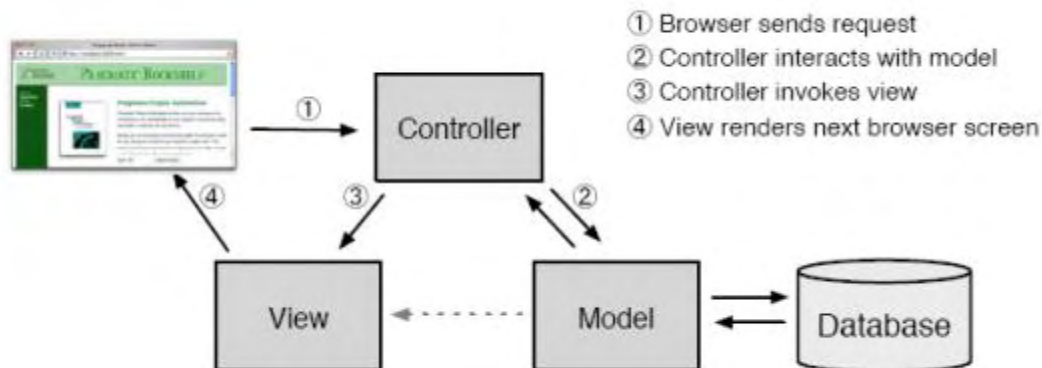
For the visualization of our data we used tools and technologies we mentioned in par 2.5 but we used also the following web technologies for designing and developing our system.

### Python

Python is a widely used general-purpose and high-level programming language. Its design philosophy emphasizes on code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. The language provides constructs intended to enable clear programs on both a small and large scale. Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library. Also, Python interpreters are available for many operating systems.

### Jingo

Jingo is a free and open source web application framework, written in Python, which follows the model-view-controller architectural pattern (called MTV in Jingo). Django's primary goal is to ease the creation of complex, database-driven web applications. Also, Jingo emphasizes reusability and "plug ability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings, files, and data models. Moreover, Jingo also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models.



**Figure 42** - How Jingo works (source: <http://ipass.wordpress.com/2009/04/06/django-note/>)

## Html/Html5

HTML (Hyper Text Markup Language) is the standard markup language used to create web pages.

HTML is written in the form of HTML elements consisting of tags enclosed in angle brackets (like <html>). The first tag in a pair is the start tag, and the second tag is the end tag (they are also called opening tags and closing tags).

## CSS/CSS3

CSS is the programming language used for formatting objects of a website. It can be embedded in HTML, and to determine the layout and appearance of elements.

## JavaScript

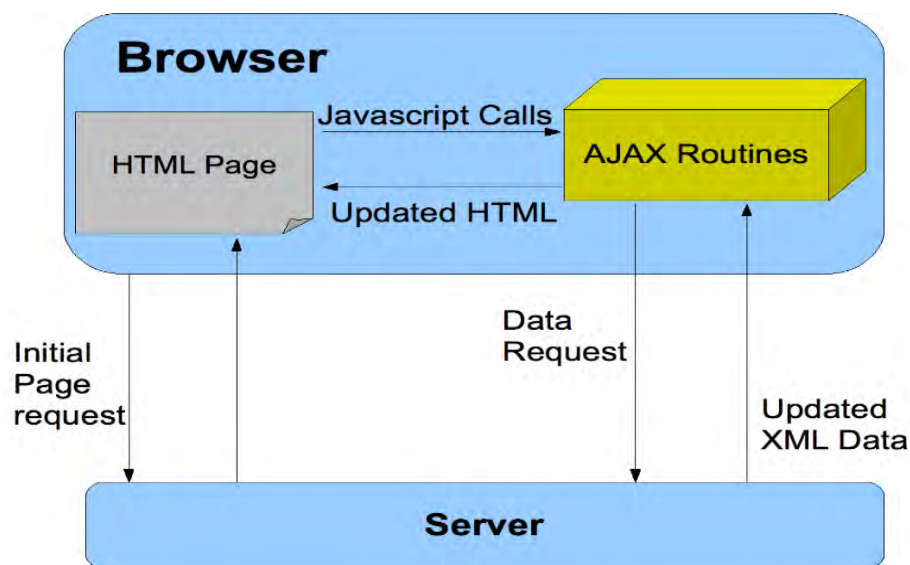
JavaScript is a dynamic programming language. It is most commonly used as part of web browsers, whose implementations allow client-side scripts to interact with the user, control the browser, communicate asynchronously, and alter the document content that is displayed. JavaScript is also being used in server-side programming (such as [Node.js](#) and [Angular.js](#)), game development and the creation of desktop and mobile applications.

## Query

jQuery is the most popular JavaScript library designed to simplify the client-side scripting of HTML. Its syntax is designed to make it easier to navigate a document, select elements from the HTML, create animations, handle events, and develop Ajax applications. The modular approach to the jQuery library allows the creation of powerful dynamic web pages and web applications. Generally, jQuery is used widely on the web, and especially over 60% of the most visited web sites are using jQuery.

## AJAX

AJAX (Asynchronous JavaScript and XML) is a group of interrelated web development techniques used on the client-side to create asynchronous Web applications. The term “asynchronous” means that we can send data to, and retrieve data from a server in the background, without interfering with the display and without the page to be refreshed. In addition, despite the name, the use of XML is not required, since Ajax is more used with JSON.



**Figure 43** - How Ajax works (source: <http://loadstorm.com/2009/02/load-testing-ajax-loadstorm/>)

## JSON

JSON is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs and it is used primarily to transmit data between a server and a web application.

## MySQL

MySQL is an open source RDBMS (Relational Database Management System) that uses SQL (Structured Query Language). SQL is the most popular language for adding, accessing and managing content in a database. It is most noted for its quick processing, proven reliability, ease and flexibility of use.

## Google Maps

Google Maps, is part of the Google Maps API, which is a free service available for use on any website and provides a range of services. In our implantation we use the Google Maps JavaScript API v3. This API, allows us to embed maps using JavaScript and provides a variety of services to handle our map. In addition, in order to use the Google Maps API, you must first register to the API you want and then get a unique API key, which you can use in order to use the services.

# Chapter 8

## 8 References

- [1] [Data Visualization for Human Perception]  
[http://www.interaction-design.org/encyclopedia/data\\_visualization\\_for\\_human\\_perception.html](http://www.interaction-design.org/encyclopedia/data_visualization_for_human_perception.html)
- [2] [An introduction to Visualizing Data by Joel Layman's]  
<http://piksels.com/wp-content/uploads/2009/01/visualizingdata.pdf>
- [3] [Visual Analysis Best Practices]  
<http://www.tableausoftware.com/asset/10-tips-to-create-useful-beautiful-visualizations?cid=70160000000w647&ls=Tableau%20Email&lsd=Tableau%20Email%20-%20Visualisation%20Guidebook%20EMEA%20-%20Remarket%20-%202014-06-26&adgroup=&kw=&adused=&distribution=Tableau&elq=96645b7f449643dabdd0eae3e25821fd>
- [4] [Web Crawler – An Overview]  
[http://www.csjournals.com/IJCSC/PDF2-1/Article\\_49.pdf](http://www.csjournals.com/IJCSC/PDF2-1/Article_49.pdf)
- [5] [What is web crawling]  
<http://blog.datafiniti.net/what-is-web-crawling/>
- [6] [Implementing an Effective Web Crawler]  
<http://www.devbistro.com/articles/Misc/Implementing-Effective-Web-Crawler>
- [7] [Bibliometric Analysis]  
<http://metrics.ekt.gr/el/epistimonikes-dimosiefseis/vivliometriki-analysi>
- [8] [Altmetrics in practice]  
<http://www.doria.fi/handle/10024/97529>
- [9] [ORCID]  
<http://orcid.org/>
- [10] [The Django Book]  
<http://www.djangobook.com/en/2.0/index.html>
- [11] [Tango with Django]  
<http://www.tangowithdjango.com/>