Searching Thousands of Learning Resources

Paul Libbrecht, Christoph Bail, and Martin Becker,
Weingarten University of Education, Germany

***Abstract*—Everyone is a publisher nowadays on the web, teachers included. The social web is growing at a fast pace, with more and more needs to *web* the content so as to make it accessible, annotated, and found. While this webbing creates a rich navigation experience, tools to access more resources are needed. Search tools are among the most used tools to discover more learning resources. However, their usage is currently rather limited and frustrating. This paper describes the challenges currently met, and the Open Discovery Space search tool, presenting how it addresses them.**

***Index Terms*—Education, search engines, web search, learning resources.**

I. Introduction

THE world wide web, having started with a model where few were in possession of the publication privilege, has evolved into the largest collaboration space of the humanity, where almost anyone is a publisher. Teachers have also taken part to this exchange: Today, multitudes of learning resources are available published by a multitude of authors, offering hints, materials, or advices to support the teaching or learning process. But how to find orientation among this multitude?

The search challenges of a teacher that "resources himself" are enormous: The learning resources are scattered across multiple sites; they employ slightly different vocabularies to describe themselves and thus to be identified; while a few teachers are able to trust a few resources, many do not. As a result it is common for teachers to spend repeated search sessions in preparing their courses, trying to identify the resources they would be able to adopt.

The Open Discovery Space portal is a large portal that harvests multiple learning object repositories within a single point of access:: with its massive amount of learning resources and indexed with a unified vocabulary, it opens the door for teachers to search through a significant amount of learning resources with a structured support. While text-search remains a central element, a structured drill-down by means of facets and taxonomies allows to remove ambiguities in search terms. Moreover, an ordering of the search result is offered that promotes resources judged relevant to the user by criteria such as the language affinity, the readiness to embrace complex file types, or the recommendation of a friend in one's network.

A. Outline

In this paper we first present an overview of existing search engines that are applicable to learning resources, we the survey current challenges that these meet (including ambiguity, multilinguality, and implicitness). The design of the search engine of Open Discovery Space is then presented. An implementation status and testing plan follows in the future works, along with open questions.

Manuscript received May 4th, 2015. This work was supported in part by they European Commission CIP PSP 297229.

Paul Libbrecht. Christoph Bail. Martin Becker, Weingarten University of Education, Informatics. Kirchplatz 2. 88250 Weingarten. Germany.

II. Search Engines for Learning Resources

Learning resources are *artifacts* that can be employed within learning and teaching processes; this very broad definition is generally endowed with one common restriction: they are *digital documents*, in the sense that they are materialized as files that can be viewed and potentially edited by a computer.

Textbooks form probably the most ubiquitous example of learning resource; they are sometimes digital. Textbooks commonly support the learning processes by guiding teachers and students, supporting their exploration and assessment. Textbooks illustrate well the *resource* nature: they can be *pulled from* in the learning. However, they are often not open.

Open Educational Resources are understood to be digital and digitally exchangeable thanks to their digital natures and thanks a license that allows anyone to receive and redistribute the resource without cost; this definition is that of the Hewlett Foundation [1]. The wave of open educational resources has grown since about a decade and has allowed the production of millions of resources, which could be applied by any teacher of the earth.

The breadth of this availability presents to many of the teachers of the earth a sea of available resources, which teachers can choose from. This makes the identification of learning resources that are relevant for the teachers' or learners practice and are of sufficient quality, quite a challenge. Among elements of the sea, one can find learning resources which are not complete, lack adaptability, employ outdated tools, use an inappropriate vocabulary, guide the teacher only partially in his attempts, or employ incompatible software. All these issues have to be recognized and coped for and thus it is important to be able to crawl among multiple resources to elect the *most appropriate* or one *requiring the least circumventing actions for it to become appropriate*.

Search tools are among the most important tools to crawl this wealth. They are commonly used to find learning resources but in a way that is not yet fully satisfactory in many cases. We describe a few typical search tools used to date to find learning resources which can be used by teachers in Europe.

A. Generic Web Search Engines

Learning resources are commonly found on the web. Thus, they can generally be fetched by web crawlers and searched for by the search engines behind the crawlers. These search engines search the broad web and thus almost only offer to employ *generic* queries; that is, queries for words in the current language, or in any language. If the normal language had special words for most topics to be learned, just as brand names are unique, the search process for learning resources would be very effective and convergence of users searching within similar topics would be found. However, the *ambiguity* is unavoidable and topic names such as *inflation*, *square*, or *reading* are not able to distinguish learning resources from other artifacts. In general, additions of query terms often bring more noise in the search results.

The teachers searching for learning resources in the broad web thus often have to allocate long repeating periods to search for learning resources, sometimes going through pages and pages of results in the hunt for a more satisfactory resource. While web search engines provide the openness that most teachers want, hoping to find open resources anywhere, and often ready to decrypt resources in a language that is not theirs, they fail to provide an enjoyable search process because of the inability to formulate queries: topic queries are difficult as expressed above, but querying for an educational level is difficult too: There does not exist a uniform query mechanism for the very many school systems and the typical age range is rarely accessible.

To our knowledge, thus far, no web search engine is leveraging the emerging metadata standards LRMI[1] to offer such a refined query mechanism.

Moreover, the broad web is made of multiple web pages that mix several languages. Thus, it is rather common to find pages in a different language than the one queried, even if limiting the results to a single language. This happens most frequently when searching for words common in several languages, the presented results appear unrelated to the query.

Thus we contend that generic web search engine pose an *implicitness problem*: they do not allow to express criteria for learning resources that reflect fine grained expectations of learning resources (such as the quality criteria, the technical affinity, or the community of practices' membership).

B. Personal Search engines

Collecting a repertoire of learning resources is among the typical activities of the professional teachers. However, because they are generally not downloaded in a single place, the regular search engine of a teacher's computer is not able to search through complete collections of learning resources that teachers meet. Such a resources' collection would be a good place to collect trusted materials, which are known relevant to the user. This could avoid the implicitness problem, but would still need to be enrichable through a discovery.

[1] The Learning Resources Metadata Initiative (LRMI) proposes a definition of a small set of *microdata* annotations, which can be encoded within web-pages for crawlers to consume. This allows information about learning resources to be harvested from any web page.

C. Platform Search engines

Platform search engines generally offer the search for learning resources contributed by the users of the platform. They leverage an information set that is asked at each contribution, using a vocabulary that is agreed upon at the design of this platform. Such vocabularies are often specific to a platform: While on i2geo.net, a platform to share dynamic geometry, the topics are fine grained mathematical concepts, those of other portals are often coarse. While several portals qualify the didactical function of a resource (e.g. being an exercise, an assignment or a scenario), others do not.

This specificity supports well the community's implicit values but make it difficult for new users to start using a new platform's search engine.

III. Design of the Open Discovery Space Search

The Open Discovery Space resources' search engine attempts to address these issues by several measures, which are made possible by the control it exercises on the learning resources it presents.

The Open Discovery Space search engine is a search tool embedded in the ODS portal, it is expected to be used by the users of the platform within such tasks as the generic search for learning resources, the selection of learning resources to be included in broader scenarios, the browsing of learning resources to explore the set of available resources.

A. Content being searched

The ODS resources' search is mandated to search through learning resources harvested from identified repositories, a broad set of repositories relevant to school education. The information about the resources is fetched during the harvesting cycles, which employ the OAI-PMH protocol, which collects Learning Object Metadata records (LOM) of each of the repositories, encoding using the vocabulary of ODS These learning resources cover most domains of school learning with inequalities in size (e.g. within physics, astronomy is richly supported, but ballistics is much less) and in languages. This diversity is somewhat similar to the broad web: for some subjects, there are far too many resources, for some subjects, only a handful.

The search engine is designed to search for resources for text queries as well as for more fine grained metadata facets such as: educational level, typical age, date, language, or learning resource type.

B. Multilinguality

The search tool is designed to be multilingual: currently, resources are in more than 23 languages with very different counts.

The multilinguality challenge that teachers meet is resolved by employing content sources whose language is explicitly qualified and by employing a user interface where changing

the language is as simple as a click. This allows the application of classical stemming mechanisms at indexing time and at query time, which are generally not applied in a safe fashion otherwise:

- At indexing time, the LOM records being exchanged within the harvesting mechanism make sure that each text that is not a person's name (titles, descriptions, tags…) is surrounded by an element that carries an `xml:lang` attribute. The indexing process converts the words of these texts to tokens in separate fields, which employ different *tokenizers*[3]. For each such text, three versions are converted: the whitespace tokenizer preserves full words, the stemmer converts words to their roots, while the phonetic tokenizer converts words to their phonetic equivalent. It is important to note that learning resources often have multiple languages.
- At query time, the same tokenization processes apply but they are given a different weight. Thus a search for the word *directing*, queried in English, will prefer documents that contain this word, while still brining in the search results, documents containing such words as *direct* or *direct*.

This simple query system allows a fairly tolerant search, as it allows, for example, to query a singular word and still obtain documents containing plural forms, while still avoiding as much as possible the confusion of search matches between different languages (e.g. matching the French *directe*).

Moreover, the search tool supports the multilingual users in a limited fashion: they can easily change language so as perform the same query in a different language. While web browsing tools easily allow the formulation of multiple languages, for example by adding supported languages in the web-browsers' preferences, it has been the experience of the authors that users easily forget about these settings and express poor search results quality feelings, whereas an adjustment of their preferences may have made the difference.

C. Ambiguity

To cope for the difficult challenge of imprecise concepts denomination, the Open Discovery Space project has defined a taxonomy of topics relevant to schools, which extends the classification of the EUN LRE.[4] Contrary to this classification, the taxonomy that Open Discovery Space has introduced is a fine grained classification which allows to express fine grained topics as fine as *planet inflation*. Such taxonomy is not always available within resources where some contributors are happy to just indicate that the resource is part of *algebra*, for example. The different granularity of the topics annotations, while they are not completely parallel, can cope with each other, since they use the same vocabulary.

This approach solves the ambiguity problem, because it forces the contributors and searchers to use the same taxonomy terms for topics, that are close to each other. While the search users are sometimes lost in browsing such a taxonomy, for example many teachers have a difficulty to figure out, that *history* is considered to be part of *social sciences*, they can explore the annotations space by interactive searching. A sample strategy to do so can be to employ text search to find typical subjects and observe the topics that could still be used to refine the search (e.g. searching *queen Elizabeth* and observing that *social studies* is among the facets which represent topics of resources which would be relevant).

D. Implicitness

The implicit expectations that users have from a search engine are elicited from the users mostly through their user-profile: this set of information about the user is first entered at registration, then incrementally refined as the users progress, e.g. through request prompts. They include the users' language, country, and ICT competencies. Based on this information, one can assert a probable preference for resources of a particular language, or for resources that involve more or less technical competencies in the use of ICT.

This preference is encoded within the same process that converts the searched text to tokens: the query expansion decorates the queries with *preferring queries,* which change the weight of resources matching particular patterns.

The search tool could even employ the social network created by the user in his or her interactions with other users; indeed the platform supports building a network of followers and it would be thinkable to prefer resources of users in one's own network, or prefer less resources, which network members have rated negatively. However, this feature has been left as a plan.

E. Synthesis Example

When searching for the words *invade France* using the English language, the query expander converts these words to the mandatory query part:

```
+(title_ws:(invade france)^60 title_en:(invad
    franc)^40 title_phon:(INVD FRNS)^10)
+text_ws:(invade france)^30 + text_en:(invad franc)
    ^25 text_phon:(INVD RFNS)^12 )
```

enriched by the preference parts, in case of a fairly low ICT-technical competency profile:

```
language:en^1.5 (cTyp:image/*
 cTyp:application/ppt^0.9)
```

This example uses the query-parser syntax[5] and indicate the weights (superscript) and wildcards (followed by star).

F. Implementation Status

At time of writing, the search engine of Open Discovery Space is in an alpha stage and can be reached at from portal.opendiscoveryspace.eu. Its current weaknesses include a shallow control on the metadata quality (e.g. mixing of languages or lack of topic metadata), and the adaptivity to technical competencies, as it has not been sufficiently tested to

[3] The *tokenization* process is generally understood to be the conversion process from strings of characters to streams of tokens. It is described in [2].

[4] The LRE Thesaurus is a classification of topics for learning resources realized by EUN for the LRE Resource Exchange platform. It aims at representing a broad spectrum of topics without going too much in details.

[5] The Lucene query parser syntax is documented here: https://lucene.apache.org/core/2_9_4/queryparsersyntax.html.

be deployed. The server code is available open-source from http://github.com/OpenDiscoverySpace/.

Basic testing of the search tool has shown simple errors in the multilinguality: among others, searching for simple words such as the German word *verstehen* (understand) yields multiple resources, which do not seem to contain that word. This is explained by multilingual keywords on a mono-lingual resource. Similarly incomplete stemming has been met. The solutions sketched in this paper, once applied strictly, will solve these issues.

This makes the ODS search engine a unique point of access to query for learning resources among a vast pool and using a query vocabulary that is far more precise than that of generic web search engine.

IV. Conclusion

In this paper, we have described the current landscape of searching learning resources and have shown the multiple challenges normal teachers meet. We have described the search tool of Open Discovery Space, a tool which combines several features to answer these features, including a significant amount of resources (more than 900'000 at time of writing). Basic testing is currently showing issues which seem easily solvable, however, it is not yet clear, if more issues will appear. The following sections describe ongoing and proposed future works related to this search engine.

A. Testing

In order for the search tool refinements to be grounded on tangible quality criteria, the project's last steps will include the involvement of a wide range of search testing experts. Cultivating the diversity of teachers in Europe, it will enroll experts in their fields of teaching, which will propose queries and evaluate the search results' list. Based on a simple bookmarklet approach which any web user can activate, the voluntary users will provide their feedback while they work, assessing the quality of a search result by simple check-boxes and comment boxes injected within the page. This should allow us to gather test suites and evaluate retrieval in a quantitative manner using approaches such as those described in [3]. The feedback will then be used by developers to identify weaknesses and tune data filtering and weighting.

We expect such feedback as the inappropriate appearance of a resource in the results, the buggy ordering among the results, or the lack of particular resources among the search results. These can be answered by an analysis of the metadata being searched through and the intermediate query processing steps. It may lead to adjustments in the processing of the harvested data, e.g. applying natural language processing techniques to support (semi-)automatic classification, in the relative weight of queries (in particular preferring queries), or in the stemming methods.

Having a significant and culturally diverse test base is probably the only method to refine the search tool in a way that respects, on the long run, the very diverse expectations of teachers in Europe. It might also uncover reasons to adopt (accept to re-use) and to adapt (modify to make more fit to the purpose) which are not yet sufficiently explored transform the platforms into comprehensive exchange marketplaces (see [4] for an early study in this direction). Such a test base might also enrich the vision of cross-lingual search engines, whose pioneering works, as reported in [5] seem not yet appropriate to the world of learning resources.

B. Suggestions

Following the quick search and refine process typical of information retrieval, e.g. [6], a mechanism to suggest *related queries* should be studied and tested, similarly to [7].

C. More detailed user information

While technical competencies are followed through, the personal interests of teachers are not yet collected in the Open Discovery Space platform. Such collection may bring richer preference queries while it runs the risk to add more forgotten context information. A mechanism to make the search process more transparent, allowing the user to understand that a given search result has not respected all of his or her criteria for example, seems not yet explored.

V. Acknowledgment

We thank Ard Lazonder, Thanasis Hadzilacos, Anna Mavroudi, Katarina Riviou, and Noortje Janssen for a productive early design elaboration, which this document sketches. We also thank Panagiotis Zervas, Lamprini Kolovou, Luis Lalueza, Salvador Sanchez, George Thanos, and Maricruz Vallente for their technical collaboration in preparing the Open Discovery Space search engine.

References

[1] Hewlett Foundation, *Open Educational Resources*, available at http://www.hewlett.org/programs/education/open-educational-resources, accessed in May 2015.

[2] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze:,*Introduction to information retrieval*. Cambridge University Press 2008, ISBN 978-0-521-86571-5.

[3] Kavi Mahesh. *Text retrieval quality: A primer*. Technical report, Oracle Coproration, 2006. See http://www.oracle.com/ technology/products/text/htdocs/imt_quality.htm.

[4] Paul Libbrecht, *Adaptations to a Learning Resource*. To appear in Acta Didactica Napocensia.

[5] Carol Peters, Martin Braschler, Paul D. Clough:,*Multilingual Information Retrieval - From Research To Practice*. Springer 2012, ISBN 978-3-642-23007-3, pp. I-XVII, 1-217.

[6] Cornelis Joost van Rijsbergen. *Information Retrieval*. Butterworths, 1979. Available at http://www.dcs.gla.ac.uk/~iain/keith/.

[7] Paul Libbrecht, Escaping the Trap of too Precise Topic Queries, *Intelligent Computer Mathematics*, Proceedings of CICM 2013, LNCS 7961, Jacques Carette, David Aspinall, Christoph Lange, Petr Sojka, Wolfgang Windsteiger (Eds.), ISBN: 978-3-642-39319-8.