# *Texty, a visualization tool to aid selection of texts from search outputs*

## Abstract

**Introduction**.
The presentation of the results page in a search system plays an important role in satisfying the information needs of a user. The usual performance management criteria and tools to organise results have limitations that may hinder the satisfaction of those needs. We present Texty as a new approach that can help improve the search experience of users.

**Method**.
The corpus of texts to which we applied Texty were papers from Information Research. To filter the texts we have build five groups of words or vocabularies on concrete fields of knowledge: conceptual approach, experimental approach, qualitative methodology, quantitative methodology and computers/IT.

**Results**.
We show how Texty, intrinsically, is capable of encoding or offer its users information about the text that other alternative classic representations (bar or lines charts, mainly) are not able to offer.

**Conclusions**.
Texty is a complementary tool that improves intellectual interaction with a lists of texts, allowing users to choose texts more effectively knowing their structure before reading them.

# Introduction

Information retrieval is a critical factor in an environment characterised by excess of information (Baeza-Yates & Ribeiro-Neto 2011). When a user conducts a search, the information retrieval systems normally respond with a list of results. In many cases, the presentation of those results play an important role in satisfying user information needs. A bad or inadequate presentation can hinder the satisfaction of the information needs (Shneiderman 1992, Baeza-Yates 2011, Hearts 2009, Baeza-Yates *et al.* 2011).

Typically, information retrieval systems present the results of a query in flat, one dimension lists. Usually, these lists are opaque in terms of order, i.e. the users do not know why the list has a particular order. To refine their search, the users have to interact again, normally by filtering the first output of results.

The four main criteria used to organize a list of results are order, relevance, recommendation and clustering (Morville & Rosenfeld 2006, Pérez-Montoro 2010). The order organises the list of results by alphabetical or numerical order of some of the features (name of the author, date of creation) of the retrieved document. The relevance ranks the retrieved documents considering the relevance of the content of the document to the user's query. The recommendation can sort the results by using the number of recommendations suggested by other users who have previously used this result. The clustering presents the results grouped into a number of subsets formed by documents that deal with the same topic and/or addressing the topic with a similar approach (Larson 1991, Tryon 1939).

All these forms of organizing results, although used by most systems of information retrieval, have important limitations. The list of results organized by alphabetical or numerical order provides no extra information to help the users decide which of the listed documents can adequately meet their needs. When organising by relevance, the system places documents that could satisfy the information needs of the user at the top, but no extra information on the approach or on the internal structure of the document is provided. In the case of an organisation on the basis of recommendation, the top of the list provides documents recommended by other users and it does not provides extra information on the approach or the internal structure of the document. Finally, clustering provides extra information about the topic of the retrieved document, but it does not guide the user on the distribution and thematic structure of the document.

## Visual presentation of search results

In recent years, apart from these more standard presentation of results and with the aim to overcome some of its limitations, a number of visual proposals have been developed to improve user interaction with search results. Most of these proposals can be articulated in three main groups: the clustering visualizations, the visualization of query terms and the visualzations using thumbnail images or miniaturized images of documents.

Visualization of clusterings intends to represent the categories and relations between those categories of the retrieved documents The main trends in these representations are based on the use of, among others, treemaps, tag clouds or network graphs.

The treemaps represent hierarchical relationships of a set of categories using nested rectangles and optimizing the space used for the visualization (Shneiderman 1992, Shneiderman & Plaisant 2009). Each rectangle's size is proportional to the number of retrieved documents under that category. Normally the rectangles are coloured according to the category they belong to, for easy reading by users (see Figure 1).

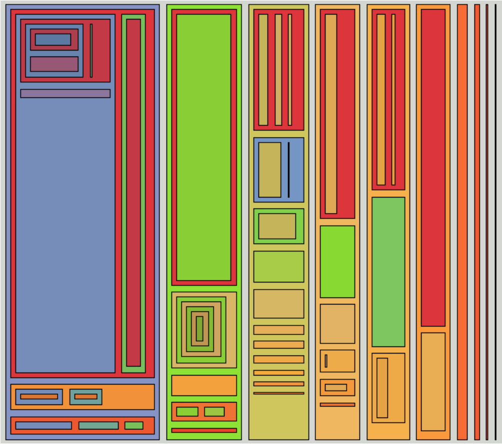


*Fig 1. Treemap* of a file system

The tag clouds represent the categories in a group of words, where the color and size of each word are proportional to the number of documents retrieved for each category (Begelman, Keller and Smadja, 2006). The labels that appear in the cloud are usually hyperlinks that lead to the list of documents that have been retrieved under that label.

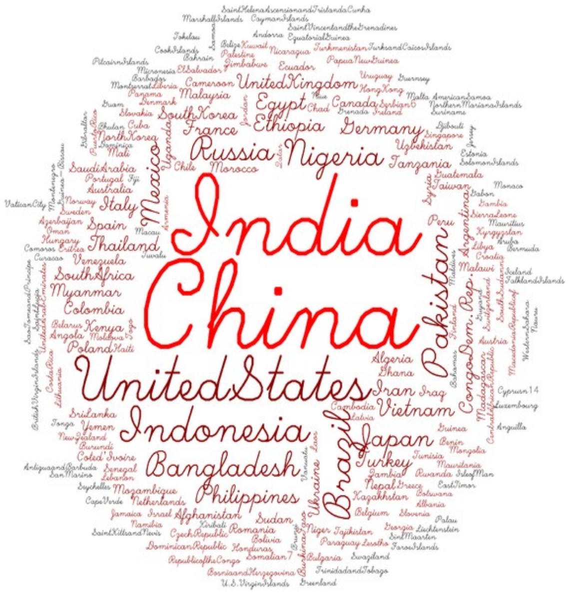


*Fig 2. World population tag cloud. Image by Seb951 under Creative Commons Attribution-Share Alike 3.0*

The network graphs represent each category as objects or nodes and their relationships using lines or curves. According to the most common interaction, if the user clicks on an object or category s/he will get the list of retrieved documents classified under a category. There are many examples applying this type of visualization, see: Moya-Anegón *et al.* (2004), Granitzer *et al.* (2004), and Brandes *et al.* (2006).

The visualization of query terms given by the user tend to follow two strategies: visualization of the terms within the document or in a page of results (Hearst 2009). In the first case, the system outputs the document highlighting those words that literally match the terms of the query (Landauer *et al.* 1993, Egan *et al.* 1989). Some studies indicate that users prefer to see this technique implemented by using colour in highlighted words that match the query terms (Hornbæk & Frøkjær 2001). In the second case, each document is represented on the results page as a horizontal bar proportional to extension of the document and small squares are added for eaxh query terms that appear in the text (Hoeber & Yang 2006). As in the previous case, some studies indicate that this representation improves with the introduction of a colour scale proportional to the frequency of the query terms in the document (Anderson *et al.* 2002).

Another technique for visualising query terms in a results page is to add thumbnail images or miniaturized images of the retrieved documents on the page. This technique is based on the fact that the human visual system captures the essentials of an image in 110 milliseconds or even less, just what it takes to read a word or two (Woodruff *et al.* 2001). Some studies claim that adding these images in the search results allows the search to function as visual summaries of documents (Jhaveri & Raiha 2005).

All these new proposals can improve the search experience of users, but they all have important limitations.

Compared to visualizations of clustering, the treemaps provide extra information on the thematic focus of the retrieved documents and the semantic relationships among them. However, the treemaps do not provide information on the distribution and thematic structure of each document. Tag clouds also provide extra information on the thematic focus of the retrieved document but, they do not provide information on possible semantic relationships between documents, nor orientation on the distribution and thematic structure of each of these documents. Finally, the network graphs provide extra information on the thematic focus of the retrieved documents and possible semantic relationships between documents, but do not provide the distribution and thematic structure of each document. N etwork graph are also difficult to explore in a comfortable way when they include many nodes and edges. Then the use of a zoom to get a global and the partial views of the network is needed (Viegas and Donath, 2004). Some authors advocate for combined strategies centring the graph on the node that interests the user (Yee et al. 2001) or eliminating those nodes that are not selected by the user (Fellbaum 1998).

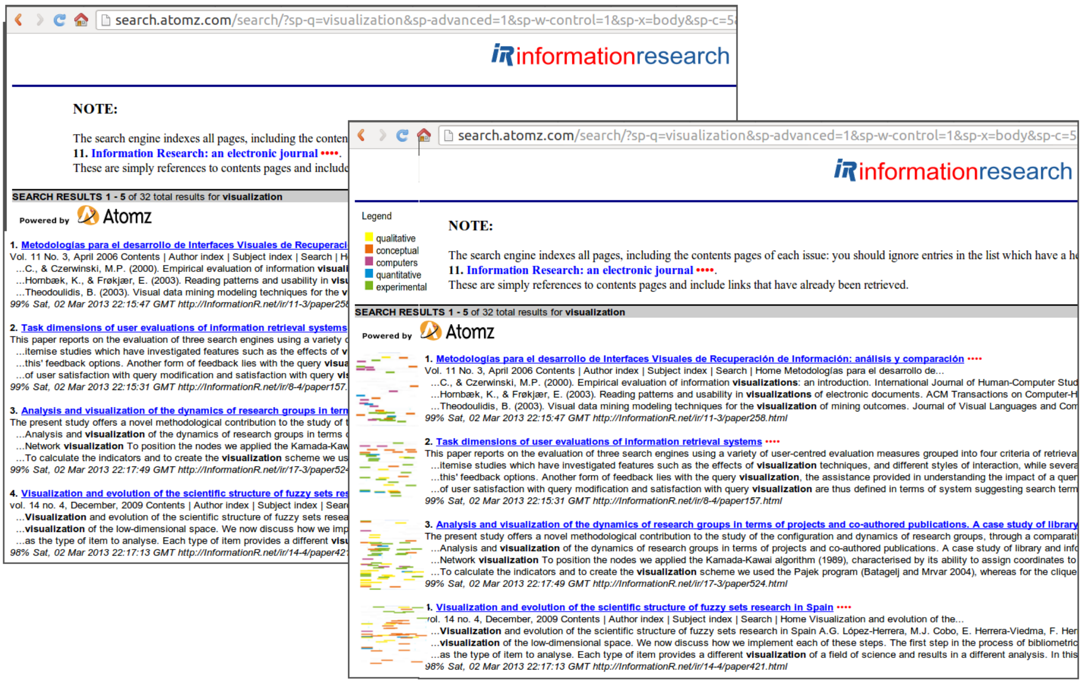
Visualizations based on the query terms also have important limitations. On the one hand, they only provide documents in which the query terms appear. They do not provide extra information on the thematic focus of the retrieved documents, nor possible semantic relationships between retrieved documents.

They do not give any orientation on the distribution and structure of those terms unrelated to each of these retrieved document either.

Finally, the visualization strategy which involves completing the list with thumbnail images or miniaturized images of retrieved documents also has important limitations. These visualizations, though complementary, do not provide extra information on the thematic focus of the content of the retrieved documents or on semantic relationships between retrieved documents. They do not show to the user the distribution and thematic structure of each document either. Along these lines, studies show that the thumbnails images strategy does not significantly improve the search experience of users (Czerwinski *et al.* 1999, Dziadosz & Chandrasekar 2002), although they can be helpful in part if the images are enlarged (Kaasten *et al.* 2002)

These limitations lead us to seek new forms of visualization that can help to improve the search experience of users in information retrieval systems and any other case where the user has to choose or select documents from one dimenssion lists of documents.

The proposed tool presented in this paper aims to face these limitations when deployed as a complement to traditional one dimension list of documents or to a list of results of information retrieval systems, such as clustering or sorting by relevance. This tool shows the essential parts of the contents of each item on the retrieved list and it helps the users in identifying the structure of the content of text documents, without having to tackle each one of the results intellectually..


search.atomz.com/search/?sp-q=visualization&sp-advanced=1&sp-w-control=1&sp-x=body&sp-c=5

**NOTE:**

The search engine indexes all pages, including the conten

**11. Information Research: an electronic journal ••••.**

These are simply references to contents pages and include

**SEARCH RESULTS 1 - 5** of 32 total results for **visualization**

Powered by Atomz

1. Metodologías para el desarrollo de Interfaces Visuales de Recuperaci
Vol. 11 No. 3, April 2006 Contents | Author index | Subject index | Search | H
...C., & Czerwinski, M.P. (2000). Empirical evaluation of information **visuali**
...Hornbæk, K., & Frøkjær, E. (2003). Reading patterns and usability in **visu**
...Theodoulidis, B. (2003). Visual data mining modeling techniques for the **v**
*99% Sat, 02 Mar 2013 22:15:47 GMT http://InformationR.net/ir/11-3/paper258*

2. Task dimensions of user evaluations of information retrieval systems
This paper reports on the evaluation of three search engines using a variety c
...itemise studies which have investigated features such as the effects of **v**
...this' feedback options. Another form of feedback lies with the query **visua**
...of user satisfaction with query modification and satisfaction with query **vis**
*99% Sat, 02 Mar 2013 22:15:31 GMT http://InformationR.net/ir/8-4/paper157.*

3. Analysis and visualization of the dynamics of research groups in term
The present study offers a novel methodological contribution to the study of t
...Analysis and **visualization** of the dynamics of research groups in terms c
...Network **visualization** To position the nodes we applied the Kamada-Kaw
...To calculate the indicators and to create the **visualization** scheme we us
*99% Sat, 02 Mar 2013 22:17:49 GMT http://InformationR.net/ir/17-3/paper524*

4. Visualization and evolution of the scientific structure of fuzzy sets res
vol. 14 no. 4, December, 2009 Contents | Author index | Subject index | Sear
...**Visualization** and evolution of the scientific structure of fuzzy sets resea
...**visualization** of the low-dimensional space. We now discuss how we imp
...as the type of item to analyse. Each type of item provides a different **visu**
*98% Sat, 02 Mar 2013 22:17:13 GMT http://InformationR.net/ir/14-4/paper421*

search.atomz.com/search/?sp-q=visualization&sp-advanced=1&sp-w-control=1&sp-x=body&sp-c=5

Legend

🟨 qualitative
🟧 conceptual
🟪 computers
🟦 quantitative
🟩 experimental

**NOTE:**

The search engine indexes all pages, including the contents pages of each issue: you should ignore entries in the list which have a he

**11. Information Research: an electronic journal ••••.**

These are simply references to contents pages and include links that have already been retrieved.

**SEARCH RESULTS 1 - 5** of 32 total results for **visualization**

Powered by Atomz

1. Metodologías para el desarrollo de Interfaces Visuales de Recuperación de Información: análisis y comparación ••••
Vol. 11 No. 3, April 2006 Contents | Author index | Subject index | Search | Home Metodologías para el desarrollo de...
...C., & Czerwinski, M.P. (2000). Empirical evaluation of information **visualizations**: an introduction. International Journal of Human-Computer Stud
...Hornbæk, K., & Frøkjær, E. (2003). Reading patterns and usability in **visualizations** of electronic documents. ACM Transactions on Computer-H
...Theodoulidis, B. (2003). Visual data mining modeling techniques for the **visualization** of mining outcomes. Journal of Visual Languages and Com
*99% Sat, 02 Mar 2013 22:15:47 GMT http://InformationR.net/ir/11-3/paper258.html*

2. Task dimensions of user evaluations of information retrieval systems ••••
This paper reports on the evaluation of three search engines using a variety of user-centred evaluation measures grouped into four criteria of retrieva
...itemise studies which have investigated features such as the effects of **visualization** techniques, and different styles of interaction, while severa
...this' feedback options. Another form of feedback lies with the query **visualization**, the assistance provided in understanding the impact of a quer
...of user satisfaction with query modification and satisfaction with query **visualization** are thus defined in terms of system suggesting search term
*99% Sat, 02 Mar 2013 22:15:31 GMT http://InformationR.net/ir/8-4/paper157.html*

3. Analysis and visualization of the dynamics of research groups in terms of projects and co-authored publications. A case study of library
The present study offers a novel methodological contribution to the study of the configuration and dynamics of research groups, through a comparati
...Analysis and **visualization** of the dynamics of research groups in terms of projects and co-authored publications. A case study of library and info
...Network **visualization** To position the nodes we applied the Kamada-Kawai algorithm (1989), characterised by its ability to assign coordinates to
...To calculate the indicators and to create the **visualization** scheme we used the Pajek program (Batagelj and Mrvar 2004), whereas for the clique
*99% Sat, 02 Mar 2013 22:17:49 GMT http://InformationR.net/ir/17-3/paper524.html*

4. Visualization and evolution of the scientific structure of fuzzy sets research in Spain ••••
vol. 14 no. 4, December, 2009 Contents | Author index | Subject index | Search | Home Visualization and evolution of the...
...**Visualization** and evolution of the scientific structure of fuzzy sets research in Spain A.G. López-Herrera, M.J. Cobo, E. Herrera-Viedma, F. Her
...**visualization** of the low-dimensional space. We now discuss how we implement each of these steps. The first step in the process of bibliometric
...as the type of item to analyse. Each type of item provides a different **visualization** of a field of science and results in a different analysis. In this
*98% Sat, 02 Mar 2013 22:17:13 GMT http://InformationR.net/ir/14-4/paper421.html*

*Fig 3. Use of texty as a complement in a list of results of a search query*

We used graphic techniques that were very similar for those used for authorship recognition by Keim & Oelke (2007) and for those used by Hearst (1995) in TileBars . Keim's technique represents the length of the phrases in each text as little squares with colour grading. In this paper this technique is applied very differently. TileBars also show the distribution of terms along the text as Texty does, but the terms come from a search query and the colour intensity is prportional to the frequency of the queried term in each document. In Texty, colour dots show the density of concepts referring to a particular linguistic field, which we call in this paper *vocabulary*. In Texty, the human eye analyses visualizations as it would do in Keim and Hearst. Visual coincident factors are colour zones, density of dots and the position and distribution of dots on the plane.

A third tool that graphically is similar to our work is Table Lens (Rao & Card 1994). In our case, we are not representing table structured data (columns, rows, data in each cell), like Table Lens does. Also Texty is not interactive as Table Lens is. Texty is simpler tool and it does not allow accurate data browsing neither.

Technically, a Texty is an image, an icon that represents the physical distribution of keywords of a text as a flat image. These keywords are grouped in vocabularies, to each of which a colour is linked (see fig. 4). Texty reveals~~,~~ the structure, conceptual density and subject matter of a text. Texty is a non-intrusive technique, in that an eventual implementation it does not necessarily interfere with the original information system that stores the documents. In this paper we show that this text representation tool enriches the one-dimensional lists that result from searches or from any other static list of documents.

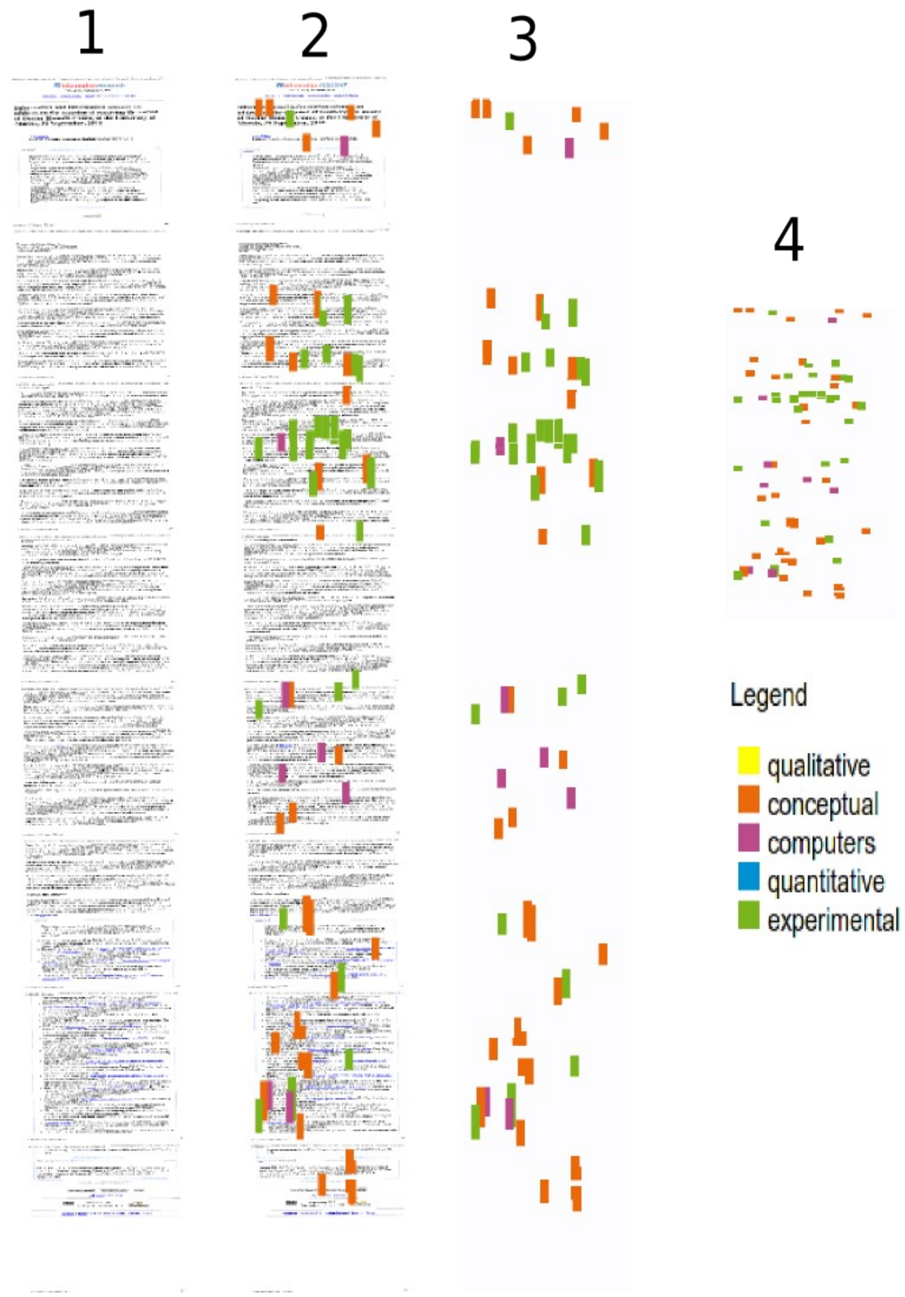


*Fig 4. Texty: the process and the colour's legend*

It is important to note that the human brain is capable of detecting variations of dots' density (Burguess & Barlow 1983) independently of the used colours (Nelson & Halberg 1979). Each array of dots of colour may represent a concrete linguistic field or vocabulary. The human vision can differentiate between colours quite well, especially when green and red are not present at once (Few 2008).

## Method

In 2008 and 2009 at the Ludwig Boltzman Institute (LBI, Linz, Austria), the challenge was laid down to make visualization tools with the data from the archive of *Ars Electronica*, a file of digital culture, media art and technology that had been collecting data since 1987. A lack of representation of collections of texts with the same linguistic register was identified. The research was to find the way of representing a text before reading it: a way to distinguish texts on a list and be able to compare them. Initially, as well as texts, there were five high-quality vocabularies on the history of media art (art work, person or institution, date, keyword and award). These were worked out by G. Dirmoser (Offenhuber & Dirmoser 2009) who provided the basis for developing a tool that showed these five vocabularies by five different colours in a proportional, representative image of the actual text. This work at the LBI gave the first intuitive representations with the Texty technique.

In this paper we present a more elaborate study of this technique, analysing its featuresin comparison with classic techniques of representation. The aim of the study was to develop and improve this technique as a complement to information search and retrieval systems.

The stages of the research were: data selection, choice of semantic categories, selecting and identifying the sources for the vocabulary corpus, the processing of terms for each vocabulary, the design of the corpus of texts for representation and the creation of Textys, for each text of the collection.

## Data selection

For the study we looked for a controlled collection of texts with a similar register and a specific semantic field. In addition, to assist the study, the texts had to be freely accessible.

For all these reasons we chose the papers published in *Information Research*. These papers belong to the same document collection, have unity, share the academic register, have a similar structure (*intro, method, analysis, results*) and have standardized quality (*peer-reviewed*).

The Information Research Website has a search system, by theme, by number or by author. It has a separate list of reviews, along with two retrieval systems: *Atomzsite search* and *Google*. The aim of this paper is to present Texty, a tool for information retrieval representation that goes further than the above resources for locating information.

## Choice of semantic categories

Once the corpus of texts had been chosen, we identified the following subject categories that could help to classify their contents: c*onceptual approach, experimental approach, qualitative methodology, quantitative methodology* and c*omputers/IT*

We could have chosen other alternative categories, however, from our personal perspective, as researchers working mainly in IT related issues, the five categories we have chosen are the main criteria for selecting the literature we use to perform the state of the art of the discipline: approach, methodology and degree of technology employed.

The election of the semantic categories, though related to the corpus of texts, is not unique and could be different without affecting the presentation of Texty as a possible helpful tool.

### Sources for the corpora of the vocabularies

The next step was to identify the sources of information from which the vocabularies that would be developed subsequently could be extracted. The choice of these sources was based on two complementary criteria. One was the intellectual prestige of the source. This criterion led us to select the Stanford Encyclopedia of Philosophy (http://plato.stanford.edu) and the Encyclopaedia Britannica (http://www.britannica.com). A second criterion was the popularity of the source, which led us to choose Wikipedia (http://www.wikipedia.org). The distribution of sources by subject matter is given in Table 1.

| | Definitions | | | | |
|---|---|---|---|---|---|
| | Qualitative Methodology | Conceptual Approach | Computers/IT | Quantitative Methodology | Experimental Approach |
| Stanford Encyclopedia of Philosophy | Aristotle's Categories<br>Intrinsic vs. Extrinsic<br>Properties | Concepts<br>Category<br>Theory | | Mathematics<br>Statistics | Experiment in physics |
| Britanica | Qualitative states<br>Qualitative tests to distinguish alternative theories | | | Mathematics<br>Statistics | |
| Wikipedia | qualitative data<br>Quantitative property<br>Qualitative properties<br>Qualitative research<br>Quality (philosophy) | Terminology<br>Theory<br>Vocabulary<br>Concept | list of programing languages<br>list of popular computers<br>list of hardware componets, software glossary | | Test method<br>Case study<br>Experiment |

*Table 1. Concepts and sources of the concepts or the five vocabularies chosen*

Importantly, although this has not been implemented in this study, it would be interesting to create, for each of the 5 vocabularies, a thesaurus (controlled vocabulary) which would spell out the different types of terms (preferred terms, variant terms, broader terms, narrow terms and related terms) and semantic dependencies (equivalence, hierarchy and association) between terms. This solution would solve the problems of silence and noise in indexing derivatives of synonymy and polysemy of terms.

### Processing of the terms for each vocabulary

Then the five vocabularies, based on the five corpora of texts for the concepts chosen, were defined (see Table 1). First, a *stopword* filter was used, to take out the empty words. Then the words occurring fewer than 4 times were deleted, as they were considered of little significance for each subject. Then the words occurring in more than one vocabulary were deleted, i.e. we removed interference between vocabularies. Thus we obtained a number of terms for each vocabulary (Table 2).

| | Number of terms | Terms /1,000 words |
|---|---|---|
| Conceptual Approach | 610 | 21.16 |
| Experimental Approach | 510 | 23.29 |
| Qualitative Methodology | 451 | 19.71 |
| Quantitative Methodology | 700 | 22.24 |
| Computers/IT | 312 | 18.91 |

*Table 2. Words selected for each vocabulary before the intellectual review*

Finally, there was an intellectual review to detect terms that were inconsistent with the subject matter, ambiguous terms and terms that were not coherent with each vocabulary. The experimental vocabularies were configured as in Table 3.

It should be said that the objective of this paper was to introduce the potential of the Texty tool. It is not a goal of this paper to study the best strategies to define the words that best represent a concrete field of knowledge or, as we call it in this paper, a *vocabulary*. In Machine Learning there are very powerfull methods like building topic models. This is an interesting possibility for future research.

| | Number of terms |
|---|---:|
| Conceptual Approach | 65 |
| Experimental Approach | 53 |
| Qualitative Methodology | 74 |
| Quantitative Methodology | 86 |
| Computers/IT | 410 |

*Table 3. Final number of terms for each vocabulary*

The c*omputers/IT* vocabulary is descriptive, which is why we left a large number of terms, as they all clearly refer to c*omputers and information technology*. The final list of words in all the vocabularies can be found online (Texty terms 2012)

Regarding the possible overlapping of colour dots we recommend to set a number of terms per vocabulary that avoids an excess of terms per line of the text.

### Corpora of texts to represent

After choosing the papers from Information Research as corpora of texts to which the Texty tool can be applied, then a private replica of the journal Information Research was made to conduct the study in laboratory comfort and also to show clearly how Texty can be implemented in an existing system. Information Research is made out of static HTML pages and Texty has been introduced in each issue's index and in subject index (link in red on the left column). Please, find the Texty representations of journal Information Research papers (2011) in the bibliography. To avoid an accessible online copy of the journal, the access has been restricted; it is necessary to introduce the user 'texty' and the password 'texty' to enter.

### Creation of the Textys

There are a lot of ways to, technically, create Textys. The choice it will depend on the required level of production and concrete conditions of each case. Here we describe the simple automated method that has been used to create the almost 500 Textys required for this study.

The initial format of the texts taken from the Information Research website was HTML. The HTML files were parsed and an specific class for each vocabulary was applied to all vocabulary's words found. Then it was defined an specific colour's style using a Cascade Style Sheet (CSS), to vocabulary's words; the rest of text was defined as white. Finally an screenshot of the HTML page was taken and the size was adjusted to 300x450px for each Texty representing each text.

In this study we created Textys with five different colours, as shown in Figure 5. When choosing the colours, the main restrictions usually recommended for this kind of graphic attribute were taken into account. One restriction was the use of basic colours that most of humans can distinguish (Kay & Maffi 2008); A second restriction was that humans have very litle difficulty identifying three to five colours schemes; and for seven to nine colours schemes the identification becomes significantly more dificult (Healey 1996).

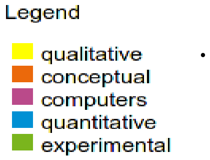
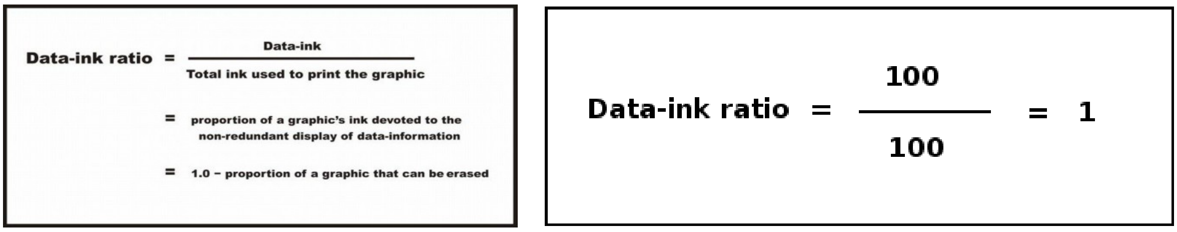


*Fig. 5. Colours of the vocabularies of Texty*

At this point we should stop a moment to analyze the information contained in the white areas of a Texty. Since the Texty is a physical representation of data, i.e. the colour dots appear in positions that reflect the real positions of terms in the text, the absence of ink gives relevant information about the text represented. Bearing in mind the theory of Tufte on the ratio of Ink and Data (Tufte 2001), for Texty we would have to say that there are *data without ink*. The absence of colours means a lower density of terms of the proposed vocabularies along the text. If we view the white zones as zones with data, Tufte’s formula in the case of Texty would be as follows, with the maximum proportion of ink devoted to representing data:



$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

$$= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information}$$

$$= 1.0 - \text{proportion of a graphic that can be erased}$$

$$\text{Data-ink ratio} = \frac{100}{100} = 1$$

*Fig 6. Tufte's data-ink ratio equation (left and the application to Texty (right)*

## Results

We created Textys for all the papers in the *Information Research*, from Volume 1, No. 1 (1995/96) to Volume 15 No.4 (2010), with a total de 454 Textys.

Below, Figure 7 gives the Textys of the 17 papers in volume 15, no. 4 (December 2010) of the journal Information Research. (http://informationr.net/ir/15-4/infres154.html).

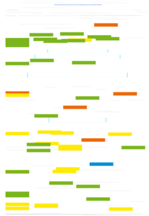
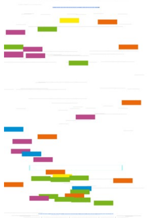
| | | | |
|---|---|---|---|
| *1. Proceedings of ISIC*<br> <br>*Cultural differences in the health information environments and practices between Finnish and Japanese university students*<br>*Graeme Baxter, Rita Marcella and Laura Illingworth* | *2. Proceedings of ISIC*<br><br>*Organizational information behaviour in the public consultation process in Scotland*<br>*Leanne Bowler* | *3. Proceedings of ISIC*<br> <br>*Talk as a metacognitive strategy during the information search process of adolescents*<br>*Jenny Bronstein* | *4. Proceedings of ISIC*<br> <br>*Selecting and using information sources: source preferences and information pathways of Israeli library and information science students of your paper*<br>*Donald O. Case* |
| *5. Proceedings of ISIC*<br> <br>*A model of the information seeking and decision making of online coin buyers*<br>Kreetta Askola,Toshimori Atsushi and Maija-Leena Huotari | *6. Proceedings of ISIC*<br> <br>*Local versus global information relevance in Website use: a case study with the information literacy portal AlfinEEES*<br>*Francisco Javier García Marco and María Pinto* | *7. Proceedings of ISIC*<br> <br>*Information behaviour research and information systems development: the SHAMAN project, an example of collaboration*<br>*Elena Maceviciute and T.D. Wilson* | *8. Proceedings of ISIC*<br> <br>*Avoiding health information in the context of uncertainty management*<br>*Anu Sairanen and Reijo Savolainen* |
| *9. Proceedings of ISIC*<br> <br>*A study of labour market information needs through employers' seeking behaviour*<br>*Sonia Sanchez-Cuadrado, Jorge Morato and Yorgos Andreadakis* | *10. Proceedings of ISIC*<br> <br>*'Information in context': co-designing workplace structures and systems for organizational learning*<br>*Mary M. Somerville and Zaana Howard* | *11.cProceedings of ISIC*<br> <br>*"We have a lot of information to share with each other". Understanding the value of peer-based health information exchange*<br>*Tiffany C. Veinot* | *12. Proceedings of ISIC*<br><br>*Information sharing: an exploration of the literature and some propositions*<br>*T.D. Wilson* |
| *13. Proceedings of ISIC*<br> <br>*Applying McKenzie's model of information practices in everyday life information seeking in the context of the menopause transition*<br>*Alison Yeoman* | *14. Regular paper*<br> <br>*Double or nothing: is redundancy of spatial data a burden or a need in the public sector of Uganda?*<br>*Walter T. de Vries and Beatrice Winnie Nyemera* | *15. Regular paper*<br> <br>*Analysis of automatic translation of questions for question-answering systems*<br>*Lola García-Santiago and María-Dolores Olvera-Lobo* | *16. Regular paper*<br> <br>*Dietary blogs as sites of informational and emotional support*<br>*Reijo Savolainen* |
| *17*<br> <br>*Information and information science: an address on the occasion of receiving the award of Doctor Honoris Causa, at the University of Murcia, 30 September, 2010 T.D. Wilson* | | | Legend<br>🟨 qualitative<br>🟧 conceptual<br>🟪 computers<br>🟦 quantitative<br>🟩 experimental |

*Fig. 7: Textys for volume 15, No 4 (December 2010) of the Information Research and legend.*

We presented Texty as a simple and complementary tool to enrich lists of texts. In this respect, a first glance to figure 7 can help the reader to select papers to read as follows:

- The predominant tone in this issue is the experimental one (green), though followed closely by the qualitative approach (yellow).
- Papers 3, 11 and 13 look clearly experimental (green), while paper 7 looks like one that requires more computers/IT's readers knowledge (violet).
- Five of the seventeen papers (38.5%) have a notable presence of computer/IT (violet).
- The paper with the biggest conceptual load is the 9th, though the 7th, 8th and 16th also have a conceptual content (orange).
- The more generalist paper seems to be the 15th.
- This issue does not involve quantitative methodologies much (blue).

Here we can see how Texty can be used for the exploration and navigation of texts before they are read.

Starting from this development we want to see what would happen if we try to represent the same data (terms and vocabularies from papers) using traditional techniques, like bar and lines charts. We are not proposing bar and line charts to be used as Texty, i.e., to enrich lists of texts but we are comparing formally how the same data set would look under these techniques compared to Texty technique.

**Texty and the bar charts:**

To illustrate this comparison, we chose papers from Volume 15, No. 4 (December 2010).

Case one, paper 441: A study of labour market information needs through employers' seeking behaviour. Sonia Sanchez-Cuadrado, Jorge Morato, Yorgos Andreadakis and Jose Antonio Moreiro (http://informationr.net/ir/15-4/paper441.html)

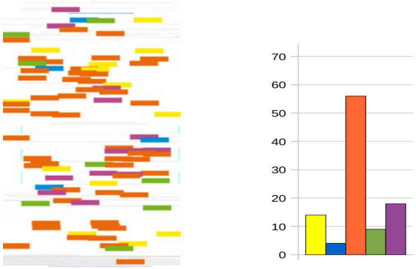


*Fig. 8: Texty and bar chart for paper 441 (Information Research)*

Both methods identify the most common vocabulary. In this case it is the conceptual one (orange). This paper describes knowledge representation techniques with computer support, which the two representations also show us. However, in the case of Texty, it can be seen that these techniques are discussed in the middle part of the paper (violet colour), whereas this was not seen with the bar chart.

Case two, paper 445: Information behaviour research and information systems development: the SHAMAN project, an example of collaboration. Elena Maceviciute and T.D. Wilson (http://informationr.net/ir/15-4/paper445.html)

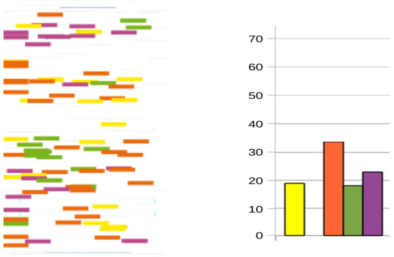


*Fig 9: Texty and bar chart for paper 445 (Information Research)*

This paper has a conceptual tone (orange). Initially, in the background on *long-term digital preservation,* we can say that techniques that require computers are being discussed (for example: *e-mail, word-processed documents and spreadsheets, as well as e-books, sound recordings, films, scientific data sets, social science data archives*, are terms used in this paper). In the middle of the paper we saw a concentration of green points belonging to the *experimental* vocabulary. This coincides with the explanation of the data used by the SHAMAN program on the basis of interviews with users. Not all this information can be deduced from the bar chart.

Case three, paper 450: Analysis of automatic translation of questions for question-answering systems. Lola García-Santiago and María-Dolores Olvera-Lobo (http://informationr.net/ir/15-4/paper450.html)

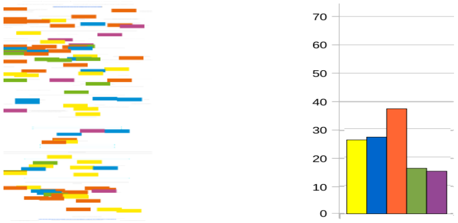


*Fig 10: Texty and bar chart for paper 450 (Information Research)*

In this case we have a paper with a considerable presence of the five vocabularies. Here, the importance of being able to see the physical distribution of terms in the paper can be seen perhaps in greater clarity. Thus we can say that the paper starts with a conceptual tone, to then explain the method in an experimental tone. The paper does not require too much knowledge of data processing, although there are references to it in the first half. At the end there are references of a conceptual kind. In general, the paper has a qualitative approach, as yellow is distributed throughout. Again, none of this information can be extracted by the bar chart.

### Texty and the line charts:

We used a line chart with the following coordinates: Y axis represents the position of the first character of the term. The X axis represents the number of terms for each vocabulary. Figure 11 gives an example of this representation.

Case four, paper 438: Dietary blogs as sites of informational and emotional support, by Reijo Savolainen (http://informationr.net/ir/15-4/paper438.html)

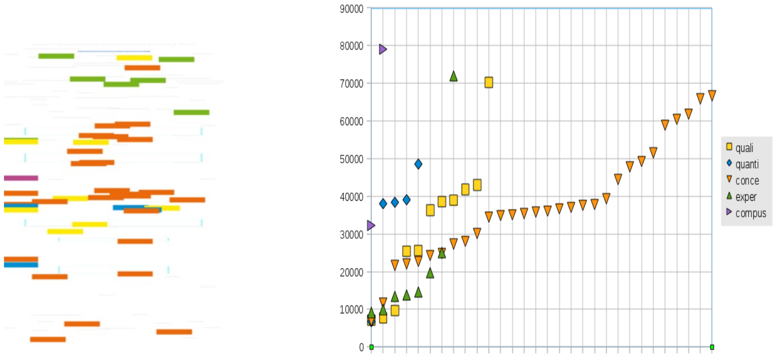


*Fig 11 Texty and line charts for paper 438 (Information Research)*

The reading of the line chart contributes more information on the structure and distribution of terms in the text than the bar chart does. Line chart shows very well the number of terms of each vocabulary. Been the Line chart more accurate in number of vocabularies than Texty. However, Texty is more suitable for everywhere use because it doesn't requite the use of axes and has a bigger range of readible sizes. For small sizes the line chart axes scales becomes unreadible.

### Texty versus bar and line charts

The objective of our work is not to quantitatively study Texty's performance against other alternative visualizations, taking this into account, both options, Texty and the charts, show the number of terms in each vocabulary, i.e. the general focus of a paper at a glance. The improvements introduced by Texty are:

1. Texty shows the distribution of terms along the text.

2. With Texty the conceptual structure of the paper can be seen: e.g. at the start there is a conceptual explanation; then the experimental part is developed; finally, the calculations in which there is intensive use of technology related and/or computer related operations.

3. Texty doesn't need axes or coordinates and scales.

## Conclusions

The development of this work allows us to point out a number of learned lessons about the complementary nature of Texty, its ability to encode information, and its non intrusive structure and technology.

Because of its complementary nature, Texty enriches lists of texts adding an image that physically represents the distribution of five conceptual fields along the text. Texty is not a replacement for classic search systems, but is proposed as a complement.

Texty's ability of encoding means that it is able to present distribution and structure of a text using only coloured dots that represent the ext itself.

Another conclusion we can draw is that Texty is not an intrusive solution from the point of view of the architecture of information. In this sense, Texty can be implemented without affecting the organizational criteria (e.g. order, relevance, recommendation or clustering) used to produce the retrieved list of documents.

Texty is a tool that can be implemented in an existing collection of texts and in is non-intrusive from a technological point of view. That means that it is not necessary to change or reprogram the storage system where the collection of texts lies. This easy implementation is presented as a critical advantage for future Texty implementations.

Finally, strictly speaking, from the point of view of information retrieval the use of Texty is not adding any advantage (not improved indexing and search algorithms, for example). What Texty can improves is the presentation of results: complements the traditional list of results (generally based on a title and a short summary) providing information on the content and structure of the retrieved document without having to interact directly with the document itself (see figure 3).

### Future developments

We want to round off these conclusions by mentioning some future lines of development derived from Texty.

Texty can be exported to other backgrounds and other vocabularies, adapted to each case and it can be personalized to the extent that it shows us other vocabularies (colours) depending on the reader preferences or the texts represented. Representation can be expanded and texts sections separators added, which indicate, for example, the customary sections of an paper (intro, method, analysis, results, conclusions, in the case of the papers of Information Research).

Dynamic, personalized and folk-vocabularies can increase the efficacy of Texty, as can the use of different layers to represent any vocabularies, as wanted. The use of interactive images (sensitive to clicks on the mouse) allows Texty to navegate through the text in question.

As noted, the use of thesaurus would improve the representative capacity of the vocabularies used in texty.

The adaptation of Texty for texts in a number of languages is another possible use: all you need are translations of the vocabularies.

## Bibliography

- Anderson, T. J., Hussam, A., Plummer, B., & Jacobs, N. (2002). Pie Charts for Visualizing Query Term Frequency in Search Results. Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology (pp. 440–451). London, UK, UK: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=646228.681545

- Baeza-Yates, R. (2011). Tendencias en recuperación de información en la web. BiD: textos universitaris de biblioteconomia i documentació, desembre, núm. 27. Retrieved from http://www.ub.edu/bid/27/baeza2.htm on 22-01-2013.

- Baeza-Yates R. A., & Ribeiro-Neto, B. (2011). Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- Baeza-Yates, R., Broder, A., & Maarek, Y. (2011). The New Frontier of Web Search Technology: Seven Challenges. In S. Ceri & M. Brambilla (Eds.), Search Computing (Vol. 6585, pp. 3–9). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-19668-3_1

- Begelman, G., Keller, P., Smadja, F., & others. (2006). Automated tag clustering: Improving search and exploration in the tag space. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland (pp. 15–33).

- Brandes, U., Hoefer, M., & Lerner, J. (2006, January). WordSpace: visual summary of text corpora. In *Electronic Imaging 2006* (pp. 60600N-60600N). International Society for Optics and Photonics.

- Burgess, A., & Barlow, H. B. (1983). The precision of numerosity discrimination in arrays of random dots. Vision Research, 23(8), 811–820. Elsevier. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6623941

- Czerwinski, M., Van Dantzich, M., Robertson, G., & Hoffman, H. (1999, August). The contribution of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3D. In *Proc. Interact* (Vol. 99, pp. 163-170).

- Dziadosz, S., & Chandrasekar, R. (2002, August). Do thumbnail previews help users make better relevance decisions about web search results?. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 365-366). ACM.

- Egan, D. E., Remde, J. R., Gomez, L. M., Landauer, T. K., Eberhardt, J., & Lochbaum, C. C. (1989). Formative design evaluation of superbook. ACM Transactions on Information Systems (TOIS), 7(1), 30–57. ACM.

- Fellbaum, C. (2010). WordNet. *Theory and Applications of Ontology: Computer Applications*, 231-243.

- Few, S. (2008). Practical Rules for Using Color in Charts. Visual Business Intelligence Newsletter, 11. Retrieved from http://www.perceptualedge.com/library.php

- Granitzer, M., Kienreich, W., Sabol, V., Andrews, K., & Klieber, W. (2004, October). Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on (pp. 127-134). IEEE.

- Healey, C. G. (1996). Choosing effective colours for data visualization. Proceedings of the 7th conference on Visualization ’96 (p. 263–ff.). Los Alamitos, CA, USA: IEEE Computer Society Press. Retrieved from http://dl.acm.org/citation.cfm?id=244979.245597

- Hearst, M. 1995. “TileBars: visualization of term distribution information in full text information access.” Proceedings of the SIGCHI conference on Human http://dl.acm.org/citation.cfm?id=223912 (March 26, 2013).

- Hearts, M. (2009). Search User Interfaces. Cambridge University Press.

- Hoeber, O., & Yang, X. D. (2006). A comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google. Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on (pp. 866–874).

- Hornbæk, K., & Frokjaer, E. (2001). Reading of electronic documents: the usability of linear, fisheye, and overview+ detail interfaces. Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human factors in computing systems (Vol. 2001, pp. 293–300).

- Jhaveri, N., & Räihä, K. J. (2005). The advantages of a cross-session web workspace. CHI’05 extended abstracts on human factors in computing systems (pp. 1949–1952).

- Kaasten, S., Greenberg, S., & Edwards, C. (2002). How People Recognise Previously Seen Web Pages from Titles, URLs and Thumbnails. People and Computers, 247–266.

- Kay, P. & Maffi, L., (2008). Number of Basic Colour Categories. In: Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, Bernard (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 133.

- Keim D. A. & Oelke D. (2007). Literature Fingerprinting: A New Method for Visual Literary Analysis. In: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology (VAST '07). IEEE Computer Society, Washington, DC, USA, 115-122. DOI=10.1109/VAST.2007.4389004 http://dx.doi.org/10.1109/VAST.2007.4389004

- Larson, R. R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog. The Library Quarterly, 133–173. JSTOR.

- Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics, 61(1), 129–145. Springer.

- Morville, P., & Rosenfeld, L. (2006). Information Architecture for the world wide web: designing large-scale web sites. O’Reilly Media, Incorporated.

- Nelson, M., & Halberg, R. (1979). Visual contrast sensitivity functions obtained with colored and achromatic gratings. … : The Journal of the Human Factors and …. Retrieved January 24, 2013, from http://hfs.sagepub.com/content/21/2/225.short

- Offenhuber D. & Dirmoser G. (2009) Semaspace: graph editor for large knowledge networks.URL:http://residence.aec.at/didi/FLweb/. Accessed: 2013-01-19. (Archived by WebCite® at http://www.webcitation.org/6DmoMu1n3)

- Rao, R., & Card, S. (1994). The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. Proceedings of the SIGCHI conference on Human …. Retrieved February 7, 2013, from http://dl.acm.org/citation.cfm?id=191776

- Shneiderman, B. (1992). Designing the user interface (2nd ed.): strategies for effective human-computer interaction. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

- Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on graphics (TOG), 11(1), 92–99. ACM,

- Shneiderman, B. (1998). Treemaps for space-constrained visualization of hierarchies.
- Texty representations of Journal Information Research papers (2011). URL: http://o.subvideo.tv/. Restricted access: user:'texty', password: 'texty'.

- Texty terms (2012) URL:http://vis.subvideo.tv/texty/terms.php. Accessed: 2013-01-19. (Archived by WebCite® at http://www.webcitation.org/6DmY1LEKU)

- Tryon, R. (1939). Cluster analysis. New York: McGraw-Hill:2, 133-173.

- Tufte E. R. . 1986. The Visual Display of Quantitative Information (pp 93). Graphics Press, Cheshire, CT, USA,

- Viegas, F. B., Wattenberg, M., Van Ham, F., Kriss, J., & McKeon, M. (2007). Manyeyes: a site for visualization at internet scale. Visualization and Computer Graphics, IEEE Transactions on, 13(6), 1121–1128. IEEE.

- Woodruff, A., Faulring, A., Rosenholtz, R., Morrsion, J., & Pirolli, P. (2001). Using thumbnails to search the Web. Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 198–205).

- Yee, K. P., Fisher, D., Dhamija, R., & Hearst, M. (2001). Animated exploration of dynamic graphs with radial layout. Presented at IEEE Symposium on Information Visualization.