Using NLP to Support Terminology Extraction and Domain Scoping: Report on the H2020 DESIRA Project

Manlio Bacco
WN Lab, CNR-ISTI, Pisa, Italy
manlio.bacco@isti.cnr.it

Gianluca Brunori
PAGE, DISAAA, University of Pisa, Italy
gianluca.brunori@unipi.it

Felice Dell'Orletta
ItaliaNLP Lab, CNR-ILC, Pisa, Italy
felice.dellorletta@ilc.cnr.it

Alessio Ferrari
FMT Lab, CNR-ISTI, Pisa, Italy
alessio.ferrari@isti.cnr.it

Abstract

The ongoing phenomenon of digitisation is changing social and work life, with tangible effects on the socio-economic context. Understanding the impact, opportunities, and threats of digital transformation requires the identification of viewpoints from a large diversity of stakeholders, from policy makers to domain experts, and from engineers to common citizens. The DESIRA (*Digitisation: Economic and Social Impacts in Rural Areas*) EU H2020 project[1] considers rural areas, with a strong focus on agricultural and forestry activities, and aims at assessing the impact of digital technologies in those domains by involving a large number of stakeholders, all across Europe, around 20 *focal questions*. Given the involvement of stakeholders with diverse background and skills, a primary goal of the project is to develop domain-specific and interactive reference taxonomies (i.e., structured classifications of terms) to facilitate common understanding of technologies in use in each domain at today. The taxonomies, which aims at easing the learning of the meaning of technical and domain-specific terms, are going to be exploited by the stakeholders in 20 Living Labs built around the focal questions. This report paper focuses on the semi-automatic development of the taxonomies through natural language processing (NLP) techniques based on context-specific term extraction. Furthermore, we crawl Wikipedia to enrich the taxonomies with additional categories and definitions. We plan to validate the taxonomies through field studies within the Living Labs.

1 Introduction

The DESIRA project aims to assess and anticipate the impact of digital transformation in rural areas, with a specific focus on the fields of agriculture and forestry. The activities within the project see the creation of a common terminology in a very multi-disciplinary consortium, the assessment of past and present game changing

In: M. Sabetzadeh, A. Vogelsang, S. Abualhaija, M. Borg, F. Dalpiaz, M. Daneva, N. Fernández, X. Franch, D. Fucci, V. Gervasi, E. Groen, R. Guizzardi, A. Herrmann, J. Horkoff, L. Mich, A. Perini, A. Susi (eds.): Joint Proceedings of REFSQ-2020 Workshops, Doctoral Symposium, Live Studies Track, and Poster Track, Pisa, Italy, 24-03-2020, published at http://ceur-ws.org

[1]Project website available at: `http://desira2020.eu`

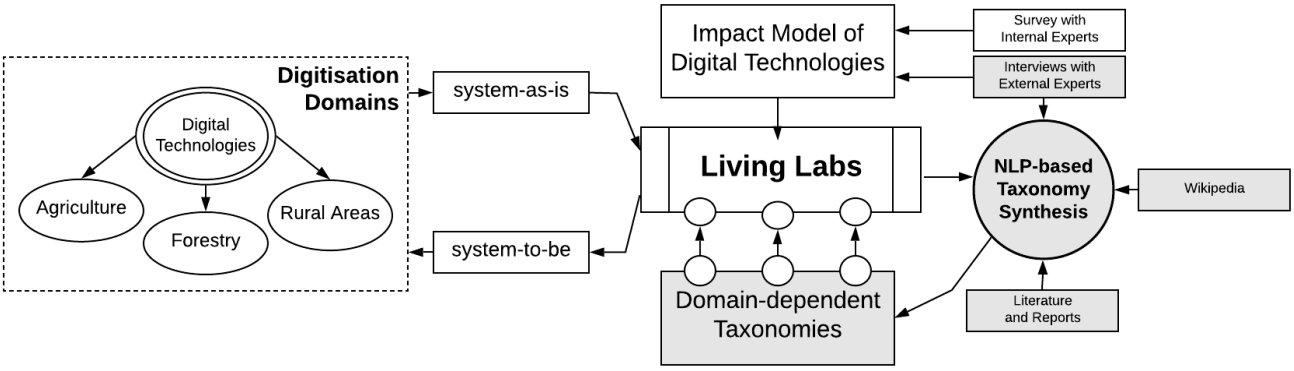

Figure 1: Overview of the DESIRA Project. The NLP-based part of the project is colored in grey (for "Interview with External Experts", transcripts of the interviews are used as input for the process of NLP-based Taxonomy Synthesis).

effects due to digital technologies, and the set-up of a methodology to anticipate future effects. Those activities are going to be performed within 20 Living Labs (LLs) in different geographical areas of Europe, composed of several stakeholders, being either part of the project consortium or providing external expertise. LLs can be defined as *user-centered, open innovation ecosystems based on systematic user co-creation approach, integrating research and innovation processes in real-life communities and settings.*

As shown in Fig. 1, LLs are at the core of the activities in DESIRA. Each LL is associated to a specific digitisation domain. In our case, three main domains are considered: agriculture, forestry, and rural areas. This latter domain includes all the application environments for digitisation that do not belong to agriculture and forestry, such as, for example, management of small towns, rural communities, roads, etc. Each LL embodies a *focal question*, i.e., a set of primary goals to be addressed in the specific context of the LL: for instance, in central Italy, the focal question is related to forest management activities, and to the need to counteract both illegal logging and lack of information about wood provenance.

The main purpose of a LL is to assess the past and present situation in the geographical area regarding its focal question (*system-as-is* in Fig. 1), identifying both drivers and obstacles in the current socio-technical system, and then agree on a desired future situation (*system-to-be*), highlighting the role that the introduction of digital technologies may play in enabling it. For instance, reference [BBF+19] focuses on novel digital technologies ready for use in the agricultural field and on existing non-technical barriers holding a larger adoption. During the DESIRA project, the stakeholders of each LL—about 15-20 people per LL—will physically meet in four workshops, and will continuously interact through Virtual Research Environments based on gCube [ACC+19], a collaborative online platform.

The objective of the first two workshops is to gain a clear picture of the system-as-is within the LL, i.e., deepen the description of the context around the focal question and clearly identify socio-economic pros and cons of the specific digital solutions already in use, if any.

The next two workshops will focus on the actions to be undertaken to transition into a more desirable digital-enabled scenario, activity to be strongly supported by the domain-dependent taxonomies and the impact model of digital technologies, as in Fig. 1.

The activities of each LL are managed by an appointed LL moderator, which facilitates the discussion by leveraging (*i*) an impact model of digital technologies (common to all LLs), and (*ii*) a set of domain-dependent taxonomies. The model, built upon a structured survey with internal experts, interviews with external ones, and literature analysis, provide guidelines and examples to try and assess the socio-economic impact of digital technologies, ultimately referring to the UN Sustainable Development Goals (SDGs)[2]. For instance, recalling the focal question of the LL in central Italy, the SDG under consideration is the 12th, i.e., *responsible consumption and production.*

The impact model is used by LL participants to brainstorm on how certain technologies are and will influence their specific context. The domain-dependent taxonomies (one for each digitisation domain) are used to learn about the different technologies in use in the domain, and the meaning of the technical terminology. These are synthesised with the support of natural language processing (NLP) tools. Specifically, the information

[2]Details on the SDGs available at: `sustainabledevelopment.un.org`

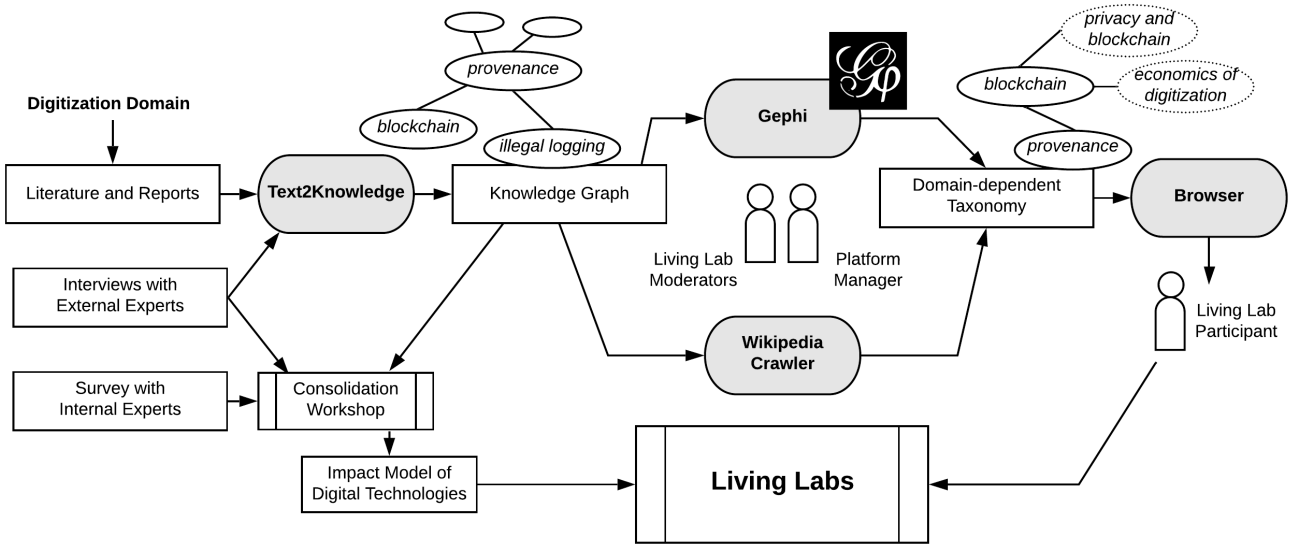

Figure 2: Synthesis of the domain-dependent Taxonomy and relation with the impact model. Software tools are colored in grey.

extraction tool Text2Knowledge [DVCM14][3], developed by the ItaliaNLP Lab[4], is used to identify relevant terms and relations from different knowledge sources, namely literature and reports concerning the application of digital technologies in each considered domain, and interviews with external experts. Furthermore, Wikipedia is used to enrich the taxonomies with additional concepts and descriptive content, to facilitate domain scoping and independent learning by the LL participants.

In this paper, we focus on the description of the approach for NLP-based taxonomy synthesis in the context of DESIRA, which can be relevant for RE community interested in NLP. The presentation of the paper at NLP4RE may also serve as a trigger to discuss with the workshop participants potential RE methodologies and tools to best manage the LLs.

2 Terminology Extraction and Domain Scoping

We describe here the proposed approach for NLP-based taxonomy synthesis and domain scoping. Fig. 2 provides an overview of the approach: given a Digitisation Domain, such as, e.g. forestry, a set of documents and reports are selected around the topics of digital technologies and digitisation. Furthermore, a set of experts' interview on such topics are performed and transcribed. The tool Text2Knowledge (T2K) is used to extract a knowledge graph that represents relevant terms and relations as in the input documents. The knowledge graph is expected to include technology-related terms, as for instance "blockchain", along with domain-specific terms, such as "illegal logging" or "provenance" (see the examples in italic in Fig. 2). The knowledge graph is used for two goals: (*i*) generation of a domain-dependent taxonomy; and (*ii*) support for consolidation of the impact model of digital technologies.

(*i*) To generate a domain-dependent taxonomy, the knowledge graph is first loaded into Gephi[5], an open source graph visualisation and manipulation tool. An appointed Platform Manager, in collaboration with the LL Moderators, will edit the knowledge graph to produce the domain-dependent taxonomy. This is performed with the support of a Wikipedia Crawler, which allows the users to identify informative Wikipedia pages related to the various technological terms in the knowledge graph, and enrich the graph with links to those pages. The resulting domain-dependent taxonomies are exported through the Sigma.js[6] tool, and can be visualised and navigated by LL participants by means of a common web browser.

(*ii*) To support the consolidation of the impact model of digital technologies, which will be used in the LLs, a consolidation workshop is foreseen. This workshop will use the information collected from internal experts and

[3]Tool accessible upon request at: `http://www.italianlp.it/demo/t2k-text-to-knowledge/`
[4]Lab website available at: `http://www.italianlp.it`
[5]Download the tool at: `https://gephi.org`
[6]See: http://sigmajs.org

external ones concerning the impact of digital technologies, according to their previous experience in technological applications in the different domains considered. Furthermore, the workshop will use the knowledge graph to identify potentially missing relations between technologies and socio-economic impacts due to digitisation.

In the following, we focus on generation of the knowledge graph with T2K (Sect. 2.1), and on the definition of the domain-dependent taxonomies with the support of Wikipedia (Sect. 2.2).

2.1 Generation of the Knowledge Graph with T2K

T2K is a tool to generate knowledge graphs from unstructured natural language documents. A knowledge graph is composed of nodes, representing relevant terms in the documents, and edges, representing relevant relations among the terms. Below, we briefly describe the principles used by T2K to extract terms and relations.

2.1.1 Identification of Relevant Terms

The NLP method for term extraction was developed by the second author and is named *contrastive analysis* [BDMV10]. In this context, a *term* is a conceptually independent linguistic unit, which can be composed by a single or multiple words. The *contrastive analysis* technology aims at detecting those terms in a document that are *specific* for the context of the document under consideration [BDMV10, Del09]. In our case, the context is given by the specific domain (e.g., forestry) combined with the topics of digitisation. Roughly, contrastive analysis considers the terms extracted from context-generic documents (e.g., newspapers), and the terms extracted from context-specific documents under analysis. If a term in the context-specific document highly occurs also in the context-generic documents, such a term is considered as context-generic. On the other hand, if the term is not frequent in the context-generic documents, the term is considered as context-specific.

In our work, the documents from which we want to extract context-specific terms are the input documents that better represent the digitisation domains involved, namely agriculture, forestry and rural areas. The proposed method requires two main steps. First, conceptually independent expressions (i.e., *terms*) are identified. Then, contrastive analysis is applied to select the terms that are specific for the context of the document. The overall process includes the following four tasks.

1. POS Tagging: Part of Speech (POS) Tagging is performed with an English version of the tool in [Del09]. With POS Tagging, each word is associated with its grammatical category (*noun*, *verb*, *adjective*, etc.).

2. Linguistic Filters: after POS tagging, we select all those words or groups of words (referred in the following as *multi-words*) that follow a set of specific POS patterns (i.e., sequences of POS), that we consider relevant in our context. For example, we will not be interested in those multi-words that end with a preposition, while we are interested in multi-words with a format like <*adjective*, *noun*> (such as "wearable device").

3. C-NC Value: terms are finally identified and ranked by computing a "termhood" metric, called C-NC value [BDMV10]. This metric establishes how much a word or a multi-word is likely to be conceptually independent from the context in which it appears. The computation of the metric is rather complex, and the explanation of such computation is beyond the scope of this paper. The interested reader can refer to [BDMV10] for further details. After this analysis, we have a ranked list of words/multi-words that can be considered *terms*, together with their ranking according to the C-NC metric, and their frequency (i.e., number of occurrences). The more a word/multi-word is likely to be a *term*, the higher the ranking.

4. Contrastive Analysis: The previous step leads to a ranked list of terms where all the terms might be context-generic or -specific. With the contrastive analysis step, terms are re-ranked according to their context-specificity. This is done by comparing the extracted terms with the terms extracted from the Penn Treebank corpus, which collects articles from the Wall Street Journal. The final ranking is analysed by the LL moderators, and non-representative terms are discarded.

2.1.2 Identification of Relevant Relations

In order to identify relevant relations among terms, we first select all the terms extracted in the previous step. Then, we search for possible relations among such terms. We state that there is a relation between two terms if such terms appear in the same sentence or in neighboring sentences. In order to give a rank to such relation, we use the Log-likelihood metric for binomial distributions as defined in [Dun93]. The explanation of such metric is beyond the scope of this paper. Here, we give an idea of the spirit of the metric. Roughly, a relation holds between two terms if such terms frequently appear together. Moreover, the relation is stronger if the two terms do not often occur with other terms. In other words, there is a sort of *exclusive relation* among the two terms. The relevant terms and relations are used to produce a knowledge graph, which can be visualised with T2K.

2.2 Domain Scoping with Wikipedia Crawling

The taxonomy is then enriched by crawling the Wikipedia pages associated to technologies that are relevant for the specific domain. Specifically, the technology-related terms that appear in the taxonomy are searched in the Wikipedia pages, and the links to the pages are attached to the graph. This can help the participants of the LL to learn about those concepts that they are not familiar with. Furthermore, LL Moderators can look upon those concepts that do not have a Wikipedia definition, and add the links to other informative webpages on the topic. The software for Wikipedia crawling in DESIRA—currently under development—is freely accessible[7]. Such software supports automatic exploration and comparison of Wikipedia hierarchical categories. The graph is then exported with Sigma.js and can be visualised through a common web browser.

3 Conclusion and Research Plan

The DESIRA project started in June 2019, and, although most of the components of the toolchain, i.e., T2K, Gephi, Wikipedia Crawler and Sigma.js are available, we still have to test their integration for our purposes. In fact, we are carefully proceeding at this time, in order to evaluate the approach in itself. Furthermore, we are discussing how to ease the integration of the knowledge graph into the impact model, ultimately a key tool for LL workshops.

To this end, we are currently experimenting the usage of the tools for the generation of the domain-dependent taxonomies. If the output meets the expectations of the project, the taxonomies will be made accessible to the LL participants. Specifically, the participants will be able to navigate the taxonomies through a web browser, to learn about the different technologies, and to use the acquired knowledge within their LL. To validate the usefulness of the taxonomies, we plan to: (a) retrieve quantitative information in terms of number of accesses to the web pages associated to the taxonomies, and (b) gather qualitative feedback of the participants on the practical usefulness of these taxonomies within the project. This will give an indication of the applicability of the considered NLP technologies for knowledge extraction in the context of DESIRA.

Acknowledgements

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 818194.

References

[ACC+19] Massimiliano Assante, Leonardo Candela, Donatella Castelli, Roberto Cirillo, Gianpaolo Coro, Luca Frosini, Lucio Lelii, Francesco Mangiacrapa, Valentina Marioli, Pasquale Pagano, et al. The gcube system: delivering virtual research environments as-a-service. *Future Generation Computer Systems*, 95:445–453, 2019.

[BBF+19] Manlio Bacco, Paolo Barsocchi, Erina Ferro, Alberto Gotta, and Massimiliano Ruggeri. The Digitisation of Agriculture: a Survey of Research Activities on Smart Farming. *Array*, 3–4:1–11, 2019.

[BDMV10] Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. A contrastive approach to multi-word extraction from domain-specific corpora. In *Proc. of LREC'10*, pages 19–21, 2010.

[Del09] Felice Dell'Orletta. Ensemble system for part-of-speech tagging. In *Proc. of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, 2009.

[Dun93] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[DVCM14] Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. T2kˆ 2: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2062–2070, 2014.

[7] Public repository available at: `https://github.com/alessioferrari/DESIRA-WikiAnalysis-Repo`