Skills Text Box A Tool to Access Resources by Mathematical Concepts

Paul Libbrecht
CERMAT, Karlsruhe University of Education

ABSTRACT

Searching for mathematical concepts being indicated in a document is not easy with the existing technologies because the current information retrieval technology has been designed for words and mathematical concepts are often made of more than single words.

Skills-text-box is an approach to this retrieval problem: it lets users use the classical search by words paradigm to search for the concept then identify it by choosing through a finite list. Skills-text-box is the device used to support the search engine of i2geo.net: at contribution and search time.

In this paper for MathUI 2012, we sketch the current technical development Skills-text-box, and present its advantages and limits as have been experimented by multiple mathematics teachers in Europe.

Introduction

Classical search engines are often insufficiently precise to search for mathematical documents. The approach presented in this paper attempts bridging the gap between a collection of learning resources which one knows well (such as a text-book in use for that year or an own collection of online resources), and a world wide web which tends to be a bit too wide (repeating itself often and rarely matching the exact expectations).

The way we tackle this enterprise is by a community platform, called http://i2geo.net, which proposes to share, annotate, and search for learning resources. This platform is meant to be easy to contribute and search and supports the fairly multilingual nature of geometric constructions by searching across the barriers of languages: e.g. a teacher is Spain should be able to find a resource contributed by a teacher in the Czech Republic without needed to speak that language.

This search and contribution approach supports practicing mathematics teachers which means that the annotations and queries are sufficiently fined grained so that *topics of next weeks course* can be expressed. For example, it is possible to differentiate resources training *right-angled triangle* and resources training *right-angles* and *triangles*: the precise nature of the mathematical concepts. We seem to observe that the mathematical science is very rich of these composite concepts and thus need a special treatment.

To this effect, the i2geo project, which finished in 2010, has encoded an ontology with the following concepts:

- a hierarchy of *topic*s relevant to the domain of mathematics
- a set of *competency*s object
- educational levels for each of the classes of Europe

Based on these annotations, the searches described above are feasible. This paper is a description of the central enabler of this search and annotation tool: a browser-component that supports the input of concepts of this ontology efficiently called Skills-Text-Box.

In this paper, we describe the detailed implementation and user-experience of this auto-completion component. It has grown between 2008 and 2010 in the Inter2Geo EU project. Papers about it include a description of the overall platform (I2Geo-DML-2009), of the overall project (inter2geo-CERME6). Its use has been successful at times but has also triggered reactions. We present the achievements and issues that have emerged after this experience, some of which were entirely unforeseeable during the planning phase.

Outline

We start with a description of the techniques used to achieve the objectives expressed above. This is followed by a detailed description of the user-experience in the current skills-text-box. Then we present achievements and limitations that we have met and open research questions that try to address these limitations.

Technical Underpinnings

The Skills-text-box library is a client and server software written in the Java programming language exploiting an ontology.

On the client, Skills-text-box is run in JavaScript: the usage of the Google Web Toolkit library (GWT) allows this compilation from Java into JavaScript code that works on the web-browser's object model. The client component displays a text-field. After about 200ms after the text-field has been modified, queries to the server are sent to suggest the nodes matching what has been typed. Once the suggestions have arrived, as an XML file, and are still valid, the result is displayed below the text-field offering the user to choose the node.

The data being searched is the GeoSkills ontology: a knowledge structure that represents topics as a hierarchy, competencies as a complex object (with verb and objects), and educational levels, pathways, and regions. GeoSkills is described in (GeoSkills-SemWeb-Handbook). Technically, GeoSkills is developed as an OWL ontology (an OWL DL ontology) which can be easily parsed and queried for. The ontology has grown along the years, reaching thousands of nodes which made it big enough to render unusable several editors. While Protégé was used, at the beginning, to edit the ontology, its usage has been

replaced a web-based tool that receives the contributions of curriculum experts in multiple languages: CompEd, a web-application that is running aside of Skills-text-box and is updating the ontology on a regular basis, see (CompEd-SWEL-2009) and the GeoSkills page.

The nodes of GeoSkills, including topics, competencies, and levels, all carry names so that they can be displayed to the user and can be queried for. The names are classified by frequency so that it is possible to rank their importance when searching through them: common-names are frequently used, less-common-name and rare-names are rather exception (but one should still find it if entering such a name), and false-friend-names represent names that should not be matched. Together, these names permit a quantitive ranking of the matches for a sequence of words as input in the text-box.

The server component of Skills-text-box is an index, based on Apache Lucene. The index stores all the names of each of the nodes. This indexing is performed by crawling through the ontology using the owlapi library and the pellet reasoner.

The server component is running as a separate web-application than the XWiki web-application which serves most of the i2geo platform's sharing needs (itself based on the collaborative asset management XCLAMS. Communication between the two web-applications allows the nodes to be rendered (by their name) and the auto-completion to trigger the search or contribution.
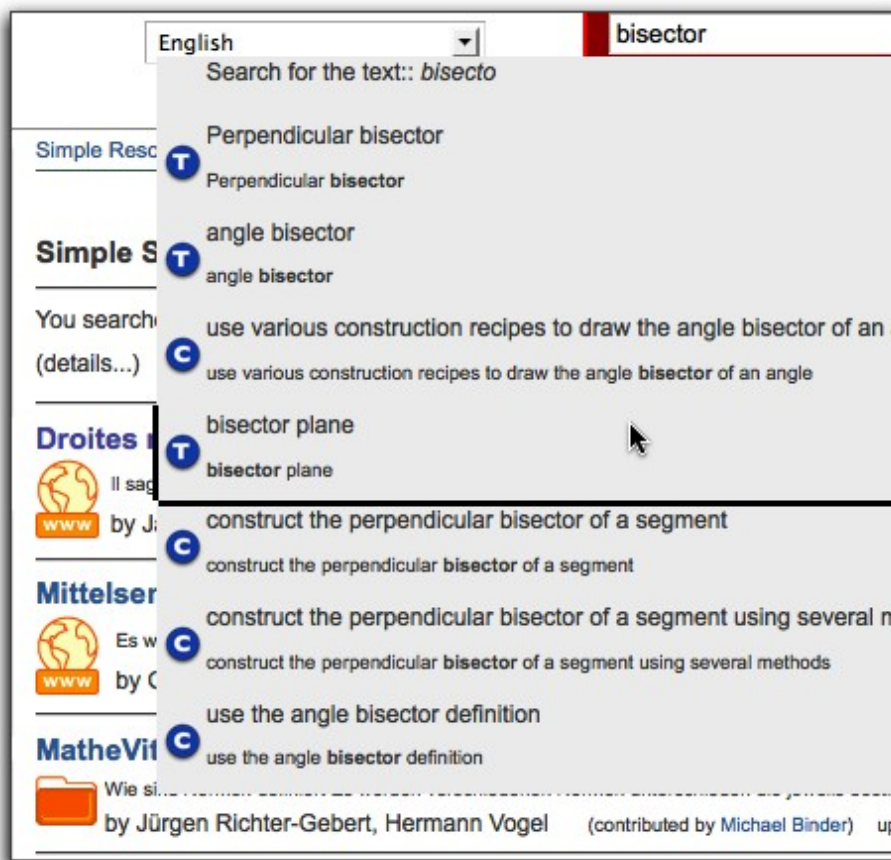
The SearchI2G server and client components, just as the CompEd API, and the i2geo customization to the XCLMS extension of XWiki are all available under the Apache Public License from the projects' pages. They are deployed and used on http://i2geo.net/ hosted by the Universy of Halle.

Current User Experience

When entering i2geo.net, users can search for resources by using the inputting a few characters in the search box on the top right. Soon after stopping to type, a waiting wheel indicates that auto-completions are being searched but do not prevent the user to keep typing: HTTP requests sent to the Skills-text-box auto-completion index are sent. When returned, after a duration of 1685 milliseconds in average, and the textfields hasn't been changed, the suggestions' popup is returned as on the left. The user can then choose the appropriate node of GeoSkills and trigger a search for the indicated node with either the mouse or the curosr and enter keys.

The auto-completions popup is made of the nodes of the GeoSkills ontology which, in order of preference:

English ▼ | bisector

Search for the text:: *bisecto*

T Perpendicular bisector
Perpendicular **bisector**

T angle bisector
angle **bisector**

C use various construction recipes to draw the angle bisector of an a
use various construction recipes to draw the angle **bisector** of an angle

T bisector plane
bisector plane

C construct the perpendicular bisector of a segment
construct the perpendicular **bisector** of a segment

C construct the perpendicular bisector of a segment using several m
construct the perpendicular **bisector** of a segment using several methods

C use the angle bisector definition
use the angle **bisector** definition

Simple Resc

Simple S

You searc

(details...)

Droites
Il sag
www by J

Mittelser
Es w
www by C

MatheVit
Wie si
by Jürgen Richter-Gebert, Hermann Vogel   (contributed by Michael Binder)   up

- have a
  name equal to the the text searched for, end to end,
- have a name which contain words of the text searched for,
- have a name which contain words which start with the words of the text searched for

Each time, a name is searched for, one searches first the URI, the common-name, or the uncommon-name (providing a positive score contribution), or the false-friend-name (providing a negative-score-contribution). Moreover, one does this in each of the languages of the user: the language of i2geo being used (which can be changed on top of the page), the list of supported languages that the browser transmits through the Accept-Language header.

Each of the nodes are displayed by type: an icon indicates the type (level, topic, capacity), a small text indicating the name that was matched, and the *default common name* of the node. The default-common-name is a single name per node and language which allows a complete identification of the concept of the node even if the user is out of context: for example, *net of a solid* allows a complete identification enough while it is often searched for or named *net*. Similarly, the French naming of *4ème* (meaning *fourth class*) is insufficiently expressive, but *4ème du collège (France)* or *4ème primaire (Fribourg)* is precise enough; Skills-text-box allows the user to choose between both. Competencies are, typically, expressed as full sentences that contain a verb and its subjects in a way that

mimics the involved topics (for example: *calculate the slope of a line*) but queries search for them would typically not be the exact sentence but a few words approaching it, typically made of the "ingredients" of the competency (e.g. the concepts being manipulated).

If the user is unable to find the node, more typing is needed so that the results' list is made smaller: it seems to be impossible for an auto-completion pop-up to browse several pages of results. Moreover, only particular devices (those with a scroll-wheel or equivalent) allow the screen to be scrolled to view suggestions beyond the current screen. We shall see below that this challenge is a curation challenge that remains open.

Applications of Skills-Text-Box

Skills-text-box is used to *speak GeoSkills*, that is to let the user express concepts encoded in GeoSkills: this is not an objective in itself, but it serves two objectives:
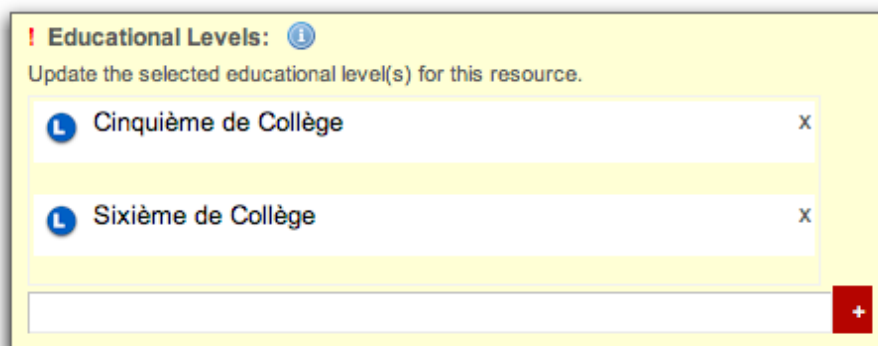
- let the user express a search query for a topic, competency, or level encoded in GeoSkills: this is done either as the sole clause in the simple search or as one of the clauses in the advanced search
- let the user annotate the resources when contributing them using the same language. In this case, as displayed in the screenshot on the right, the user's choices add subjects, competencies among the "trained topics and competencies" or levels among the "target educational levels".

The search exploits the ontological nature of the GeoSkills queries further: if queried for a topic node of the ontology, it also queries for any topic node that is an instance of the class of this topic. This way an *equilateral triangle* is found when a *regular polygon* is searched, however resources annotated with *regular polygon*, in their generality, are preferred. Searching for competencies does a similar generalization: it searches for resources which are also annotated for the same competency-verb and the same objects.

Finally, the ontological nature is used in the subjects' search: subjects are encoded using the ontology editor Protégé by the usage of axioms which allow a fine description of the collections of topics and competencies. The screenshot below shows the editing of the

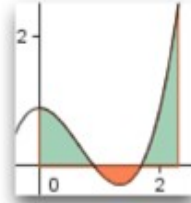subjects as axioms using the Protégé 4 editor and the choice of subjects in the i2geo platform:
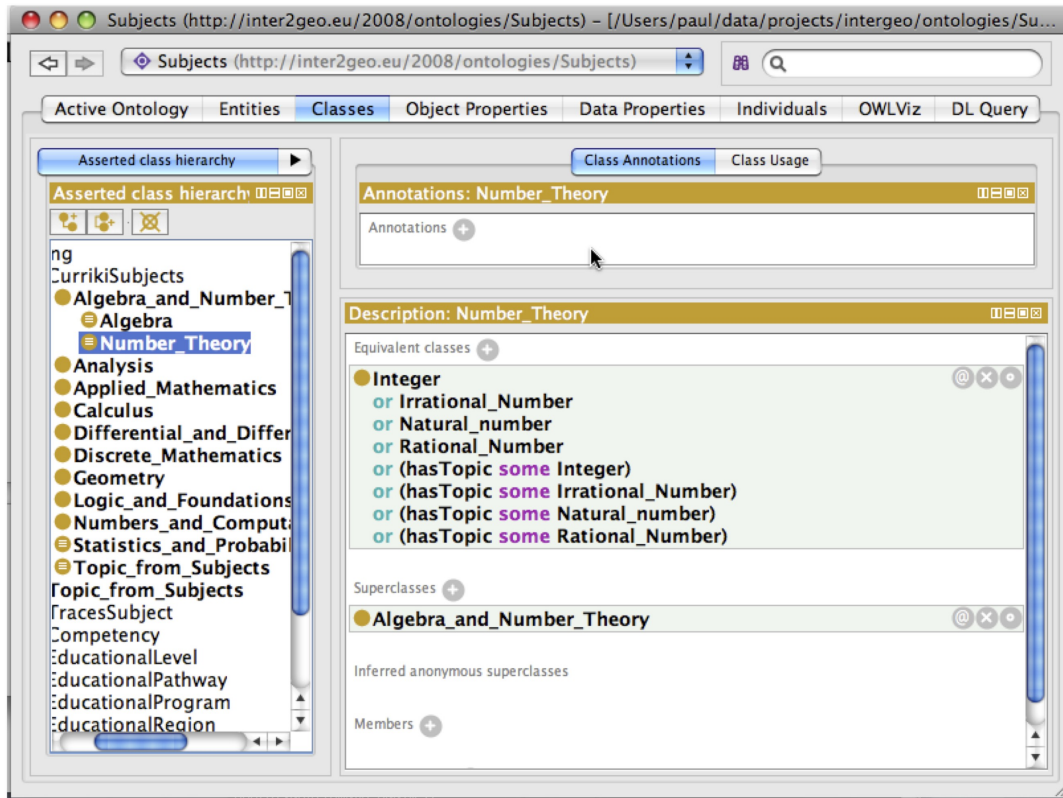
Browse Resources by Subject

$$a^p \equiv a$$
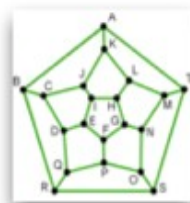$$\mathrm{mod}(p)$$

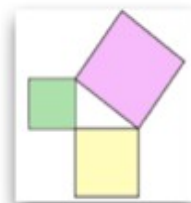Algebra and Number Theory (7)

Applied Mathematics (24)

Calculus (57)

$$\frac{\partial^2 u}{\partial x^2}$$

Differential and Difference Equations (0)
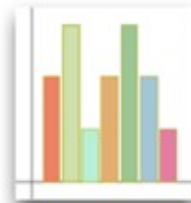
Discrete Mathematics (0)

Geometry (739)

Logics and Foundations (5)

Numbers and Computation (121)

Statistics and Probability (17)

Analysis (0)

Successes and Challenges

The approach to auto-completion and the challenges of cross-language and cross-curriculum search have been sketched at the very start of the Inter2geo project, together with all stakeholders. Its implementation, including the development of the ontology and its knowledge input by curriculum experts in each country, has been realized during the project.

The search and contribution tool has been in regular uses by multiple teachers, in France, Spain, Germany, and the Czech Republic during the Inter2Geo project. This gave rise to feedback, sometimes close to rejection and sometimes enthusiastic. For some, this feedback was described as a log-book describing the multiple attempts at searching and the following evaluation of the applicability of a learning resource for subsequent teaching. This feedback has lead to incremental enhancements the final state of which is described above. Among the crucial feedbacks that came is that plain text search is still an essential feature that should not be discouraged.

Between February and June 2012, after the i2geo project finished and its usage was mostly spontaneous, 3398 auto-completions requests were responded while 2417 simple searches and 303 advanced searches were performed.

Successes

The skills-text-box function has succeeded at least under the perspective of its original missions to provide means to express annotations and queries for elaborate multi-words concepts: indeed, one can search for the phrase ~~angle droit~~ yielding 2 results, search for the words ~~angle droit~~ yielding 498 results (among which a fair amount which are not about right angle), search for the concept ~~angle droit~~ yield 1 result that is precisely an exercise about this elementary mathematical topic.

A query by concept is the way to formulate a query and find results in multiple languages. For example, querying for the text *calculate areas* and selecting the suggested competency, gives rise to two matching resources, both of which are in a different language than English. A screenshot of the result is below:

Limitations

Current users, however, also noted imperfections in the approach. Skills-text-box was often indicated to be insufficiently easy to use for the following reasons:

The set of terms that were most difficult to work with are educational levels: in countries where many educational levels exist (e.g. one per state is Germany). This certainly is due to the fact the states have not yet been enriched with common names, an issue which is related to the fact that the inter2geo consortium, which ran the project between 2007 and 2010, have long expected government agencies to provide us the list of educational levels. However, even if provided such a list of names, it is not clear that it will be easy to select levels because of the large overlap in naming between a *siebte Klasse* of the Gymnasium of Baden-Württemberg and Hamburg,

A topic is useful if it helps to find a category of learning resources which otherwise would not be possible. What to do with a GeoSkills' node that gives no matches? Currently, many nodes of GeoSkills match no learning resources in i2geo. They have been contributed to GeoSkills following an analysis of the learning standards, using the same words that these texts use. However, no-one has contributed a resource about it. A potential strategy is to hide such terms from the search (but not from the contributions' forms) but it has not yet been attempted because of the cross-application nature of such a implementation.

Different displays of the ontology are probably also needed and have been partially implemented:

- for example the tree of topics or the tree of competencies (present in CompEd but too slow to be useful)
- annotated displays of the curriculum standard for a few countries, either in the faithful reproduction of the official text, or as a javascript tree. They allow the teachers using them to find resources annotated with topics and competencies related to parts of the standard. we could not conclude on the utility of this approach.

Ongoing Research

Having described the technical foundations, the achieved user experiences and their limitations, we describe here open investigations which we intend to tackle during the Open Discovery Space project that just started and will federate multiple repositories of learning content through Europe, including i2geo.net.

Formal User Testing

So as to raise the utility of the search engine, it should be possible to apply classical testing methodologies of information retrieval such as those presented in Manning-Prabakhar-Schütze, chapter 8: through a formal approach, it is possible to obtain quantitative measures of the utility of the search engine and reproduce this approach having addressed issues reported in a qualitative fashion. The application of a formal testing should be done related to the "utility" of the search engine for the day-to-day of teachers who often measure the quality of a learning resource by criteria unexpected to computer-scientists (see the quality approach in I2Geo-Quality as one of the evaluation methodologies).

This practice would support, for example, refining the exploitation of the ontology structure into the search by guaranteeing that an added tolerance does not introduce too much visible noise.

How to better use the context?

There is a strong potential to make better use of the context of a user of the i2geo platform.

Indeed, it would be normal for a registered user to indicate his country of origin, and thus avoid to make it precise that the *4ème* (fourth class) is that of France. Such a refinement would be an extra step in the query expansion and should probably not exclude other types of *4ème*.

Beyond elementary query additions, one could "make closer" terms that are in curriculum standards of interest to the user. This can be done because of the property `belongsToCurriculum` which can be computed from the curriulum standards, a set of html pages that link to search queries for the given nodes.

Such an approach could be the right approach to respond to a request that we have never been able to implement: respect the different wording of the educational standard. Examples include the wording of the same competencies in different countries such as Luxembourg and France: for many, the competencies are equivalent, however the wording is different.

Strategies for the Maintenance of a Classification

An area where we have found surprisingly little support from the broad semantic web or digital library research infrastructure is on the long-term maintenance of the ontology: while several methodologies for the development of (new) ontologies can be found, little literature is available about maintaining an ontology: we have learned the hard way such rules as to avoid to change identifiers as soon as the ontology may been referenced elsewhere: for example, URIs are kept readable so that they run the risk of containing typos which one considers natural to fix for example. We would expect infrastructure and best practice markup to indicate that a node is kept there only for the sake of completeness but should not be referenced in new annotations. Such deprecations could, for example, be "sufficiently documented" so that user-interfaces such as i2geo.net would suggest replacement nodes the next time the user edits the resource.

Curation of The Search Results

Should the maintainers of the platform run regular gardening activities on the search results? One of the strategy could simply be to go through each of the nodes of GeoSkills, search for it and compare this search result with a search result for plain-text. A strong difference there, provided a clash of the meanings is not occurring (such as *angle droit* and *angle* and *droit*), could mean one of the following issues:

- that there has been insufficient annotations contributed
- that contributions have been done to other topics but this one was not considered necessary
- that the search engine is not doing the right generalization

This curation process is a community process as it involves raising the quality of the results for the use of everyone. Would sharing the search results-pages, since they can be exchanged by a URL, leverage this aspect and let people talk about it and invite them for action?

Curators could, following tests above:

- request a change to the programmes (e.g. change the way generalizations are made or displayed)
- request a change of the ontology (for example add the common names of educational levels or suggest to differentiate too unprecise concepts)
- mail the contributors of resources considered insufficiently tagged and propose an annotation enrichment
- publish the search as one of the test cases for future testing
- share a good search result as part of a demonstration action

Among such features, the i2geo project attempted to accept suggestions: on places where the user can contribute, a little "+" sign allows the users to formulate a suggestion so that

the editors of the ontology, which have the overview, can incorporate or invite to use another term. 34 suggestions were formulated during the project, with better results achieved when using direct communication to the curriculum-encoders, the community of workers that edit GeoSkills.

Finally, it is certainly the role of a community curator to listen to requests for content and evaluate if it is relevant and applicable to the platform and, if yes, start the appropriate contributions showing best practice that others can follow: indeed, quite often external contributors come to i2geo.net and expect to find particular topics but they do not leave a visible trace of such quests, especially if unanswered. Sharing such a quests, in the form of search URLs and accompanying texts in social networks or emails is a way to raise awareness both about the platform and the platform potential. This should be stimulated.

It is interesting to note that most of these actions are triggered only because of a particular state of the available data (the ontology and the annotated resources) and have been probably not identified as requirement in the development phases: the search paradigm is, indeed, strongly influenced by its available data; it is easy to make corpora which are impossible to search with ease because the words one would use as queries are not discriminating enough.

Fuzziness Approaches

Probably one of the hardest curation situation is when two GeoSkills nodes are meaning something similar to each other, but not exactly the same. Different communities will tag different resources with different nodes leading to the isolation of communities.

One way to exploit fuzziness is to let the user walk around: using graphical displays of the relationships between the GeoSkills nodes can support the user into generalizing or specializing his or her request. While this is currently available today, going through CompEd and navigating the tree, it is not smoothly integrated yet. This could be done by embedding the navigation graph in a small portion of the screen of the search results, including the navigation to "weak synonyms".

Beyond formally authored relationships, statistical methods could also be leveraged to detect relatedness of nodes of GeoSkills. Approaches such as Latent Semantic Analysis, based on the concepts' names or on corpora of mathematical definitions for each of the concepts are likely to create methods to find nodes close-by.

Ideally, such neighbours strategy should also exploit the ontological nature (generalizing Germany's 9. Klasse of Saarland into 9. Klasse of any state of Germany.

Acknowledgement

Space.

Bibliography

- (Manning) Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008. Available from http://nlp.stanford.edu/IR-book/.
- (i2geoCERME6) Kortenkamp, U., Blessing, A. M., Dohrmann, Ch., Kreis, Y., Libbrecht, P., Mercat, Ch.: Interoperable Interactive Geometry for Europe – First Technological and Educational Results and Future Challenges of the Intergeo Project. (2009); published at CERME 6 - Sixth Conference of European Research in Mathematics Education (FR – Lyon) in the working group 7 http://ife.ens-lyon.fr/editions/editions-electroniques/cerme6/working-group-7
- (I2Geo-Quality) Trgalova, J., Jahn, A. P., Soury-Lavergne, S. "Analyse de ressources pédagogiques pour la géométrie dynamique et évaluation de leur qualité : le projet Intergeo ». Actes du Colloque Espace Mathématique Francophone 2009. Paper
- (GWT) Google Inc., Google Web Toolkit, http://developers.google.com/web-toolkit/ (accessed 2012)
- (GeoSkills-SemWeb-Handbook) Libbrecht, P., Desmoulins, C.: A Cross-curriculum Representation for Handling and Searching Dynamic Geometry Competencies, Handbook of Semantic Web for E-learning, 2009. See http://www.iospress.nl/book/semantic-web-technologies-for-e-learning/
- (OWL) Smith, M., Welty C., McGuinness D., Web Ontology Language, W3C Recommendation, 2004 http://www.w3.org/TR/2004/REC-owl-guide-20040210/
- (CompEd) Libbrecht, P., Desmoulins, C: Comped, a Web-based Competency Ontology Editor for Dynamic Geometry. (2009); published at SWEL'09 @ AIED'09 - Ontologies and Social Semantic Web for Intelligent Educational Systems (UK – Brighton). See http://compsci.wssu.edu/iis/swel/SWEL09/
- (Lucene) http://lucene.apache.org/. Accessed in 2012
- (owlapi) OWLAPI, a library to manipulate OWL ontologies. http://owlapi.sourceforge.net/. Accessed 2009.
- (pellet) Clark & Parsia Inc., Pellet: OWL reasoner for Java. http://clarkparsia.com/pellet/
- (xwiki) XWiki SàRL, XWiki, entreprise wiki for Java. http://www.xwiki.org/
- (XCLAMS) XWiki SàRL and Curriki Inc., the XWiki Collaborative Learning Assets Management System.