Metadata of the chapter that will be visualized in SpringerLink

Book Title	Research Challenges in Information Science: Information Science and Global Crisis	
Series Title		
Chapter Title	KEOPS: Knowledge ExtractOr Pipeline System	
Copyright Year	2021	
Copyright HolderName	Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Martin**
	Particle	
	Given Name	**Pierre**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	CIRAD, UPR AIDA
	Address	34398, Montpellier, France
	Division	
	Organization	AIDA, Univ Montpellier, CIRAD
	Address	Montpellier, France
	Email	pierre.martin@cirad.fr
	ORCID	http://orcid.org/0000-0002-4874-5795
Author	Family Name	**Helmer**
	Particle	
	Given Name	**Thierry**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	CIRAD,DSI
	Address	34398, Montpellier, France
	Email	thierry.helmer@cirad.fr
Author	Family Name	**Rabatel**
	Particle	
	Given Name	**Julien**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	CIRAD,DSI
	Address	34398, Montpellier, France
	Email	jrabatel@gmail.com
	ORCID	http://orcid.org/0000-0002-4742-923X

Author	Family Name	**Roche**
	Particle	
	Given Name	**Mathieu**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	CIRAD,UMR TETIS
	Address	34398, Montpellier, France
	Division	
	Organization	TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE
	Address	Montpellier, France
	Email	mathieu.roche@cirad.fr
	ORCID	http://orcid.org/0000-0003-3272-8568

Abstract	The KEOPS platform applies text mining approaches (e.g. classification, terminology and named entity extraction) to generate knowledge about each text and group of texts extracted from documents, web pages, or database. KEOPS is currently implemented on real data of a project dedicated to Food security, for which preliminary results are presented.

Keywords (separated by '-')	Knowledge management system - Text mining

KEOPS: Knowledge ExtractOr Pipeline System

Pierre Martin[1,2(✉)], Thierry Helmer[3], Julien Rabatel[3], and Mathieu Roche[4,5]

[1] CIRAD, UPR AIDA, 34398 Montpellier, France
pierre.martin@cirad.fr
[2] AIDA, Univ Montpellier, CIRAD, Montpellier, France
[3] CIRAD,DSI, 34398 Montpellier, France
thierry.helmer@cirad.fr
[4] CIRAD,UMR TETIS, 34398 Montpellier, France
mathieu.roche@cirad.fr
[5] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France

Abstract. The KEOPS platform applies text mining approaches (e.g. classification, terminology and named entity extraction) to generate knowledge about each text and group of texts extracted from documents, web pages, or database. KEOPS is currently implemented on real data of a project dedicated to Food security, for which preliminary results are presented.

Keywords: Knowledge management system · Text mining

1 Introduction

Many tools and platforms are available for exploring textual data and highlighting new knowledge. TyDI (Terminology Design Interface) [5] is a collaborative platform for manual validation and structuring of terms from existing terminologies or terms extracted automatically using dedicated tools [1]. Other tools like NooJ [8] use linguistic approaches to build and manage dictionaries and grammars. NooJ integrates several NLP (Natural Language Processing) methods like named entity recognition approaches. Other platforms integrate text mining components like CorTexT [2]. CorTexT allows the extraction of named entities and advanced text mining approaches (e.g. topic modeling, word embedding, etc.) are integrated in this platform. Finally some platforms such as UNITEX rely on dictionaries and grammars [6].

KEOPS (Knowledge ExtractOr Pipeline System) is a platform that contains various indexing and classification methods to be applied to the texts that come from databases, documents, or web pages. As output, KEOPS combines classification and indexing results to generates knowledge about each text and group

Supported by Leap4FNSSA H2020 project, SFS-33-2018, grant agreement 817663.

S. Cherfi et al. (Eds.): RCIS 2021, LNBIP 415, pp. 1–7, 2021.
https://doi.org/10.1007/978-3-030-75018-3_36

of texts. KEOPS is currently implemented in different projects dedicated to agriculture domain, e.g. LEAP4FNSSA[1].

This paper introduces the four-step process of the KEOPS plateform in Sect. 2 and presents first results of KEOPS applied to LEAP4FNSSA documents in Sect. 3. Section 4 concludes this paper.

2 The KEOPS Platform

The KEOPS platform is made up of autonomous modules linked together to perform the comprehensive processing of the text, extract knowledge, and make it visible to users. KEOPS process is based on 4 main phases presented in Fig. 1 and summarized in the following subsections.

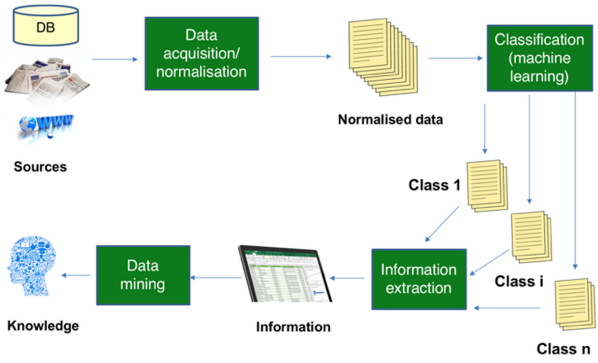

Fig. 1. The KEOPS process.

2.1 Step 1: Data Acquisition

The textual data is collected using three successive tasks. The first one consists in importing a document or a set of documents (i.e. txt, doc, html, pdf). In the case of a website address, the website is copied up to the crawling depth indicated by the user. The second task transforms the imported documents into normalized plain text (deleting images, menus, etc.). The third one identifies the language used in each text.

2.2 Step 2: Document Classification

This step aims to classify the documents according to classes predefined by the user (e.g. Food production, Processing, Distribution, and Consumption). The adopted process is based on a supervised learning method (machine learning - See Sect. 3.1). To proceed, some documents are associated with the predefined

[1] https://www.leap4fnssa.eu/.

classes and a model is learned from this training data. For each new text, the learned model predicts the class to be assigned. This classification is based on the premise that "if documents have a common vocabulary then they can be grouped into common classes (themes)". It should be noted that the texts are first represented in vector form (i.e. bag-of-word representation).

2.3 Step 3: Indexing and Information Extraction

This step aims to extract information from documents through the semantic features and to position a tag next to each occurrence of indexed terms (e.g. keywords, named entities, etc.) in the document (see Fig. 3) according to the following methods and resources:

- Terms of the thesaurus Agrovoc[2] dedicated to agriculture;
- Keywords provided by the users (identified during specific workshops [7]);
- Expert keywords provided by reference databases;
- A terminology (thematic entities) acquired automatically (using BioTex[3]);
- A set of named entities (people, places, organizations, etc.) extracted by SpaCy[4].

Note that the spatial information identified using SpaCy is managed through Gazetteers (e.g. Geonames).

The extracted information, used by step 4, can also be useful for the document indexing task as follow:

- Examples of generic information: location, organization, etc.
- Examples of information related to a Food security domain: water management, food security, crops, etc.

2.4 Step 4: Extraction and Visualization of Knowledge

Using the information obtained in the previous steps (i.e. types of documents in step 1, classes in step 2, and indexed terms in step 3), some results based on data-mining and visualisation algorithms are proposed. The extracted knowledge is then aggregated and presented as maps, graphs, curves, and Venn diagrams in order to enable their validation by experts. An example of output is presented in Fig. 2.

[2] http://www.fao.org/agrovoc/.
[3] http://tubo.lirmm.fr/biotex.
[4] https://spacy.io/.

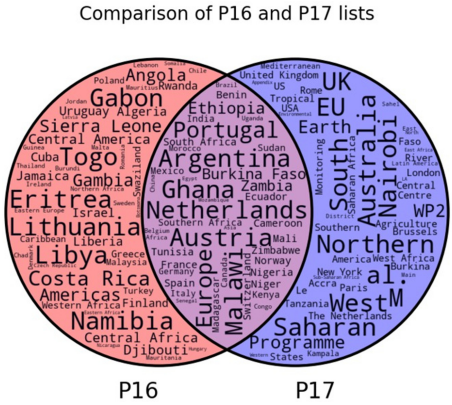

Fig. 2. Example of Venn diagram result from a subset of projects explored by KEOPS. This representation compares spatial meta-data related to the projects (called P16) and locations extracted with SpaCy in the description of the projects (P17). For instance, this visualisation highlights specific locations (e.g. Nairobi, Australia) cited in the contents of the documents but not as meta-data.

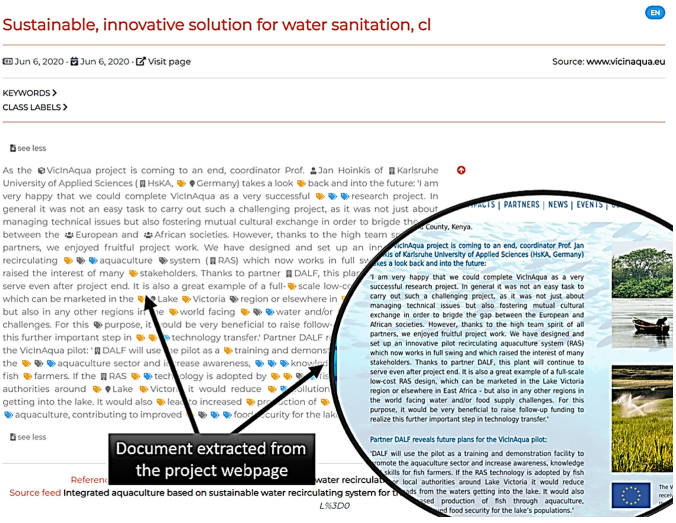

Fig. 3. KEOPS screenshot: example of tags on a LEAP4FNSSA project document. An orange, black, or blue label corresponds respectively to an Agrovoc concept, a user term, or a term extracted using BioTex or Spacy.

3 Case Study

Applied to the LEAP4FNSSA project, the objective of KEOPS is to enable the analysis of the corpus of the projects described in the FNSSA database[5],

[5] https://library.wur.nl/WebQuery/leap4fnssa-projects/.

using associated documents (e.g. report, publication, etc.). The current corpus includes 208 project descriptions, 1227 documents, and 156 website references. The analysis output, based on the documents, can be provided at the project level.

3.1 Classification

In this case study, two classification levels of the documents are considered, i.e. document type and thematic. Classes are identified through an iterative evaluation process taking into account the corpus content.

For the document type level, 8 classes were initially identified during a workshop [7], i.e. Case studies, Facts sheet, Policy brief, Presentation/news/poster, Project report, Publication, Thesis, and Workshop report. Their evaluation on 386 documents, obtained using the multi-layer perceptron classifier[6], provided a best score of 51.75% (average accuracy in a 5-fold cross validation protocol). Merging some of them to get 3 classes (i.e. Information document, Report, Publication) then provided a best score of average accuracy of 82.29% using the Random Forest Classifier (with 50 trees with maximal depth 15). The average precision, recall, and f1-score for this model respectively were 0.80, 0.79, and 0.80.

For the thematic level, 4 classes describing agri-food system were initially suggested, i.e. Food production, Processing, Distribution, and Consumption. An initial evaluation of these classes on 15 documents during a workshop, conducted by 10 experts, showed the need to group 3 classes (i.e. Processing, Distribution, and Consumption), and to add an additional one, i.e. Health. This new set of classes is under testing.

3.2 Indexing

Terminology of KEOPS for indexing is extracted using generic parameters of the BioTex tool [4]. BioTex uses both statistical and linguistic information to extract terminology from free texts. Candidate terms are first selected if they follow defined syntactic patterns (e.g. adjective-noun, noun-noun, noun-preposition-noun, etc.). After such linguistic filtering, a statistical criterion is applied. This measures the association between the words composing a term by using a measure called C-value [3] and by integrating a weighting (i.e. TF-IDF - Term Frequency - Inverse Document Frequency). The goal of C-value is to improve the extraction of multi-word terms while the TF-IDF weighting highlights the discriminating aspect of the candidate term. In our experiments with the LEAP4FNSSA corpus, a terminology is extracted by applying 5 strategies:

- **M1**: Frequent Agrovoc keywords;
- **M2**: Words and multi-word terms extracted with disciminative criteria (i.e. F-TFIDF-C) of BioTex;

[6] The multi-layer perceptron uses the default configuration available in the scikit-learn library.

- **M3**: Multi-word terms extracted with disciminative criteria (i.e. F-TFIDF-C) of BioTex;
- **M4**: Words and multi-word terms extracted with C-Value;
- **M5**: Multi-word terms and multi-word terms extracted with C-Value.

In this context, 10 participants of the project analyzed the first terms according to these strategies. The results (see Table 1) highlight that multi-word terms extracted with discriminative measures (i.e. M3) are more relevant for indexing tasks.

Table 1. Quality of terms evaluated by 10 participants of the LEAP4FNSSA project.

	Very relevant	Relev. but too general	Not really relevant	Irrelevant	I don't know
M1	21 (19.0%)	53 (48.1%)	21 (19.0%)	12 (10.9%)	3 (2.7%)
M2	34 (33.0%)	42 (40.7%)	15 (14.5%)	12 (11.6%)	0 (0.0%)
M3	73 (73.0%)	17 (17.0%)	3 (3.0%)	3 (3.0%)	4 (4.0%)
M4	24 (22.6%)	56 (52.8%)	11 (10.3%)	15 (14.1%)	0 (0.0%)
M5	64 (64.0%)	19 (19.0%)	7 (7.0%)	8 (8.0%)	2 (2.0%)

4 Conclusion and Future Work

The classification and indexing steps of the KEOPS process have been implemented and evaluated on real data from the LEAP4FNSSA project. KEOPS is continuously developed in order to meet users' objectives. Within the framework of other European projects, multi-language processing is being integrated.

Acknowledgement. We thank the WP3 members of the LEAP4FNSSA project for their contribution to the indexing and classification tasks. We thank Xavier Rouviere for his contribution to the development of the user interface.

References

1. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) FinTAL 2006. LNCS (LNAI), vol. 4139, pp. 380–387. Springer, Heidelberg (2006). https://doi.org/10.1007/11816508_39
2. Barbier, M., Cointet, J.P.: Reconstruction of socio-semantic dynamics in sciences-society networks: Methodology and epistemology of large textual corpora analysis. Science and Democracy Network, Annual Meeting (2012)
3. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. Int. J. Digital Libraries **3**(2), 115–130 (2000)

4. Lossio-Ventura, J., Jonquet, C., Roche, M., Teisseire, M.: Biomedical term extraction: overview and a new methodology. Inf. Retr. J. **19**(1–2), 59–99 (2016)
5. Nedellec, C., Golik, W., Aubin, S., Bossy, R.: Building large lexicalized ontologies from text: a use case in automatic indexing of biotechnology patents. In: Proceedings of EKAW, pp. 514–523 (2010)
6. Paumier, S.: Unitex - Manuel d'utilisation, November 2011. https://hal.archives-ouvertes.fr/hal-00639621, working paper or preprint
7. Roche, M., et al.: LEAP4FNSSA (WP3 - KMS): Terminology for KEOPS - Dataverse (2020). http://doi.org/10.18167/DVN1/GQ8DPL
8. Silberztein, M.: La formalisation des langues : l'approche de NooJ. ISTE, London (2015)

Author Queries

Chapter 36

Query Refs.	Details Required	Author's response
AQ1	Please note that as per Springer guideline private email address of the author will not be displayed.	
AQ2	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	