



A taxonomy generation tool for semantic visual analysis of large corpus of documents

Belen Carrion¹ · Teresa Onorati¹  · Paloma Díaz¹ · Vasiliki Triga²

Received: 13 December 2018 / Revised: 17 April 2019 / Accepted: 10 June 2019 /

Published online: 1 July 2019

© The Author(s) 2019

Abstract

Taxonomies are semantic resources that help to categorize and add meaning to data. In a hyperconnected world where information is generated at a rate that exceeds human capacities to process and make sense of it, such semantic resources can help to access relevant information more efficiently by extracting knowledge from large and unstructured data sets. Taxonomies are related to specific domains of knowledge in which they identify relevant topics. However, they have to be validated by experts to guarantee that its terms and relations are meaningful. In this paper, we introduce a semiautomatic taxonomy generation tool for supporting domain experts in building taxonomies that are then used to automatically create semantic visualizations of data. Our proposal combines automatic techniques to extract, sort and categorize terms, and empowers domain experts to take part at any stage of the process by providing a visual edition tool. We tested the tool's usability in two use cases from different domains and languages. Results show that all the functionalities are easy to use and interact with. Lessons learned from this experience will guide the design of a utility evaluation involving domain experts interested in data analysis and knowledge modeling.

Keywords Knowledge modelling · Semantic visualization · Taxonomy development process · Big data

✉ Teresa Onorati
tonorati@inf.uc3m.es

Belen Carrion
bcarrion@pa.uc3m.es

Paloma Díaz
pdp@inf.uc3m.es

Vasiliki Triga
vasiliki.triga@cut.ac.cy

Extended author information available on the last page of the article

1 Introduction

Living in the so-called *Information Age*, we find ourselves constantly surrounded by ever-growing volumes of information [6]. Every second, large volumes of information are available online in platforms like news channels, social networks or blogs, making it rather time consuming to access, collect and process.

Such platforms are a great source of public information that is generated by a wide range of heterogenous sources. Being able to monitor this information might be useful in different scenarios, including political elections, emergency response, and marketing. For instance, monitoring political elections, it could reveal the voters' opinions on the candidates and parties and provide more accurate predictions on electoral outcomes [3, 11, 24, 42]. It would be also useful during emergency situations in which the users often turn to social networks to gather and share information about a crisis [22, 29]. Another relevant scenario is monitoring social media for marketing and human resources purposes, which could augment our understanding of public opinion regarding products or services [14].

These scenarios require new ways to process and visualize information published online in order to extract useful knowledge that could assist users in understanding the situation and taking informed decisions. In previous research [29, 30], we have proposed the use of semantic analysis to identify and categorize concepts in a specific domain of interest with a view to giving meaning to unstructured data. This approach is based on modeling the domain knowledge using structures like ontologies or taxonomies. These semantic models, which contain representative terms for the considered domain [37], are used in our research to identify which terms extracted from large volume of data are semantically relevant in a specific domain. From the semantic taxonomies, semantic visualizations are generated where data are organized according to their meaning and semantics. The quality of domain specific semantic resources will determine the utility of the visualization, since semantic categories need to be close to the user understanding of the domain.

The domain-specific nature of taxonomies makes it necessary to build new ones whenever the domain of interest changes [7] and involve the domain experts to know which concepts have to be included. As a result, building these knowledge models becomes a very time-consuming and resource-demanding task. For this reason, there is an increasing interest in proposing semiautomatic approaches to create domain-specific taxonomies, where part of the process is automatized but the domain experts continue to play a crucial validation role.

In this paper, we propose a semiautomatic tool for building domain-oriented taxonomies from a corpus of documents chosen by the experts from any source of information (e.g. blogs, social network messages, news channels) in multiple languages (e.g. English and Spanish). We combine automatic techniques for extracting, sorting, and categorizing concepts based on their semantics, with manual filter mechanisms that the experts can apply to review, adapt and change the resulting taxonomy. The usability and the applicability of the tool have been evaluated in two case studies, varying both the domain and the language: the politics domain in Spanish and biotechnology domain in English. In the first case study, the tool generated a taxonomy from a collection of documents in Spanish about the 2017 Catalan constitutional crisis. The unilateral independence referendum of Catalonia in October 2017 was an extremely polarising political event in Spain. As we know from the literature, staging such contested sovereignty referendums can rapidly spiral into conflict escalation and outbreaks of violence, as was the case in Catalonia or generate military conflict as was the case in Chechnya. Since these political events can polarise society and generate competing political discourses, many

stakeholders, including journalists, politicians, political scientists or citizens need to be able to analyse all the available information. Since this option is not affordable giving the quantity of data generated, the use of a semantic taxonomy as the one we propose here might help to focus the analysis on specific issues and separate the wheat from chaff. The second case study focuses on a different application of taxonomies as interactive dictionaries to learn more about a subject. In particular we used the tool to generate a taxonomy about New Plant Breeding Techniques (i.e. NPBTs) in order to transfer knowledge on this scientific and controversial topic to the general public. In this context, it is crucial to offer an engaging experience to foster the curiosity and the interest of the users. For this reason, the taxonomy generation tool offers both a textual and a visual representation. A visualization can offer a more intuitive way to explore the concepts included in the model and how they are semantically grouped into categories. Whilst in the first use case the corpus is built in Spanish, the second one deals with documents and data in English.

The structure of the paper follows with an overview on the literature about development tools for taxonomies, focusing in particular on the semiautomatic approaches (Section 2). Next, we introduce our proposal for generating taxonomies (Section 3) and we present the results obtained from the usability evaluation (Section 4). Finally, some conclusions and future works are drawn (Section 5).

2 The development of taxonomies

As the volume of information published on online platforms (i.e. blogs, news channels, social networks, to name a few) keeps growing, the need to find ways to take advantage of these unstructured data in order to make sense of them and exploit their value becomes more important [2]. Mining techniques, like natural language processing, information retrieval, data mining, or machine learning, are applied for organizing, navigating, retrieving, and summarizing content shared on online platforms [2, 13, 36]. These techniques can be applied and combined to analyze different aspects of a large dataset. One of them is the *semantics* to identify and extract the most meaningful information and build a sort of dictionary with representative concepts related to a specific subject of interest. An approach to semantically analyze a dataset is based on modeling the domain using structures like taxonomies or ontologies. These knowledge models contain a set of domain-relevant concepts and relations that can be used to categorize and filter the most relevant information in the dataset [16, 17, 37].

The key challenge of using knowledge models lies in their development, in particular in the extraction of the right concepts and relations from unstructured datasets and in measuring their relevance in the domain. Zilli has identified three main approaches to deal with this challenge [47]: *manual*, *semiautomatic*, and *automatic*. The *manual* techniques involve domain experts to organize and structure the knowledge based on their expertise. They can start from scratch filtering and structuring manually the most relevant concepts from the data, or reuse and adapt existing taxonomies about similar topics. *Semiautomatic* methodologies reduce the human effort required by manual techniques and involve the experts in limited scope actions, like collecting the data to analyze, or refining the selected categories. Finally, the *automatic* taxonomy generation avoids the participation of domain experts and employs a top-down approach. Starting from clustering the dataset depending on its specificity or generality, these tools focus on identifying an initial set of labels as possible candidate nodes for the taxonomy,

and successively filtering them applying different criteria (e.g. term frequency, document frequency, or Tf-Idf).

In this work, considering that our aim is to develop taxonomies that could be used for a wide variety of applications (e.g. social media mining, interactive dictionaries, or knowledge exploration), our focus is on developing a semiautomatic taxonomy generation tool. In the literature, it is possible to find several examples of this kind of tools, where end users play a crucial role in the development process contributing with their expertise in the domain of interest. The semiautomatic tools combine the advantages of automatizing part of the process of building taxonomies, with the active contribution of the domain experts interested in designing models that could suit their needs. Consequently, the end users are involved in different phases of the development process mainly to improve the quality of the generated taxonomies. Kolterman et al. propose a method for building taxonomies based on the Wikipedia Category Hierarchy with an interactive tool to support the users in the post-editing [19]. In this way, they can remove unnecessary categories, modify their content or add new ones, improving the quality of the initial taxonomy.

Similar to Kolterman's contribution, in [45] the author proposes several techniques to model the document collection based on the tasks performed by the users. In this case, the users not only contribute to the improvement of the generated taxonomies, but they are also a source of information for the algorithms to learn about their preferences to organize concepts. Another way to involve the experts is during the term extraction to include not considered entities from other sources or taxonomies before starting the categorization [12].

Focusing on the considered sources for extracting relevant knowledge for the taxonomy, end users can also contribute indirectly as Tsui et al. propose [9], with a tool that generates a hierarchical structure of terms based on a folksonomy, built real-time by people collaborating on the categorization of the same documents using tags.

To the best of our knowledge, there is a lack of tools where the end users are involved as active contributors in the whole knowledge modelling loop. Mainly, the research works in this direction limit the human participation to a specific phase or a set of actions within the development process. Our proposal offers end users the possibility to explore the results obtained from each phase (i.e. from the collection of the document corpus to the selection of the term candidates and the categorization) and contribute with their expertise. To support their involvement and to help them in understanding the defined categorization, the tool provides both a textual and a graphical representation of the defined taxonomy as described below.

3 The taxonomy generation tool

As introduced in [46], the taxonomy generation tool has been designed as a 4-step process (see Fig. 1): *document corpus*, *term extraction*, *term categorization* and *taxonomy visualization*. Each one of these steps combines different well-known techniques in the area of data mining, taking into account that the final result has to work for any domain and in multiple languages. The first step of the architecture (see *step 1 – Document corpus* in Fig. 1) consists of collecting a set of documents (i.e. corpus) from different sources, like domain-specific material (e.g. handbooks, brochures, blog posts, or news articles), or social network messages. The corpus is analyzed in the second step (see *step 2 – Term extraction* in Fig. 1) to extract the most relevant terms using parsing techniques. For each term N , the relevance is determined based on two

measures: the frequency F_N (i.e. how many times a term occurs in the corpus) and the domain consensus or weight W_N (i.e. whether a term is highly frequent in each document of the corpus). In the third step (see *step 3 – Term Categorization* in Fig. 1), the terms are grouped in categories using several semantic relations, like synonyms, antonyms, co-occurrences, and multiword terms, generating the domain-specific taxonomy. Finally, the taxonomy is visually represented to provide users with an alternative way to explore it (see *step 4 – Taxonomy visualization* in Fig. 1).

The architecture in Fig. 1 combines the advantages of an automatic process for extracting and categorizing knowledge with the experience of the domain experts. The proposed semiautomatic approach consists of automatizing the first three steps of the process to reduce the effort of the experts while still allowing them to get involved in the upload of the document corpus, the selection of the most meaningful terms and the categorization of the extracted terms. Involving the domain experts in this kind of applications has shown to provide major control and a higher precision over the definition of concepts and relations among them [37].

The taxonomy generation tool has been implemented as a web application to guarantee its portability to different operating systems and devices. More details about each step of the architecture are explained in the next subsections.

3.1 Step 1 – Document corpus

The first step of the architecture aims at gathering a set of documents, called corpus, that the experts consider significant for their domain of expertise. These documents are going to be used for modeling the taxonomy based on the extraction of meaningful concepts for the considered domain.

The experts can use a great variety of sources, including handbooks, brochures, blog posts, social network messages, news articles and magazines, and upload them to the web application as pdfs or web addresses (see options *File* and *URL* in Fig. 2). Moreover, they can keep track of the imported documents, checking their content and eventually removing them from the corpus (see options *preview* and *delete* for each item of the file list in Fig. 2).

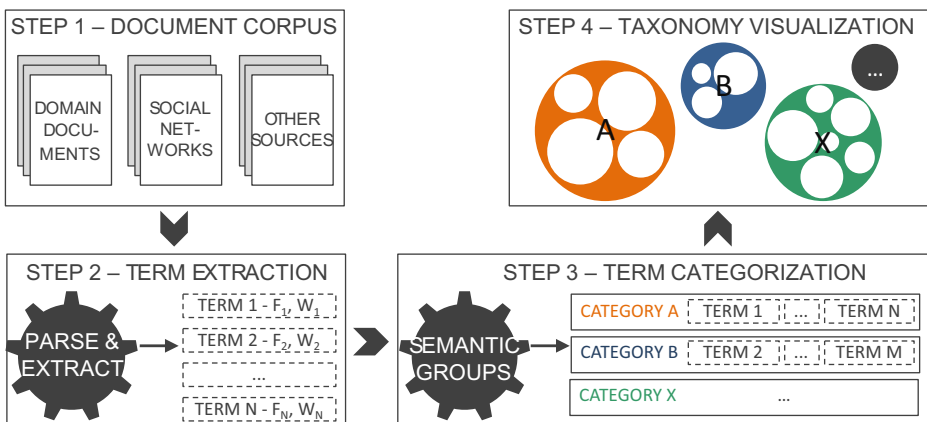


Fig. 1 The architecture for the Taxonomy Generation Tool

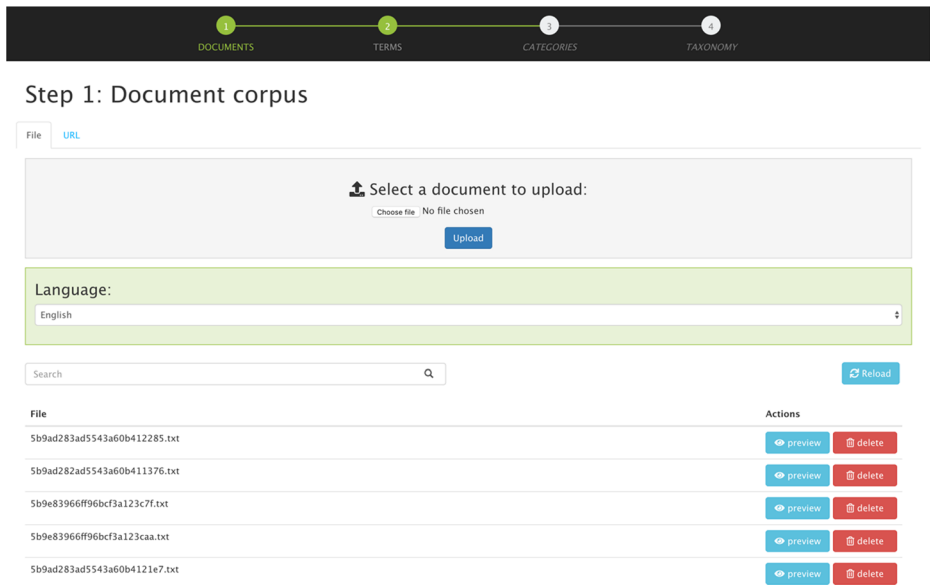


Fig. 2 Document corpus (step 1): the users can upload a document from a local file or a web address, list the imported documents, preview or delete them

The cost in terms of time of collecting a corpus depends on the number of documents that the experts want to include. To speed up this step, the architecture also integrates an automatic tool called BootCat [1] to gather reference texts about a topic, running ad-hoc queries in a search engine.

3.2 Step 2 – Term extraction

Once the corpus has been uploaded, the user can move to the second step of the architecture, organized into two phases (see Fig. 3): the parsing, and the relevance.

3.2.1 Phase 1: Parsing

Parsing consists of extracting the domain lexicon from the corpus. A domain lexicon is defined as a list of semantically relevant terms for a given community of interest [41]. Considering that the collected corpus could contain a great variety of unstructured data, a preprocessing phase is crucial to prepare the texts to improve the quality of further analysis [4].

The parsing phase follows the four-stage approach by LeTs [39]: tokenization, part-of-speech tagging, lemmatization, and chunking. In particular, the proposed architecture

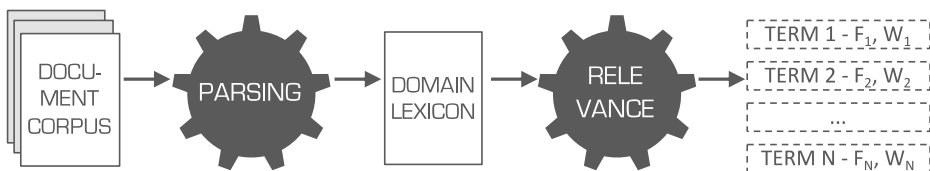


Fig. 3 The architecture for the step 2 – Term extraction

integrates the Stanford POS tagger [17] to tokenize and tag the words with the corresponding part of speech (POS) or syntactic category (e.g. noun, verb, adjective, adverb, or determiner); the Stanford CoreNLP [23] to obtain the root forms of the tagged words; and the Dias et al.’s definition of multiword [9] to identify sequences of nouns occurring together as chunks (e.g. *Donald Trump*, *Eiffel Tower*, or *concert hall*).

Among the tagged terms, the lexicon will include just a reduced set of syntactic categories. The selection of the POS to consider is based on the assumption that not all of them carry out the same semantic value: some (e.g. nouns) can have more semantic value than others (e.g. determiners). This issue has been extensively discussed in the literature on knowledge modeling, and even if no consensus has been reached, the focus is on nouns [9], nouns and adjectives [34], nouns, adjectives, verbs and adverbs [25]. Our tools collects the four POS like in [25], and we let the users choose which ones they wish to keep in the taxonomy (see the green filter panel in Fig. 4).

3.2.2 Phase 2: Relevance

The second phase of the term extraction is the relevance where the lexicon is filtered based on the meaning of the terms respect to the domain of interest (see Fig. 3). The proposed architecture offers two different selection criteria to measure the relevance: the frequency and the weight. The frequency is defined as the number of times a term appears in the corpus: the most frequent terms can be considered as the most representative ones for the domain [26]. The weight or domain consensus is an indicator of the consensus of the corpus on using frequently a term [28]: a high weight means that the term is highly frequent in each document included in the corpus. The user can choose the criteria to establish a hierarchy among the terms and the threshold to determine the number of candidates for generating the taxonomy (see the green filter panel in Fig. 4).

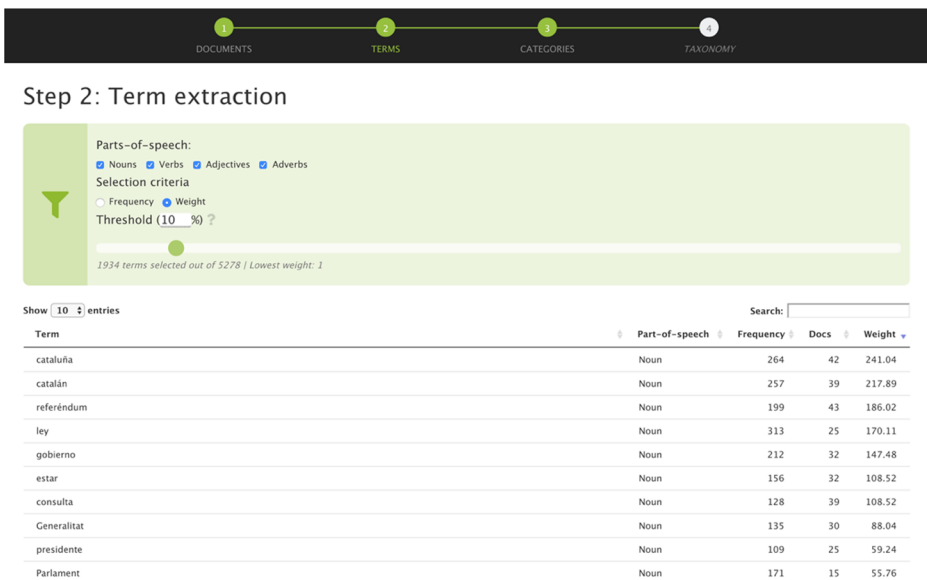


Fig. 4 Term extraction (step 2): the users can select the parts of speech to include, the selection criteria to use for the relevance and how many terms to filter (i.e. threshold)

3.3 Step 3 – Term categorization

The third step of the architecture consists of categorizing the domain lexicon previously extracted and filtered. The categorization is performed in three different phases (see Fig. 5): the NER tagger, the semantic relations, and the expert categorization.

The Stanford Named Entity Recognition (NER) Tagger [17] identifies words with special meanings referring to known entities, like people, nationalities, institutions, time, or events. These entities and the tagged terms or groups of terms are recognized as categories by the tool.

The remaining categories are defined based on the semantic relations established among the terms extracted from the corpus. Lefever has identified five different approaches to detect relations between terms [20]: pattern-based [15] (i.e. the related terms matches a previously defined set of patterns), statistical and machine learning [35] (i.e. the terms are related following a probability distribution), distributional [40] (i.e. the terms are related depending on their distribution in the dataset), morpho-syntactic [38] (i.e. the related terms have similar syntactic properties) and word class lattices (i.e. a directed acyclic graph where the words are edges and the vertices are weights, used by Navigli and Velardi to model textual definitions) [27]. Our method follows the morpho-syntactic approach, taking into account both the syntactic function and the thesaurus relations to categorize the terms. If two terms are semantically related, they will be included in the same category. In particular, we consider four semantic relations: synonyms, antonyms, co-occurrences, and multiword terms. The co-occurrences have been identified using the Wilks's method [44]: if two words co-occur in the dictionary (i.e. one appears in the definition of the other), they are related. The semantic relations and the definitions have been extracted using the API provided by the Oxford Dictionary [33]. The multiword relation is based on the chunks identified during the step 2 as sequences of nouns stored as single terms.

Once the taxonomy is generated automatically from the NER tagger and the semantic relations, the architecture offers a visual tool for exploring and eventually modifying collected categories and terms (see Fig. 6).

The interface has three panels: the working space on the right, the categories on the left, and the actions in the center. The working space contains the initial categorization generated automatically, while the categories panel is used to structure the final version of the taxonomy. The categories are also associated to a color and the same color will be used in the next step for the visual representation of the taxonomy. Interacting with the different options provided by the central panel, the user can move a group of terms or a single term from the right to the left to be included to the final version of the taxonomy. It is also possible to add a new term or category, modify an existing one or merge two or more categories. A detailed list of the available actions is shown in Table 1.

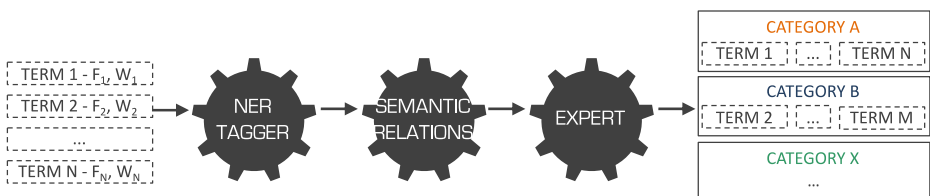
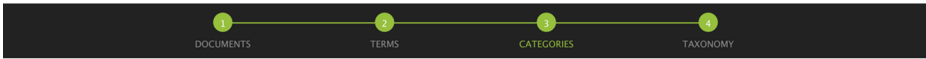


Fig. 5 The architecture for the step 3 – Term categorization



Step 3: Categories



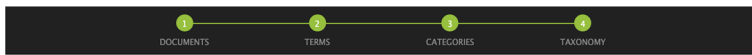
Fig. 6 Term categorization (step 3): the users can explore the groups of terms generated automatically (right panel), modify them (central panel), and include them to the final version of the taxonomy (left panel)

3.4 Step 4 - Taxonomy visualization

The last step of the architecture offers a visual representation of the generated taxonomy. We have designed a Zoomable Circle Packing (see Fig. 7), a technique for representing hierarchical datasets with a tree structure, where the branches and the leaves are circles of different sizes [43]. The main categories of the taxonomy are colored circles containing white smaller circles (i.e. the individual terms). The size of the categories depends on the number of contained terms, while the size of the individual terms is a measure of their relevance in the domain. Zooming on a category, it is possible to explore the contained terms and, clicking on them, the list of documents where they appear.

Table 1 List of all the buttons available in the action panel described by the icon, the label and the associated action

Icon	Label	Action
+	New category	Create a new category
<	Move left	Move a group to the panel of categories
>	Move right	Move a category to the panel of groups
⇕	Merge	Merge selected groups or terms into a category if selected in the left panel, or a new group otherwise
🗑️	Delete	Delete the selected terms, groups or categories
📄	Download	Download the taxonomy in JSON format
💾	Save	Save all the changes made in the categorization
👉	Select	Highlight the selected groups, terms or categories
💡	Doubt	Highlight the terms marked with doubt.



Step 4: Taxonomy

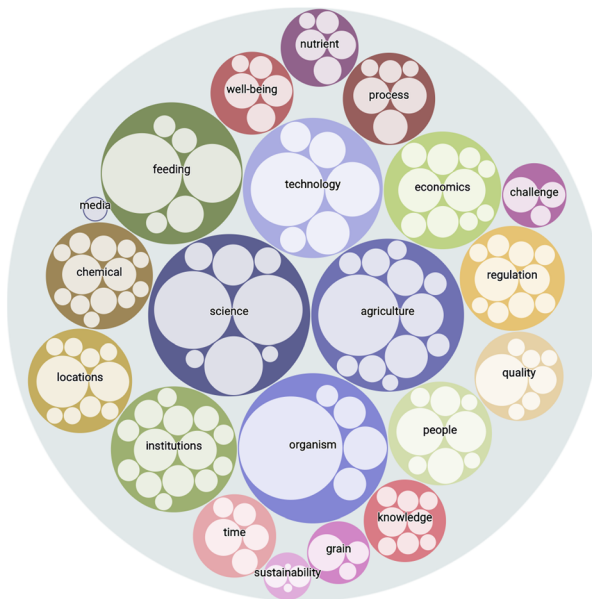


Fig. 7 Taxonomy visualization (step 4) for the NPBTs use case of the evaluation (see Section 5): colored circles are the main categories of the taxonomy, while the white circles are the individual terms

In case of representing a knowledge model like a taxonomy, the Circle Packing gives a visual overview of the main elements in the representation, their relevance and how they are related, keeping visible even the smallest node (see the node *media* in Fig. 7). The colored circles represent the main categories, while the white ones are the individual terms. The size of each circle depends on their relevance (i.e. frequency or weight, as selected by the end users in the second step), and the color can be changed by the end users in the third step. In this way, users can visually check whether the resulting taxonomy is adequate or not, and eventually go back and improve it until it fits their needs.

4 Uses of the taxonomy

Taxonomies are semantic structures used to model knowledge about a domain of interest. In the Big Data era, where information is generated massively from a wide range of sources, these models can represent a useful exploration tool serving different purposes. Here, we describe three scenarios where the taxonomies play a key role: visual analytics, interactive dictionaries, and semantic searches.

Nowadays, people are used to sharing information, opinions, and experiences through online platforms like social networks, blogs or news channels. This practice generates large amounts of unstructured data that are difficult if not impossible to process. For instance, in an emergency like a typhoon, an earthquake or a terrorist attack, publishing a message on social media can be the only way that victims and witnesses have to communicate [22]. Emergency

operation centers could find helpful information on these citizen generated information as far as they can quickly identify relevant information and integrate it into their protocols [10]. Similarly, during an election campaign both political parties and political analysts could take advantage from the content shared by citizens to have a clearer picture of their preferences. In these kinds of situations, taxonomies can be used as part of visual analytics tools to semantically analyze large volumes of data and filter out the most relevant knowledge about a topic of interest [30]. The resulting knowledge is then represented as a visualization that users can explore according to their needs, whether to support sense-making, decision-making or even learning on a specific domain.

In previous research, we have developed a tool for semantically analyzing and visualizing large datasets of social networks content based on the usage of domain-specific taxonomies [29]. The tool has been tested with real datasets in the domains of emergency [31] and politics [32]: changing the domain means the definition of different knowledge models. A semi-automatic tool for taxonomy generation can be a solution to reduce the effort, but still keeping the end users as active contributors of the modeling process.

Another interesting usage of the taxonomies is building interactive dictionaries. In this case, the domain is modeled applying the same semantic approach of the generation tool presented in this work, extracting terms and categories from a representative corpus of documents (e.g. blogs, news channels, or handbooks). The users can explore an interactive visualization of the resulting taxonomy to see the included concepts and how they are semantically related. The aim is to foster the curiosity and create more engaging learning experiences in multiple contexts: an audience interested in a topic of general interest, a student doing research about a subject, or a group of collaborators that needs to share a common language.

For instance, the European Community is investing in New Plant Breeding Techniques (i.e. NPBTs) that will guarantee a more sustainable application of biotechnologies in agriculture. Even if this topic is of great interest for the general public, it could be rather demanding for the general public to understand while there is also a lot of controversial and fake information on smart crops. The generation tool described previously made it possible to generate a NPBT taxonomy from a collection of news articles and blog entries, that was presented to a general audience during a science fair as part of the H2020 project CHIC [8]. In this case, the visualization chosen to represent the taxonomy was a bubble chart designed for a virtual reality space [32]. The immersive experience surrounds the users with concepts and categories from the taxonomy, isolating them from the reality and fostering the exploratory task. As a result, the attendees could easily learn about this particular biotechnology, focusing on the most relevant information needed to understand its advantages and disadvantages.

The third example of scenario where the taxonomies can represent a helpful instrument, is the development of personalized search engines [18]. In this case, taxonomies are used to match the organization of the results with the users' preferences. Consequently, the users improve their expressional power in terms of search queries, and the engine offers a more precise retrieving mechanism.

5 Evaluation

The taxonomy generation tool has been designed to work for multiple languages and any domain of interest. For this reason, we have analyzed its applicability in two different real scenarios, varying both the domains and the languages: the 2017 Catalan constitutional crisis

in Spanish, and the implementation of New Plant Breeding Techniques (i.e. NPBTs) in English. We run a usability evaluation on both case studies, as detailed in the next subsections.

5.1 First use case: Politics

In collaboration with experts in political science from the Department of Communication and Internet Studies of the Cyprus University of Technology and in the context of the NOTRE H2020 project, we tested the tool with a collection of documents in Spanish about the 2017 Catalan constitutional crisis. This event has polarized the Spanish society generating a growing volume of publications on several online media platforms that we have selected as sources for our tool.

In terms of the methodology used for our data collection, we first implemented a web scraper in Python to retrieve the most relevant articles from the Google Custom Search (GCS) service. The search inputs were a set of keywords related to the event of Catalonia's independence referendum 2017 (in both Catalan and Spanish) and a series of media sources. The latter were chosen to reflect the territorial diversity in Spain. Thus, we had the following number of sources per territorial group: 5 National media, 4 Catalan media, and 5 regional media. We also paid attention to ideological diversity by incorporating media from the left and right of the political spectrum. The end result was a corpus of 120 documents in Spanish containing one article from each source for every month in the year 2017.

5.2 Second use case: Biotechnology

In the H2020 project CHIC [8] transferring knowledge on NPBTs using creative and novel ways is established as a main goal to be able to engage citizens, and in particular teenagers, in this scientific challenge. In this context, we used the tool to create a taxonomy for generating semantic and engaging visualizations about the NPBTs and biotechnologies. To avoid too technical documents whose vocabulary will not be understood by a general audience, we selected two main sources: the CHIC website and Google News. From the CHIC website, we collected several brochures and blog posts aimed at presenting the scope of the project and its impact on the society. From Google News, we have searched for the keyword NPBTs and stored articles and general-purpose documents from the last month. Considering the specificity of selected sources, in this second use case we have limited the volume of the corpus to 30 documents in English.

5.3 Usability evaluation

The goal of this evaluation is to measure the usability of the taxonomy generation tool. The evaluation was carried out by 16 participants in two groups: 10 of them worked on the first use case, and the remaining on the second. The participants were not required to have any previous knowledge on taxonomy generation or knowledge models, but they were required to be fluent in Spanish or in English, depending on the language of the use case.

The evaluation had three phases:

1. *Pre-test questionnaire*: a questionnaire to gather the demographic data
2. *Test*: list of tasks to perform with the tool
3. *Post-test questionnaire*: a questionnaire about the usability of the tool

5.3.1 Pre-test questionnaire

In the first place, the participants were asked to fill out a questionnaire to collect their demographic data. The majority were PhD researchers, Master or PhD students, between 25 and 44 years old, with a high technological background and a low level of expertise using knowledge models (e.g. taxonomies, ontologies).

5.3.2 Test

The participants were briefly introduced to the tool and its main functionalities before starting the test. Later, they were asked to complete a set of tasks designed to evaluate the usability of the different actions supported by the tool and to identify strengths and weaknesses. The complete set of tasks is listed in Table 2: each task is related to the stage of the development process where it is performed.

After performing each task, participants were asked to give feedback about their experience filling out the post-test questionnaire, focusing on the interaction with the elements in the interface. Mainly, we were interested in collecting opinions about the interface design and how to improve it.

5.3.3 Post-test questionnaire

The last stage of the usability evaluation consisted of answering a post-test questionnaire based on the SUS (System Usability Scale) [5]. The questionnaire had a total of 9 statements about the tool, alternating between negative and positive ones. Participants were asked to specify their agreement (or disagreement) with the statements using a 5-point Likert scale, where 5 corresponds to *strongly agree* and 1 to *strongly disagree*. We omitted one of the items from the original questionnaire: “*I think that I would like to use this system frequently*” since the participants of the evaluation were not experts and therefore categorizing information is not part of their routine. The final list of statements is as follow:

Table 2 List of tasks performed during the test in relation to the stage of the development process

STAGE	TASK
Document corpus	Upload a document to the tool Upload a document from a URL See the content of a document Remove a document from the corpus
Term extraction	Modify which parts of speech are included in the taxonomy Adjust the value of the threshold Order the terms alphabetically
Categorization	Move a group of terms to the panel of categories Create a new category Add new terms to the new category Merge several groups or categories Mark several items with doubt Review the terms marked with doubt Remove a group or category Save all changes

1. I found the system unnecessarily complex.
2. I thought the system was easy to use.
3. I think that I would need the support of a technical person to be able to use this system.
4. I found the various functionalities in this system well integrated.
5. I thought there was too much inconsistency in this system.
6. I would imagine that most people would learn to use this system very quickly.
7. I found the system very cumbersome to use.
8. I felt very confident using the system.
9. I needed to learn a lot of things before I could get going with this system.

To evaluate the obtained results for each one of the questions, we have compared them with the regression equations (R^2) defined by Lewis and Sauro to understand how each SUS item is related to the overall SUS score [21]. They have built a curved grading scale from F (the lowest) to A+ (the highest) analyzing 241 SUS studies and modeled two regression equations to associate the SUS score 68 to C, and 80 to A-. We have used these equations to evaluate the results obtained from the post-test questionnaire. Table 3 shows the mean value (μ) for each statement of the questionnaire, the standard deviation (σ), the values of the regression equations (R^2) for the SUS score 68 and the SUS score 80.

Overall, the participants showed a positive response to the tool, with all the mean values included in the limits defined by the regression equations. In particular, the highest comparison with a SUS score of 80 has been obtained for the first statement about the complexity of the tool, the fourth about the integration of offered functionalities, the fifth about the inconsistencies, the eighth about the users' confidence in carrying out the tasks on their own, and the ninth about the needed learning effort. This means that the participants have found the interface of the tool simple to interact with, visually consistent and intuitive.

The remaining statements have reached the limits for a SUS score greater than 68, meaning that the participants could easily use the tool and understand the different options offered in the interface without requiring an excessive learning effort. The only statement with an outlier value respect to the two regression equations is the third item suggesting that the users can have some initial difficulties in learning all the possible functionalities offered by the tool and they could appreciate the support of a technical person. This is mainly related to the low experience they have working with knowledge modeling and data analysis processes.

Table 3 Mean and standard deviation for the nine statements in the post-test questionnaire

STATEMENT	μ	σ	R^2	R^2
			SUS = 68	SUS = 80
1. I found the system unnecessarily complex.	1.75	0.77	≤ 2.44	≤ 1.85
2. I thought the system was easy to use.	4.19	0.75	≥ 3.67	≥ 4.24
3. I think that I would need the support of a technical person to be able to use this system.	1.94	1.06	≤ 1.85	≤ 1.51
4. I found the various functionalities in this system well integrated.	4.00	1.03	≥ 3.55	≥ 3.96
5. I thought there was too much inconsistency in this system.	1.44	0.73	≤ 2.20	≤ 1.77
6. I would imagine that most people would learn to use this system very quickly.	4.00	0.89	≥ 3.71	≥ 4.19
7. I found the system very cumbersome to use.	1.75	0.58	≤ 2.25	≤ 1.66
8. I felt very confident using the system.	4.25	0.77	≥ 3.72	≥ 4.25
9. I needed to learn a lot of things before I could get going with this system.	1.56	0.73	≤ 2.09	≤ 1.64

5.3.4 Learned lessons and future works

Analyzing the results obtained from the usability evaluation, we have learned useful lessons for improving the tool and defining the next steps to take.

The involvement of participants with no particular experience in modeling knowledge and working with data analysis has shown that the tool and the available options are easy to use and interact with in both use cases. Moreover, the participants have found useful the possibility to interact with the interface using both direct manipulation (i.e. drag-and-drop mechanism) and buttons. They have also appreciated the usage of feedback (i.e. progress bar and inspection window) to be informed about the current status of the building process and, in particular, how the tagging of the terms works.

Based on this experience, we are going to further evaluate the tool focusing on its utility in a real scenario. We want to involve domain experts interested in data analysis and knowledge modeling to analyze the applicability and effectiveness of the proposed solution for generating taxonomies. This kind of evaluation will require an extra effort, in particular in terms of time. For this reason, to assess the usability of the tool we have decided to involve non-expert users working on a predefined set of tasks.

Domain experts will use the tool to create a taxonomy from scratch, deciding which documents to include in the corpus, which criteria to apply for selecting the terms, and how to categorize the resulting groups of terms. Another result we will expect to obtain from the expert evaluation is the accuracy of the taxonomy resulting from the automatic categorization. In particular, we will focus on measuring missing or redundant terms and categories, as well as terms in the wrong category or categories with non-representative names. Finally, the domain experts will be also involved in the design of the taxonomy visualization. Based on the proposed circle packing and using a participatory approach, we want to understand what they expect from a graphical representation of the taxonomy and which technique best fits with their needs.

6 Conclusions and future works

In the Big Data era, a key challenge is represented by the growing interest in collecting, analyzing and extracting useful knowledge from the large volumes of unstructured data published every day on online platforms. Among proposed approaches, knowledge models aim at adding semantics to data and focusing on the most meaningful information. For that purpose, what is meaningful in a domain has to be identified. Automatic and intelligent tools can build semantic taxonomies but to be adopted in decision making or knowledge exploration processes they need to be validated by human experts. The collaboration of end users is crucial to guarantee the quality of the resulting model and its value as a tool to transmit and acquire knowledge about a subject. In this paper we have described a tool that put experts and end users in the knowledge modelling loop. In our semiautomatic generation tool, users can contribute at any stage of the semantic modeling process: from deciding which documents have to be included in the corpus, to the selection of the most relevant terms and their organization in categories. The ability to check the taxonomy using a visualization helps end users to understand better the impact of their decisions in the previous stages (selection of documents, terms and categories, and relevance metrics). For instance, if after a first round using the tool, a very relevant topic for the end user is presented in the visualization as a marginal topic, she can go back and include more documents related with that term or check

whether it is correctly categorized. Therefore, end users can act upon all the knowledge modeling processes and quickly get visual feedback on the consequences of their actions.

The tool here presented works for multiple languages and any domain of interest, as shown by the two use cases of the evaluation: the documents collected during the 2017 Catalan constitutional crisis and the information on smart crops and NPBTs. After a usability evaluation, we can conclude that even though some initial effort is required to learn how to operate with the tool, participants felt confident and comfortable using it and found the tool useful to create meaningful relations among terms. Since participants had no experience in knowledge modeling and data analysis, we are planning to perform further evaluations to test the applicability and the effectiveness of the tool in real scenarios. We are also interested in involving domain experts in the design of new visualizations and scenarios for interacting with the taxonomy. Following a participatory approach, we are specifically interested in understanding what the users look for in a graphical representation of a knowledge model.

Acknowledgements This work was supported by the project PACE grant funded by the Spanish Ministry of Economy and Competitiveness [TIN2016- 77690-R]. Authors also acknowledge travel funding from the European Union's Horizon 2020 Framework through NOTRE project (H2020-TWINN-2015, GA Number: 692058).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

7. References

1. Baroni M, Bernardini S (2004) BootCaT: Bootstrapping corpora and terms from the web. *Proc Lr* 4:1313–1316
2. Bello-Orgaz G, Jung JJ, Camacho D (2016) Social big data: Recent achievements and new challenges. *Inf Fusion* 28:45–59. <https://doi.org/10.1016/j.inffus.2015.08.005>
3. Bendle N, Ryoo J, Nastasoiu A (2018) The 2016 US Primaries: Parties and Candidates in a World of Big Data. In: *Political Marketing in the 2016 U.S. Presidential Election*. Springer International Publishing, Cham, pp 65–80
4. Biemann C (2005) Ontology Learning from Text: A Survey of Methods. *LDV-Forum* 20:75–93
5. Brooke J, others (1996) SUS-A quick and dirty usability scale. *Usability Eval Ind* 189:4–7
6. Castells M (2007) An introduction to the information age. *City* 2:6–16. <https://doi.org/10.1080/13604819708900050>
7. Centelles M (2005) Taxonomies for categorization and organization in web sites. *Hipertext net*
8. CHIC European Project. <http://chicproject.eu/>. Accessed 9 Apr 2019
9. Dias G, Guilloré S, Bassano JC, Pereira Lopes JG (2000) Combining linguistics with statistics for multiword term extraction: A fruitful association?. In: *Content-Based Multimedia Information Access, Volume 2*, pp 1473–1491
10. Díaz P, Carroll J, Aedo I (2016) Coproduction as an Approach to Technology-Mediated Citizen Participation in Emergency Management. *Futur Internet* 8:41. <https://doi.org/10.3390/fi8030041>
11. Dokoochaki N, Zikou F, Gillblad D, Matskin M (2015) Predicting swedish elections with twitter: A case for stochastic link structure analysis. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. pp 1269–1276
12. Feldman R, Gresko M, Kinar Y, Lindell Y, Liphstat O, Rajman M, Schler Y, Zamir O (1998) Text mining at the term level. In: *European Symposium on Principles of Data Mining and Knowledge Discovery*, Springer, Berlin, Heidelberg, pp 65–73
13. Frakes WB, Baeza-Yates R (1992) *Information retrieval: Data structures & algorithms*. Prentice Hall, Englewood Cliffs
14. Gundecha P, Liu H (2012) *Mining Social Media: A Brief Introduction*. 2012 TutORials Oper Res:1–17. <https://doi.org/10.1287/educ.1120.0105>
15. Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. *Proc 14th Conf Comput Linguist* 2:539. <https://doi.org/10.3115/992133.992154>

16. Hoang HH, Cung TN-P, Truong DK, Hwang D, Jung JJ (2014) Retracted: Semantic Information Integration with Linked Data Mashups Approaches. *Int J Distrib Sens Networks* 10:813875. <https://doi.org/10.1155/2014/813875>
17. Jung JJ (2010) Integrating Social Networks for Context Fusion in Mobile Service Platforms. *J. UCS*, 16(15), pp 2099–2110
18. Kerschberg L, Kim W, Scime A (2001) A semantic taxonomy-based personalizable meta-search agent. In: *Proceedings of the Second International Conference on Web Information Systems Engineering*, Volume 1, IEEE Computer Society, pp 41–50
19. Kotlerman L, Avital Z, Dagan I, Lotan A, Weintraub O (2011) A Support Tool for Deriving Domain Taxonomies from Wikipedia
20. Lefever E (2015) LT3: A Multi-modular Approach to Automatic Taxonomy Construction. *9th Int Work Semant Eval (SemEval 2015)* 944–948
21. Lewis JR, Sauro J (2018) Item Benchmarks for the System Usability Scale. *J Usability Stud* 13:158–167
22. Lindsay BR (2010) Social Media and Disasters: Current Uses, Future Options and Policy Considerations. *J Curr Issues Media Telecommun* 2:287–297
23. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D The Stanford CoreNLP Natural Language Processing Toolkit
24. McGregor SC, Mourão RR, Molyneux L (2017) Twitter as a tool for and object of political and electoral activity: Considering electoral context and variance among actors. *J Inf Technol Polit* 14:154–167. <https://doi.org/10.1080/19331681.2017.1308289>
25. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to WordNet: An On-line Lexical Database. *Int J Lexicogr* 3:235–244
26. Murthy K, Faruque TA, Subramaniam LV, Prasad KH, Mohania M (2010) Automatically generating term-frequency-induced taxonomies. *Proc. ACL 2010 Conf. Short Pap.* 126–131
27. Navigli R, Velardi P (2010) Learning Word-Class Lattices for Definition and Hyponym Extraction. *Proc 48th Annu Meet Assoc Comput Linguist (ACL 2010)* 1318–1327
28. Navigli R, Velardi P, Gangemi A (2003) Ontology learning and its application to automated terminology translation. *IEEE Intell Syst* 18:22–31. <https://doi.org/10.1109/MIS.2003.1179190>
29. Onorati T, Díaz P (2015) Semantic visualization of twitter usage in emergency and crisis situations. In: *Lecture Notes in Business Information Processing*, pp 3–14
30. Onorati T, Díaz P (2016) Giving meaning to tweets in emergency situations: a semantic approach for filtering and visualizing social data. *Springerplus* 5:1782. <https://doi.org/10.1186/s40064-016-3384-x>
31. Onorati T, Díaz P, Carrion B (2018) From social networks to emergency operation centers: A semantic visualization approach. *Futur Gener Comput Syst.* <https://doi.org/10.1016/j.future.2018.01.052>
32. Onorati T, Díaz P, Zarranandia T, Aedo I (2018) The Immersive Bubble Chart: a Semantic and Virtual Reality Visualization for Big Data. In: *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings - UIST '18 Adjunct*. ACM Press, New York, pp 176–178
33. Oxford Dictionaries API. developer.oxforddictionaries.com. Accessed 29 of June 2019
34. Park Y, Byrd Roy J, Beguraev BK (2002) Automatic glossary extraction: beyond terminology identification. *Assoc Comput Linguist* 22421:1–7. <https://doi.org/10.1007/s13398-014-0173-2>
35. Ritter A, Soderland S, Etzioni O (2009) What Is This, Anyway: Automatic Hyponym Discovery. In: *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pp 88–93
36. Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval. In: *Proceedings of the international communication of association for computing machinery conference*, p 260
37. Sujatha R, Rao BRK (2011) Taxonomy construction techniques – issues and Challenges. *Indian J Comput Sci Eng* 2:661–671
38. Tjong Kim Sang E, Hofmann K, de Rijke M (2011) Extraction of Hyponymy Information from Text. *Interact Multi-modal Quest Theory Appl Nat Lang Process* 223–245
39. Van de Kauter M, Coorman G, Lefever E, Desmet B (2013) LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit Lieve Macken Véronique Hoste. *Comput Linguist Netherlands J* 3:103–120
40. Van Der Plas L, Bouma G (2005) Automatic acquisition of lexico-semantic knowledge for QA. *Proc OntoLex 2005-Ontologies Lex Resour*
41. Velardi P, Cucchiarelli A, Petit M (2007) A taxonomy learning method and its application to characterize a scientific web community. *IEEE Trans Knowl Data Eng* 19:180–191. <https://doi.org/10.1109/TKDE.2007.21>
42. Vepsäläinen T, Li H, Suomi R (2017) Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections. *Gov Inf Q* 34:524–532. <https://doi.org/10.1016/j.giq.2017.05.004>
43. Wang W, Wang H, Dai G, Wang H (2006) Visualization of large hierarchical data by circle packing. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*. ACM Press, New York, p 517
44. Wilks Y (1993) *Providing Machine Tractable Dictionary Tools*. Springer, Dordrecht, pp 341–401

45. Yang H (2012) Constructing Task-Specific Taxonomies for Document Collection Browsing. Association for Computational Linguistics
46. Zamir O, Etzioni O (1998) Web document clustering: A feasibility demonstration. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp 46–54
47. Zilli A (2009) Semantic knowledge management: an ontology-based framework. Information Science Reference

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Belen Carrion is a research and teaching assistant at Universidad Carlos III de Madrid, Spain where she graduated in 2016 in Computer Science and Engineering. She is currently a M. Sc. student in Computer Science and Technology and she also collaborates with the DEI research group. Her research interests focus on information visualization.



Teresa Onorati is an assistant professor at Universidad Carlos III de Madrid, Spain where in 2013 she received her PhD in Computer Science, and a member of the DEI research group. She also holds a B. Sc. and a M. Sc. in Computer Science from La Sapienza Università di Roma. Her research interests focus on semantic modelling, visual analytics and information visualization.



Paloma Díaz is a full professor at Universidad Carlos III de Madrid and head of the DEI research group. Her research interests include Web engineering methods, combining usability principles and practices as well as their practical application to different kinds of interactive systems, including e-learning and e-gov. She holds a PhD in computer science from Universidad Politécnica de Madrid, Spain.



Vasiliki Triga is an Assistant Professor at the Cyprus University of Technology. Her current research lies in the field of Southern politics and the economic crisis, social movements, internet politics and referendums. She is co-author of the book *Referendums and the European Union: A Comparative Inquiry*, Cambridge University Press 2014.

Affiliations

Belen Carrion¹ • **Teresa Onorati**¹ • **Paloma Díaz**¹ • **Vasiliki Triga**²

¹ Department of Computer Science, Universidad Carlos III de Madrid, Leganes, Spain

² Department of Communication and Internet Studies, Cyprus University of Technology, Limassol, Cyprus