Cross-Domain Crawling for Innovation

Pierluigi Assogna and Francesco Taglino

Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti" - IASI-CNR
Viale Manzoni 30, 00185 Roma, Italy
pierluigi.assogna@mensa.it, taglino@iasi.cnr.it

Abstract. Innovations, in any field, originate in the mind of people, on the base of mechanisms not yet completely understood. There have been many studies relevant to thinking techniques that have been proven to favor creativity, like for instance those studied by De Bono. A general characteristic of these techniques is the recommendation of avoiding usual thinking paths, habitual mind frames: this is facilitated by putting oneself in unusual physical settings, or introducing absurd concepts, and the like. The use of metaphors is another recognized enabler of creativity, by bridging different conceptual domains.

A Knowledge Base (KB) structured around an Ontology can be seen as a close simulation of the conceptual structure that, according to Constructivism, supports a person's thinking processes, and the Web can be seen as the corresponding world to be explored and that contributes to that person's culture. This kind of domain specific KBs is being organized and used as support for advanced enterprise information systems. This paper presents a technique for extending the working domain (WD) of an organization with concepts belonging to other domains, obtained by retrieving documents that discuss both concepts of this WD and "foreign " ones. These documents, proposed to the KB editors, are considered candidates for innovative problem solving activities and considerations.

Keywords: Ontology, Web Crawling, Innovation, Creativity.

1 Introduction

Innovations, in any organization, can of course only come from the people involved there, and not from the systems. Systems can support all the stages of the process of innovation, as for instance analyzed in the BIVEE project[1]. The first spark that triggers an innovation has not yet found effective support: there are no established ICT methods and tools specifically focused on this starting discontinuity.

As an example of the SotA in the field of creativity support, we can see that within the European Objective 8.1, 'Technologies and Scientific Foundations in the field of creativity' [1] we can find, at their start, the development of 3 platforms (i-Treasures, CULTAR, RePlay) for supporting the use of cultural resources, 2 projects targeted at promoting creativity: Idea Garden [2] (planning to develop hard and software

[1] http://wordpress.bivee.eu

X. Franch and P. Soffer (Eds.): CAiSE 2013 Workshops, LNBIP 148, pp. 286–297, 2013.

technologies that assist designers during all phases of the creative process), and Collage [3] (planning to develop an innovative Social Creativity Service-Set), plus a number of systems aimed at capturing, maintaining, providing a structured view of cultural heritage.

The implicit assumption for these projects is: the main ingredient for supporting the process of creation is the availability of existing knowledge, that needs to be adequately proposed to (human) designers.

The method proposed in these notes aims at providing a pre-digested form of this ingredient, simulating the processes allegedly going on (according to recent studies) in our minds when we are confronted by new issues. Along this line the European Commission launched in 2009 the Human Brain Project, aimed at analyzing the thinking processes of the mind. One of the objectives of it is to "build revolutionary new computing technologies" [4].

A Knowledge Base (KB), together with the supporting semantic services, should simulate as close as possible the architecture and workings of our mind, in order to be an effective support to our thinking and decision making processes, like a sort of super-fast Thinking Assistant. There are two big issues related to this approach: a) we do not really know how a mind works at this abstract level, and b) what we know for sure is that our mind is probably the most complex system that we know of in the universe. Luckily both issues, rather than discourage research and commercial efforts, keep driving a substantial body of researchers and stakeholders, and promoting an interesting convergence of neural scientists and ICT experts.

This paper presents a research hypothesis that combines accredited theories of mental processes with state of the art knowledge -management and -mining technologies, as a support for creative, innovative thinking.

2 Cognitive Issues

2.1 Construction of Conceptual Structures

From Wikipedia we get a good definition for Constructivism: "… is a theory of learning and an approach to education that lays emphasis on the ways that people create meaning of the world through a series of individual constructs". According to this theory, that is one of the most accredited, our mind, our personal culture, is a dynamic (generally growing) network of concepts and models, that starting from a limited set genetically inherited, keeps getting more and more complex during our life. The "father" of Constructivism is generally considered Jean Piaget [5].

The models are used to organize experiences, that in turn are used to create more models. This paradigm is very distant from Locke's *tabula rasa*, as it presumes the existence of a first set of models, and is in tune with Kant's *a priori* categories. An exaggeration of this conception is the stereotype that you find easily just what you are looking for.

An interesting theory, Memetics, now dwindling out of popularity, is a reductionist approach related to knowledge acquisition, proposed by Richard Dawkins in his book "the Selfish Gene" [6], where concepts and methods that are experienced, exchanged,

learnt by people migrate from mind to mind like genes across DNA's strands, so that Darwinian mechanisms can be applied to them.

It is of particular interest that two theories stemming from two fiercely opposed approaches (teleology vs causality) can both provide hints for the proposed simulation of conceptual processes related to creative thinking. Constructivism because of its focus on the importance of models for knowledge acquisition and integration, Memetics because of its "objectification" of concepts and models, and of the underlining idea that "memes" are capable of auto-organization.

The stage of the proposed simulation is composed of:

- Knowledge Bases and services seen as a computerized version of minds, in terms of evolution and usage
- The conceptual structure, represented by one or more ontologies, populated with topics (and related content) linked to each other
- The net as the world to be experienced (searched)
- The auto-organizing capability of mind, represented by the application of specific methods for the semantic annotations (connections) that characterize each conceptual node
- Conscience, that In orthodox Memetics has no (or very scarce) place but in our case has a strong role, represented by the stakeholders (contributors/users) of the KB. The semantic annotations that each contributor is asked to apply to an inserted topic (concept, class, method, process, etc.) make up the overall structure, the KB's culture.

The main challenge is the possibility of simulating the (supposed) auto-organizing capability of memes in minds. In this way a KB could exploit the enormous corpus of knowledge items present in the net, taking advantage of the searching speed of bots.

The focus of these notes is innovation, so that as first step we need to consider the mind's functionalities that seem to promote innovation, and then proposing how to simulate them.

2.2 Lateral Thinking and Metaphors

Lateral Thinking is an expression coined in 1967 by Edward de Bono [7], an author that has devoted much of his work to problem solving methodologies. Even if criticized, his ideas have had considerable impact on the methods aimed at finding creative solutions to all sort of problems. The basic approach is that of figuratively "stepping out" of a top-down or bottom-up rational thinking process, when faced by an issue, and looking at the situation from a lateral view point, that is from a "path" of the conceptual structure that is not strictly connected to the ones facing the issue. An example of this approach is opening a dictionary at a random page, taking a random item, and seeing how it applies to the issue at hand.

If we consider the "world of ideas" represented by the sum of all human cultures as an interconnected conceptual structure, this technique means associating, on an ad-hoc mode, topics pertaining to remote sections of the structure, while each singular mind hosts connected memes that belong to specific limited domains.

It is widely accepted that innovations, from art to technology to any other field, come by trodding scarcely or never used paths connecting concepts, methods, processes, situated in distant branches of knowledge structures.

An interesting experiment of a semi-automated proposition of heterodox connections is represented by the "metaphorical search engine" experimentally provided by the organization YossarianLives![2]: the idea of their approach is that while traditional search engines tend to retrieve documents related more or less closely to the keywords entered, consolidating in this way the implied conceptual domain, their metaphorical search retrieves documents that are metaphorically connected, that hence could help the searcher to address "laterally" the issues he/she is affording.

Metaphors are a very powerful way for triggering innovative and out-of-the-box ideas in relation to an issue. Their use can be seen as a form of lateral thinking that starts from a neighborhood that, even being different, has some subtle relations with the one where the issue starts.

The first authors to point out the fact that our common language is ripe of metaphors have been Lakoff and Johnson [8]. They analysed the characteristics of a metaphor. It connects two conceptual domains, termed "source" and "target", and exercises a conceptual transfer between the two. In order to be effective, that is to trigger new perspectives and conceptual paths, it has to comply to the so-called "Invariance Hypothesis": the mapping can exercise its suggestion if it is applied to similar "image-schemas" [9] on both sides of the connection. These image-schemas (examples proposed by Johnson are: Merging, Matching, Iterating, Splitting, and the like) can be interpreted as processes.

Many phenomena share similar evolution processes, in particular those related to complex environments. Their direct comparison can provide useful hints: if you draw parallels between drivers, trends, obstacles, enablers, anything that characterizes the stages of these processes, you may find that specific situations, that are clearly outstanding in process X, can help in shading light to a parallel situation in process Y, where these are not easy to pinpoint, and this can boost innovative considerations and actions.

2.3 Automatic Brain Activities

Neurological researchers, particularly through the recent use of brain imaging techniques, have demonstrated that in experiments (see for instance [10]) where a person is asked to perform simple actions requiring a rational choice (like pushing a button every time a specific sound scheme is perceived) the conscious decision related to the action is always registered in the brain seconds after the action has been automatically performed. This outcome has been erroneously used by many as a proof that free will (or conscience) does not exist. A person who plays a musical instrument, let's take a piano, has always been aware of this situation: you have to pay attention to the keys you press only in the early study stage of a specific piece; when you have learnt it, if while you are playing you happen to think of the keys you are pressing, a mistake is

[2] www.yossarianlives.com

guaranteed. This simply means that a task that initially requires a conscious decision process, is by and by delegated to lower level processes. In this case the functions that typically maintain the request for attention are those like the loudness, and the expression you want to communicate to the audience.

In any case we know that most of the activities of our brain are automatic. In the case of a creative thinking process you have a conscious tension, that leads you to walk many paths, some of which have been built automatically. There are many stories of solutions found while sleeping, so also the exploration can go on automatically.

Artists, for instance, keep integrating into their conceptual structure bits and pieces of esthetic memes, and this happens because they consciously at first, and automatically with practice, know that any of these could become useful sometime. These are then used, when creating an artifact, as if they were there by chance or by an overall automatism.

2.4 Extensions of Domain Knowledge Bases

A person can master a very limited part of the global knowledge, and along the same line an organization maintains, in order to provide a semantic structure for its activities, a domain-oriented Knowledge Base, as it would be unreasonable and unuseful to maintain a "universal" base.

In this case *domain* means not only a bounded set of concepts, but also a local, subjective interpretation and appreciation of each concept. In fact the KB is mainly used by humans, and in this respect the ambiguity of natural language, that is the main communication mode among minds, rather than be considered as a disadvantage, is to be seen as the main source of creativity and explosion of the global culture.

In this respect the considerations related to the role of *boundary objects* within a specific organization (see for instance [11]) underline the fact that different functional units can assign different semantic value to the same concept, and the "boundary" documents that are exchanged between them have the potential at the same time of causing misunderstandings or of promoting innovative ideas. The considerations presented in this Paper explore the value of crossing more drastic boundaries, those existing between disconnected domains, where the melting and contrasting of heterogeneous concepts can trigger innovation.

Most, if not all, of the topics that populate the typical domain KBs are relevant to the organization's mission. Additional and "diverging" knowledge is typically imported by its users, either by following specific searches related to any new situation or important event, or through the use of tools like RSS subscriptions or "follow X". In any case these "institutional" enrichments do not generally tap into different domains of interest. This mode of using the existing knowledge is analogous to a person using his/her settled personal models, skills, experience, to manage events. There are cases when a person has to update his/her models, when for instance an experience is not easily processed using them, and it is anyway considered important. In these cases the flexibility of modifying one's models can be very important for taking benefit from these experiences. The creation of new models requires the integration, into

one's conceptual structure, of new memes, that are concepts, processes, and so on, pertaining to different domains. And this has to be done ad-hoc, when required, as it is by definition impossible to forecast the unexpected.

It would be preposterous for a person to think of enriching his/her culture in all directions, and with enough depth, to be prepared for any unexpected situation. What is needed is a) the predisposition to change, b) the availability of tools for mining useful pieces of knowledge in domains that are related to the new situation that needs to be afforded (the "know who knows what" approach), and c) the capability of using them efficiently. The same works for organizations: their KBs have to be focused on their usual business, and the related ontologies management systems must include mechanisms ready for supporting ad-hoc mining activities.

A way that a person has, in order to be prepared for these possible sudden requirements for a fast integration of substantial chunks of new topics and models (like for instance when moving from a city to another one, may be in another Country) is to maintain a basic knowledge related to domains "neighboring" to the one this person is managing, as the probability of having to explore a different conceptual sector is higher for contiguous rather than for totally removed ones. This knowledge not focused on one's daily interests can be called lateral, also in view of its potential capability of triggering lateral thinking in day-by-day tasks. The problem here is that even limiting the number of candidate contiguous domains, this number can be very high, so that subjective selection methods are generally applied.

An organization, on its part, needs only tools for exploring the net when the need to expand its knowledge arises, that is when a creative solution for an issue is required.

3 Ontology-Guided Crawlers

On the base of the considerations that ontologies can represent the "conscience" of an organization, and that the net represents the world to explore, we propose to automate some of the steps of the process that we have described above, in order to support the search of documental resources in extended and loosely related domains.

An option that we are exploring is based on the organization of a crawler capable of searching documents that in specific ways (controlled by the organization's stakeholders) are related to an issue, or are considered candidate for triggering new ideas relevant to the organization business.

In particular, here we outline a semantics-based search method that integrates automatic and human controlled steps (Fig. 1). For knowledge extraction from text, we are currently using Alchemy[3], a set of APIs able to analyze textual documents by using sophisticated statistical algorithms and natural language processing technologies. A sketchy architecture of the infrastructure that implements the method is reported in Fig 1 and described in terms of its main components as follows.

[3] http://www.alchemyapi.com/

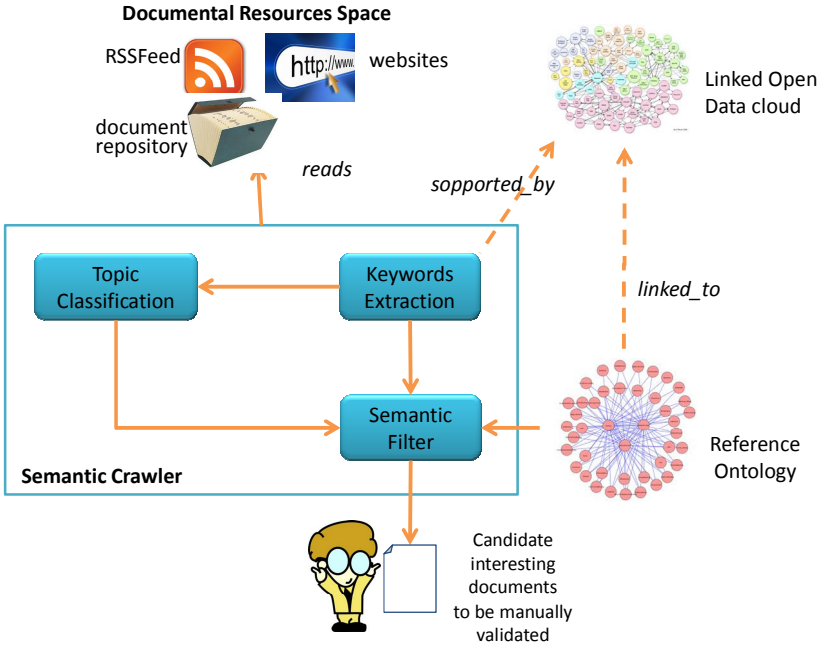

Fig. 1. Semantics-based cross-domains crawling

Documental Resources Space: This is the space where we search for interesting documents. Here, we can have different kinds of resources, for instance websites, rss feed, and public documents repositories. As illustrated later in the examples, our reference domain is Robotics and Machine Vision (R&MV). So, we could search in very technology- and innovation-oriented sources (e.g., the section of the MIT website on innovations[4]), as well as in generalist websites like the BBC news portal[5].

Bridging the Reference Ontology and the Linked Open Data Cloud. As a preliminary step we assume to have a reference ontology (or at least a vocabulary) specific for our domain of interest, e.g., Robotics and Machine Vision (R&VM) and to associate it to contents in the Linked Open Data[6] cloud. For instance, in the BIVEE project, we currently have built a glossary of almost 600 entries about Robotics and Machine Vision. In this first implementation, for each entry in our glossary we tried to identify a corresponding entry in the DBpedia dataset (DBpedia is a project aiming to extract structured content from the information created as part of Wikipedia). If the entry exists we created a link between the entries via the *owl:sameAs* property. For instance, the *Photodiodes* and *Camera* entries in our R&MV glossary have been

[4] http://web.mit.edu/newsoffice/topic/innovation.html
[5] http://www.bbc.co.uk/news/
[6] http://www.w3.org/DesignIssues/LinkedData

linked to the *Photodiode* (http://dbpedia.org/page/Photodiode) and *Camera* (http://dbpedia.org/page/Camera) entries in DBpedia, respectively.

Terms Extraction: This module is in charge of extracting relevant terms from a given document. Relevant terms are those which are intended to be representative and somehow synthesize the document's content. For each extracted keyword, the *URLGetRankedConcepts* method from the AlchemyAPI, which has been used for the keywords extraction, reports the reference to the corresponding DBpedia entry, and a numeric value in the range [0..1] representing the relevance of the keyword with respect to the document.

Topic Classification: This module is in charge of classifying candidate interesting documents with respect to their main topic. The *URLGetCategory* from the AlchemyAPI is able to classify documents in one of the following categories: *Arts & Entertainment*, *Business*, *Computers & Internet*, *Culture & Politics*, *Gaming*, *Health*, *Law & Crime*, *Religion*, *Recreation*, *Science & Technology*, *Sports*, and *Weather*. Since they are high level categories, extracted terms could help in refining the classification.

Semantic Filter

Once keywords have been extracted, the semantic filter is in charge of proposing if the analyzed document is an interesting resource. In this proposal, we define a metrics based on the relevance value of each extracted keyword. First we need to identify extracted keywords that are related to our domain of interest (e.g., Robotics and Machine Vision). For collecting keywords related to the specific domain of interest, for each keyword, one of the two criteria must be satisfied:

- it exists an entry in the domain specific glossary that is *owl:sameAs* the DBpedia entry representative of the extracted keyword. For instance, considering the example related to Document 1 in Table 1, the extracted keyword *Camera* is considered in the domain of R&MV because the *Camera* entry is in the R&MV glossary and it has been previously linked to the *Camera* entry in DBpedia.
- it exists an entry in the domain specific glossary that has been defined *owl:sameAs* a given DBpedia entry, and this DBpedia entry has in common with the DBpedia entry representative of the extracted keyword a subject. For most of the DBpedia entries a set of subjects is defined through the Dublin Core *subject* property. For instance, both *Photodiode* and *Light-emitting-diode* entries in DBpedia have *optical_diodes* among their subjects. For this reason, considering the example related to Document 1 in Table 1, the extracted keyword *Light-emitting-diode* is considered to be related to the R&MV.

Table 1. Example of anlyzed articles from the BBC news web portal

Document 1	http://news.bbc.co.uk/2/hi/science/nature/1542588.stm			
Classification	science_technology		**R&MV**	**Other**
Text and **Relevance**	Light-emitting diode(*)	0.913929	**0.37**	**0.48**
	Slug	0.858322		
	Power (*)	0.855041		
	Foot-and-mouth disease	0.854825		
	Mucus	0.832959		
	Camera(*)	0.831103		
	Soil	0.830792		
Document 2	http://www.bbc.co.uk/news/technology-10687701			
Classification	computer_internet		**R&MV**	**Other**
Text and **Relevance**	Female body shape	0.967518	**0.11**	**0.40**
	Body shape	0.635835		
	Clothing	0.476781		
	Human body	0.467204		
	Robotics(*)	0.447413		
	Robot(*)	0.441914		
	Fashion	0.342003		
	Physical attractiveness	0.331898		
Document 3	http://www.bbc.co.uk/news/health-21965092			
Classification	Health		**R&MV**	**Other**
Text and **Relevance**	Obesity	0.976225	**0**	**0.54**
	Bariatric surgery	0.597715		
	Gastric bypass surgery	0.535516		
	Weight loss	0.479716		
	Bacteria	0.464552		
	Nutrition	0.45946		
	Bariatrics	0.45838		
	Medicine	0.374071		
Document 4	http://www.bbc.co.uk/news/business-20800118			
Cassification	arts_entertainment		**R&MVs**	**Other**
Text and **Relevanc**	Robot(*)	0.971036	**0.42**	**0.14**
	Robotics(*)	0.691485		
	White-collar worker	0.615792		
	Industrial robot(*)	0.509681		
	Humanoid robot(*)	0.418013		
	Manufacturing	0.37688		
	Automaton(*)	0.347331		

Once we have identified those extracted keywords realted to the specific domain of interest, a document is considered a valid candidate if the sum of the relevance values for the extracted keywords belonging to our specific domain divided by the number of extracted keywords is greater than 0.1 and less than 0.4. At the same time, the sum of the relevance values for the extracted keywords not belonging to our specific domain divided by the number of extracted keywords must greater or equal to than 0.4. The first condition ensures that the analyzed document deals with our reference domain, but in a small manner, while the second constraint ensures that the analyzed document deals with other topics in a considerable measure. According to that, we report in Table 1 the results related to four analyzed articles from the BBC news web portal. Keywords extracted from the documents are marked with an asterisk (*) if they are considered to belong to the R&MV domain.

- Document 1 and Document 2 are considered relevant. This meets exactly our expectations since they consider Robotics and Machine Vision in very singular applications: one for extracting energy from insects and the other for supporting to help shoppers get the right fit when buying clothes online.
- Document 3 is not considered relevant because it does not consider Robotics and Machine Vision at all.
- Document 4 is too much Robotics oriented, so it can be surely useful for experts in the Robotics field, but it does not appear too inspiring for lateral thinking activities.

4 Related Works

Slug[7] is a web crawler designed for harvesting semantic web content. Implemented in Java using the Jena API, Slug provides a configurable, modular framework that allows a great degree of flexibility in configuring the retrieval, processing and storage of harvested content. The framework provides an RDF vocabulary for describing crawler configurations and collects metadata concerning crawling activity. Crawler metadata allows for reporting and analysis of crawling progress, as well as more efficient retrieval through the storage of HTTP caching data.

LDSpider[8] includes handlers to read RDF/XML, N-TRIPLES and N-QUADS. Simple interface design to implement own handlers (e.g. to handle additional formats). Different crawling strategies: Breadth-first crawl; Depth-first crawl; optionally crawl schema information (TBox). Crawling scope: crawl can easily be restricted to specific pages e.g. pages with a specific domain prefix. The crawled data can be written in various ways, such as RDF/XML or N-QUADS. The crawler can write all statements to a Triple Store using SPARQL/Update. Optionally uses named graphs to structure the written statements by their source page. Optionally, the output include provenance information.

APERTURE[9] crawls file systems, websites, mail boxes and mail servers. It extract full-text and metadata from many common file formats.

The major concern about these crawlers is the fact that the searching strategy cannot be guided in terms of a domain specific ontology. However, they will be better investigating for understanding the opportunity to integrating them.

[7] http://ldodds.com/projects/slug/
[8] http://code.google.com/p/ldspider/
[9] http://aperture.sourceforge.net/index.html

5 Conclusions and Future Works

The outlined method wants to be a very preliminary work showing that, with existing semantic tools, it is possible to support people in exploiting our mind's capability of cross-fertilizing processes typical of a specific domain with concepts pertaining to a different one, obtaining in this way new insights and viewpoints.

The key concept of the proposed foray for new ideas in the Web is searching for knowledge resources that are not completely centred on one's principal domain of interest, but concerned about both this domain and one or more other ones: the assumption is that a document with these characteristics has a high probability of talking about applications that span these domains, or of using interesting analogies or metaphors connecting them.

The work is in a very early stage and several feature can be improved. About knowledge extraction from documents, we are investigating additional solutions with respect to AlchemiAPI, such as Open Calais[10], Zemanta[11], and Fise[12], considering also the opportunity to integrate more than one single tool. About bridging a domain specific ontology and the contents in the Linked Open Data cloud, we are investigating on using additional ways and not only the *owl:sameAs* property. For instance we could use the generalization/specialization relationships (via the *rdfs:subClassOf* property), and weighting types of links and distances between DBpedia (and not only) entries. Then we also need to massively test the method on a significant number of documents and evaluating at least the precision of the method.

Acknowledgments. This work has been partly funded by the European Commission through the ICT Project BIVEE: Business Innovation and Virtual Enterprise Environment (No. FoF-ICT-2011.7.3-285746). The authors wish to acknowledge the Commission for its support. We also wish to acknowledge our gratitude and appreciation to all BIVEE project partners for their contribution during the development of various ideas and concepts presented in this paper.

References

1. Technologies and Scientific Foundations in the field of creativity (March 01, 2013), http://cordis.europa.eu/fp7/ict/creativity/creativity-objectives_en.html
2. Idea Garden (March 01, 2013), http://idea-garden.org/
3. Collage (March 01, 2013), http://projectcollage.eu/
4. Human Brain Project (March 01, 2013), http://www.humanbrainproject.eu/vision.html
5. Jean Piaget (March 01, 2013), http://en.wikipedia.org/wiki/Jean_Piaget

[10] http://www.opencalais.com/
[11] http://www.zemanta.com/
[12] http://wiki.iks-project.eu/index.php/FISE

6. Dawkins, R.: The Selfish Gene. Oxford University Press. New York City (1976) ISBN 0-19-286092-5
7. de Bono, E.: The Use of Lateral Thinking (1967) ISBN 0-14-013788-2
8. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago (1980)
9. Johnson, M.: The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason, University of Chicago (1987)
10. Unconscious determinants of free decisions in the human brain (March 01, 2013), http://www.nature.com/neuro/journal/v11/n5/full/nn.2012.html
11. Carlile, P.R.: A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development. Organization Science 13(4), 442–455 (2002), doi:10.1287/orsc.13.4.442.2953