数理解析研究所講究録
第1978巻 2015年 12-21

Challenges and Opportunities when Crossing Languages in the Search for Mathematics Open Educational Resources

Paul Libbrecht

Weingarten University of Education, Germany
paul@hoplahup.net

Abstract. The Web we use every day has been built to be international and it is not rare that we meet documents in other languages and handle it as well as out translation capacities go. Mathematics learnning resources are no exceptions to this actions. They include the very many explanations and exercise texts that mathematicians make available and that we can (partially) re-use for us to prepare courses.
However, mathematics learning resources have challenges that resources of other domains do not have: While their underlying "message" is considered universal, it often differs in its expression forms, in a way that currently no automatic translator can handle.
In this paper we present the current approaches to searching for learning resources, how they can be published and found, and how crossing language barriers can open new avenues but presents difficult challenges.

Keywords: open educational resources, translations, sharing, searching

1 Introduction

Learning resources are medias that can be used in learning activities. While blackboards, textbooks, and leaflets can all be considered to be such, one generally considers learning resources to be electronic documents that can be duplicated digitally so as to enter a learning process. Open educational resources are such, but they are, moreover, *open* in the sense that they are available through a license that allows free reproduction and adaption.

The advent of the world-wide-web has made it possible for teachers across the earth to exchange open educational resources. They often make widely available their work on web-sites and others find them, adjust them, and use them in their teaching. If there is a motivation, the recipient may publish a learning resource with adjusted material corresponding to the needs he or she encountered. The chain of actions described above is called the resourcing cycle.

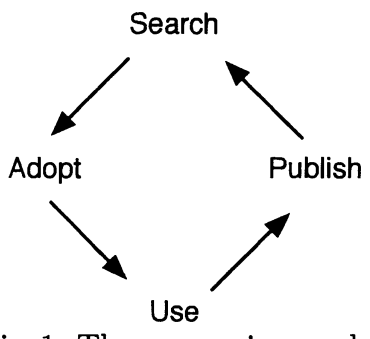

Fig. 1: The resourcing cycle.

The resourcing cycle is depicted in figure 1 The resourcing cycle is common to all open educational resources and it can include other operations of communication such as comments, quality evaluations, and team development. Also, only some parts of it can be regularly applied. For example, the search action is likely to be run considerably more often than an adoption and even a use in classroom.

The focus of this paper is on the search possibilities. However, this action is connected to the adoption (which is conditionned by the quality of the search) and by the publication (which impacts the search considerably). Therefore we shall present the search in connection to these two aspects as well as to the aspects of the information entered when using the search tool.

The search actions for learning resources make use of a search tool which is manpulated by the user, typically a teacher who wishes to find resources that can enrich his teaching. This search tool is used by the input of selected terms, which are expected to match terms or attributes found in the published learning resources. Resolving such a match may apply various techniques to enhance the facility to find resources:

- automatic translation between languages
- synonyms dictionaries
- automatic analysis of the learning resource (e.g. to extract topics)
- match to approximate terms to cope for typos

The search activity for learning resources can become a fairly long process as it involves finding something that can be adopted and as the search process only lets you specify a set of queries that cannot be fully disambiguating. One often sees, thus, teachers spend time to go through all the search-results of a given query so to be sure that a desired resource does not exist and, as a result, it is worth investing time to create something new.

For example, using a text search engine, the usage of such query terms as the French words *sommet d'une courbe* made of *sommet* (which means summit or top, but also is the name of a vertex in a graph or a polygon) and *courbe* (which means curve) matches documents which carry these two words even if they are not successive and thus are used in different contexts (e.g. in business where curves and top often appear) and miss document with similar meanings (for example *maximum of a curve*). Other search engines that operate using a thematic dictionary might score better, if the user is able to use it e.g. by drilling down till the domain of calculation of extremal points.

The search activity will invariably aim at searching the complete World-Wide-Web: while the set of learning resources made by people that speak a common language and practice similar learning methods will be of greatest interest, teachers will often want to lurk out and see how other worlds have approached the subject. In particular for teaching topics which classically display challenges, there will be a will to lurk out and see if other cultures have solved it differently. The ways of expressing the mathematical concepts, the ways of operating with their notations, the ways of explaining concepts and building on each others. Teachers aware of the importance of semiotic mediation [MM12] will want to follow the representation transfers described as psychologically benefic

by [GH97]: For a teacher, this allows him or her to look at the subject and its didactical methods with a different perspective; this allows students that are confirmed into a subject to strengthen their understanding of the subject by articulating other relationships which become available when changing representations.

The automatic translation methods used above could, in principle, be applied in search engines: resources in another language would be indexed after being automatically translated. This approach has been tested in the Organic EduNet portal.[1] For the mathematics domains, however, no such initiative exists and it seems more difficult to achieve as the vocabulary of every is more common in mathematics than in agriculture.

Issues that arise when using automatic translators in mathematics include the differences in semantic fields: e.g. the word *droite* in French and *line* in English have quite some overlap, so they should be translated to each other in most of the mathematics, but not always, e.g. not when speaking about a telephone line (which is not *droite* which means, in this case, *straight*) or une *maison droite* where *droite* describes the verticality of a house. Other issues will be described in section 4. Searching for these terms will inevitably mix the concepts and will bring more noisy search results which will require more results sorting.

This paper attempts to shortly describe the challenges met when translating mathematical learning resources, especially relevant to the search. It attempts to answer to the question *How much can I be surprised when re-using a resource coming from another language?* or the more productive question *What can be done once I find a learning resource that seems to match my expectations so I am confident re-using it?*

Outline The paper first presents learning resources tools that can cross languages (sec. 2) and examples of multilingual mathematics learning resources (sec. 3). It then attempts to classify the mismatches specific to the translation mathematics learning resources (sec. 4). Future perspectives conclude the paper.

2 Learning Resources' Tools that Cross Languages

Automatic translators certainly form a strong basis to decipher a text. Current experiences with the automatic translators show that mathematical texts are weakly translated: false friends' such as the Spanish *Teorema de Tales*, which should be translated to the *intercept theorem*, seem to have not yet reached them. Nonetheless, mathematicians that read an automatically translated mathematical text can often make some sense of it and probably use it as a basis before appropriating it. However, if search tools are to benefit of these services, such false friends may well bring too big a noise making the search results poor.

[1] Organic EduNet is a learning resources' portal for the education around organic food and agriculture. See `http://organic-edunet.eu/`.

Several knowledge representations exist to encode mathematical documents in a *semantic* way, that is according to a fixed meaning that does not depend on a language.

This includes OpenMath [BCC+04] and MathML-content [CIM10] which allow users to exchange complex mathematical formulæ between different systems without, in principle, concerns for language specificities. Similarly, the i2geo format [ABE+09] proposes a common syntax to describe dynamic geometry constructions. Moreover, OMDoc [Koh06] describes a structure of mathematical documents allowing synchronously-multilingual statements. The standards-based nature of these formats promise a more faithful search, provided the learning resources searched for are encoded using them, and indeed very early attempts in this direction have been started see [HQ14] and the emerging NTCIR Math search tasks.[2] Moreover, various retrieval mechanisms for mathematical knowledge are described in [GSC15].

More user oriented tools can offer cross-lingual access to learning resources. E.g., most learning resources' sharing platforms work in multiple languages.

All of them employ vocabularies for several of the properties of the learning resources. These include elementary vocabularies such as the license (a choice between a handful of supported licenses) or more specific ones. Notably, for mathematics learning resources:

- the educational function, expressing the typical method of use of the resource. Large parts of these vocabularies seem to offer no challenge in being translated (e.g. *exercise*, *reference*, *demonstration*...).
- the educational level that the learning resource is aimed at: depending on the intent, this property can be as specific as aiming at a particular year in a particular school form. Because of the diversity of the educational offers, there seems to be only the possibility of a universal language of the age-group of the target learners.
- the topic and maybe trained competencies: this property can be expressed in a very shallow way (simply describing *algebra* or *calculus*) or in a very precise way (e.g. *the L'Hospital's Rule* or the *roots of a polynomial*).

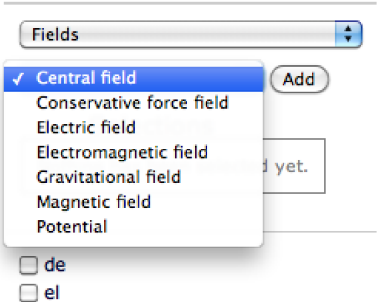

Fig. 2: Selecting a topic in two steps.

These vocabularies allow users that use tools using them to cross languages in a way that avoids almost all of the semantic fields issues: indeed, if the vocabulary is sufficiently fine grained, the creators and translators of the vocabularies can ensure that the translations avoid false friends and are using widespread namings.

Applications of these vocabularies include the use of taxonomies in the choice of topics as in the

[2] The classical NTCIR competition for the experimental validation of search engines has a track for mathematical searches call MathIR. See `http://ntcir-math.nii.ac.jp` .

picture on the left in a portal in astronomy.[3] Users drill down the hierarchy by a sequence of choices. The hierarchical nature allows an accessible display of a limited size even if the set of topics is fairly big. However, it assumes that users know the hierarchy e.g. know that, in the picture on the left, the *potential* is a form of a *field*.

Another approach to cross languages by using a multilingual thesaurus is to access its entries by auto-completion: users type parts of the concepts' names and choose the concept from the suggestions. This is the approach used in the i2geo.net portal[4] and is depicted in the figure on the right. Advantages of the auto-completion approach include the fact that multiple names are allowed by thesaurus entry (for example, the *intercept theorem* can also be found by typing *intercepting lines theorem*) including names in other languages that the user may be mastering as well (e.g. *Vierstreckensatz* in German or *théorème de Thalès* in French).

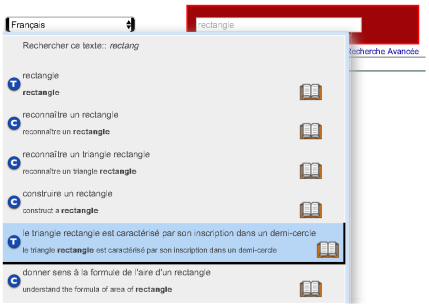

Fig. 3: Choosing a topic in the i2geo portal.

To conclude this section, we see that some tools allow precise cross-lingual access but they are quite partial in their function (covering topics only, age only, or requiring an input that is not widespread) and that, on the other hand, some tools such as the automatic translators allow very shallow cross-lingual access and need a constant proofing by eyes that understand the domain. Fortunately, these tools are distinct and, users know when to search for multiple terms (e.g. when having precise search terms) or when to quickly exclude search results (e.g. when having shallow matches).

3 Translatable Resources

Mathematics is universal... to some extent.

For some learning resources, multilingual learning resources do exist. Resources at the PhET repository[5] indeed are considered software projects with internationalization dictionaries which multiple contributors offer. It is not rare to see resources in more than 20 languages. Dictionaries specify the text messages but can also specify colours.

[3] The Cosmos portal is a learning objects repository to share resources pertaining to astronomy in classroom. `http://portal.discoverthecosmos.eu/`

[4] The `http://i2geo.net` portal is a sharing platform for learning resources with dynamic geometry.

[5] PhET is a repository centered on a few physics and mathematics animations (less than 200) around which thousands of scenarios are proposed. See `https://phet.colorado.edu`.

Resources of dynamic geometry can also often be easily translated: a teacher that found the resource can edit the document using the same software that the author used and change the texts and probably much more.

This multilingual feature of learning resources, or... their readiness to be translated, is rather rare and is concentrated on fairly small artifacts.

4 Mismatches when Translating Mathematics

Translating mathematics learning resources can be seen as a relatively straightforward task for a mathematician with a good knowledge of the source language and that practices in the target language. In this section we identify a few types of issues which make the translation challenging and can only be resolved by an expert choice of translation in the expected context of use.

4.1 Incompatible Semantic Fields

The first type of issues are the **incompatible semantic fields**, the set of meanings that a given word can have. This imposes a translator the perpetual attention to the interpretation of the learning resource's text. Examples include the translation of *line* to *droite*, the translation of *intercept theorem* to *théorème de Thalès* (indeed, the *intercept theorem* is also attributed to Thalès). An interesting experience to discover the incompatibility of semantic fields is to employ Wikipedia's offer to see an entry in another language employing these links several: it is not rare, doing that, to experience a much wider navigation spectrum than simple cycles. The area of incompatible semantic fields, because of its requirement to understand the terms and their contexts, is an area where automatic translators are likely to fail.

4.2 Varying Relevance of Learning Content

The second issue is in the **relevance of learning resources' content** for the learners: in particular in the connection to the real world, the same reality (say, a mountain hike) can become very relevant for a learner (for whom this would be common) but very far away for others (for which a mountain hike starts with a long trip in the flat surroundings); other examples include the strong relevance of the geometry of paper folding in some cultures and the very weak one in others. While this obvious challenge seems natural it is crucial for learning since the connection to the real world is well known to support mathematics learning.

The only way for a translator to address this is to reformulate the content to other application domains, a work that is considered more an authoring work.

4.3 Translation of more than text

The third issue is the requirement to translate more than just text and in particular the **mathematical notations**: Even though mathematical notations

appear to carry a universal semantic to the broad public, they show quite some divergences. A simple example is the notation of the half-open interval displayed in figure 4.

Fig. 4: The half open interval in English, German (and French), and Dutch: scans of textbooks displayed in the notation-census.

While many of these differences are bound to language (e.g the sine function being written as sin in English but as sen in Spanish), many are the results of quite different evolutions and are, sometimes, even disconnected from the mere language associations. For example in [DL08], one sees that the set $\mathbb{N}$ of natural numbers is considered with or without the number 0 depending on the school tradition of the mathematician; similarly the root of -1 is expressed as i or j depending on the domain one works in (mathematics or electricity). An attempt to snapshot mahtematical notations across different cultures is done in the Notation Census.[6]

It should be noted that the notation variations is strongly bound to the memorability and ease of reading of its elements. Thus, while it is common in many languages to use P or Q for the names of points in geometry, polygons are rather named from the start of the alphabet (A, B, C, ...) and particular points often have their names as the initial of their particularity (e.g. the summit of a mountain would be writte S in English but G in German (for *Gipfel*). Such differences imply that learning resources' translation needs to go as far as graphics and include an understanding of semantic of the graphical ingredients.

Similarly, as mentioned in [Mar09], several graphical differences exist in the use of colours and symbols in the regular documents. There seem to be no systematic study of these differences yet but some can be quite relevant for documents around learning, including the systematic value of the red colour to denote *wrong* in Europe but to denote corrections (positive or negative) in Japan. The same requirement is imposed on translations.

4.4 Diverging Learning Practices

Finally, challenges for translators appear when teaching or operative methods diverge from one language and another. For example, the Japanese school system cultivates the difference of concept of a *proportion* (written as $5 : 3 = 20 : 15$) but this concept is expressed using proportionality tables in French. Translating one to the other is almost impossible as the set of operations are radically different (one lays tables to compute the transitions whereas inline simple operations are common in Japanese books). Similarly, in the effort to translate the concept of

[6] The notation census is available at `http://notations.hoplahup.net/` and has been introduced in [Lib10].

instant slope, the French team of the ActiveMath-EU project failed to identify a corresponding concept which would be connected in the same way to its prerequisites and followups: indeed, this concept borrows from a mechanical approach of calculus which is rarely done in the French language where one finds more often geometric descriptions: *instant slope* should be translated to *pente de la tangente* but the two concepts cannot be articulated in a similar fashion, e.g. they cannot have the same prequisites or examples.

5 Perspectives

In this paper we have presented challenges in the translation and in the now regular activity of viewing learning resources that are in another language. The world wide web has empowered all teachers of the earth to view and re-use learning resources in other languages.

Can they take advantage of it? Certainly, it can help them discover other teaching practices, other representation and other operative means. The computer-based tools can support this discovery and, more generally, learning in multilingual environments as sketched in [LG16].

The challenges that teachers meet are at the same time an opportunity of enrichment: Incompatible semantic fields represent different ways of perceiving a concept and connecting it to the real world: meeting these connections allows a teacher to provide alternative explanations which may enrich his or her learning. Similarly, different notations are linked to different operative modes. Demonstrating the ability to use several notations is a demonstration of a strong conceptual understanding.

Searching the web for learning resources in mathematics will meet these differences in stronger way than *just* meeting texts in other languages. Through a choice of words, one searches the complete semantic field of that word, including the non-mathematical ones. The word *field* for example, takes you to agriculture, to differential geometry, to physics, and to algebra. Through the use of thesauri (e.g. in learning resources' sharing platforms), search can become more precise (as, for example, the *field* concept in astronomy is only the physical field) and multilingual (matching resources with this topic in other languages).

5.1 Traveller Recommendations

From the analysis above of incompatibilies, one can gather the following *recommendation* to teachers aiming at re-using resources done in another language: leverage multiple search strategies, going from a word search to a thesaurus search and back so that one can adjust one's search term and discover the terms in other languages and possibly broader thesaurus categories; accept subject ambiguities as a didactical feature. Finally, and probably most important, take the time to edit entirely a re-used resource from a different language so as to make sure that notation traditions of the target language are fully followed thus avoiding an extraneous cognitive load.

5.2 A Unified Language?

To diminish the disorientation effect of crossing languages advocates of an international language, which include many researching mathematicians, would often prefer to unify the language as much as possible, e.g. using the same notation for the same concepts. And indeed a growing range of courses present to the students that the notions of the half-open interval in figure 4 simply have multiple notations. Similarly, many teachers in countries such as France or Germany are forced to teach that the period sign is also the decimal separator (e.g. that $\pi = 3.14159...$) because available calculators apply this but financial systems (online banking, accounting systems, the default display of spreadsheets) all consider the same character as the thousands separator (that twenty three thousand and one is written as 22.001) and refuse the period as decimal separator. Only interpretations, as far as a priori estimates, can disambiguate these differences.

Such a uniformization can only be done gradually and comes at a price which has not been yet properly evaluated since each of the specificities is bound to explanations, traditions, and memory-hints which would also need to be changed. As an example, each of the 17 ways of doing long division described in [CIM10, sec 5.3] has an operation sequence which many thousands of persons have learned.

5.3 Richer Learning Resources Exposure

Learning Resources can be the hub of multiple other documents which show how they have been used (e.g. by traces of learning analytics) or how they are assessed (e.g. by quality evaluations). Meeting such aspects is likely to help potential

5.4 Combining Thesauri and Text Search Engines

Can both of these worlds be combined? This is at least what R. Steinberger stated when presenting the architecture of the news search engine of the Joint European Research Center.[7] which employs automatic translation massively to access news to government executives of the European Union. Among the core ingredients of this service, an entity recognition is supported by multilingual thesauri; this includes a navigation along these entities but does not seem to let users support the disambiguation using such an approach as auto-completion.

[7] The work described in this talk of the Multilingual Information Access Technology Transfer Day in Berlin in 2009. It includes infrastructures such as `http://www.newsbrief.eu/` or `http://medusa.jrc.it` and are part of a family of services of the European Union aimed at early crisis detection.

References

ABE+09. M. Abanades, F. Botana, J. Escribano, M. Hendriks, U. Kortenkamp, Y. Kreis, P. Libbrecht, D. Marques, and Ch. Mercat. The Intergeo File Format in Progress. In *Proceedings of OpenMath Workshop 09*, July 2009. Available from `http://www.openmath.org/meetings/22/`.

BCC+04. Stephen Buswell, Olga Caprotti, David Carlisle, Mike Dewar, Marc Gaëtano, and Michael Kohlhase. The OpenMath Standard, version 2.0. Technical report, The OpenMath Society, June 2004. Available at `http://www.openmath.org/`.

CIM10. David Carlisle, Patrick Ion, and Robert Miner. Mathematical markup language, version 3.0. W3C Recommendation, October 2010. Available at `http://www.w3.org/TR/MathML3/`.

DL08. James H. Davenport and Paul Libbrecht. The freedom to extend openmath and its utility. *Journal of Computer Science and Mathematics*, 59:1–19, 2008.

GH97. J. Greeno and R. Hall. Practicing representation: Learning with and about representational forms. *Phi Delta Kappan*, 78(5), 1997.

GSC15. Ferruccio Guidi and Claudio Sacerdoti Coen. A survey on retrieval of mathematical knowledge. In Manfred Kerber, Jacques Carette, Cezary Kaliszyk, Florian Rabe, and Volker Sorge, editors, *Intelligent Computer Mathematics*, volume 9150 of *Lecture Notes in Computer Science*, pages 296–315. Springer International Publishing, 2015.

HQ14. Yannis Haralambous and Pedro Quaresma. Querying geometric figures using a controlled language, ontological graphs and dependency lattices. In Stephen M. Watt, James Davenport, Alan Sexton, Petr Sojka, and Josef Urban, editors, *Intelligent Computer Mathematics*, volume 8543 of *LNCS*, pages 298–311. Springer International Publishing, 2014.

Koh06. Michael Kohlhase. *OMDoc – An Open Markup Format for Mathematical Documents.* Springer Verlag, 2006.

LG16. Paul Libbrecht and Leila Goosens. Using icts to facilitate multilingual mathematics teaching and learning. In Richard Barwell, Philip Clarkson, Anjum Halai, Judit Moschkovich Mercy Kazima, Núria Planas, Mamokgethi Setati-Phakeng, Paola Valero, and Martha Villavicencio Ubillús, editors, *Mathematics Education and Language Diversity*, volume 21 of *New ICMI Series.* Springer Verlag, Berlin, Germany, 2016.

Lib10. Paul Libbrecht. Notations around the world:census and exploitation. In *Intelligent Computer Mathematics*, volume 6167/2010, pages 398–410. Springer Verlag, July 2010.

Mar09. Aaron Marcus. Global/intercultural user interface design. In Andrew Sears and Julie Jacko, editors, *Human-Computer Interaction: Design Issues, Solutions, and Applications*, chapter 18, pages 355–381. CRC Press, 2009.

MM12. Maria-Alessandra Mariotti and Mirko Maracci. Resources for the teacher from a semiotic mediation perspective. In Ghislaine Gueudet, Birgit Pepin, and Luc Trouche, editors, *From Text to 'Lived' Resources*, volume 7 of *Mathematics Teacher Education*, pages 59–75. Springer Netherlands, 2012.