Integration of automatic indexing system within the document flow in grey literature repository

Jindřich MYNARZ[2]
Ctibor ŠKUTA[3]

December 7th, 2010

Abstract

The Web empowered the authors of grey literature to publish their work on their own. In case of self-published works their author is also their indexer. And because not many of the grey literature authors are professional indexers, this may result in poor or no indexing.

Even though the Web made publishing easier, indexing is still hard. Nevertheless, we believe that the web technologies and machine learning algorithms may help to reduce the cognitive overhead involved in indexing, and make it eventually as easy as publishing on the Web is.

To help overcome the issue of quality and consistency of subject indexing automatic indexing systems can be used. Given enough full-texts already equipped with the terms from the controlled vocabulary that is to be used, machine learning algorithms can be employed.

Our aim is to provide *human-competitive* automatic indexing to authors and producers of grey literature. We demonstrate how an automatic indexing system based on machine learning can be integrated into the document flow in an open source digital repository of grey literature. We build upon open source tools and a controlled subject headings vocabulary available in an open standard format.

We will be using *Maui Indexer* as an automatic indexing system, *CDS Invenio* as a digital repository software, and *Polythematic Structured Subject Heading System* (PSH) as a knowledge organisation system. Both Maui Indexer and CDS Invenio are open source, and CDS Invenio's modular architecture makes it possible to extend it with new functionality. Maui Indexer works with controlled vocabularies expressed in Simple

[2] National Technical Library, Technická 6, 160 00 Prague 6 – Dejvice, Czech Republic, e-mail: `jindrich.mynarz@techlib.cz`
[3] National Technical Library, Technická 6, 160 00 Prague 6 – Dejvice, Czech Republic, e-mail: `ctibor.skuta@techlib.cz`

Knowledge Organisation System format in which the PSH is available.

From these components combined we will try to put together a solution for automatic indexing aimed at grey literature in the Czech language environment. Maui Indexer is domain and language independent so it is possible to adapt it for the field of Czech grey literature. The document samples we will test on will come from the *National Repository of Grey Literature* which is maintained by the *National Technical Library* of Czech Republic.

In the end, we will discuss integration of the automatic indexing component from the user perspective and sketch out how the user can interact with it through the user interface. We will also provide details around the actual implementation of the proposed system. The conclusion will deal with the evaluation of benefits of the implemented system for grey literature authors.

1 Introduction

The Web offers a publishing model that empowers masses of users to publish their works on their own. As it was stated in the previous literature, grey literature *"does not imply any qualification, is merely a characterisation of the distribution mode"*[3], and the Web can be considered as the single most significant distribution mechanism for grey literature. The sheer volume of documents available on-line constitutes a significant part of the grey literature publishing landscape. The Web has made *self-publishing* easier by lowering both the financial hurdles and the amount of know-how necessary for the publication process and thus it enabled a kind of *"do it yourself"* publishing.

While this is a tremendous benefit without which the *open access* movement would not have been established as firmly as it was, there are also drawbacks to it. By contrast to this mode of making documents accessible, the traditional publication models have procedures in place that go along with publishing that are not replicated well in grey literature publishing on the Web. The part of traditional publishing process that can be neglected in publishing on the Web is *subject indexing*. While it is now for the most part clear how to publish documents on the Web, the approaches to subject indexing are less established and available for use to non-professional users. This topic will be discussed in our paper and our focus will be to show how *self-indexing* of documents published on the Web via a digital repository can be accomplished; much in the same way users were endowed with the ability to *self-publish* their works.

2 Indexing of Grey Literature

Grey literature is characterized by a way of publishing that outputs documents with limited visibility. It may be hard to find such documents because they are distributed in a way that does not use established document access mechanisms, such as commercial databases and the like. This aspect of grey literature makes it difficult to be searched for either through libraries or web search engines. Also, the field of grey literature is closed tied to the *open access* way of publishing. However, if a document *cannot be found* there is no use of it being released in the *open access* way.

As we will argue, additional *subject indexing* terms can make grey literature documents searchable in a meaningful way so that they eventually become more prominent in the search results of library or web search engines. Subject indexing can be seen as an essential requirement to make documents findable [8], even though there are powerful search engines that enable documents to be found even without carefully crafted indexing. It can help in making grey literature documents more visible. Moreover, there is a consensus that it is necessary for useful navigation interfaces that can be built on top of digital document repositories (e.g., a faceted navigation) [7].

Subject indexing metadata enriches documents with *affordances* that allow to do more with them. It can support navigation interfaces that make it possible to browse the document collection in a useful manner based on *navigation paths* taken from the structure of the knowledge organisation system used for indexing. In this way the connections within the subject indexing system, such as hierarchical or associative relationships, may be harnessed as a *"map"* to the document collection. The subject indexing places the documents in a logical space constructed by the knowledge organisation system's structure and allows the user to browse the documents organized in such way by following the relationships between indexing terms. Every subject indexing term that is assigned to a document constitutes an *entry point* through which the document can be found and accessed.

Thus, we see subject indexing as an important enrichment to the document that enables to build interfaces for document repositories that can be navigated in a meaningful way. In this paper, we are interested in subject indexing done by the grey literature authors, and we have to admit that there are barriers that can make subject indexing a difficult task for them.

Grey literature is often published directly by the documents' authors, which may imply that, if it contains any subject indexing terms at all, it is the indexing done by the *authors themselves*. Subject indexing may be difficult for non-professionals and therefore this situation may result in no or poor indexing.

The established best practise for subject indexing is to use a *controlled vocabulary*. Authors know best about the contents of their works but they might not be familiar with the controlled vocabulary that they are supposed to use to express their works' content. They may not know how best to use the subject indexing system to describe their documents and therefore it constitutes a barrier for them, because first they have to learn how to use it.

This is why professional indexers are able to produce better indexing - they know how to *use* the subject indexing terms, especially with respect to the whole document collection. This background knowledge of the indexing system and the document corpus is a fundamental requirement for high-quality indexing.

In automatic indexing this background knowledge is in a sense captured in the *indexing model* on which the automatic system operates. And thus it turns out that, if configured properly, automatic indexing might come up with subject terms of reasonable quality and consistency. In the following sections we will describe how to reach that goal, while we will deal in detail with an application of automatic indexing to grey literature.

3 Automatic Indexing

In this paper we will investigate the option of non-professional *self-indexing* of grey literature on the example of an automatic indexing system for a digital document repository. We propose a semi-automatic indexing system that incorporates human feedback for the final selection of indexing terms. The system suggests a set of pre-selected indexing terms from a controlled vocabulary that may be used to describe document's content. These terms can be refined in the next step by the user interacting with a selection via interface that enables to remove or add new indexing terms.

The main help of our proposed approach is to lessen the cognitive overhead involved in intellectual indexing. Rather than completely automating the process of subject indexing we decided to use the automation for suggestions of indexing terms that can be amended and validated by the human user.

In this way we strive to provide grey literature authors with a tool that makes it possible for them to come up with non-professional indexing that is of high quality and consistency. The reason for such a goal is that the inconsistency of *user-generated indexing*, often done with freely created keywords, is one of its main drawbacks. This is what we try to alleviate by preprocessing the document and suggesting indexing terms in an automated manner. Also, we argue that this approach might help to increase the *scalability* of subject indexing.

The intellectual indexing carried on by professional indexers does not *scale*. Given that the size of the grey literature published on the Web is ever-increasing, there is a need for indexing system that is able to scale. The traditional solutions for subject indexing based on manual examination of each processed document do not provide a way to scale them up and therefore they are not the best answer for the requirements of the current grey literature publishing.

Scaling is a difficult problem in any situation and we do not strive to provide a definitive answer on how to scale up subject indexing in general or even in the case of grey literature. We propose that this issue can be alleviated by the use of automatic indexing. This solution scales because the time that is being used for subject indexing is *machine time* instead of human time. The processes that are carried out automatically with computers can be scaled up by assigning more computational power to them. Although in fact, our focus is not to achieve full automation of the indexing process so that it may be scaled on demand, but rather *to scale up the number of people* that are able to do the indexing while retaining a reasonably high quality.

Now that we have described the aim we try to achieve with automatic indexing, we will continue with a basic description of what automatic indexing is. It is a process of assigning indexing terms to a document in an automated fashion. It can use techniques based on analysis of language corpora. A common form of automatic indexing relies on “simple" statistical analysis of the full-text. In the case of our indexing system, we have used automatic *term assignment* that selects terms from a controlled vocabulary for a particular document based on the analysis of the document’s content. We have employed a *machine learning* approach that is based on computing conditional probabilities.

The approach we have chosen employs *supervised machine learning* that gathers feedback from users of the indexing system. Machine learning modifies the automatic indexing functionality with regard to the set of training data on which it “learns" how to do good indexing. The approach of supervised learning adjusts the indexing algorithm based on the newly acquired information about the way the indexing system is being used. Each time user approves of a set of automatically generated subject terms this decision is fed back into the indexing model and makes it more aligned with the specifics of the document collection.

One of the fundamental requirements of automatic indexing is the access to the *full-text* of the examined document. However, this is not a problem in the use case we have described. If the indexing of a document is done by its author, the access to the document’s full-text is granted.

Automatic indexing consists of a sequence of processes. During the course of automatic indexing the full-text is processed by a number of procedures that are collectively referred to as the machine processing *pipeline*. The full-text is sent through a sequence of processes that take the text as their input and pass their output to the process that is directly after them in the pipeline’s sequence.

In our case we start with a procedure that yields a *plain-text* of the document in question which might have been in another format, such as PDF or MS Word. Once we have acquired a plain-text embodying the content of the document a series of normalizations and helper procedures are run on it.

One of them removes the *stop words* - the words that do not affect the meaning of the document, such as prepositions or conjunctions. The system contains a list of stop words that are automatically excluded from further processing.

Another common technique that we take advantage of is *stemming* which reduces words to their root forms. In this way, we get rid of inflections, plural suffixes, and other characters that differentiate among the derivatives of the same root form. This method is supposed to collapse the different forms that refer to the same meaning to one word form so that a more effective computation can be done with it.

After these pre-processing steps automatic indexer carries on with its main suite of functions; it analyses the full-text and outputs a set of suggested indexing terms that can be assigned to the processed document. Since this paper is not about automatic indexing itself, but rather its application for grey literature, we will not discuss this part in detail and instead, we will move to the actual implementation of the indexing system.

4 Implementation

After having described the field of automatic indexing in general we will now proceed to provide an overview of the way we have implemented the automatic indexing system and put the methods of automatic indexing for grey literature into practice.

The guiding principle of our implementation was *re-use* of existing components, which we combined together in a document processing pipeline, or extended them in the cases where there was a need for it. This way of development would not have been possible if the parts we were building with had not provided access to their source code. Hence, their *open source* nature enabled us to modify them and extend their functionality. The combination of the constituent parts was possible due to their modular architecture that enabled them to be joined in a chain of processing procedures, which are applied on the examined document.

Not only the software that we have used in the automatic indexing system was open source, the data is communicated in this system in *open formats*. To illustrate this point, the subject headings system we have used was already available in RDF[1] data format expressed with SKOS,[2] an established standard for representing knowledge organisation system, such as thesauri, subject headings systems, or systematic classifications. It has gone through the standardization process and has reached the status of a recommendation of the World Wide Web Consortium.[3] This open standard is well supported by the indexer we have used, which eliminated the necessity of data conversion to a suitable format.

In order to manage the flow of control in the system a unifying data communication format is used. For this purpose we have adopted *JSON*, a light-weight data communication format.[4] The parts of the system exchange short JSON messages to pass the data needed for the indexing process to another part of the system. In this way we have harnessed another standard format to glue the components of the system together.

We wanted to preserve high modularity of the individual components in the system as a whole as well. Therefore we have exposed most of the functionality of the resulting system as a *web service*, which encourages loose coupling and re-use of the system's parts.

Now that we have described the overall architecture and design of the system we will move to the discussion of the individual parts. In the section that follows we will present an overview of the components that are involved in the automatic indexing system we have built.

[1] Resource Description Framework. <`http://www.w3.org/TR/rdf-concepts/`>.

[2] Simple Knowledge Organisation System. <`http://www.w3.org/TR/skos-reference/`>.

[3] http://www.w3.org/

[4] `http://www.json.org/`

4.1 Components

We will briefly describe each of the components that together make up for the whole automatic subject indexing system. These are not strictly limited to *software* but they also include *data* that is used in the process of subject indexing – the subject headings system from which the indexing terms are drawn, and the full-text corpus which serves for machine learning algorithms.

According to the design goal we have stated previously, the automatic indexer pipeline is composed mainly of already existing applications that we have re-used for this purpose. The parts that are new serve to connect the re-purposed components. We have written the *“glue code"* that ties the parts that are used in the process of automatic indexing. In order to connect all the components together and set up the indexer for processing Czech language, only a few additional functions had to be implemented.

4.1.1 Subject Headings System

One of the core components of the system is the controlled vocabulary of subject headings we have used. It provides indexing terms that are assigned to documents, which enables to maintain a degree of indexing’s consistency by referring only to indexing terms that are *authorized* by the subject headings system. The use of a controlled vocabulary implies that the suggested indexing terms are more consistent compared to the keyword extraction techniques.

The subject headings system we have used is the *Polythematic Structured Subject Heading System* (further abbreviated as PSH).[5] PSH is a bilingual Czech-English controlled vocabulary maintained and used at the *National Technical Library.*[6] It is a universal system and it consists of headings describing all major aspects of human knowledge. Its structure is similar to thesauri with hierarchical, associative, and equivalence relationships. PSH is primarily expressed in the *MARC 21 Format for Authority Data*,[7] but it was also converted to RDF data format, expressed with SKOS, which is more suitable for the automatic indexer we have employed.

4.1.2 Digital Repository

The *“host environment"* in which the system of automatic indexing is built in is the digital repository software. The software we have used for this purpose is *CDS Invenio*.[8] This software’s modular architecture enabled us to extend its functionality with a new plug-in that does the automatic indexing.

Invenio processes newly submitted documents in a series of step that are referred to as the document workflow. We have inserted the automatic indexing pipeline into the workflow so that a document can go through this additional suite of procedures to be enriched with subject indexing terms.

[5] The on-line version is available at `http://psh.ntkcz.cz/skos/`

[6] `http://www.techlib.cz/en/`

[7] `http://www.loc.gov/marc/bibliographic/`

[8] The project’s website is at `http://invenio-software.org/` and our installation of Invenio is available at `http://invenio.ntkcz.cz/`.

The user interface of the automatic indexing system is included in the repository. To achieve this level of integration not only we had to extend the functionality of the modified software but also alter the presentation interface to enable the user to access the added new functionality. This was possible due to the clear separation of the code responsible for the repository's core functions and the templates that build up the interface the user is interacting with.

4.1.3 Automatic Indexer

The component that is responsible for the main function of the system is the automatic indexer. We have chosen to use *Maui Indexer*.[9] The author of this software claims that it produces *human-competitive indexing*[4] and it seems correct from the results of the comparative studies done with this indexer.[10] This assertion is based on the implicit presumption that the subject terms assigned by humans are the standard with which the quality of automatic indexing is compared. The indexing produced by Maui Indexer is thus comparable to human indexing both in terms of quality and consistency, which is precisely the result we are looking for in our automatic indexing system.

The software is described as being independent of the domain and language for which it is used. However, to achieve the best precision some adjustments need to be done. Because of the language of the document collection, for which the automatic indexer was intended, we wanted to adapt Maui Indexer for the Czech language. The modifications involved changing of the indexer's parts: the list of stop words and the stemmer code.

In the ideal case, stop words would be based on the corpus of documents that we want to index. We chose to use the *Czech National Corpus*[11] instead to create a list of the most frequent words that may be used as stop words. Czech National Corpus is a vast document collection reflecting the contemporary written Czech [2]. Thus, it served well to establish a good baseline with respect to our document collection.

To reduce words in an analysed full-text to their root forms we have taken over the *aggressive Czech stemmer* [1], which we have adapted in a way so that it can be plugged in the Maui Indexer's source code.[12] The aggressive nature of the stemmer is based on the approach it takes to stemming non-root word forms. It addresses the morphological characteristics of the Czech language to normalize the irregularities of inflection, consonant alterations and the like. However, in some situations, it may remove characters that are necessary for the distinction of the word sense and thus create the same root form from multiple words that do not share such a root. This feature may compromise the quality of resulting indexing and it is the reason why we consider it as an immediate target for further refinement of our indexing solution.

[9] `http://code.google.com/p/maui-indexer/`

[10] Examples can be found at `http://code.google.com/p/maui-indexer/wiki/Examples`

[11] `http://ucnk.ff.cuni.cz/english/index.php`

[12] The stemmer we have used is available at `http://members.unine.ch/jacques.savoy/clef/CzechStemmerAgressive.txt`.

4.1.4 Text Corpus

The procedure built by combining the previously mentioned components was applied to a text corpus of the document collection of grey literature documents stored and maintained in our digital repository. In our case, we have applied the automatic indexing system we have built to the *National Repository of Grey Literature*.[13] This repository, maintained at the *National Technical Library*, collects grey literature from the network of cooperating partner institutions, ranging from the institutes of the Academy of Sciences of Czech Republic to public universities [6].

The contents of the documents included in this repository are mostly in Czech. This was the primary drive behind the decision to enhance the functionality of the automatic indexer towards the Czech language. We have also taken into account that the contents of the repository are produced in collaboration with the partner institutions. Its long-time goal is to have the co-operating institutions produce the document's descriptive metadata, including subject indexing, on their own without a need for central co-ordination. This intention made for an adequate use case of the author-generated subject indexing.

4.2 User interface design considerations

While the quality of the underlying procedures is certainly crucial to produce results of a reasonable quality, the part of the indexing system that has a comparable significance is the *user interface*. The importance of the user interface design stems from the necessity of user feedback. The way how the users provide feedback on the automatically generated set of indexing terms needs to be designed carefully to take advantage of the author's knowledge about the indexed document. The resulting design has some notable features that may have a significant influence on the user experience with the indexing system.

We have decided to provide the automatic indexing as an *opt-in* procedure, which means that the user has to actively declare that the document entered into the repository should be processed with the automatic indexer. If the user checks in a box for automatic indexing, during the next step in the document workflow there will be an additional screen containing the suggested indexing terms and the functionality that allows to modify them.

The primary functionality of the automatic indexing system we have developed is to suggest a list of subject headings that in some way describe the processed document. Its objective is to facilitate non-professional indexing while maintaining a high level of consistency. It is not meant to serve as a *replacement* for the person doing the indexing, but in fact, it is a start for the indexing process controlled by the human user.

We still see the main value being added to the processed documents by the user of the system rather than generated by the system itself. For this purpose we have enhanced the user interface with helper functions that are meant to facilitate more effective indexing.

[13] `http://nrgl.techlib.cz`

One of these functions is the *auto-complete* feature. To bridge the gap between the language of the user and the language of the knowledge organisation system used for indexing the user interface displays a list of suggested headings based on a heading's fragment supplied by the user.

To assist the user in deciding on the subject heading's adequacy we have added a utility that shows citations of the example documents indexed with the subject heading user considers to use. When user highlights certain subject heading a short list of links to the documents which have the same heading attached is presented in the interface. In this way, the user can *learn by example* to consider the applicability of a given subject heading for a particular document.

We had to make alterations to the *search interface* as well. There would be no value in added indexing terms if one could not use them to search and navigate the document collection in which they are used. To reflect this added structure a change to the user interface needs to be made so that it is possible to harness the indexing terms to access documents. In our case, responding to this requirement consisted in adding a new search index built for the subject headings and appending a new field to the search form to access this index.

5 Future Possibilities and Challenges

The work we have done with the automatic indexing system for grey literature is by no means finished and we are aware that there are further possibilities for improvement and challenges that have to be solved to deliver a better system. In the specific use case we have described in this paper there are certain aspects of the automatic indexing that are worth underscoring.

It is important to note that there is no use in adding subject indexing terms if such metadata cannot be harnessed via the search or navigation interface for the document collection. If the indexing is not reflected in end-user interfaces it does not provide any additional value. Thus, the indexing must be represented in user interfaces to have an impact on the overall functionality of a digital repository. As we have mentioned in the previous section, we have extended the digital repository with a new search field to access the documents through subject headings to address this issue.

We have to take into account that the indexing we are dealing with here is still a *user-generated indexing*, even though it is somehow refined by the system we have implemented. This implies that the resulting indexing terms might need further verification by a professional if we want to have a quality control in place. As we have sketched in the paper, this is the way our system works and will work for the near term future. Nonetheless, our goal is eventually to get rid of this necessity once we will have a higher level of confidence in the non-professional subject indexing that is fed into the repository.

Due to the modular nature of the whole system, there is a plethora of ways how it can be enhanced and developed further. Every part of the document processing pipeline can be considered for an improvement. Our aim is first to focus on the parts that affect the quality of the results of indexing the most. After every change we want to check if it leads to an increase in the system's precision by comparing it with the results the system had on the same document with the previous configuration.

We argue that the system not only benefits grey literature authors and maintainers of digital repositories, but, moreover, it can also benefit the individual components it is made of because it reflects on the way the component is being *used*. This can be applied on the knowledge organisation system from which the indexing terms are drawn. The data about its usage in practise can be crucial for its development and further evolution reflecting the changing needs of the user community and a shift in the way its concepts are perceived.

6 Conclusions

This project would not have been possible if there were not *open standards* that govern the field of subject indexing. They enabled the re-use of existing components adhering to certain standards and their combination in a novel way. The layer provided by open standards constituted an environment for interoperability and systems built with an open architecture in mind.

All the parts we have put together in this open framework are *open source*. This means they are open to modifications and extensions, and those were necessary for the system to work as a whole. Moreover, due to the their modular character it was possible to switch one part for another or plug in a new component.

The indexing system that came out of this way of development is applied to the grey literature documents. It was designed to reflect the nature of grey literature. We have argued that the situation of subject indexing of grey literature is unsatisfactory and we have expressed our view of the causes for such state. Our motivation was to react to the current conditions and propose an approach that may lead to an improvement of the way subject indexing is done for grey literature.

References

[1] DOLAMIC, Ljiljana; SAVOY, Jacques. Indexing and stemming approaches for the Czech language. *Information Processing & Management*. November 2009, vol. 45, iss. 6, p. 714 – 720. ISSN 0306-4573. DOI 10.1016/j.ipm.2009.06.001.

[2] KUŘERA, Karel. The Czech National Corpus : principles, design, and results. *Literary and Linguistic Computing.* 2002, vol. 17, no. 2, p. 245 – 257. ISSN 0268-1145.

[3] MACKENZIE OWEN, John S. The expanding horizon of Grey Literature. In *Perspectives on the design and transfer of scientific and technical information : proceedings of the 3rd international conference on grey literature*. Amsterdam : TransAtlantic, 1998, pp. 9–13. Also available from WWW: <http://cf.hum.uva.nl/bai/home/jmackenzie/pubs/glpaper.htm>.

[4] MEDELYAN, Olena. *Human-competitive automatic topic indexing.* Waikato, 2009. 214 p. Dissertation thesis (PhD.). University of Waikato,

Department of computer science, 2009. Also available from WWW: <http://www.cs.waikato.ac.nz/~olena/publications/olena_medelyan_phd_thesis_July2009.pdf>.

[5] PEPE, A. [et al.]. CERN Document Server Software : the integrated digital library. In DOBREVA, Milena; ENGELEN, Jan (eds.). *9th ICCC International Conference on Electronic Publishing : from author to reader : challenges for the digital content chain.* Leuven : Peeters, 2005. ISBN 90-429-1645-1.

[6] PEJŠOVÁ, Petra (ed.). *Grey literature repositories*. Zlín : VeRBuM, 2010, 156 p. ISBN 978-80-904273-6-5.

[7] RIBEIRO, Fernanda. Subject indexing and authority control in archives : the need for subject indexing in archives and for an indexing policy using controlled language. *Journal of the Society of Archivists.* April 1996, vol. 17, iss. 1, p. 27 – 65. ISSN 0037-9816.

[8] SYKES, Jan. *The value of indexing : a white paper prepared for Factiva, a Dow Jones and Reuters Company* [online]. February 2001 [cit. 2010-12-02]. Available from WWW: <http://4info-management.com/pdf/indexingwhitepaper.pdf>.

[9] VLACHIDIS, Andreas [et al.]. *Excavating grey literature : a case study on rich indexing of archaeological documents by the use of natural language processing techniques and knowledge based resources* [online]. 2009 [cit. 2009-09-30]. Preprint for ISKO UK 2009 : Content Architecture: Exploiting and Managing Diverse Resources. Available from WWW: <http://www.iskouk.org/conf2009/papers/vlachidis_ISKOUK2009.pdf>.