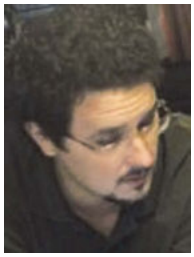
Evaluating the Practical Applicability of Thesaurus-Based Keyphrase Extraction in the Agricultural Domain: Insights from the VOA3R Project†

David Martín-Moncunill*, Elena García-Barriocanal**,
Miguel-Angel Sicilia***, Salvador Sánchez-Alonso****

Information Engineering Research Unit, Computer Science Department,
University of Alcalá, Ctra. Barcelona km. 33.6 28871 Alcalá de Henares (Madrid), Spain,
*<d.martin@uah.es,> **<elena.garciab@uah.es,>
<msicilia@uah.es,> *<salvador.sanchez@uah.es>

David Martín-Moncunill (University of Alcalá, Spain) is a researcher and Ph.D. candidate at the Computer Science Department. He is a computer management engineer from University of Alcalá, he also has a degree in Information Systems from the same university and a M.Sc. in e-learning and social networks from International University of la Rioja. He joined the Information Engineering Research Unit in 2012, since then he has collaborated in several European Funded Research projects mainly on the topics of metadata, e-learning, repositories, semantic web and semantic interoperability. His preferred areas include usability, user-centered design, user experience, accessibility, knowledge representation.



Miguel-Angel Sicilia (University of Alcalá, Spain) is a full professor at the Computer Science Department. He holds a University degree in computer science from the Pontifical University of Salamanca and a Ph.D. from Carlos III University in Madrid, Spain. Head of the Information Engineering research unit, he has a sound background in project management as well as a clear research profile in information systems, author of more than 50 *JCR* publications in the last 10 years. He has been involved in the last ten years in different Semantic Web and metadata research projects.

Salvador Sanchez-Alonso (University of Alcalá, Spain) is an associate professor at the Computer Science Department. He previously worked as an assistant professor at the Pontifical University of Salamanca. He holds a Ph.D. in computer science from Polytechnic University of Madrid and a degree in library science from University of Alcalá. Author of more than 30 high impact factor publications in the last 10 years, he has participated or coordinated in several EU-funded projects in the last 5 years on the topics of learning object repositories and metadata.

Elena García-Barriocanal (University of Alcalá, Spain) is an associate professor at the Computer Science Department. She obtained a university degree in computer science from the Pontifical University of Salamanca in Madrid and a Ph.D. from the Computer Science Department of the University of Alcalá. Her research interests are centered in Knowledge representation and Semantic Web. In the last few years Elena has supervised several Ph.D. works on those areas, and has authored a good number of papers published in *JCR* -indexed journals. She also has extensive experience in EU-Funded projects.

Martín-Moncunill, David, García-Barriocanal, Elena, Sicilia, Miguel-Angel, Sánchez-Alonso, Salvador. **Evaluating the Practical Applicability of Thesaurus-Based Keyphrase Extraction in the Agricultural Domain: Insights from the VOA3R Project.** *Knowledge Organization. 42(2)*, 76-89. 43 references.

Abstract: The use of Knowledge Organization Systems (KOSs) in aggregated metadata collections facilitates the implementation of search mechanisms operating on the same term or keyphrase space, thus preparing the ground for improved browsing, more accurate retrieval and better user profiling. Automatic thesaurus-based keyphrase ex-

traction appears to be an inexpensive tool to obtain this information, but the studies on its effectiveness are scattered and do not consider the practical applicability of these techniques compared to the quality obtained by involving human experts. This paper presents an evaluation of keyphrase extraction using the KEA software and the AGROVOC vocabulary on a sample of a large collection of metadata in the field of agriculture from the AGRIS database. This effort includes a double evaluation, the classical automatic evaluation based on precision and recall measures, plus a blind evaluation aimed to contrast the quality of the keyphrases extracted against expert-provided samples and against the keyphrases originally recorded in the metadata. Results show not only that KEA outperforms humans in matching the original keyphrases, but also that the quality of the keyphrases extracted was similar to those provided by humans.

† *Acknowledgments:* The research within the project VOA3R leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 4—"Open access to scientific information," grant agreement n° 250525. We would also want to thank the FAO-AGRIS team, especially to Imma Subirats, Stefano Anibaldi and Fabrizio Celli for their implication in the VOA3R project and their quick and precise reply, every time we raised an issue to them.

Received: 18 February 2015; Revised 19 March 2015; Accepted 23 March 2015

Keywords: keyphrases, keyphrase extraction, documents, KEA, AGROVOC

1.0 Introduction

Metadata describing information resources usually includes some form of classification or description using some type of knowledge organization system (KOS) (Hjørland 2003), be it a controlled vocabulary, thesaurus, classification system, ontology, etc. This facilitates a digital collection with a homogeneous and consistent description of its resources, unlike a collection with free keyphrases (Zeng and Chan 2004). However, when larger, aggregated digital collections contain source collections that use different KOS, the descriptions need to be aligned.

Mapping KOSs provides a solution to this problem. In some cases, a source KOS is completely mapped to many others (Anibaldi et al. 2013) providing a high quality mapping solving the problem. However, in other cases, either the mappings are unavailable, or their coverage is limited to small parts of the original KOS. Chen and Chen (2012) stated that the research on the interoperability of KOSs under specific domains is mostly related to mathematics or life science but, on the contrary, there is less research on the interoperability of KOS in different languages under the domain of humanities, art, or cultural assets. This incentivizes the use of automated keyphrase extraction techniques that produce terms in the common terminology of choice, thus providing the homogeneous description needed.

A number of techniques for keyword or keyphrase extraction have been proposed to date; GenEx (Turney 1999), KP-Miner (El-Beltagy 2006) and KEA (Medelyan and Witten 2005) are some of the most popular and most referenced (Lim et al. 2013). However, information on the quality and applicability of these keyphrase extraction techniques is limited and uneven, as few studies have addressed them.

Evaluation of the effectiveness of these kinds of algorithms typically has been done using precision and recall measures, which compare the list of keyphrases generated to author-supplied lists; most of such studies have been limited in terms of sample size. Furthermore, this approach only gives an assessment of the accuracy of the system for matching the keyphrases provided by the original author (or curator), which raises some important questions. For example, it is not always accurate to assume that the keyphrases originally selected were the most appropriate ones, or that human experts would obtain better results than algorithms when trying to match original author keyphrases. In addition, from the viewpoint of practical applicability, we need to know the relative quality of human-assigned keyphrases compared to machine-assigned ones.

This makes collection owners uncertain about the quality and the practical applicability of these techniques to their integrated services. This paper reports on an attempt to partially fill the gap by describing additional evidence on the outcomes of keyphrase extraction algorithms applied in the context of the VOA3R (Goovaerts 2011) project. Specifically, the KEA open source framework (Witten et al. 1999) is used to extract keyphrases from large samples of the AGRIS[1] collection—a comprehensive thematic collection of scientific information in the field of agriculture—using the AGROVOC vocabulary (Sini et al. 2008), a controlled vocabulary that covers all areas of interest to FAO[2].

"VOA3R - Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research related to Agriculture, Food, and Environment", which launched in June 2010, is a project funded by the European Commission under the CIP PSP program. The general objective of VOA3R is to improve the spread of European agriculture and aquaculture research results by

using an innovative approach to sharing open access research products. VOA3R connects several publication systems in the agriculture and aquaculture domain—which includes AGRIS—under a strict open access policy. VOA3R platform aims to reuse existing and mature metadata and semantics technology—such as AGRIS application profile and AGROVOC controlled vocabulary—to deploy an advanced, community-focused integrated service for the retrieval of relevant open content.

The proper functioning of VOA3R services requires not only the integration, but often the enrichment of metadata; which motivated us to make a pilot exploratory study to provide insights about the evaluation of the practical applicability of keyphrase extraction techniques for this purpose. The motivation for the choice of KEA for this pilot exploratory study is described at the end of the "Background" section.

Results show that KEA performance when trying to match original keyphrases was similar to that seen in previous experiments done in the field and slightly better than the human average. Evaluators participating in a blind test considered the quality of the extracted keyphrases, judged by how well the keyphrases represented the paper, to be at a level similar to human performance. The rest of this paper is structured as follows. We provide a brief background on keyphrase extraction evaluation and the context of use of the present research, followed by a description of the data preparation. Next, we detail the analysis and discuss the results, and end with conclusions and outlook.

2.0 Background

Keywords and keyphrases could be defined as word sequences which characterize the topic of a document and its content (Turney 1999). They have proven to be valuable tools in the information science and knowledge organization (Dahlberg 2006) context, particularly in tasks such as annotation (Frank et al. 1999; Park et al. 2013), indexing (Hjørland 2011), summarization (Al-Hashemi 2010) or to ascertain the subject—or "aboutness" (Hjørland 2001)—of a document; thus allowing to improve the retrieval and facilitate the categorization and browsing of information (Jones and Paynter 2002; Hjørland 1998; Hjørland 2011).

Keyphrase extraction consists of the selection of the most important topical phrases from within the body of a document (Turney 1999), which could be achieved either automatically—by the use of automatic keyphrase extraction techniques—or manually—by human experts. Manual keyphrase extraction using a controlled vocabulary is a very time consuming task which requires not only knowledge in the topic, but also experience in the task itself. Current information and data growth (Lord et al. 2004) makes manual keyphrase extraction not suitable for the vast majority of the cases, due to both time and resources reasons. In this context, automatic thesaurus-based keyphrase extraction (Lim et al. 2013) appears to be an inexpensive and fast tool to face the problem.

The assignment of keyphrases to documents—keyphrase indexing (Erbs et al. 2013)—can be done using a controlled vocabulary or freely choosing representative phrases appearing in the body of the document, which is known as "free indexing." Controlled vocabulary keyphrase extraction techniques have shown better results in areas as medicine (Névéol et al. 2007) or agriculture (Medelyan and Witten 2005), however, it is clear that results will vary according to the completeness and overall quality of the vocabulary used.

In fact, as stated by Joorabchi and Mahdi (2013) "the main weakness of the keyphrase indexing approach is that it assumes there exists a comprehensive domain-specific thesaurus for the target domain, which is not always a feasible assumption." To address these cases, Arash and Mahdi (2013) propose a keyphrase indexing approach which uses Wikipedia and a supervised ranking function based on Genetic Algorithms, following previous research carried out by Milne et al., (2006) and Medelyan et al. (2008, 2009b). As previously stated, this is not our case, since we use the highly comprehensive Agrovoc vocabulary, which is presented in section 3.1.

2.1 Keyphrase Extraction Techniques

Keyphrase extraction techniques can be classified according to two criteria (Lim et al. 2013):

> The learning approach (Turney 2000), which may be "supervised," "unsupervised," or "non-learning," the type of problem, which may be a classification problem (Witten et al., 1999) or a ranking problem (Frantzi et al. 2000). Some of the most relevant keyphrase extraction techniques, are introduced below. Additional information on automatic keyphrase extraction techniques could be found on the review elaborated by Lim et al. (2013), who provide an extensive review of the most relevant automatic keyphrase extraction techniques, describing their strengths and weaknesses.

GenEx (Turney 1999) is a Keyphrase extraction technique which combines two modules: "Extractor"—the automatic keyphrase extraction system and "Genitor"—an external system used to calibrate the keyphrase extractor, which uses 12 parameters. The extractor manages a decision-tree-like process to perform both candidate selection and

weighting, based on three attributes: term frequency (TF), first occurrence information, and phrase length. Term frequency is used as the base score in Extractor. GenEx does not use any domain-dependent attributes for classification and weighting purposes. Thus, although it requires a long initial training period, there is no need to re-train GenEx for every new domain. The most valuable contribution of GenEx is the ability to retain its performance across different domains (Lim et al. 2013; Frank et al. 1999).

The *multilayer perceptron (MLP)* technique (Sarkar et al. 2010) was conceived to address the problem of having fewer generated keyphrases than the requested number when treating keyphrase extraction as a classification problem. This technique assumes that keyphrase extraction should be treated as a ranking problem rather than a classification problem. MLP processes are similar to that of Extractor in GenEx (Turney 1999), but use more attributes: TFxIDF (Salton et al. 1975) and the combination of phrase length and word length. Also, in order to address the issue of selecting an insufficient number of keyphrases, MLP attaches the classified non-keyphrase list to the end of the keyphrase list.

KP-Miner (El-Beltagy 2006) is a non-learning, ranking-based approach, which means that no training is needed for the system. This keyphrase extraction technique uses three attributes: TFxIDF (Salton et al. 1975), First Occurrence Position, and a boosting factor, focusing on both candidate selection and weighting process. KP-Miner has outperformed techniques with a learning approach, such as GenEx and KEA (El-Beltagy and Rafea 2009), but has been criticized for the complexity of its structure, the use of unnecessary processes, and the existence of a bias in the calculation of term frequency, where smaller n-grams tend to achieve higher scores because of potential presence in both parent phrase and sub-phrase (Lim et al. 2013). In fact, Kumar and Srinathan (2008) suggested to improve KP-Miner by introducing the N-gram Filtration Technique in the weighting process, using LZ78 data compression techniques to generate the candidate keyphrase list and ignoring IDF when choosing phrase attributes for weighting.

KEA (keyphrase extraction algorithm), by Frank et al. (1999), is based on a supervised learning approach in a classification problem context employing naïve Bayes as the machine learning algorithm. KEA focuses on candidate selection, term weighting and classification/ranking process, using two attributes for selecting candidate keyphrases: first occurrence and TFxIDF. The naïve Bayes learning algorithm allows KEA to require far less training time than GenEx, while still performing at about the same level (Sarkar et al. 2010). KEA allows improving precision and recall of keyphrase extraction by incorporating domain dependence, used to calculate IDF attributes based on the collection of domain corpus used in trainings.

2.2 Evaluation Techniques

As discussed above, the assignment of keyphrases to documents (keyphrase indexing) can be done using a controlled vocabulary or without one, which is known as "free indexing." In this paper we focus on the first approach as we aim to provide information systems with a uniform keyphrase space for different tasks. Controlled vocabulary keyphrase extraction techniques have been subject to evaluation in the literature. Previous work has grouped evaluation techniques into the following two categories:

> Automatic evaluation: compares automatically machine-generated indexes with the originally human-assigned ones, establishing a gold standard. This approach has some problems, notably that the choice of index terms has been criticized as subjective (Pouliquen et al. 2006).

> Manual evaluation: human evaluators compare the set of machine-generated indexes with the source text, a process in which human evaluators usually perform qualitative analysis (Mendelyan et al. 2009; Névéol et al. 2007; Ruiz and Aronson 2007).

2.3 Motivation for the Choice of KEA

In this work, we are interested in applying pre-trained models to new collections from a practical perspective, so that available extractor models can be used in a straightforward way in a digital collection, avoiding the burden of the bootstrapping and supervised model-building process. We are also interested in controlled vocabulary indexing, as it provides a homogeneous terminological space for a digital collection built through the aggregation of heterogeneous ones.

KEA has been widely used along with AGROVOC to extract keyphrases from documents on the agricultural domain, the most relevant ones will be presented in this section. Results of the evaluations carried out in previous experiments sound promising, but they focus on the automatic evaluation not providing the "practical applicability" information we were looking for. Also, the number of analyzed documents was reduced and our collection of documents seemed to be much more heterogeneous than the selections considered in previous experiments. Following this reasoning, and building on previous experiments using the AGROVOC thesaurus, we

have selected KEA as the keyphrase extractor technique for our experiment.

Medelyan and Witten (2005) used KEA to automatically extract index terms from documents relating to the domain of agriculture using the AGROVOC thesaurus developed by the FAO as a controlled vocabulary and as a knowledge base for semantic matching, evaluating their algorithm in a corpus of 200 documents. Variations of keyphrase extraction for particular types of resources have been proposed by authors like Nguyen and Kan (2007), who introduced features on top of KEA that capture the positions of phrases in a document with respect to logical sections found in scientific discourse, evaluating their algorithm in a corpus of 120 documents.

Human evaluations of automatic keyphrasing have been done in some experiments, such as the one carried out by Mendelyan et al. (2009), which demonstrated that documents could be tagged automatically with an accuracy comparable to that of assignments by human taggers. This experiment analyzed the tagging and keyphrase consistency of the CiteULike.org service for organizing academic citations, using a set of 180 documents indexed by 332 taggers. Results showed that the algorithm's consistency could compete with and even improve upon humans' consistency.

Jones and Paynter (2001) evaluated KEA based on subjective evaluations of the quality and appropriateness of extracted keyphrases by human experts. A set of six papers from the *Proceedings of the ACM Conference on Human Factors, 1997*—with the authors' keyphrases removed—were used as test documents. Then 28 subjects were asked to evaluate how well the extracted keyphrases represented each paper, ranking them from 0 to 10. Using the Kendall Coefficient, the analysis demonstrated that there were significant and sometimes strong levels of agreement between the subjects in assessing keyphrases, and that most of these phrases were rated positively.

As we will detail in the following section, the "human-evaluation approach" of our experiment is different as we compare matches between KEA and original keyphrases (assigned by curators using the AGROVOC vocabulary) with matches between human domain-related experts and these AGROVOC-based keyphrases. We have found no previous experiments that take this approach, which seems promising as a way to generate additional evidence about the outcomes of keyphrase extraction algorithms versus manual extraction techniques. Finally, in our evaluation of KEA's performance, our set was considerably larger than those employed in the previously mentioned experiments with AGROVOC; furthermore, it was heterogeneous and comprehensive, with documents not all sourced from the same collection, the same topics, or the same years.

3.0 Materials and Methods

The experiments reported herein were conducted using the set of documents from the United Nations Food and Agriculture Organization (FAO) AGRIS collection, which was selected to be integrated in VOA3R and the AGROVOC vocabulary.

The data gathering and selection process was complex, involving different stages, including the elaboration of a software tool to analyze and classify the records. First, we introduce AGRIS collection and provide information about the number of documents classified by year and by type. Then we analyse the set of AGRIS documents selected by FAO and the VOA3R consortium to be integrated in the VOA3R collection in order to properly select and gather a suitable set for the experiment. Here we provide a briefing of the main steps in order to facilitate readers' understanding:

First we extracted the number of AGROVOC keyphrases of every document. The analysis of this data allowed us to start the selection process according to the number of keyphrases.

Then we aimed to look for full-text English documents suitable to be processed by KEA software. We realized that only about 3% of the entire AGRIS collection had a full-text link available in its metadata, furthermore, not all of these links pointed to suitable full-text documents, but to abstracts or to scanned documents (image format) which were not suitable to be processed by KEA.

We also aimed to elaborate a set of documents as comprehensive as possible, as part of this task, we had to work with the AGRIS Resource Number (ARN) which contains information about the provenance of the resource: the country, the year of inclusion and the subcenter code. In order to ensure the adequacy within the previous points, a sample of 2000 full-text documents was manually selected. Finally a selection of five documents for comparing expert-generated and automatically extracted keyphrases was made. This point is explained in section 3.4.

3.1 The AGRIS Collection

AGRIS is a collection of more than 7.6 million bibliographic records on agricultural science and technology topics. It is one of the most important worldwide information systems in the agricultural domain, serving a million pages a month, with more than two hundred fifty thousand users accessing the system every month.

These records cover the various fields of agriculture in a broad sense, including forestry, animal husbandry, aquatic sciences and fisheries, and human nutrition from 1975 to the present. Metadata on several types of docu-

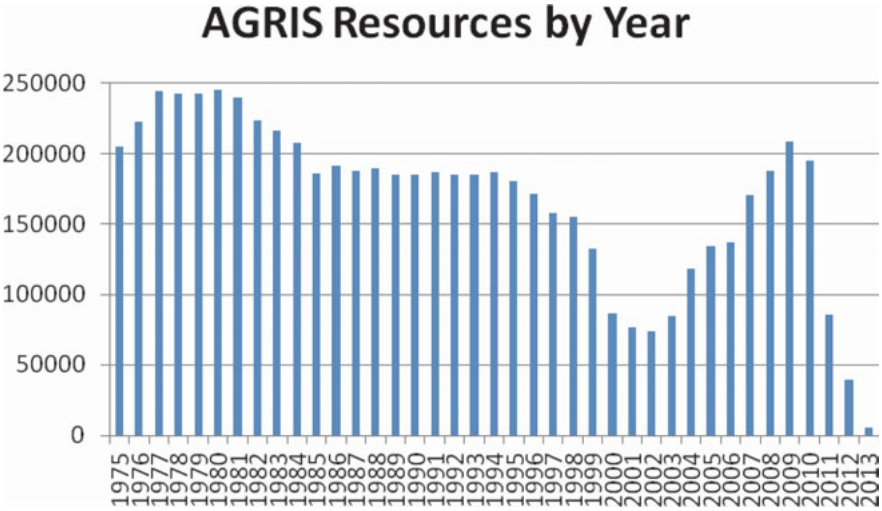


Figure 1. AGRIS resources by year (November 2013)

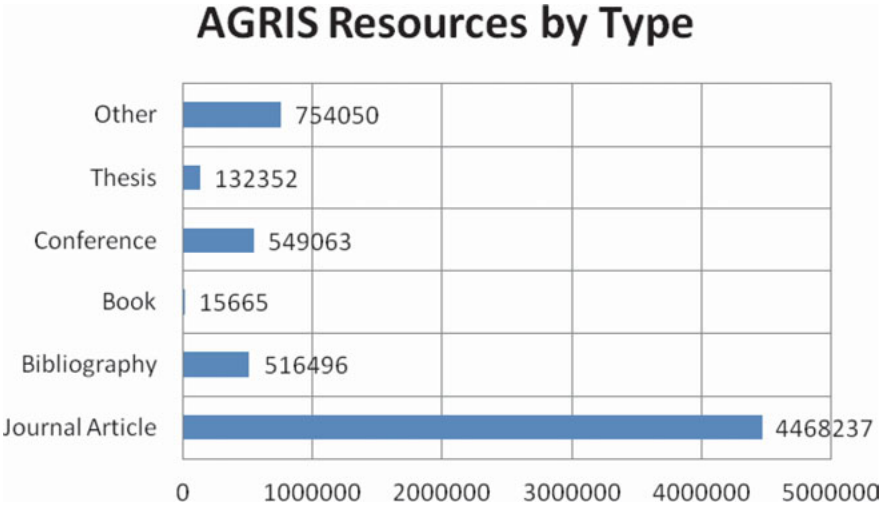


Figure 2. AGRIS resources by type (November 2013)

ments and resources can be found, including scientific and technical reports, government publications, theses, and conference papers, among others. This makes AGRIS a highly comprehensive database, as can be seen in figure 1, which shows the distribution of AGRIS resources by year (element 4.5.2 "Date of publication" of the AGRIS AP) and figure 2, which shows the distribution of AGRIS resources by type of publication (element 4.9 "Type" of the AGRIS AP).

AGROVOC[3] is the FAO corporate thesaurus. It covers topics related to the interests of this organization, including agriculture, forestry, fisheries, environment, and related domains. It was first developed in the 1980s to standardize the indexing process for the International System for Agricultural Science and Technology (AGRIS) database in order to make searching simpler and more efficient, and to guide the user to the most relevant resources. Today AGROVOC contains over 40,000 concepts organized in a hierarchy; each concept may have labels in up to 22 languages. It is widely used by researchers, librarians and information managers for indexing, retrieving and organizing data in agricultural information systems and web pages around the world (Šimek et al. 2013).

The AGRIS Application Profile (AGRIS AP) is the standard metadata schema created by the FAO to define resources listed in AGRIS. The AGRIS AP element "4.6.3 Subject Thesaurus" is used to provide keyphrases that describe the content of the resource and indicating descriptors that are part of a controlled vocabulary. Figure 3 shows an example of AGRIS AP metadata.

As can be seen from the previous analysis, the size, number of resources, and distribution of AGRIS represent a comprehensive thematic collection of scientific information in the field of agriculture. In the same way, AGROVOC's vocabulary coverage, number of terms and

```
<ags:subjectThesaurus xml:lang="en" scheme="ags:AGROVOC">Mentha pulegium</ags:subjectThesaurus>
<ags:subjectThesaurus xml:lang="en" scheme="ags:AGROVOC">chemical composition</ags:subjectThesaurus>
<ags:subjectThesaurus xml:lang="en" scheme="ags:AGROVOC">essential oils</ags:subjectThesaurus>
<ags:subjectThesaurus xml:lang="en" scheme="ags:AGROVOC">antimicrobial properties</ags:subjectThesaurus>
```

Figure 3. A fragment of a resource description using AGRIS AP. The code shows four keyphrases from AGROVOC (in boldface).

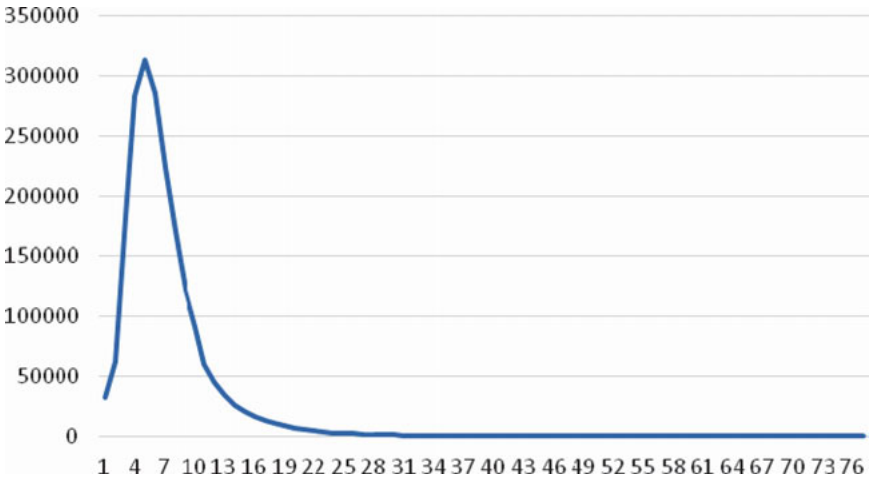


Figure 4. Number of resources in the AGRIS collection for VOA3R by number of keyphrases.

hierarchy represents a highly comprehensive vocabulary. This makes AGRIS and AGROVOC ideal tools for accomplishing the objectives of this study.

3.2 Sample Selection

As described in the introduction, the need to integrate and enrich metadata for the VOA3R platform brought us the opportunity to experiment and analyze the practical applicability of the KEA automatic indexing service in a real-world context, with a document collection that is significantly larger than the one used in the AGROVOC-related experiments mentioned in the background section. Among its collections, VOA3R contained a selection of documents coming from AGROVOC. The selection of these documents was made in the context of the VOA3R project's "Content Providers Work Package" by FAO and VOA3R consortium experts.

The set of documents chosen as the sample for this experiment was selected from the AGRIS collection in VOA3R, which is a collection of almost 3 million AGRIS papers and journal articles provided by FAO to be integrated into the VOA3R platform. The selection of a sample from this collection required devising a system to extract the necessary information to prepare all the materials needed for the experiment, aiming to ensure the applicability of the results to the collection.

This system allowed us to gather valuable statistical data, analyzing the AGRIS collection selected for VOA3R and creating one file containing the keyphrases for every resource—AGRIS bibliographic records are manually created by cataloguers and sometimes suffer from incompleteness; e.g. some of them do not have a reference to the full text of a document. This step was also useful in classifying the resources and establishing the environment needed to properly run KEA.

Using this system, we realized that only about 3% of the entire AGRIS collection had a full-text link available in its metadata, so we extended the system so that it could help us search for full-text documents on Google Scholar (Harzing and Van der Wal 2007) using the information contained in the AGRIS records. This was done by submitting the title of each document and then manually cleaning the data for matches with full text available.

The total number of VOA3R-AGRIS collection records was 2,939,982. We evaluated the number of AGROVOC keyphrases in English, which ranged from 0 to 77 keyphrases per document. Figure 4 shows the overall distribution of keyphrases per document, starting with documents with one keyword.

In the distribution above, the mode of keyphrases was 5 and the mean about 6.9 keyphrases per resource. Figure 5 provides further detail in the range from one to ten keyphrases.

Documents containing no keyphrases (905,451 or 31%) were discarded for the study. In order to ensure the applicability of the results to the collection we tried to find a set of documents as comprehensive as possible

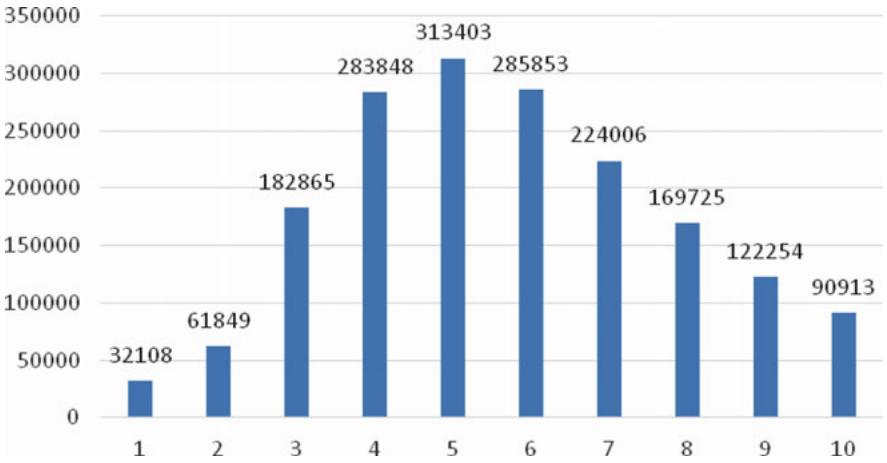


Figure 5. Number of resources in the AGRIS collection for VOA3R by number of9 keyphrases. Range from 1 to 10 keyphrases.

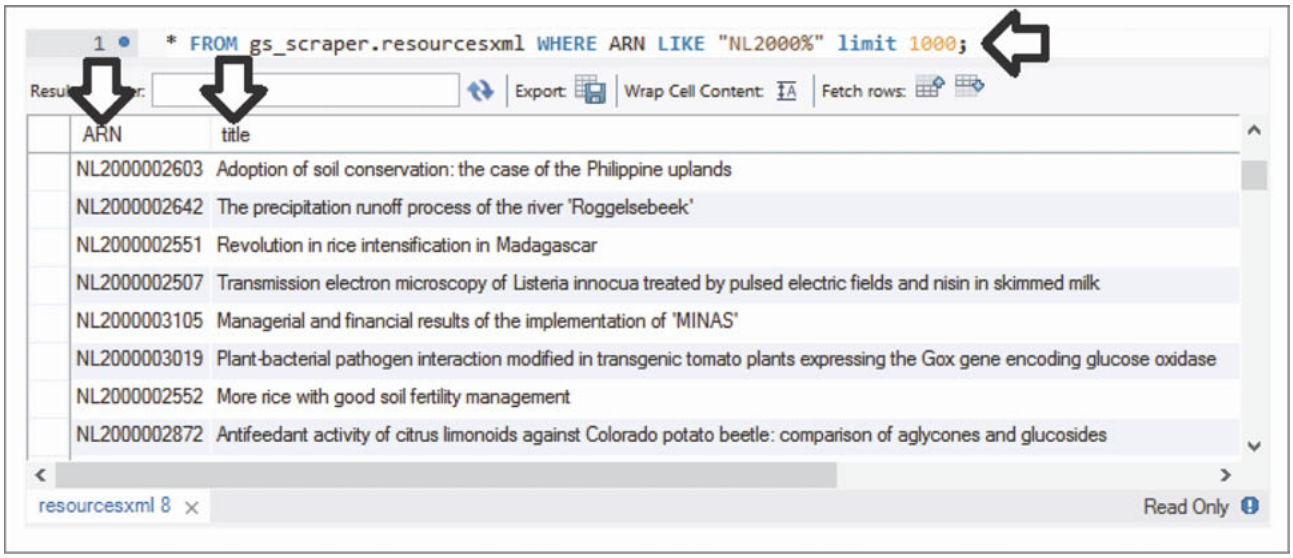


Figure 6. The relational database, showing resources from the Netherlands for the year 2000.

Figure 7. AGRIS Resource Number

with regard not only to the number of originally assigned keyphrases but also to attributes like the year of publication, country or topic. To facilitate this, we also dumped the ARN (AGRIS Resource Number) and title of every resource to a relational database (figure 6). ARN is a unique alphanumeric identifier for AGRIS resources, the structure of which directly provides information about the country the resource comes from, the year of inclusion, the sub-center code, and a serial number, as depicted in figure 7.

Documents were downloaded and revised by hand in order to check the appropriateness of each one, avoiding directly scanned documents (which contain no text, just images; a common problem with old documents), documents containing only an abstract, and other documents with similar problems, thus ensuring the validity of the experiment to the best of our ability. KEA processes TXT documents, so all the downloaded articles (in PDF) were converted to TXT (UTF-8). In this way, we were able to manually select a sample of 2000 full-text documents that were suitable for the experiment, according to the previously exposed criteria. The training corpus contained 1200 documents.

3.3 Measuring KEA's Effectiveness

We assessed KEA's effectiveness using the "precision/recall" criterion espoused by van Rijsbergen (1979), which consists of comparing the algorithm's keyphrase sets with keyphrases assigned manually. The number of matching keyphrases is expressed as a proportion of all extracted keyphrases (*precision "P"*) and of the number of originally (manually) assigned phrases (*recall "R"*) for each document separately.

$$P = \frac{\#correct_{extracted_{keyphrases}}}{\#all_{extracted_{keyphrases}}} \quad R = \frac{\#correct_{extracted_{keyphrases}}}{\#manually_{extracted_{keyphrases}}}$$

For the automatic indexing evaluation and the human indexing evaluation, we define a KEA keyphrase to be correct if it is an exact match for the original, extracted keyphrase. Precision is calculated by dividing the number of correct extracted keyphrases by the total number of extracted keyphrases. Recall is calculated by dividing the number of correct extracted keyphrases by the number of originally assigned ones. Then, a balanced combination of the two is expressed by the *F-measure*, as follows:

$$F - measure = \frac{2PR}{P + R}$$

3.4 Comparing Expert-generated and Automatically Extracted Keyphrases

The second part of the experiment compared KEA results with a sample of keyphrases extracted by human experts and validated using rater agreement. As all the experts invited to participate in the experiment were familiar with VOA3R domains, the documents selected for this part of the experiment were extracted from the same corpus used for the automatic evaluation.

We took into consideration that keyphrase extraction using a controlled vocabulary is a really time consuming task which requires experience and knowledge in the field. Thus we looked for research articles between 6 and 12 pages, manually checking that they were suitable for the experiment, aiming to ensure that all the invited experts would had enough knowledge in the topic to extract keywords, that the articles were complete and that there were no other factors which would affect the evaluation. The final sample of five documents was randomly selected from documents complaining the just mentioned characteristics and checked again to omit direct references to the original keywords.

Six human evaluators were asked to manually assign AGROVOC terms as keyphrases. Their effectiveness was evaluated in the same way as KEA, then the average "human effectiveness" was calculated. It was recommended that the experts extract a minimum of six keyphrases, taking into account that, as shown in figure 4, the mode of originally assigned keyphrases in the VOA3R-AGRIS collection was 5 and the mean was 6.9. No time limit was set for the experts to complete this tasks. They reported an average time of two and a half hours for the whole analysis, keyphrase extraction took from 20 to 40 minutes for each article.

However, the choice of index terms is subject to a degree of subjectivity. By only accepting exact matches—with the original assigned keyphrases—as relevant terms do not provide accurate information to assess their practical applicability, since in some cases, terms semantically close to or with morphological similarity could also be useful from a practical point of view.

To assess this practical applicability, we asked all the subjects to evaluate how well different sets of keyphrases described the five papers they had previously analyzed, ranking them from 1 ("Completely wrong representation") to 5 ("Perfectly well represented"). To prevent bias in the evaluation, subjects in the study did not know how these keyphrases were generated. In this way, evaluators were asked to rate the extracted keyphrases according to how well they represented the document. As a complement to the approach used by Jones and Painter (2011), in our study, each human rater evaluated not only the keyphrases automatically assigned by KEA, but also the ones manually assigned by her colleagues. The blind evaluation process prevented the evaluators from learning how these keyphrases were assigned (whether automatically or by a fellow expert). This kind of blind evaluation has been lacking in previous studies, as noted in the recent study carried out by El-Haj et al. (2013).

4.0 Results and Discussion

4.1 Automatic Indexing Evaluation

As previously described, during the automatic indexing evaluation, KEA generated a number of keyphrases that were compared with the ones manually assigned by human experts. Table 1 shows the results in terms of precision and recall.

The results obtained were similar to those obtained in previous experiments evaluating thesaurus-based automatic indexing, such as those reviewed by Lim et al. (2013). Table 2 shows the average calculated results for precision, recall, and F-measure according to this literature review:

Our averages for these measures are slightly lower, but considering the number of resources and the homogeneity of the collections used in the abovementioned experiments, our results look promising.

4.2 Human Indexing Evaluation

Six experts were asked to extract keyphrases from five selected documents coming from the same corpus used for the automatic evaluation. Table 3 contains the results obtained by these experts, expressed in the same way as the results of the automatic indexing evaluation.

These results show that when humans try to match original keyphrases exactly, their performance average is slightly lower than KEA's. Although two experts (namely C and F) outperformed KEA, the deviations between the best and worst human results cannot be considered significant.

4.3 Manual Expert-oriented Evaluation

Following the procedure detailed in section 3.4, each expert rated the keyphrases extracted by the other five, plus those extracted by KEA, plus the original keyphrases assigned by curators for each paper. Table 4 presents some of Evaluator A's ratings of the keyphrases representing the selected papers. Each evaluator reported her evaluation in a form similar to Table 4.

Table 5 shows the rankings that each evaluator assigned to the KEA-generated keyphrases. Table 6 summarizes the overall results: the frequency of rankings assigned to the keyphrases selected by human evaluators, KEA, and the original indexers.

Precision	**Recall**	**F-measure**
0.18	0.20	0.19

Table 1. Results for the automatic indexing evaluation, in terms of precision and recall using the selected sample of 2000 documents (the training corpus had 1200 documents).

Sources	**Average Precision, Recall and F-measure**		
	Precision	Recall	F-measure
Turney (2002)	0.18	N/A	N/A
Frank et al. (1999)	0.23	N/A	N/A
El-Beltagy et al. (2009)	0.16	0.33	0.21
Kumar and Srinathan (2008)	0.21	0.31	0.25
Sarkar et al. (2010)	0.13	0.44	0.20
Average	0.182	0.36	0.22

Table 2. Average calculated results for the automatic indexing evaluation, in terms of precision, recall, and F-measure. Derived from Table 1. in Lim et al. (2013).

Expert	**Precision**	**Recall**	**F-measure**
A	0.13	0.12	0.12
B	0.16	0.17	0.16
C	0.21	0.22	0.21
D	0.16	0.14	0.14
E	0.14	0.16	0.15
F	0.21	0.23	0.21
Human Average	0.168	0.173	0.171

Table 3. Results for the human indexing evaluation, in terms of precision, recall, and F-measure.

Evaluator A's Ratings					
Items	Completely wrong rep.=1	Not well rep.=2	Partially rep.=3	Well rep.=4	Perfectly rep.=5
Paper B1				×	
Paper B2				×	
Paper B3			×		
Paper B4				×	
Paper B5				×	
Paper C1		×			
(…)	**(…)**	**(…)**	**(…)**	**(…)**	**(…)**

Table 4. First rows of the results reported by Evaluator A.

Results for KEA					
Evaluator	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5
A	4	3	3	4	4
B	4	4	4	4	3
C	4	2	4	4	2
D	3	3	4	4	4
E	4	3	4	4	3
F	4	4	4	4	3

Table 5. Results for the KEA-generated keyphrases—ranked by the six evaluators.

Evaluator	Completely bad rep. (1)	Not well rep. (2)	Partially rep. (3)	Well rep. (4)	Perfectly rep. (5)
A	0	0	9	16	0
B	0	0	5	20	0
C	0	1	3	21	0
D	0	0	3	22	0
E	0	0	6	19	0
F	0	0	1	23	1
KEA	0	2	8	20	0
Original	0	1	5	24	0

Table 6. Overall results of the manual evaluation.

The subjective assessment and reliability of agreement among the subjects was evaluated using the Intraclass Correlation Coefficient (ICC), which represents agreements between two or more raters or evaluation methods on a set of subjects, handling multiple observers and multiple factors with ease. The equivalence of kappa—which has also been used to assess reliability in previous experiments on this topic—and the intraclass correlation coefficient as measures of reliability was described by Fleiss and Cohen (1973). The ICC is adjusted for the effects of the scale of measurements, and as we were working with a five-point Likert scale, we decided to employ it to analyze the agreement, as suggested by Norman (2010). Table 7 shows the analysis of variance of the overall results of the manual evaluation.

With these values, we calculated the ICC=0.965, which indicates a very strong, indeed, almost perfect, agreement.

5.0 Limitations of the Approach

As previously described, we tried to compose an adequate collection of documents suitable for the experiment to the best of our ability, selecting them by hand. For further validation, the approach should be applied in larger sets of documents, which would require the elaboration of tools capable of automatically choosing ade-

Variance	df	Ssq	msq	F	p
B. Rows	4	2502.75	625.68	191.99	<0.0001
W. Rows	35	98.75	2.82		
B. Cols	7	7.5	1.07	0.32	0.8563
Residual err.	28	91.25	3.25		
Total	39	2601.5			

Table 7. Analysis of variance

quate documents, download them and check their validity—i.e. avoid directly scanned-as-image documents, documents containing only an abstract, and other similar problems. All the subjects participating in this experiment were experts in the field. Due to time and resource restrictions—keyphrase extraction using a controlled vocabulary is a really time consuming task which requires experience and knowledge in the field—we only managed to gather information from 5 subjects.

6.0 Conclusions and Outlook

The main objective of this experiment was to analyze the practical applicability of KEA's automatic indexing service in a real-world, practical context. As described, a good number of experiments with KEA have been conducted to date, but most assess effectiveness taking the set of original manually assigned keywords—which may or may not be the "best" ones—as the gold standard. In some cases (El-Haj et al. 2013) this approach is quite problematic, and represents a clear limitation of the experiments. Also, the samples selected for previous experiments have been small subsets of collections, containing documents on similar topics and/or from the same time period. This is not the case for the AGRIS collection in VOA3R, from which the sample in this study was collected.

Considering this, we selected a set of documents from the AGRIS collection in VOA3R to assess KEA's effectiveness according to the precision/recall criterion. We obtained similar results to those of previous experiments, which provides additional evidence of KEA's effectiveness in this kind of collection. Our experiment comparing human performance with KEA when trying to exactly match originally assigned keyphrases showed that KEA "outperformed" humans, even when using a small training corpus. However, this result cannot be used to conclude that KEA is better than humans, since the quality of the original keywords can be questioned. Nonetheless, this finding clearly illustrates that KEA provides similar results to those of humans.

Evaluating exact matches as relevant terms does not provide complete information for purposes of practical applicability, since terms semantically close or with morphological similarity could also be useful from the practical point of view. For this reason, we proposed a new approach to validate the keyphrases on another dimension, asking humans to evaluate keyphrases as to how well they represented a document. The results are promising, as KEA obtained results similar to the human ones, including the original assigned keyphrases.

Our initial investigations with KEA lead us to believe that it has practical applicability as an automatic indexing service for a repository containing highly comprehensive data in the domain that AGROVOC covers. As the choice of index terms is time-consuming for humans, especially when using a large thesaurus such as AGROVOC, our study shows that KEA has practical applicability to supporting indexers' work as a function for recommending terms in the metadata creation process of curation, or even as a quality assurance tool. Finally, we have proposed a blind evaluation, filling a gap in previous research with KEA.

Further work should extend the current pilot study towards a systematic investigation of all the VOA3R collections (not only AGRIS) which includes documents in several different languages. This will entail the development of a system capable to search and download full text documents and then transform them to a format suitable for KEA. In order to increase the number of suitable documents, the system would have to use optical character recognition to allow working with scanned documents, in which the text couldn't be directly extracted.

Automatic evaluation could be used for all the documents containing keywords as part of their metadata. The rest of them should be manually analyzed following the approaches described in this paper, with full detail about the experts work process to extract the keyphrases as the time needed, language proficiency, confidence in their work, as well as other qualitative data; and further information about the document characteristics: number of words, document type, year, quality of the communication, etc.

Finally, we believe that additional efforts should focus on user-centered evaluations employing the approach described in this paper, since the evaluation of KEA's effec-

tiveness appears to be consistent within all the experiments done in the field. Also, as we previously mentioned, the main weakness of the keyphrase indexing approach is that it assumes that there exists a comprehensive domain-specific thesaurus for the target domain. Even this is our case with AGROVOC, it would be very interesting to repeat the experiment and contrast the results employing free indexing algorithms like the ones in the research carried out by Arash and Mahdi (2013).

Notes

1. AGRIS: International Information System for the Agricultural science and technology—Food and Agriculture organization of the United Nations—http://agris.fao.org/content/about
2. FAO: Food and Agriculture organization of the United Nations—http://www.fao.org/
3. AGROVOC: http://aims.fao.org/standards/agrovoc

References

Al-Hashemi, Rafeeq. 2010. "Text Summarization Extraction System (TSES) Using Extracted Keywords." *International Arab Journal of e-Technologies* 1, no. 4: 164-8.

Anibaldi, Stefano, Yves Jaques, Fabrizio Celli, Armando Stellato and Johannes Keizer. 2013. "Migrating Bibliographic Datasets to the Semantic Web: The AGRIS Case." In *Proceedings of Semantic Web EFITA conference, 23 -27 June 2013, Torino, Italy.*

Chen, Shu-jiun and Hsueh-hua Chen. 2012. "Mapping Multilingual Lexical Semantics for Knowledge Organization Systems." *The Electronic Library* 30, no. 2: 278-94.

Dahlberg, Ingetraut. 2006. "Knowledge Organization: A New Science?" *Knowledge Organization* 33: 11-9.

El-Beltagy, Samhaa R. 2006. "KP-Miner: A Simple System for Effective Keyphrase Extraction". In *Proceedings of the Innovations in Information Technology, November 2006 Dubai, Dubai,* 1-5.

El-Beltagy, Samhaa R. and Ahmed Rafea. 2009. "KP-Miner: A Keyphrase Extraction System for English and Arabic Documents." *Information Systems* 34, no. 1: 132-44.

El-Haj, Mahmoud, Lorna Balkan, Suzanne Barbalet, Lucy Bell and John Shepherdson. 2013. "An Experiment in Automatic Indexing Using the HASSET Thesaurus." In *Proceedings of the Computer Science and Electronic Engineering Conference (CEEC), 17-18 November 2013, Colchester, UK,* 13-18.

Erbs, Nicolai, Iryna Gurevych and Marc Rittberger. 2013. "Bringing Order to Digital Libraries: From Keyphrase Extraction to Index Term Assignment." *D-Lib Magazine* 19, no. 3.

Frank, Eibe, Gordon W. Paynter, Ian H. Witten, Carl Gutwin and Craig G. Nevill-Manning. 1999. "Domain-Specific Keyphrase Extraction." In *Proceedings of 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, July 31-August 6 1999*. San Francisco, USA: Morgan Kaufmann Publishers, 668-73.

Frantzi, Katerina, Sophia Ananiadou and Hideki Mima. 2000. "Automatic Recognition of Multi-Word Terms: The C-Value/NC-Value Method." *International Journal on Digital Libraries* 3, no. 2: 115-30.

Fleiss, Joseph L. and Jacob Cohen. 1973. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." *Educational and Psychological Measurement* 33: 613-9.

Goovaerts, Marc. 2011. "Virtual Open Access Agriculture & Aquaculture Repository: Sharing Scientific and Scholarly Research Related to Agriculture, Aquaculture & Environment." In *Proceedings of the European Association of Aquatic Science Libraries and Information Centres, 18 May 2011, Lyon, France,* 58-9.

Harzing, Anne-Wil and Ron Van der Wal. 2007. "Google Scholar: The Democratization of Citation Analysis." *Ethics in Science and Environmental Politics* 8, no. 1: 61-73.

Hjørland, Birger. 1998. "Information Retrieval, Text Composition, and Semantics." *Knowledge organization* 25, no. 1: 16-31.

Hjørland, Birger. 2001. "Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content... and Relevance." *Journal of the American Society for Information Science and Technology* 52, no. 9: 774-8.

Hjørland, Birger. 2003. "Fundamentals of Knowledge Organization." *Knowledge Organization* 30, no. 2: 87-111.

Hjørland, Birger. 2011. "The Importance of Theories of Knowledge: Indexing and Information Retrieval as an Example." *Journal of the American Society for Information Science and Technology* 62, no. 1: 72-7.

Jones, Steve and Gordon W. Paynter. 2001. "Human Evaluation of Kea, an Automatic Keyphrasing System." In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, 24-28 June 2001, Roanoke, VA, USA,* 148-56.

Jones, Steve and Gordon W. Paynter. 2002. "Automatic Extraction of Document Keyphrases for Use in Digital Libraries: Evaluation and Applications." *Journal of the American Society for Information Science and Technology* 53, no. 8: 653-77.

Joorabchi, Arash and Abdulhussain E. Mahdi. 2013. "Automatic Keyphrase Annotation of Scientific Documents Using Wikipedia and Genetic Algorithms." *Journal of Information Science* 39, no. 3:410-26.

Kumar, Niraj and Kannan Srinathan. 2008. "Automatic Keyphrase Extraction from Scientific Documents Using N-Gram Filtration Technique." In *Proceedings of the*

ACM Symposium on Document Engineering, 16-19 September 2008, Sao Paulo, Brasil, 199-208.

Lim, Vicky M. H., Siew F. Wong and Tong M. Lim. 2013. "Automatic Keyphrase Extraction Techniques: A Review." In *Proceedings of the IEEE Symposium on Computers & Informatics, 7-9 April 2013, Langkawi, Malaysia,* 196-200.

Lord, Philip, Alison Macdonald, Lyz Lyon and David Giaretta. 2004. "From Data Deluge to Data Curation." In *Proceedings of the UK E-Science All Hands Meeting, 31 August-1 September 2004, Nottingham, UK,* 371-57.

Medelyan, Olena and Ian H. Witten. 2005. "Thesaurus-Based Index Term Extraction for Agricultural Documents." In *Proceedings of the EFITA/WCCA Joint Congress on IT in Agriculture, 25-28 July 2005, Vila Real, Portugal*, 1122-9.

Medelyan, Olena, Ian H. Witten and David Milne. 2008. "Topic Indexing with Wikipedia." In *Proceeding of the Twenty-Third AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, July 13–14, Chicago, Illinois, USA,* 19-24.

Medelyan, Olena, Eibe Frank and Ian H. Witten. 2009. "Human-Competitive Tagging Using Automatic Keyphrase Extraction." In *Proceedings of the International Conference of Empirical Methods in Natural Language Processing, 6-7 August 2009, Singapore,* 1318-27.

Medelyan, Olena, David Milne, Catherine Legg and Ian H. Witten. 2009b. "Mining meaning from Wikipedia." *International Journal of Human-Computer Studies* 67, no. 9: 716-54.

Milne, David, Medelyan Olena and Ian H. Witten. 2006. "Mining Domain-Specific Thesauri from Wikipedia: A Case Study." In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence, 18-22 September 2006, Hong-Kong, China,* 442-8.

Névéol, Aurélie, Sonya E. Shooshan, Susanne M. Humphrey, Thomas. C. Rindflesch and Alan R. Aronson. 2007. "Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature." In *Proceedings of the Pacific Symposium on Biocomputing, vol. 12, 3-7 January 2007, Maui, Hawaii, USA,* 292-303.

Nguyen, Thuy D. and Min-Yen Kan. 2007. "Keyphrase Extraction in Scientific Publications." *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*. Heidelberg, Berlin: Springer, 317-26.

Norman, Geoff. 2010. "Likert Scales, Levels of Measurement and the "Laws" of Statistics." *Advances in Health Sciences Education* 15, no. 5: 625-32.

Park, Carissa A., et al. 2013. "The Vertebrate Trait Ontology: A Controlled Vocabulary for the Annotation of Trait Data Across Species." *Journal of Biomedical Semantics* 4:13.

Pouliquen, Bruno, Ralf Steinberger and Camelia Ignat. 2006. "Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus." In *Proceedings of the EUROLAN Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology-Its Potential and Practicalities. 28 July-8 August 2003, Bucharest, Romania.*

Ruiz, Miguel E. and Alan R. Aronson. 2007. *User-Centered Evaluation of the Medical Text Indexing (MTI) System.* Bethesda, MD: National Library of Medicine. http://ii.nlm.nih.gov/resources/MTIEvaluation-Final.pdf.

Salton, Gerard, Anita Wong and Chung-Shu Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18, no. 11: 613-20.

Sarkar, Kamal, Mita Nasipuri and Suranjan Ghose. 2010. "A New Approach to Keyphrase Extraction Using Neural Networks." *International Journal of Computer Science Issues* 7, no. 2: 16-25.

Turney, Peter D. 1999. *Learning to Extract Keyphrases from Text.* National Research Council of Canada, Institute for Information Technology, Technical Report ERB-1057. http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8913245.

Turney, Peter D. 2000. "Learning Algorithms for Keyphrase Extraction." *Information Retrieval* 2, no. 4: 303-36.

Šimek, Pavel, J. Vaněk, J. Jarolímek, M. Stočes and T. Vogeltanzová. 2013. "Using Metadata Formats and AGROVOC Vocabulary for Data Description in the Agrarian Sector." *Plant, Soil and Environment* 59, no. 8: 378-84.

Sini, Margherita, Boris Lauser, Gauri Salokhe, Johannes Keizer and Stephen Keizer. 2008. "The AGROVOC Concept Server: Rationale, Goals and Usage." *Library Review* 57, no. 3: 200-12.

Van Rijsbergen, Cornelis J. 1979. *Information Retrieval. 2nd ed.* London: Butterworths.

Witten, Ian H., Gordon W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning. 1999. "KEA: Practical Automatic Keyphrase Extraction." In *Proceedings of the fourth ACM Conference on Digital Libraries, 11-14 August 1999, Berkeley, California, USA,* 254-5.

Zeng, Marcia Lei and Lois Mai Chan. 2004. "Trends and Issues in Establishing Interoperability among Knowledge Organization Systems." *Journal of the American Society for Information Science and Technology* 55, no. 5: 377-95.