Exploratory portals of research data in education

Nicoletta Di Blas

Polytechnic of Milan, Italy, nicoletta.diblas@polimi.it, 0000-0002-7987-2930

Abstract

This paper presents a powerful tool to enhance research in education: 'exploratory portals', supporting effective storage, sharing and exploration of large sets of research data. The workflow is the following: data are gathered by a research group; they are then classified according to a taxonomy (the one that best fits the group's research interest); once uploaded in the portal, they can be 'explored' via a combination of faceted search (enriched by Boolean operators) and data mining techniques. The system can thus answer in a few seconds to sophisticated user's queries that otherwise would require hours; it can save a session's results and materials for sharing with other scholars or for further investigation. The paper presents a case study of exploratory portal, dealing with data on (technology-based) education. The portal has effectively been used by five different research groups, to run complex investigations of data about technology integration into schools.

Keywords: Case study; Data collection; ICT; Educational research.

Introduction

Technology can affect research, powerfully enhancing processes that 'on paper' would require hours as well as giving vent to unexpected procedures. Not only can it enhance current practices, for example, facilitating data storage and retrieval, sharing of resources and collaborative writing but through advanced techniques it can also support 'new applications/combinations, with unforeseen value and impacts emerging over time: e.g. new forms of analysis based on data mining, "mashups" of data sets that had not been previously related to each other, including machine searching for hitherto undetected correlations' (Houghton, Gruen, 2014: 15).

This paper introduces 'exploratory portals' as advanced systems to store, share, retrieve and – most importantly – *explore* research data. Exploratory portals are being developed by HOC-LAB at the Politecnico di Milano in cooperation with GSA-Lab (Graphics and Software Architectures Laboratory) of the University of Salento, since 2012, in the frame of two national research projects (Ferrari et al., 2012). From a technical point of view, exploratory portals are at the crossroad between faceted search (enriched by Boolean operators) and data mining: they can answer in a few seconds to user's queries that would otherwise require hours to be performed. One case study will be presented in details: the Learning4All portal, gathering data about 276 educational experiences of technology integration into schools.

The paper is organised as follows: in the state-of-art section, the basic elements of faceted search and data mining are introduced plus an overview of the existing portals hosting research data in education. Then, the case study is presented from design to evaluation. The 'lessons learned' will lead to the final conclusions and future developments discussion.

1. State of the art

Owing to the focus of this paper, in this section, the basic elements of only faceted search and data mining techniques are presented, in order to equip the reader with a fair understanding of both approaches and allow him/her to appreciate in what sense exploratory portals build upon (and move beyond) them. In any case, the reader must be aware that both fields are quite huge and have given vent to a vast body of literature. Furthermore, a review of the main online repositories of open research data, as well as of open educational resources, is presented.

Exploratory portals of research data in education
Di Blas

1.1. Faceted search

'Faceted search, also called faceted navigation or faceted browsing, is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters. A faceted classification system classifies each information element along multiple explicit dimensions, called facets, enabling the classifications to be accessed and ordered in multiple ways rather than in a single, pre-determined, taxonomic order. Facets correspond to properties of the information elements' (Wikipedia). Faceted search (Tunkelang, 2009; Hearst et al., 2003) has undergone a major adoption in these past years and has become 'the' standard for information exploration. It is based on a facet taxonomy, that is, a set of parameters used to classify the single pieces of information. Let us clarify with an example: if we were to classify an 'educational experience', parameters could be 'school level' (with the following values: pre-school, primary, middle school, high school), 'technology used' (with the following values: tablet, Interactive White Board, PC, etc.), etc. Facet browsing systems support dynamic access structures by means of these parameters, which the user can combine at his/her will. The outcome is typically the set of items that satisfy the criteria. The main difference amongst different implementations consists in the way the structured data and the query are syntactically and semantically modelled. Basic faceted navigation (Stefaner, Muller, 2007; Tunkelang, 2009) is based on property-value pairs that are usually combined conjunctively only. Dynamic taxonomies (Sacco, 2006; Tzitzikas, 2009) and Logical Information Systems (LIS) (Ferré, Ridoux, 2004) generalise the faceted navigation model by implementing Boolean logic operators (conjunctions, disjunctions and negations).

Exploratory portals are based on faceted search, thanks to which they can interactively create subsets and redefine them (through new selections of facets); they move a step further by introducing percentage values instead of absolute values as a way to 'make sense' of data sets (see figure 1).

1.2. Data Mining

'Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amount of data, stored in large databases, data warehouses or other information repositories' (Han et al., 2011). Data sets are analysed by data mining software, to find interesting correlations or patterns amongst 'raw' data in large relational databases. Data mining is characterised by (Grossman et al., 2001)
- Dealing with large quantities of data
- 'Patterns' discovery
- Being data driven (NON hypothesis-driven)
- Being machine driven (NON human-driven)

Amongst data mining's major tasks, we can list the following: class description (a concise summary of a collection of data that can be used for comparisons, e.g. a comparison between the sales of a company in two different countries); association (the discovery of correlations within a set, e.g. whenever a customer in a grocery store buys X, also Y is bought); prediction (of possible values of missing data); clustering (i.e. collections of 'similar' data); and so on.

Exploratory portals borrow from data mining the focus on a set's properties rather than on a set's entities (which is more typical of faceted search results: a list of items, link with a search engine); it moves a step further by introducing dynamic changes in sets driven by *human* choices and *following research hypothesis*.

1.3. Online repositories of research data and education

In spite of the increasing awareness of the importance of sharing research data, areas such as human sciences are still lagging behind, with respect, for example, to health, physics or biology (Kvalheim, Kvamme, 2014; Houghton, Gruen, 2014). There are a number of good reasons for sharing data (Houghton, Gruen, 2014): the scientific debate is enhanced; the same data sets can be re-used (thus optimizing the costs of the research); close scrutiny of the research findings is possible, leading to transparency and reduction of research misconduct; collaboration between different research groups accessing the same data set is fostered; access to data for young researchers who may have more difficulties into getting research funds is guaranteed; and so on.

As repositories of research data are spreading, some tools are being created in order to allow researchers to quickly find and access them: Re3data, for example, is an online registry created in 2012 and funded by the German Research Foundation (DFG) that 'covers research data repositories from different academic disciplines' (www.re3data.org). It currently (September 2015) hosts 1,335 data repositories in various disciplines: linguistics, computer science, medicine, physics and so on. The repositories can be accessed by browsing the registry by subject, content type or country. No repository on education is included in re3data. DataBib (http://databib.org/about.php) is a similar tool; it is meant to help people 'identify and locate online repositories of research data. Users and bibliographers create and curate records that describe data repositories that users can search'. DataShare repository, by the University of Edinburgh, is an online digital repository of multi-disciplinary research data sets. But amongst the communities through which the user can browse (economics, engineering, mathematics, health sciences, etc.), again education is missing.

Exploratory portals of research data in education
Di Blas

The education sector is apparently more keen on sharing Open Educational Resources (OER[1]) rather than 'raw' research data and materials. Here are some (famous and less famous) examples. Khan Academy (www.khanacademy.org/) is an online repository, launched in 2006, of educational videos about virtually any subject matter. MIT OpenCourseWare (http://ocw.mit.edu/index.htm) is a pioneer initiative by the Massachusetts Institute of Technology that started putting online materials (videos, lecture notes, assignments, etc. in 2001 on more than 2,000 courses in a range of disciplines. ISKME (California), an independent, education non-profit established in 2002, supports the Digital library of Open Educational Resources (https://www.oercommons.org/), for all school grades. Webquest (http://webquest.org/) is an 'historical' (having been founded in 1995) repository of inquiry-oriented lesson plans in which most or all the information that learners work with comes from the Web. It is subject driven, similar to the vast majority of repositories of educational resources. The National Science Digital Library (https://nsdl.oercommons.org) is an online repository of educational resources with emphasis on the sciences, technology, engineering and mathematics (STEM) disciplines. It can be browsed by subject area, education level, material type, conditions of use and content source. Merlot (www.merlot.org) is a repository of online resources for higher education by the California State University Center for Distributed Learning; it also includes a pedagogy-driven subsection (http://pedagogy.merlot.org/) in which pedagogical issues rather than subjects are the driving force. Its aim is to support teachers with a variety of instructional strategies, to answer questions such as 'how to effectively teach a large class' or 'how to engage students in problem-based learning'. The 'Teachers Pay Teachers' portal (www.teacherspayteachers.com) is a quite original idea by a New York public teacher, who in 2006 opened an online 'market place' for teachers to share or even sell lesson plans and course materials. "Pedagogy in action" (http://serc.carleton.edu/sp/library/pedagogies.html) is a library of pedagogic methods, complemented by a collection of activities that exemplify each method. It currently contains information on 30 different teaching methods and with 700 examples of use. 'Pedagogy unbound' (www.pedagogyunbound.com) is a 'place' to share college teaching practices (how to start the new year, how to help students write in a more concrete way, etc.) founded in 2013.

Most repositories are from the Anglo-Saxon area; a couple of Italian examples can still be quoted. 'Innovascuola' ('school and innovation') was a repository of resources created by teachers in various projects and made accessible online with a Creative Commons license (www.innovascuola.gov.it). GOLD (http://gold.indire.it/gold2/) is a database of good educational practices curated by ANSAS/Indire (an Italian agency devoted to research in education). Eventually, 'Rai Educational' (the Italian national TV channel) hosts educational videos that can be organised to create lesson plans.

To summarise, it can be noted that whilst being quite generous in sharing resources of various kinds in an open-access perspective (and it must be furthermore noted that a content-wise rather than a pedagogy-wise perspective is taken, in most cases), the education sector is paradoxically quite absent in the realm of open research data. The portal presented in this paper represents instead an attempt at introducing education into the realm of open research data.

2. The Learning4All portal

In 2012, HOC-LAB (at the Politecnico di Milano, Italy) was involved in a national research project together with three pedagogy and four technical schools (Ferrari et al., 2012). The focus of the project was on technology integration: in order to investigate the issue, researchers interviewed via Skype a number of teachers from all school grades (following a semi-structured interview schema) on their experience with technology within the classroom. Moreover, surveys were administered to the teachers both at the beginning (on expectations) and at the end (on results) of their experiences. Overall, data about 276 real-life educational experiences, in which technology had played an important role, were gathered and the following resources were available for each experience:

- Audio file of the interview
- Transcript of the interview
- Filled-in survey on expectations
- Filled-in survey on results
- A form in which a researcher had extracted the main features of the experience (FeES: Features Extraction Schema)
- A form in which a researcher had highlighted the differences between the expectations and the results

Some of the materials could be considered as 'raw' data: namely, the interviews, the transcript and the surveys; others were already a form of interpretation by researchers: the FEES forms, where the main features of the experiences were gathered and the comparison between the expectations and the results was performed.

The next crucial issue was how to make these data available and effectively usable by researchers: 'exploratory portals', on which HOC-LAB had already some experience (Di Blas et al., 2012; Spagnolo et al., 2010; Di Blas et al., 2014a), were identified as a possible answer. Exploratory portals are online repositories where objects (in broad sense) are stored without any hierarchical organisation, as if they were 'scattered in a box'. Each object, though, is tagged

---

[1] 'Teaching, learning and research materials in any medium, digital or otherwise, that reside in the public domain or have been released under an open license that permits no-cost access, use, adaptation and redistribution by others with no or limited restrictions. Open licensing is built within the existing framework of intellectual property rights as defined by relevant international conventions and respects the authorship of the work' (2012 Paris OER declaration).

(classified) according to a taxonomy (a set of facets with their values), so that by checking and selecting the proper facets and values, a user can effectively make sense of the whole data set, as well as of potentially interesting sub-sets (figure 1).
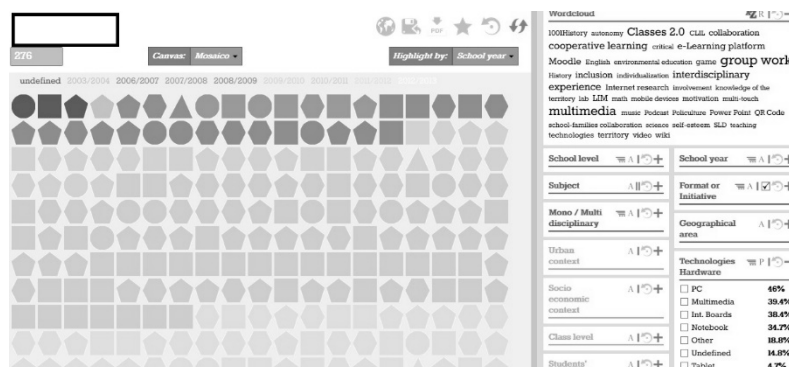

Figure 1. The Learning4All portal of educational experiences (www.l4all.it/eng). The data set is visualised as a mosaic, where each tile represents an educational experience; on the right, part of the taxonomy (facets) can be seen.

The Learning4All portal encompasses 30 facets organised into six macro-areas. Each facet entails a number of values, from a minimum of 4 to a maximum of 12. Table 1 shows the facets (according to which all the educational experiences had been tagged), organised into the six macro-areas, whilst Table 2 shows an example of values within a facet ('activities').

Table 1: The facets of the Learning4All portal.

Macro-area	Facets
Basic data	School year (in which the experience took place)
	School level
	Monodisciplinary versus multidisciplinary
	Subject (humanities vs. science and technology and then specific subject: math, music...)
	Format (a pre-defined experience)
	Program (e.g. 'Classi 2.0')
The context of the experience	Social context
	Cultural context
	Class performance (average level of the students)
	Class homogeneity (in terms of performance)
Technology	Hardware (e.g. PC, Interactive White Boards, Tablet, etc.)
	Software (e.g. authoring SW)
	Digital content
Implementation	Where (class, lab, home, etc.),
	When (during or after school hours, etc.)
	Curriculum (whether or not 'curricular')
	Programme
	Design (how defined an idea the teacher had in mind at the start of the activity)
	Human resources
	Involvement of 'helpers' (relatives, experts, etc.)
	Activities
	Organisation
	Key features (what was the experience characterised by, e.g. 'collaboration')
Inclusion	Problems (e.g. 'immigration')
	Strategies (to overcome inclusion issues, e.g. 'peer-to-peer learning')
Benefits	Cognitive benefits
	Motivational benefits
	Relational benefits
	Communication-related benefits
	Technology-related benefits

Table 2. Example of a facet ('activities') with its values and descriptions

Activities	Value	Description
	Lecture	Traditional lecture
	T-Discussions	Traditional discussions or brainstorming in class
	D-Discussions	Digitally supported discussions or brainstorming
	T-Individual	Individual study or work, in a traditional way
	D-Individual	Digitally supported individual study or work
	T-Self-assessment	Self-assessment in a traditional way
	D-Self-assessment	Digitally supported self-assessment in a traditional way
	T-Collaboration	Collaborative activities, in a traditional way
	D-Collaboration	Collaborative activities, in a digitally supported way
	T-Creation	Creation of traditional content or artefacts
	D-Creation	Creation of digital content or artefacts
	Content collection	Activities for searching and collecting large amount of content
	Problem solving	Activities oriented to problem solving
	Gaming	Game activities for learning
	Simulations	Simulations of some sort
	Other	

The portal can support various scenarios of use, ranging from very simple ('I'm interested in experiences with tablets were at primary school' or 'I'm interested in experiences where the tablets were used at primary school but NOT with immigration problems') to very sophisticated, such as the following:

Scenario: a researcher investigating inclusion and group work

A researcher in educational technology wants to investigate the relationship between inclusion and group work. He/she accesses the Learning4All portal, gathering reports and data about information and communications technology (ICT)-based educational experiences at school. Each experience is profiled according to a set of 30 'facets' (e.g. subject matter, technology used and benefits); each facet is defined through a number of 'values' (e.g. for subject matter: history, literature and math). The researcher first selects 'inclusion' within the facet 'key features'; looking now at the facet 'organisation', he/she immediately sees that group work (both homogeneous and heterogeneous) goes up; a closer look shows that heterogeneous group work emerges as the dominant strategy. This result seems to corroborate the idea that a relationship between inclusion and group work is actually there.

Now the researcher wants to further dig into this issue, checking how experiences with inclusion and group work are characterised. He/she, therefore, selects within the facet 'organisation' all the different kinds of group work; he/she also excludes (thanks to Boolean logic operators) the value 'disability' within the facet 'inclusion problems', because the focus of his/her research does not include this issue. He/she obtains a sub-set of the original set that encompasses reports about educational experiences where inclusion is an issue (but excluding disabilities) and group work has been used as organisational strategy.

He/she can now explore what values all the other facets in the portal have gained. He/he checks the educational benefits facet and notices that with respect to the initial set, 'motivation' skyrockets, as well as 'self-esteem', as may be expected. Looking at the 'inclusion – strategies' facet, he/she also notices that peer-to-peer learning seems the most used. He/she can now make a number of further steps, for example, selecting the different school levels (primary, middle, high, etc.) within the school level facet to compare the results and answer questions such as 'does the school level make a difference?'.

The researcher moves on in his/her exploration, reaching a subset of experiences he/she is definitely interested in. He/she downloads all the data to go through them in details later on and move on with his/her research.

As the reader can see, simple scenarios could be easily supported by faceted search mechanisms (though Boolean operators are almost never included in current systems) but sophisticated scenarios definitely require a more comprehensive approach, in which data mining plays a crucial role.

To summarise, exploratory portals allow

* an 'at a glance' understanding of a data set. Even by just browsing the facet values, a user can already learn something about the data set. For example (figure 1, right-hand side), the user can immediately see that PC is the most commonly used hardware (amongst the 276 experiences gathered in the portal). Similar discoveries can be done by simply opening the various facets;

* serendipitous discovery. A researcher may spot interesting correlations by creating subsets; for example, he/she may notice that if 'primary school' is selected within the facet 'school-level' tablets outnumber PCs in the facet 'technology';

* a sequential research (one step of the research acts as stepping stone for the next). Thanks to the visual feedback by the portal, the researcher can decide to iteratively select all the school levels and then compare the different distributions of technology used; furthermore, he/she can also consider what different benefits are achieved.

The reader can easily see that faceted search is used for selecting and refining the subsets whilst from data mining an interest in their 'intension' (in logic terms, their 'properties') rather than their 'extension' is borrowed. Exploration can

thus be seen as an interactive process, where the user understands a set of objects, modifies the set creating subsets, gets an understanding of the new sets and so on borrowing the 'mechanics' from faceted search and the 'understanding' from data mining (Di Blas et al., 2014b).

3. Evaluation

Five research groups with different backgrounds (technology and pedagogy) were involved in a two-month experimentation of the portal as research tool. They were asked to perform an investigation using the portal and its affordances. The following researches were performed:
- "The Unexpected Learning: How Unexpected Benefits Can Be Generated Through ICT-Based Experiences," by Polytechnic University of Marche;
- "PoliCultura & Moodle: A Blended Learning Environment," by the University of Perugia;
- "Educational Technology at Primary School Level: A Survey," by the University of Perugia;
- "Inclusion and Group Work: What Benefits?" by Polytechnic of Milan;
- "Educational Technology at Junior High-School Level: A Survey," by University of Bologna;
- "Technology@School: Analysis of an ICT-Based Format," by University of Bologna;
- "Investigation on the Relationship Between Technology and Educational Benefits," by University of Salento.

Some of the reports were turned into scientific papers (e.g. Di Blas, Paolini, 2013; Falcinelli, Laici, 2012). After the work, participants were asked (via semi-structured interviews) to evaluate the portals. The following are the main findings:
- content: researchers acknowledged the in-depth understanding of the experiences afforded by the portal's wide range of materials as well as the direct access to the 'raw' data;
- HCI (human–computer interaction): the mechanics of the interface were quickly understood;
- browsing: the most important method was the formulation of a research hypothesis with subsequent verification;
- serendipity: some groups experimented browsing around and 'stumbling' into interesting data. Researchers declared that this method was a pleasant novelty.

Conclusions and future works

Exploratory portals are a powerful tool to support researchers in education (but not only) to store, share and investigate research data. They can also be clearly used by other targets: teachers on the job (looking for inspiration) and school principals and decision makers, in charge of taking sometimes difficult decisions (e.g. would it be a good idea to adopt tablets at high-school level?).

HOC-LAB is currently developing, in the frame of a national project and in cooperation with GSA-Lab (Graphics and Software Architectures Laboratory) of the University of Salento, a new generation of portals, again in the field of education and building on the lessons learned with the Learning4All portal (Di Blas, Paolini, 2014). The main improvements (apart from the strictly technical ones that include better performing algorithms: speed is crucial to support effective exploration) regard the concept of 'relevance': if, currently, the values are just 0 versus 1 (e.g. 'tablets were used = 1; tablets were not used = 0'), in the future, a fuzzy scale will be introduce in order to (try to) express relevance. Using the same example, the use of tablets could be evaluated as 0,1 or 0,2 or 0,3 and so on according to how relevant tablets were in a specific experience.

Moreover, the interface is being changed to give more space to the facets with respect to the 'mosaic' of the experiences: in this way, the researcher can better focus on research parameters and better check the results. Tools to allow portal customisation are also being created, thanks to which, for example, a researcher will be able to add/subtract facets of interest, customising the interface according to his/her research interest. Eventually, future developments include tools to support collaborative investigation sessions amongst remotely located researchers.

Acknowledgments

References

Di Blas, N., Fiore, A., Mainetti, L., Paolini, P., Vergallo, R. (2014a). A portal of educational resources: providing evidence for matching pedagogy with technology. In *Research in Learning Technology*, vol. 22, ISSN: 2156-7069, UK: Co-Action Publishing, May 2014, 1-26

Di Blas, N., Mazuran, M., Paolini, P., Quintarelli, E., Tanca, L. (2014b), Exploratory computing: a draft manifesto", In *Proceedings of DSAA 2014 (International Conference on Data Science and Advanced Analytics)*, Shanghai, China, Oct. 30-Nov. 1, 2014

Di Blas, N., Paolini, P. (2014), Exploratory portals. The need for a new generation. In *Proceedings of DSAA 2014 (International Conference on Data Science and Advanced Analytics)*, Shanghai, China, Oct. 30-Nov. 1, 2014.

Di Blas, N., Paolini, P., Spagnolo, L. (2012), *Policultura portal: 15.000 students tell their stories about cultural heritage*. In N. Proctor and R. Cherry (Eds.), *Museums and the Web 2012, Selected Papers from an International Conference*, Archives &Museum Informatics, 2012.

Di Blas, N., Paolini, P. (2013), Technology and group work: inclusion or diversification of talents? In Parmigiani, D., Pennazio, V., & Traverso, A. (Eds.). *Learning & Teaching with Media & Technology*. ATEE-SIREM Winter Conference Proceedings. 7-9 March 2013, Genoa (Italy). Brussels: ATEE aisbl, pp. 218-231.

Falcinelli, F., Laici, C. (2012). Teaching with ICT: the policultura and moodle didactic format experimented in schools, IJCEE, Vol. 2, No. 1, January-March 2012.

Ferrari, L., Di Blas, N., Paolini, P., Arpetti, A., Lanzillotti, R., Falcinelli, F., Vergallo, R., Ierardi, M.G., Pacetti, E. (2012) Learning for all: is everyone learning?. In Theo Bastiaens & Gary Marks (Eds), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012*, pp. 1782-1792.

Ferré, S., Ridoux, O. (2004). Introduction to logical information systems. *Information Processing & Management, 40(3), 383–419.*

Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. (Eds.) (2001). *Data Mining for Scientific and Engineering Applications*. Vol. 2, Springer. ISBN: 978-1-4020-0033-1.

Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques. In *The Morgan Kaufmann Series in Data Management Systems*, Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791

Hearst, M., Swearingen, K., Li, K., Yee, K.P. (2003), Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 401-408). ACM.

Houghton, J., Gruen, N. (2014). *Open research data*. Report to the Australian National Data Service. http://ands.org.au/resource/open-research-data-report.pdf. Retrieved September 2015

Kvalheim, V., Kvamme, T. (2014). IFDO Report 2014: *Policies for Sharing Research Data in Social Sciences and Humanities*. Norwegian Social Science Data Services. http://ifdo.org/wordpress/wp-content/uploads/2015/07/ifdo_survey_report.pdf. Retrieved September 2015.

Sacco, G. (2006). Dynamic taxonomies and guided searches. *Journal of the American Society for Information Science and Technology*, 57(6), 792–796.

Spagnolo, L., Bolchini, D., Paolini, P., Di Blas, N. (2010). Beyond findability: search-enhanced information architecture for content-intensive RIAs. *Journal of Information Architecture*, vol. 2, issue 1, pp. 19-36

Tunkelang, D. (2009). "Faceted search", *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Morgan and Claypool Publishers. ISBN 1598299999

Tzitzikas, Y. (2009).Faceted Taxonomy-based Sources. *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*. 25, (pp. 19-34), Springer Publishing Company, Incorporated, (ISBN 3642023584, 9783642023583)