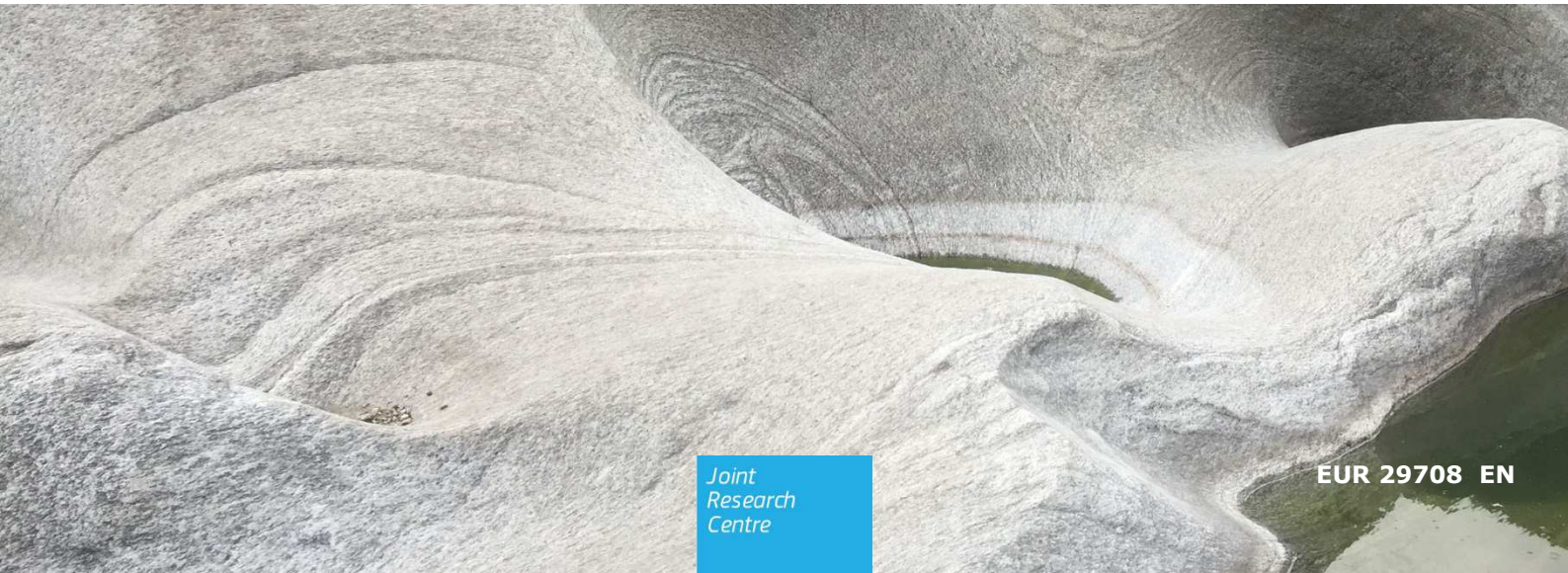
European Commission

JRC TECHNICAL REPORTS

Semantic Text Analysis tool: SeTA

Supporting analysts by applying advanced text mining techniques to large document collections

J. Hradec, N. Ostlaender, C. Macmillan, S. Acs, G. Listorti, R. Tomas, X. Arnes Novau



Joint Research Centre

EUR 29708 EN

EU Science Hub
https://ec.europa.eu/jrc

JRC116152

EUR 29708 EN

PDF ISBN 978-92-76-01518-5 ISSN 1831-9424 doi:10.2760/577814

Ispra: Publications Office of the European Union, 2019

How to cite this report: J. Hradec, N. Ostlaender, C. Macmillan, S. Acs, G. Listorti, R. Tomas, X. Arnes Novau , *Semantic Text Analysis Tool: SeTA*, EUR 29708 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-01518-5, doi:10.2760/577814, JRC116152

Contents

Acknowledgements

This digital assistant would never have materialised without people who invested their time and knowledge, and who patiently communicated their needs and use cases to the team. The close collaboration and knowledge sharing between units A.5, I.1, I.2, I.3, D.5 and B.6 made this possible.

The web application and underlying techniques were developed to satisfy concrete needs and use cases. The authors would like to thank especially Paul Smits, Jutta Thielen–del Pozo, Bettina Baruth, Marta Perez-Soba, Nicholas Nicholson, Massimo Craglia and Milan Kalaš of the Joint Research Centre (JRC) for their contributions that fuelled this work.

The system runs on JRC Earth Observation Data and Processing Platform (JEODPP) infrastructure. Open, friendly and forthcoming JEODPP experts enabled publication of our web application to the European Commission services.

The work of the European Publications Office in preparing and publishing the EUR-LEX and Bookshop collections was a key enabler for this project, and their support in facilitating access to these document collections is greatly appreciated. Requirements from other DGs and Services, namely DG DIGIT and the European Publications Office, helped making the tool much more policy and user relevant. We would like to thank them as well for their contributions.

Authors

Jiri Hradec, Nicole Ostlaender, Charles Macmillan, Szvetlana Acs, Giulia Listorti, Robert Tomas, Xavier Arnes Novau

Executive Summary

Much of the world's data is textual – in large document archives, in scientific papers, in scattered websites, in social media. The information contained in text is invaluable and yet hard to access. The sheer volume of text means that, unassisted, we cannot hope to read all available sources, nor even to keep up to date with all advances in a particular field. For example, EUR-Lex, the database of EU Legal texts, grows by over 15 000 texts per year[1] while Scopus, a database of scientific papers, has over 70 million entries.[2] The problems of scale are compounded by other challenges such as the breadth of topics covered, their jargon specific to each field and the changes in meanings of phrases over time.

The mission of the JRC is to provide scientific support to policy development, through original and applied research and knowledge management (JRC Strategy 2030). The challenges of accessing information "trapped in text" are very relevant to this mission of the JRC, as timely, relevant information is needed at all stages of the policy development process.

To help overcome the challenges posed by text the JRC has produced a new tool, SeTA – Semantic Text Analyser – which applies advanced text analysis techniques to large document collections, helping policy analysts to understand the concepts expressed in thousands of documents and to see in a visual manner the relationships between these concepts and their development over time.

A pilot version of this tool has been populated with hundreds of thousands of documents from EUR-Lex, the EU Bookshop and other sources, and used at the JRC in a number of policy-related use cases including impact assessment, the analysis of large data infrastructures, agri-environment measures and natural disasters. The document collection which have been used, the technical approach chosen and key use cases are described in this document.

[1] https://eur-lex.europa.eu/statistics/2016/eu-law-statistics.html, considering only EN versions of texts
[2] https://www.elsevier.com/solutions/scopus/how-scopus-works/content

1 Introduction

The famous quote that "We are drowning in information but starved for knowledge"[3] seems more relevant today than ever, as we struggle to keep pace with the continuing explosion of available data which is driving the information age.

Being an expert in any domain requires constant learning to keep up with the state of the art, and given the rate of new publications makes this challenging even within specific fields. But when we talk about horizon scanning, problem definition, and the design and implementation of impactful policies, we talk about the need to deepen cross-domain and cross-policy knowledge, to combine insights from interdependent fields, with possibly highly domain-specific concepts. The challenges clearly multiply, and analysts require support from automated tools to cope with and benefit from the vast volumes of available information.

Recent advances in machine learning and other applications of Artificial Intelligence (AI) have transformed the way we can analyse the world around us, allowing us to make sense of the information overload and to go way beyond the human limits of information processing. With carefully prepared data and the right choice of machine learning algorithm startling results are possible, and the challenge for the analyst becomes one of interpreting results, identifying patterns, formulating new questions and making good use of the results.

Because of the challenges given above, we believe that policy-making is one of the domains that can greatly benefit from these new approaches of "sense-making". This fits closely with the JRC mission of scientific support to policy, and so JRC has developed a pilot tool called SeTA to support policy analysis and development in any domain.

This tool uses recent developments in in big data, machine learning and especially in natural language processing.

The SeTA tool combines recent developments in big data, machine learning and natural language processing into a **knowledge exploration and recommendation engine** that supports policy analysts in:

- understanding concepts, their synonyms and the context in which they have been used in legislation across domains;
- understanding the temporal development of a term, its changing meaning and context over the last fifty years;
- centralised searching of the public knowledge corpus of the European Commission – EUR-Lex, the EU Bookshop including all technical reports, CORDIS, JRC PUBSY, EU Open Data Portal, etc.;
- using content similarity to discover the most closely related documents, duplicates or to see relevant documents through time; finding documents by its content similarity for discovery of most related documents, duplicates, or temporal cascading of relevant documents;
- transposing knowledge between domains, using a powerful and yet simple approach to answer questions suchs as "if the *WATER FRAMEWORK DIRECTIVE* is about *WATER*, how do we govern *WASTE*?", to which the results from the system would include "*eu waste framework directive*"," *waste framework directive*", "*revised waste framework directive*");
- direct discovery of most the common terms explaining relationships between two or more terms.

The tool was not designed to replace human analysts but to empower them, enabling them to achieve faster results with much deeper insight in a reproducible way.

[3] From John Naisbitt's book *Megatrends*,1982

Originally developed to support the Competence Centre on Modelling (CC-MOD) in 2017, the tool has already been extensively used to support analysts working in the area of legislative impact assessments for policy formulation, and other domains such as agri-environmental measures or natural disasters.

The tool has now been transferred to the Text Mining and Analysis Competence Centre (TMA-CC), and is provided as an easy to use web application.

2 Background

In this chapter we provide the relevant background information for this publication, namely: (1) the machine learning and text mining principles, methods and tools applied in this report, (2) the two European Competence Centres that are involved in this activity, namely the Text Mining and Analysis Competence Centre, and the Competence Centre on Modelling, and the relevant policy background, (3) the DigiTranScope project and (4) the EC knowledge management systems and document repositories used in this work.

2.1 Natural language processing and text mining

Natural language processing (NLP) is a research area focussed on how computers can be programmed to analyse human languages. Text mining is a complementary field, focussed on extracting useable, structured information from text. Techniques from both of these fields have been applied in the current work.

Some of the more common tasks in NLP include:

- Identification of sentence boundaries – detecting the start and end of each sentence in the text. This was complicated in our corpus as sentences were often split over several lines due to formatting.
- *Tokenization* – splitting the text into the different words, numbers, punctuation and other symbols. This task is rather complex due to identifiers such as (SWD(COM)2016 333), chemical formulas (benzo(a)pyrene) etc.
- *Part of speech (POS) tagging* – assignment of word role within the sentence such as designating a word as a noun or an adverb.
- *Parsing* – involves dividing the given sentence into related syntactic parts, which allows the relationships between different parts of the sentence to be explored and is indispensable for phrase identification.
- *Named Entity Recognition* – identification of entities such as persons, quantities, location and time within the documents.
- *Co-Reference resolution* – discovers the relationship of given word, often a pronoun, in a sentence with the previous and the next sentence, for example understanding that the word "it" in a sentence refers back to a previous mention of "policy measure". This is frequently used for example in question answering and definition mining.

The current research is focussed on the application of neural network models to text mining, rather than on traditional statistical text mining. Work in this area accelerated enormously with the publication of a key paper on word vector embedding by Tomas Mikolov of Google (Mikolov, 2012). Nevertheless, while phrase embedding has been known since 2013, a high-quality phrase extraction algorithm was missing to grasp the specificities of EU legal, technical and scientific language.

New scientific deep learning models appear every week yet engineering and practical usage for practical specific domain language modelling are sparse. The majority of applications are jealously protected and seldom open sourced. We share our application with all the Commission services in order to promote collaboration, re-use, and to learn from the needs.

2.2 The Text Mining and Analysis Competence Centre

The Text Mining and Analysis Competence Centre (TMA-CC) was created in late 2016, as part of a restructuring of the Joint Research Centre, to follow the JRC Strategy 2030 by bringing together expertise in text mining and make related services and tools available to the JRC and across the Commission. The rationale for the Competence Centre,

described on the JRC Science Hub,[4] is reproduced below, as it clearly shows how well the current tool fits with the aims of the TMA-CC.

Accurate, targeted, and timely information is needed by EU institutions at almost every stage of the decision making process. However, such data required by policy makers is increasingly embedded in large amounts of textual data available on the Internet, e.g. traditional or social media, or in large public or proprietary document sets. The sheer volume of this data makes it nearly impossible to extract the relevant information it contains manually. Text mining and analysis tools are necessary to address not only the problem of volume, but also of timeliness in order to provide the right information in the proper format for the decision making process, in a variety of contexts.

The number of application domains relevant to EU institutions (including the European Commission's directorates-general or DGs) where text mining and analysis (TMA) plays an important role is extensive e.g.: political current affairs media monitoring (DG Communication); targeted information for crisis rooms to improve EU's prevention, preparedness and response capabilities (DG European Civil Protection and Humanitarian Aid Operations (ECHO), European External Action Service (EEAS)); information used for security purposes (DG Migration and Home Affairs, DG Human Resources and Security); business intelligence based on framework proposals (research DGs and executive agencies); research and innovation monitoring (research DGs and agencies); monitoring of health related issues (DG Health and Food Safety, European Centre of Disease Prevention and Control (ECDC), European Food Safety Authority (EFSA)); monitoring of news in the financial sector (DG Financial Stability, Financial Services and Capital Markets Union, DG Economic and Financial Affairs).

Text mining techniques and tools are very much needed throughout the EU institutions but are highly specific and not directly accessible or useable by decision makers or policy domain experts supporting them. To use these tools and techniques so as to reliably provide decision makers with timely information requires a range of complementary skills: from analysis, through research and development of solutions based on computational linguistics, to deployment and operation of the systems, based on sound IT knowledge and practices. Each of these skills is necessary to accomplish the above. It is unlikely that small isolated groups could cover all these aspects or reach the required level of expertise.

The digital assistant SeTA is one of the projects spearheaded by TMA-CC to improve access and utility of policy analysts and policy designers to the complete COM document knowledgebase.

2.3 Cross domain policy applications in the context of the Competence Centre on Modelling

The text mining tool was **initially developed** to be used by the Competence Centre of Modelling (CCMOD) of the European Commission in support of the Better Regulation (BR) Agenda. The BR Agenda adopted by the EC in 2015 (European Commission, 2015) is recognized as a step forward towards a sound use of evidence for all policy making activities throughout the policy cycle (Broughel 2015; Radaelli 2018; Renda 2015).

The analysis of complex and cross cutting policy scenarios requires a thorough assessment of the interactions across domains and policies. Challenges faced by the EU then require finding effective solutions across policy areas, which, in turn, calls for integrated working practices finding synergies between the various activities. The EC proposes a new approach for **collaborative working to gather, manage, share and preserve data, information and knowledge**. In this context, one of the key areas for action is maximising the use of data for better policy making (Communication to the Commission: Data, Information and Knowledge Management at the European Commission (SWD(2016), 333 final)).

[4] https://ec.europa.eu/jrc/en/text-mining-and-analysis

Impact assessments (IA) of the EC are an interesting case study in this respect. According to the BR guidelines, the 'impact assessment process is about gathering and analysing evidence to support policymaking' […] 'Impact assessment promotes more informed decision-making and contributes to better regulation' (European Commission, 2017:15). In IAs, a robust analysis of economic, social and environmental impacts has to be performed. IAs have a long history dating back to 2002. With the BR Agenda, their use has been further extended and consolidated (EPRS, 2015). Their quality has also improved over time (EP cited in Golberg, 2018). However, some remaining issues are also reported. Notably, there should be more transparency on data, assumptions, methodology and results (RegWatchEurope cited in Golberg 2018; Impact Assessment Institute ,2017).

In the analysis of the impacts of the various policy options, the BR Guidelines recommend quantifying costs and benefits to the extent possible (European Commission 2017). In this respect, the EC makes extensive use of **simulation models**[5] throughout the policy cycle. Models can be used to assess the environmental, economic, and social impacts of policies. The CCMOD promotes a responsible, coherent and transparent use of modelling to support the evidence base for EU policies, and pools the Commission's competencies and best practices in building and using models[6].

CC-MOD is responsible for the development and management of **MIDAS**[7], the Commission-wide modelling inventory and knowledge management tool for modelling. MIDAS enables enhanced transparency and traceability of models in use for EC policy making. MIDAS supports a proper documentation, use, and reuse of models: it provides access to related data, modelling exercises, past and ongoing policy contributions and related publications. MIDAS also captures the links and dependencies between models, and offers a set of interactive tools for analysis, reporting and interactive design.

The initial development of the text mining tool stem from the need to answer a few simple questions related to model documentation in MIDAS: which models were used by the European Commission in legislative Impact Assessments in the past? Which Impact Assessments made use of modelling results from these models? Additional questions followed that were of a broader nature, related to the use of models in other phases of the policy cycle, the combined use of models for answering specific policy questions, and the development and use of models in the context of projects funded by the EU's framework programmes for research and innovation. Until today the Knowledge base forms one of the tools that the CCMOD is using to support their various activities.

2.4 Digital transformation and the governance of the society

The JRC Centre for Advanced Studies hosts the project DigiTranScope,[8] researching the governance of digitally transformed human societies.

This project aims to provide a deep understanding of digital transformation, to help policymakers address the challenges facing European society over the next decades.

The principal research themes are:

---

5 A model can be defined as an analytical representation or quantification of a real-world system, used to make projections or to assess the behaviour of the system under specified conditions.

6 CC-MOD contributes to the implementation of the Better Regulation policy, the Inter-Institutional Agreement on Better Law Making, and the Communication on the Management of Data, Information, and Knowledge at the Commission. Information on the Competence Centre is available at https://ec.europa.eu/jrc/en/modelling

7 Modelling Inventory and Knowledge Management System of the European Commission MIDAS can be accessed by **EC services** at http://midas.jrc.cec.eu.int. Starting in 2017, MIDAS is a tool of the Better Regulation Toolbox, and from 2019 onwards parts of the system are open to the European Parliamentary Research Service under the umbrella of the Inter-Institutional Agreement of Better Law-Making. This version is accessible to EC services and the Parliament at https://webgate.ec.testa.eu/midas-ii/ .

8 https://ec.europa.eu/jrc/communities/en/community/1286/about

- To explore the changing flows, ownership, quality and implications of digitised data and information. Data constitute the decisive ingredient of the transformed society, and data ownership, access, sharing, analysis and dissemination (for example in social, economic and political contexts) will interplay in uncertain ways as they have to date;
- To identify the key policy challenges relating to massive interconnection (Internet of Things, IOT) and the associated opportunities and risks;
- To determine what skills are needed to live fulfilling and healthy lives in a digitally transformed society, and to explore how to offer all citizens the opportunity to develop these skills;
- To explore innovative forms of governance for Europe leveraging the characteristics of digital transformation.

The project is designed to bring together a high-level group of stakeholders from across all sectors, to hear from the latest thinking, research, and commercial developments.

The project is researching how classical concept of regulation, which creates overlapping clouds of obligations, will fare against a concept of personalised governance, where entities (natural and legal persons) have profiles containing all their duties and data directly attached.

Natural language processing is a crucial tool to analyse the policy framework and project the identified duties and obligations to personal level.

2.5 EC knowledge management systems and document repositories used in this work

Hundreds of thousands of documents have been collected from different sources and analysed by the SeTA tool, and these sources are briefly described here.

The following systems and repositories are under the responsibility of the *Publications Office of the European Union* (OP) which publishes and disseminates the publications of the institutions and other bodies of the European Union[9]:

- EUR-Lex[10]: provides free access to The Official Journal of the European Union, EU case law and other resources for EU law, with documents dating back as far as 1951.
- EU Publications[11]: The EU publications website provides access to reports, studies, information booklets, magazines and other publications from the EU institutions and other bodies. This was formerly known as the EU Bookshop.
- CORDIS[12]: CORDIS stands for *Community Research and Development Information Service*. It is the European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation (Framework Programme (FP)1 to Horizon 2020).
- EU ODP[13]: EU ODP stands for EU Open Data Portal, a catalogue of datasets from the EU institutions and other bodies.

Some of the content can be accessed through the CELLAR, as the common repository of content managed by the Publications Office. The CELLAR contains the files and the metadata of various collections of documents, which can be fetched using the machine-readable SPARQL endpoint facility or the HTTP RESTful web services. The CELLAR is powered by semantic technology and enables direct access to information stored as Linked Data.

[9] https://publications.europa.eu
[10] https://eur-lex.europa.eu
[11] https://publications.europa.eu/en/web/general-publications/publications
[12] https://cordis.europa.eu/
[13] http://data.europa.eu/euodp/en/home

Another resource used is the JRC Publications Repository, PUBSY[14]. PUBSY is an online service giving access to data about research publications produced by the European Commission's Joint Research Centre. It was established to assist with central storage, management and search to our research publications that go beyond the official publications stored in the EU Bookshop. PUBSY has a native API.

[14] http://publications.jrc.ec.europa.eu/

3 Methodology

In the following chapter we describe the two distinct steps we have taken for the creation of the knowledge base: document collection, cleaning and storage in the first step and the actual text analysis and modelling in the step two.

3.1 Corpus preparation

3.1.1 The corpus creation

The corpus of the public policy-related European documents counts more than 500.000 documents, coming from the following sources described in some detail in chapter 2: EUR-LEX, EU Publications, CORDIS, the EU Open Data Portal, and JRC PUBSY.

The auxiliary document repository also contains documents from:

- Full copy of Wikipedia dump (6M articles, 13GB)
- 39 million open source scientific articles that were collected by the Allen Institute for Artificial Intelligence[15]
- A full copy (590.000 documents) of the US federal legislation since 1994[16]

Metadata in various formats are harvested from the document sources, harmonised into a common format and sent to a RabbitMQ pipeline for further processing. The actual document content is downloaded and processed in later steps. This approach, processing metadata independently from the document contents, is a key prerequisite for building a unified corpus of documents as this source abstraction layer permits a wide variety of sources to be ingested.

All the texts collected are in English only (except for some older legal texts where multiple languages are interleaved on the same page). The reasons for this decision are:

1. English language sentence dependency parsing is rather straightforward and there are several open source semantic parsers with excellent tuning for this language.
2. We are interested in extracting knowledge from plain text and as the translation of, for example, a directive into all EU languages does not create new knowledge, processing a single language well should capture the available information.

3.1.2 Document cleansing pipeline

Since there is a variety of web service document retrieval end-points and there are often more than one document per metadata record, a document harvesting process was established as the first part of the whole pipeline. Documents are retrieved through SPARQL, SOAP, OAI-PHM, FTP or HTTP protocol parsing, even though every source has its own specificities.

Transparency, repeatability and focus are the key features of any analysis and the results can only be as good as the input data. Data cleansing was an iterative learning process, and new needs for data cleaning were discovered at several points only after the neural network was trained.

The majority of sources contain information in several formats but these formats are usually unsuitable for direct and immediate information processing. All documents must be harvested and every source has its own way of being accessed and metadata formats it produces. After metadata have been harvested and interpreted, the documents must be downloaded and processed. At this stage we meet many challenges specific to automated document processing. Features designed to help human readers often add great complexity to automated processing. For example:

[15] http://labs.semanticscholar.org/corpus/
[16] https://www.govinfo.gov/bulkdata

Header and footer text – although visually separate from the main text of the document, to automated systems these are found mixed together with the text;

Numbers – stripped of formatting information such as superscripts it is hard to tell the meaning of a number found in a stream of text which could be a page number, a reference, a count, a phone number etc;

Multilinguality – legal texts from the 1960s often have multiple languages interleaved on the same page, meaning that the stream of text processed by automated systems changes language frequently.

The fully automated and repeatable data cleaning mechanism, developed over two years is not perfect and is being constantly improved. We keep learning the needs and requirements of neural networks to produce quality results. The typical process to create a general corpus involves:

- Conversion from original formats (PDF, HTML, XML, MSWord, …) to plain text
- Conversion to Unicode (often not easy), removal of text conversion artefacts, removal of non-alphanumeric characters, transposition of diacritics to ascii characters, spacing enforcement.
- De-hyphenation (a rather critical step)
- Sentence separation based on dependency parsing (allowing the reconstruction even of sentences split over several lines)

For repeatability of analysis it is important that the data collection and cleansing pipeline works with guaranteed behaviour and so the process is fully automated. Adding another corpus is very easy since the general cleansing pipeline has already been implemented, metadata are parsed through abstraction layer for consolidation and processing is very swift

The output of this step is a document repository containing the completely cleaned unified plain text, divided into sentences. This new document structure is stored within an ElasticSearch (ES) database which allows searching and reproduction for human readers. The original text is also stored in the same database.

3.1.3 The document repository

The document repository, which contains the text harvested and cleaned in the previous steps, is already of great value. The full text of all documents can be searched through a simple interface, and users are able to target their search either to the individual document collections or to search across all collections in a harmonised way.

The future evolution of the technical solution chosen, an ES database, is under consideration. One possibility is to make a central repository for documents from across the many units at the JRC, providing simple and direct access to their text.

Directly useful to human users, the database itself also enables the next steps of automated processing, by making carefully cleaned text available to the machine learning algorithms which are described below.

3.2 Neural networks training

This is the pivotal point of the whole analytical process. Neural networks can learn any function and only data availability defines how complex the function can be. Therefore, the data preparation, feature engineering and domain coverage become essential elements for obtaining meaningful and analysable results from neural network training.

The EC public knowledge corpus sports rather consistent language and thus the features could have been created from phrases instead of words like in general language.

3.2.1 Phrase compositionality

This is possibly the single most critical step where textual features in the form of phrases have been engineered. Many months were spent attempting to create "the best" phrase engine with these results:

- Word ngrams are useless in a corpus with fixed vocabulary and complicated phrases
    - "and the" is the most frequent bigram
    - We can learn better quality phrases through 2-4 iterations but will not catch longer phrases when the EC phrases often exceed this size and larger number of iterations are both highly computationally demanding and becoming erroneous
- Dependency tagging (e.g. JJ*NN* for a chain of adjectives and nouns) misses many important phrases
    - "regulatory framework" or "member states" are the typical phrases identified by dependency tagging
- We hoped that the Google Text-Rank algorithm would be able to catch compound phrases
    - It can correctly identify "legal and regulatory environment" but not "area of freedom, security and justice"
- Dependency parsing of a noun phrase as produced by the Stanford CoreNLP[17] java engine or better spaCy[18] python library can provide high quality dependency tagging and identify the noun phrase correctly, but catch many artefacts on the way
    - "the same member states" or even "the following 5 member states" must be iteratively cleansed to extract the "member states" and to get rid of the other textual information that does not provide added value and distracts the network.
- Iteratively cleaned noun phrases created from noun dependency trees can then produce even very high-quality phrases:
    - "Intellectual, industrial and commercial property rights"
    - "African, Caribbean and Pacific (ACP) group of states"

All together there are about 1 million phrases with a frequency higher than 50 out of 26 million identified cleaned phrases.

In the next step we harmonise the different variants of common phrases by taking the most commonly occurring variant as canonical, and using it to replace other variants. For example, the variants "real time", "real-time" and "realtime" are all replaced with "real-time", as this has the highest frequency in our corpus. Without this step, the network will become a spell-checker: there are literally tens of thousands of occurrences of "sustain-able development", "sustainable develop-ment", etc. The neural network will produce those variants for the similarity search instead of "environmental sustainability", "sustainable growth" and "socio-economic development".

Once the final set of phrases has been created, the separate words constituting each phrase are combined into a single token by replacing the spaces separating them with underscores. In this way the phrase "member state" becomes the token "member_state". This step increases the quality of the trained networks, as the meaning of the phrase is encoded separately from the meanings of its constituent words.

The implementation of these steps is performed by extracting titles, abstracts and identified sentences from the whole corpus, identifying, harmonising and replacing phrases and then storing as a text file outside of the ES database. An idea of the scale is given by the size of this text file, around 23 GB. For analysis of term development by

[17] https://stanfordnlp.github.io/CoreNLP/
[18] https://spacy.io/

(half)decades, separate text files were created for 1950s-1980s, 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2010-2014, 2015-2018.

3.2.2 Final text preparation

As the last step before the neural network training is further normalisation:

The only character allowed in the text are a-z, 0-9, /, -, _, (space)

All words not containing at least one character a-z are removed.

This general corpus now contains 7 billion words and phrases, about 80 million sentences and 23 GB of plain text.

3.2.3 Actual neural network training

We used a powerful python library for language modelling called gensim[19] for the neural network training.

Currently there are three topologies in use:

- Skip-gram Word2Vec for similarity queries
- Continuous Bag of Words (CBOW) with sub-word information (FastText, Bojanowski 2016) is used for calculations in the vector space
- Doc2Vec DM (distributed memory, Le and Mikolov, 2014) for document similarity calculations.

The training of one network using Word2Vec topology takes about 20 hours on a fast Linux workstation with 36 physical cores and plenty of RAM. Models persisted to hard disk take about 5GB.

The training of one network in FastText topology takes about 80 hours on the same system and the models are also of about 5GB in size.

The training of one network in Doc2Vec topology takes about 8 hours on the same workstation and the models are of about 25GB in size

3.2.4 Verification and accuracy tests

The trained model networks now represent the billions of words and phrases found in the corpus as short mathematical vectors. Derived from the positions of the words in the input corpus, these vectors not only represent each of the words, but also capture some of the meaning of each word, found from the context of surrounding words: in the famous quote from Firth, "you shall know a word by the company it keeps." This approach means that the meaning captured for each word is strongly dependent on the corpus used for training.

Each of the trained networks is tested against the same similarity test which was designed to match the characteristics of our large, single language corpus.

The first tests involve basic similarity queries to reflect our corpus: the concept most similar to "cap" is not "hat" but "common_agricultural_policy". "wfd" results in both water_framework_directive and waste_ framework_directive, "eee" in "electrical_and_electronic_equipment", "wwtp" in "waste_water_treatment_plant" but also in "wwtps" and "municipal_sewage".

The next test utilises an interesting property of the vector representation of word meanings: it is possible to perform mathematical calculations with the vectors with interesting results. The standard example used to demonstrate this, for a large, general purpose corpus of English, is to calculate the effect of taking the word vector for "king", subtracting that for "man", and adding the vector for "woman": the result is found to be

[19] https://radimrehurek.com/gensim/

(very close to) the word vector for "queen". This approach is exploited in our text analysis, for example to discover directives in particular fields. For our particular corpus, which mostly contains legal texts, we find that the word vector calculation for king – man + woman does not give queen, but the closest phrases we find are "education officer", "member of parliament" and "immigration officer". Rather than reflecting a gender bias, this (and many other examples tested) seems to reflect the lack of gender information in our corpus of technical and legal texts. The gender vector in our vector space is therefore non-representative.

This highlights an important point: we are dealing with scientific and technical reports and legal texts and their language bears completely different information from general text. The analyst must be aware of this focus when analysing the content. As we will demonstrate in next chapters, while king-man+woman does not obviously work, waste_framework_directive – waste + water truly results in water_framework_directive.

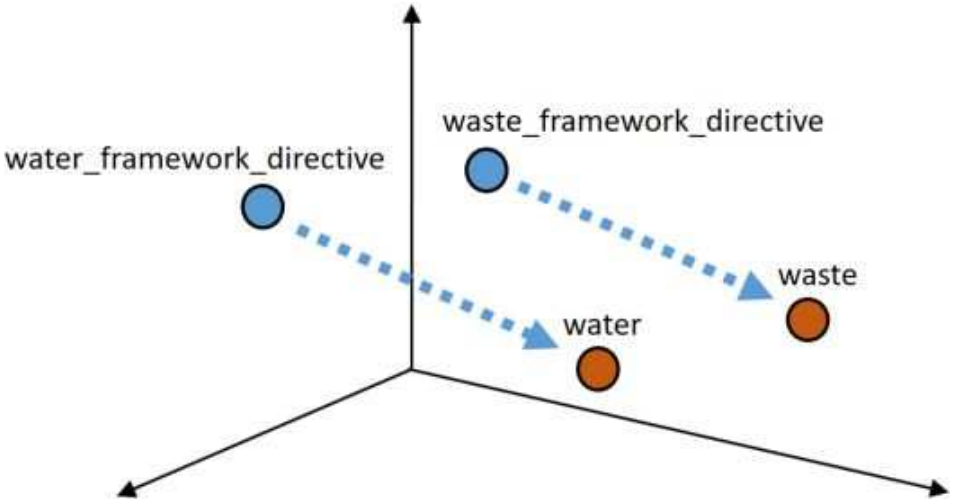


Figure 1: Similarity of meaning vectors in our vector space

4 Discussion, verification and interpretation of results

Each trained network results in a table of 300 columns where the state of the neural network's weights is being stored, and 1-2 million rows where every word/phrase has its own row. This python numpy[20] table is an excellent start for further interpretations, and we found many.

4.1 Findings with Word2Vec

There is a state-of-the-art method (until early 2018) of creating language models called *Neural Word Embedding*[21]. The principle is to feed to recurrent neural network patterns of words (e.g. a word/phrase and a window of two preceding and two successive words/phrases in the sentence to grasp the context). From billions of examples the neural network will learn the distribution function ("France to Paris" is the vector of the same length and direction as "Estonia to Tallinn", only with a different origin). As of early 2019, there are several shallow network topologies and tens to hundreds of deep neural networks. Shallow (or "not deep") does not necessarily mean bad or useless, on specialised corpora they tend to be rather powerful as we will show it in the following chapters.

Word2vec (Mikolov, 2003) is a two layer neural network that processes text. Its inputs are text split into sentences and it produces a set of word vectors as output. These methods are shallow networks that trade expressivity for efficiency. This neural network, consisting typically of one-hot encoder for every word in the corpus, feeds data into hidden layer (we used 300 neurons). The output layer is typically hierarchical softmax to preserve low cardinality words. The whole corpus, when split into sentences, can be fed into the neural network to train rules by which words emerge in the corpus. These rules are encoded into a 300-dimensional spherical vector space normalised to -1..1. Since we use approx. 2.5 million unique words and phrases, this vector space is stored as a simple matrix 2,500,000x300 for fast calculations.

The greatest problem of Word2Vec is that it does not capture meaning related to word order. The game changer for us was when we encoded phrases alongside simple words. While in general English the number of identified phrases would quickly skyrocket, our specialised legal, scientific and technical phrasal dictionary is rather limited. Since the EC has a relatively stable language with many fixed phrases (e.g. Water Framework Directive), the key challenge in our work was to discover these phrases directly from the plain corpus and not to rely on existing ontologies that were created for human use and must, therefore, be very limited in scope.

Our phrase identification technique stems from sentence dependency parsing, where for every noun the complete dependency tree has been calculated (see chapter 3.2.1). Therefore, we can embed phrases and word to encode differentiated meaning of words and phrases – words "artificial" and "intelligence" have very different meaning vectors from the phrase "artificial intelligence". And the results give fascinating insights into the whole knowledge base of the Commission without a need for any tailored algorithm.

We have successfully employed FastText rather than Word2Vec for tasks involving vector calculations, because the addition of sub-word information (character n-grams) provided higher quality representations.

While general English needs the new (post-2017) types of deep embeddings that were started with ELMo (Peters et al, 2018), for our corpus Word2Vec has a huge advantage in speed of training and unimportant decrease in quality of the encoding.

[20] http://www.numpy.org/

[21] Tomas Mikolov et al: Efficient Estimation of Word Representations in Vector Space, Tomas Mikolov et al: Distributed Representations of Words and Phrases and their Compositionality.

Word2Vec has many limitations, undifferentiating bag-of-words input to be the most important one. We tested sequential models and new embeddings (ELMo[22], OpenAI GPT/GPT-2[23], BERT[24], ULMFiT[25] etc.) to train networks, which was highly computationally intensive, but did not bring added value in our line of work. These language models change the way the word is represented and solve polysemy by encoding complete context. "go" can be both verb and old Japan game and word2vec suffers to encode both meanings if they are not equally represented in text. But our analytics is based on word representation and these novel methods are not applicable. One of the future planned features is natural questions answering and this is where e.g. BERT become indispensable.

New methods are popping up weekly with a lot of promises but the word2vec methods we use show a robustness which is much needed for transparent policy related analyses. These methods have relatively low computational demands. They do not require GPU for training and results can be interpreted even on a laptop.

4.2 Ground-truth ontology

The most important finding was that the skip-gram neural networks tend to generalize the concepts the further we go hopping over similarity clusters.

Manually created ontologies today provide an excellent means for understanding corpora. But the approach followed here employs machine learning to automatically analyse 100% of the words in each corpus. This means that we longer need high quality, manual document tagging because we can extract relations directly from the corpus and analyse documents this way.

What happens is that as we walk through the network of terms, we keep discovering relations that may help navigate our user through complicated terms that need explanation. When we construct a network of terms, we get the picture below showing the context of the term good governance:

[22] https://allennlp.org/elmo
[23] https://blog.openai.com/better-language-models/
[24] https://arxiv.org/abs/1810.04805
[25] https://arxiv.org/abs/1801.06146

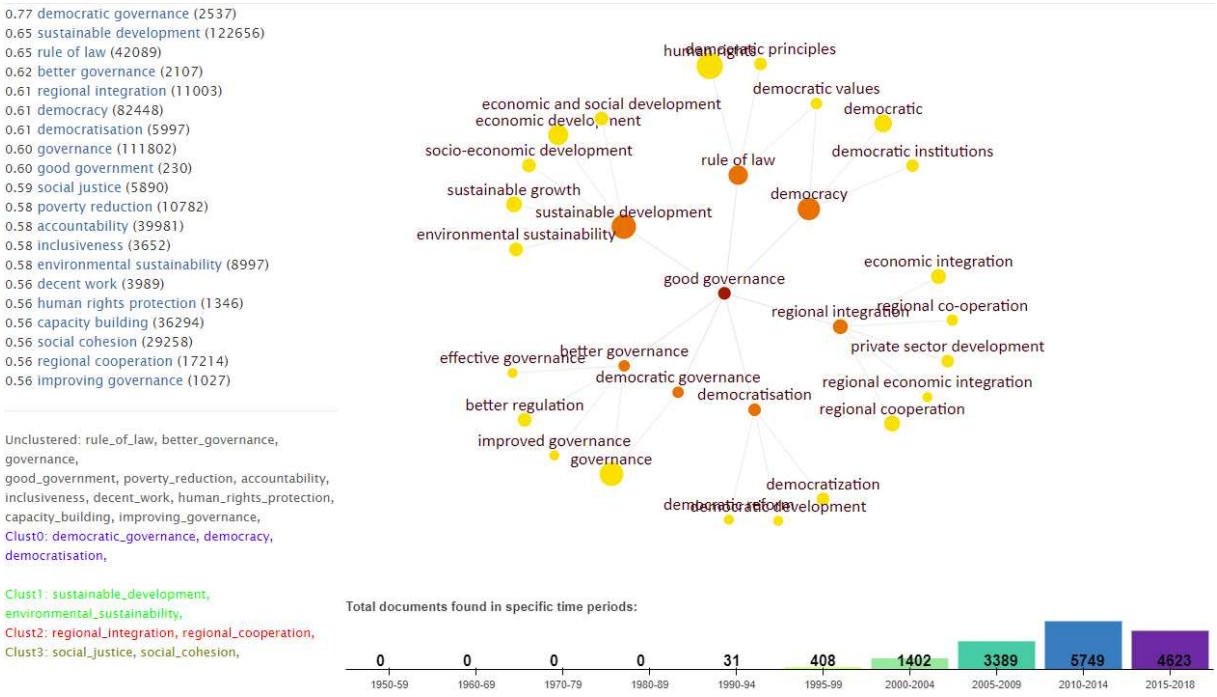


Figure 2: The context of the term good governance

In this picture we can see several automated analytical outputs. Top-left is the ranking of terms as provided by term similarity including term cardinality. Below are the same terms but grouped into several clusters using the DBSCAN clustering algorithm (EPS=0.3, min_samples=2, metric="cosine", algorithm="brute"). The lower right section shows occurrence of the terms over half-decades. And at the upper right side we can see the network graph constructed from word similarities. In most cases this network provides a very good decomposition of the chosen term. In this case, the network shows that good governance is closely linked to democratic governance, sustainable development, rule of law, etc. Technically, this network represents how the European institutions typically describe the meaning of good governance.

4.3 Term development

Another domain where neural networks provide interesting insight is when we train skip-gram networks by decades and build the "ontologies" around them. This allows us to see is how the term changed context over time. A very nice example is shown below for the phrase "impact assessments".

Table 1: Term development of the phrase Impact Assessment using skip-gram networks

1980s	Messy (low frequency) context: drafting, standards, new approach, coordinated
1990-1994	Environmental audits, emergency actions, risk assessments, assessments
1995-1999	Assessments, environmental impacts, impact studies, environmental assessment
2000-2004	Environmental assessments, assessments, impact assessment, policy proposals
2005-2009	New legislative proposals, better regulation, stakeholder consultations, policy proposals
2010-2014	Impact assessment, public consultation
2015-2018	Impact assessment, stakeholder consultations, fitness check, ex-post evaluation

When we train the same for CBOW instead, we get this:

Table 2: Term development of the phrase Impact Assessment using CBOW

1990-1994	Workplace assessment, risk assessments, nuclear emergency, dose reconstruction
1995-1999	Environmental assessments, ecological impacts, social and economic impacts, assessments
2000-2004	Evaluations, assessments, recommendations, impact assessment
2005-2009	Ex-ante evaluations, new legislative proposals, policy proposals, impact assessment
2010-2014	Impact assessment, cost-benefit analyses, policy proposals
2015-2018	Ex-ante impact assessments, fitness checks, impact assessment

We can see that both topologies identify years after 2000 as the time when EC has started implementing impact assessments in the policy cycle. Before 2000 we can see environmental impact assessments dominating the decade. While CBOW still follows other uses of impact assessments (workplace, nuclear, pharmaceutical), skip-gram sees 1990s as the decade of the environment.

Similarly, we can see word INSPIRE as verb up to 2005 and then it changed context to geospatial data, metadata and services.

One of the most fascinating results can be seen for sustainable development:

Table 3: Term development using the term Sustainable development

1990-1994	Sustainable growth, shared responsibility, environmental policy, environmental objectives
1995-1999	Sustainable urban development, environmental sustainability, social development
2000-2004	Sustainable, sustainability
2005-2009	Sustainable transport, sustainable development, sustainability
2010-2014	Environmental sustainability, socio-economic development, green economy
2015-2018	Green growth, socio-economic development, economic development, environmental sustainability

4.4 Shared term spaces

If in the 300-dimensional space vector products can produce analogy (concept-context 1-context 2 -> result), possibly the path between the two terms can contain terms that could explain the relation between these two terms. It seems it works when we intersect the spherical spaces around each term and consider only terms in a tube around the connecting line to limit the terms too far from the connecting line.

The results are encouraging and relatively robust.

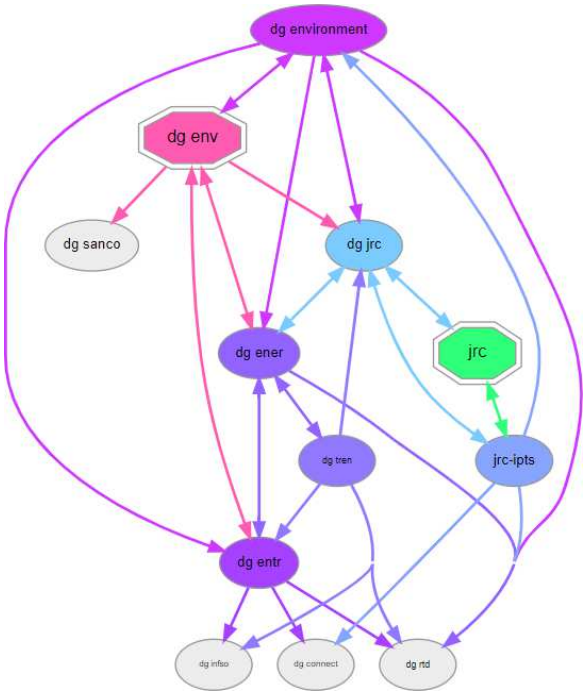


Figure 3: Exploration of densely populated vector space – linking "JRC" to "DG ENV"

The link between terms JRC and DG_ENV is very short – there is only one term directly connecting both – DG_JRC. It is both similar to JRC because it is just a synonym, and it

is also similar to DG_ENV, because it is also DG and there is vivid collaboration between these two DGs making them closer in the corpus.

But there are other DGs mentioned as well. The colourful ones are terms within the cylinder around the line connecting these two terms, the grey ones are outside but connected. The other DGs are obviously semantically highly similar but are more faraway from the connecting tube.

In two dimensions the cut would look like this. Only the red dashed area of the vector space has been used:

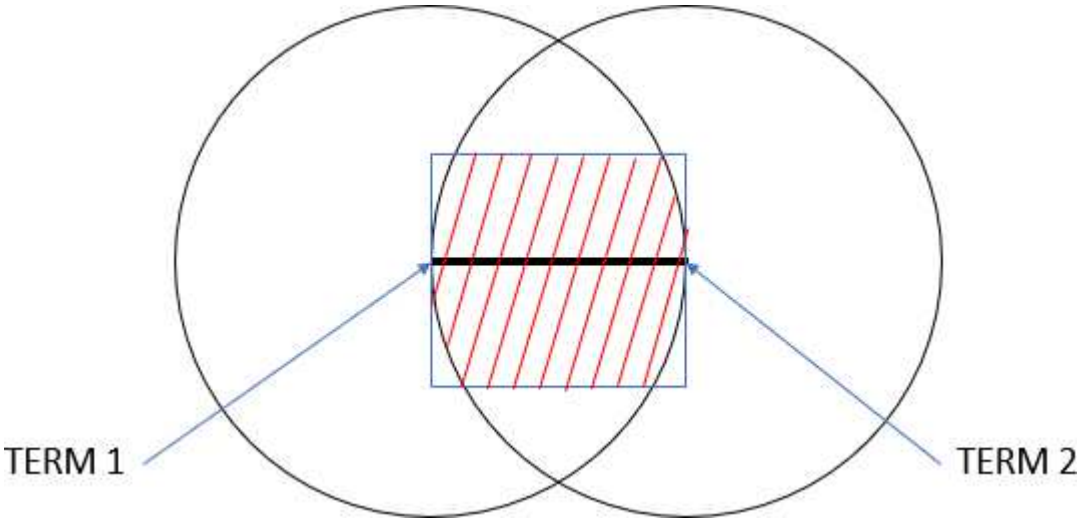


Figure 4: Geometric explanation of the shared space algorithm

The idea for this analysis is that on the connecting line going through multidimensional space there are terms that are relevant to both terms and can serve as a kind of explainer. The situation where we are cutting tubular space through 300 dimensions helps avoid terms that would be included in 2D or 3D but 300D space allows much more precise cut:

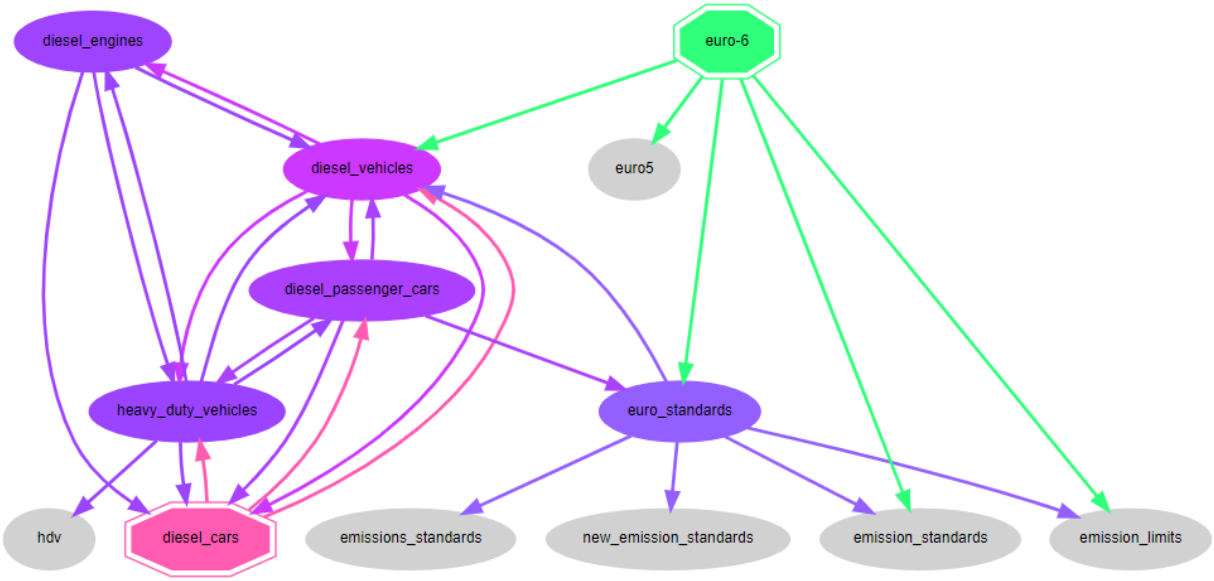


Figure 5: Cutting tubular space through 300 dimensions – Euro6/Diesel cars

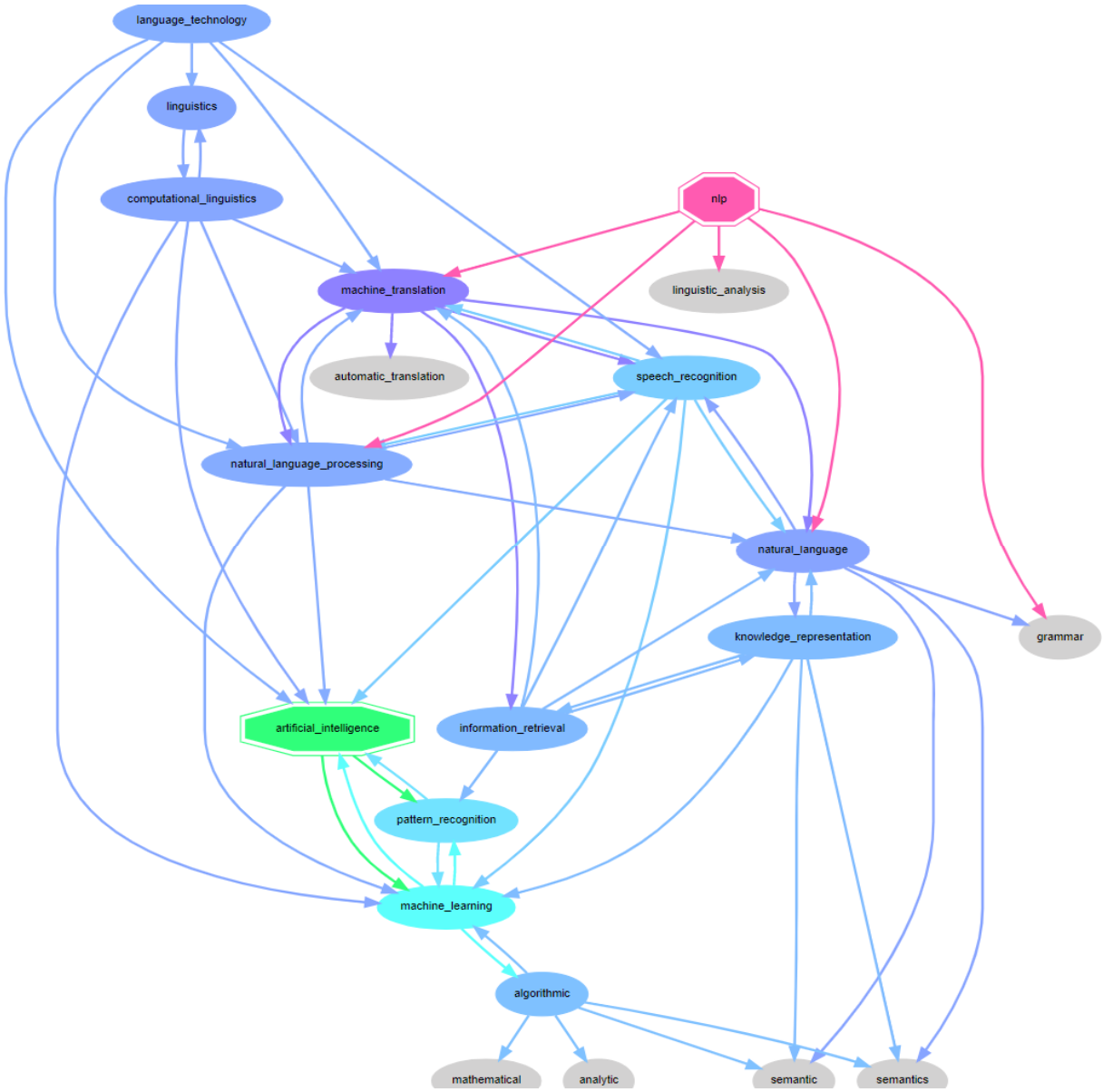


Figure 6: Cutting tubular space through 300 dimensions - the AI/NLP example

4.5 Toying with graph networks

As we have shown above, the trick of converting word similarity to a graph network helps building network explaining the terms. But this process is computationally rather intensive and can create only small networks in real time of tens of nodes.

Therefore we have generated complete networks from all topologies to see how they will behave when we apply standard network analysis. The pictures below show comparison of how behaves a network created from the cut-out vector space and what information we can extract when we apply shortest path algorithm to the generated network. This way the shortest path finds terms that can be connected through term similarities and not necessarily lie in the same space shared by both terms:

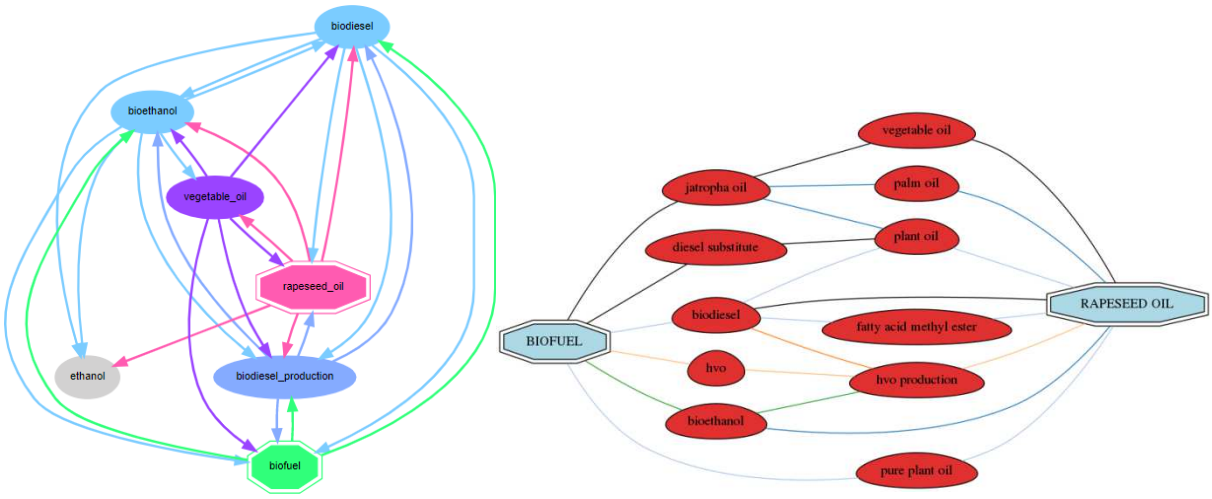


Figure 7: Comparison of shared terms from vectors space and from network of similarities

We can see that moving from singular to plural, i.e. slight shift of the cylindrical cut out in the vector space, affects which and how many terms are found. On the other hand, the shortest path will extract a different path as well:

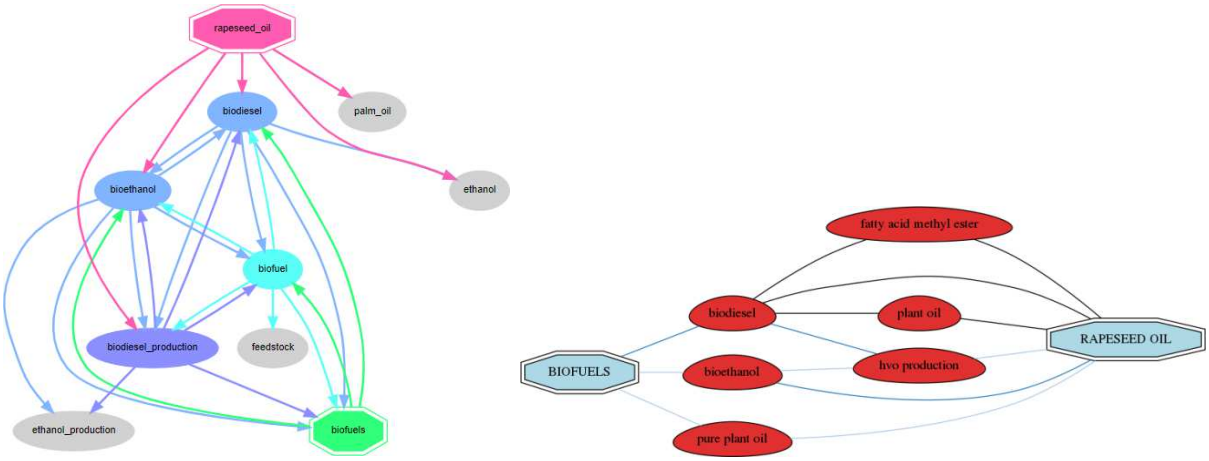


Figure 8: Comparison of shared terms from vectors space and from network of similarities – impact of semantic alteration (plural)

But as we can see, the shortest path in the left picture goes from rapeseed oil to biodiesel and then to biofuels, which is exactly what this substance is used for. The right picture shows many more relevant nodes.

There are many other examples where we can try to separate the true value the network generates from haphazard connections that emerged as a result of a huge complexity:

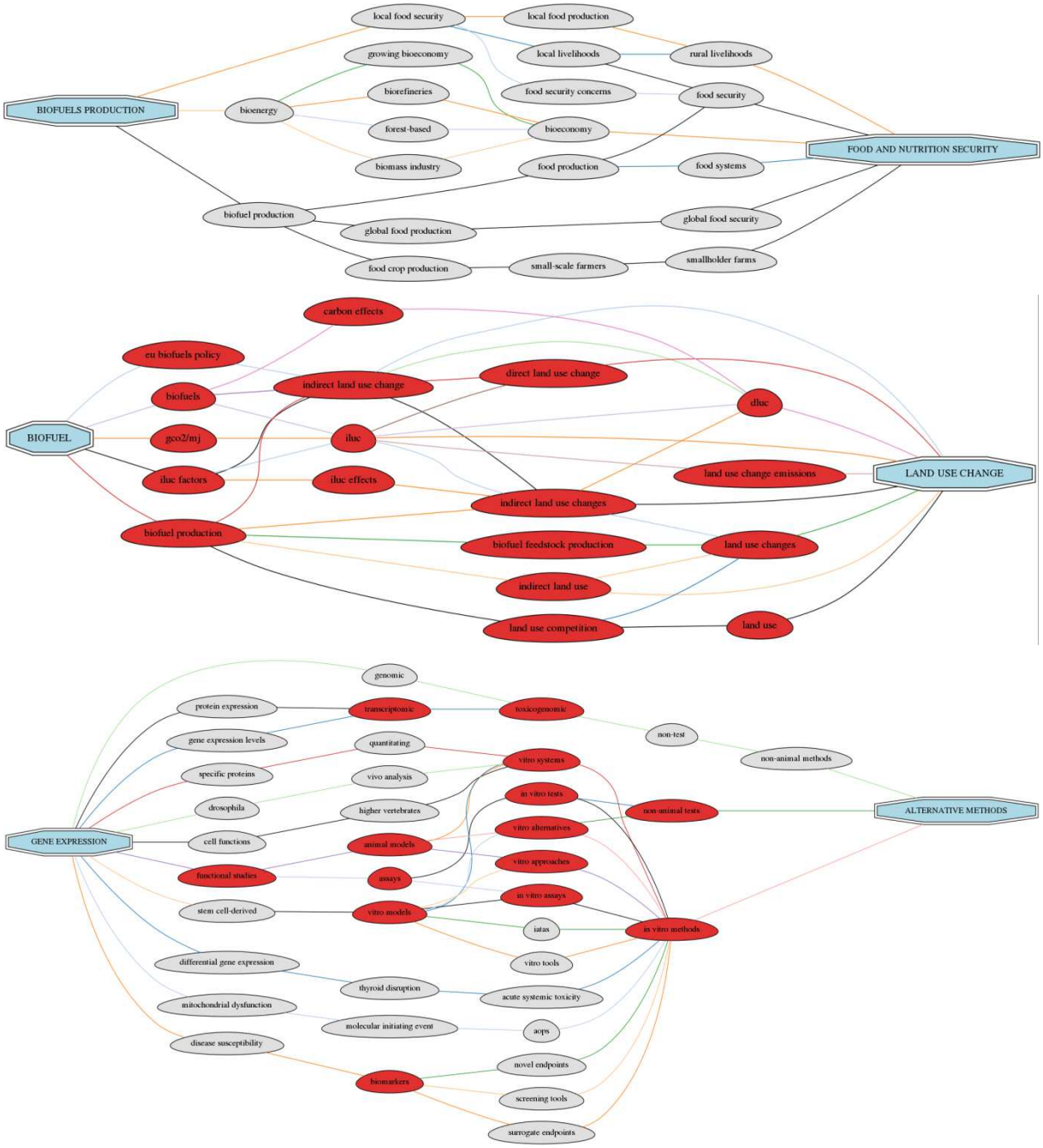


Figure 9: Extraction of shared concepts explaining causality between two policy concepts

The idea here is that word similarities in a complete graph network can still provide very good insight.

Unfortunately, this process is very computationally intensive. The creation of the network takes about 35 hours on a Linux workstation with 36 cores and the complete analysis takes some 20 minutes. Since the word similarity graph is connected, it is important to run the search in both directions. We have shifted the trade-off between too much and too little towards too much and analyse all the four network topologies – cbow, skip-gram, both either with or without hierarchical softmax.

4.6 Document networks

This is very simple doc2vec algorithm where the major speedup of training (5x) was achieved by training only on first 200KB of plain text in each document. Also, all documents were cleaned and the phrases replaced.

Same graph as for term was created for the documents in our corpus. This way we can see the most similar documents in the central graph, the same graph but with vertical split by date of publishing and list of documents found on the right. Hovering mouse of node provides the document tile, abstract, identifier and the date of publishing.

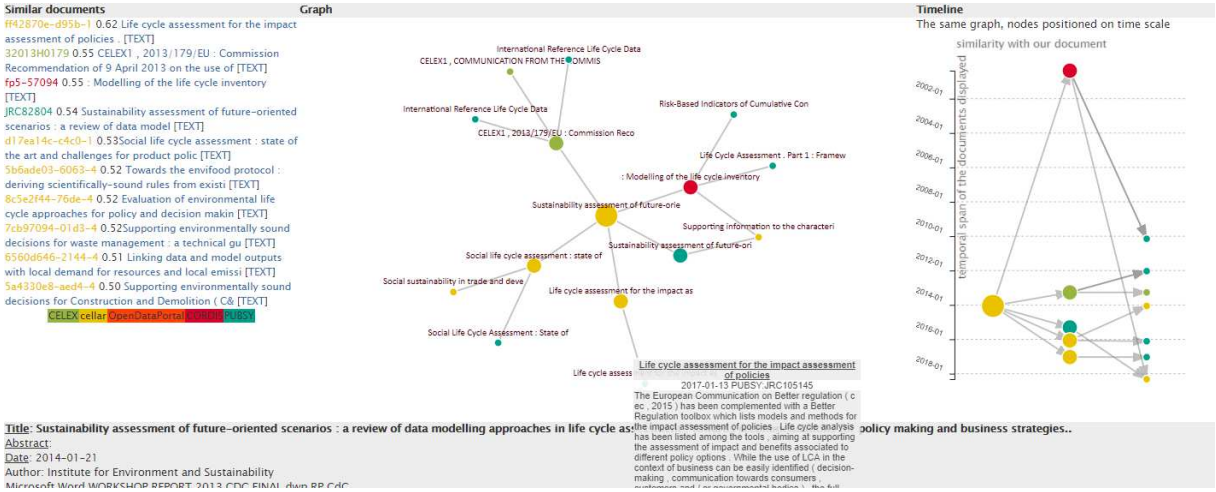


Figure 10: Example of a temporal-thematic document network

5 Policy applications

The tool was thoroughly tested and applied in policy use cases and under real-life conditions for about two years. Areas of application included impact assessments for policy formulation, the INSPIRE Directive, agri-environmental measures or natural disasters, and personalised policies. In this chapter we will illustrate some of these use cases to demonstrate advanced uses of the trained neural networks.

Some instances go beyond the capabilities of the existing web application, and required programming to obtain the desired results..

5.1 Use of models in Impact Assessments

5.1.1 Context and question

The Competence Centre on Modelling (CC-MOD) supports a proper documentation, use, and reuse of models. CC-MOD is responsible for the development and management of **MIDAS**[26], the Commission-wide modelling inventory and knowledge management tool for modelling. MIDAS enables enhanced transparency and traceability of models in use for EC policy making. MIDAS puts a particular focus on the use of models in EC **Impact Assessments** (IAs), where the former are used, for example, to assess the environmental, economic, and social impacts of policies. Such use has become increasingly important to support EU policy making throughout the policy cycle. For this reason, a better understanding of how models have been used in the past decades can then lead to improving the efficiency of model development and use.

5.1.2 Method

The present analysis focuses on the assessment of models used in EC Impact Assessments from 2003 until 2018. In order to achieve this, we applied a two-step methodology: 1) identify those Impact Assessments that mention models using the knowledge base, followed by 2) post-processing investigating the role of these models in the IAs. On the basis of these findings, we draw up and apply criteria for models to be considered for further analysis.

Step 1 was carried out using the knowledge base, to identify both the Impact Assessments and the models they mention; we also detected the models themselves:

— The cleaned text was searched, starting with a set of acronyms of models, which we knew had been used in Impact Assessments (e.g. through investigations carried out by *Petrov et al* and through entries in MIDAS by JRC modellers). However, in many cases the search was yielding words that were not models, simply because the used model acronyms also represent common words (e.g. GAINS, IMPACT, G2, SMART, etc). The workaround was a frequency-based noun chunk identification (e.g. phrase "GAINS model" can be found in the text more than 50 times). Searching the texts for complex search strings was programmed as a customised *Aho Corasick Algorithm* (Aho & Corasick, 1975) to keep the search time linear with the number of keywords.

— Word disambiguation (e.g. how to distinguish MAGNET the model from MAGNET the ferromagnetic material) is still under development. Sub-word neural embedding (AdaGram) yields the best results so far, but it requires a substantially larger corpus (several gigabytes of text) to obtain reasonable results, see box 2 on **Disambiguation**.

— New models were discovered using neural word embedding. Deep recurrent neural network allowed clustering similar word vectors (eg. PRIMES -> E3ME, GREEN-X,

[26] Modelling Inventory and Knowledge Management System of the European Commission MIDAS can be accessed by **EC services** at http://midas.jrc.cec.eu.int . Starting in 2017, MIDAS is a tool of the Better Regulation Toolbox, and from 2019 onwards parts of the system are open to the European Parliamentary Research Service under the umbrella of the Inter-Institutional Agreement of Better Law-Making. This version is accessible to EC services and the Parliament at https://webgate.ec.testa.eu/midas-ii/ .

GEM-E3, TREMOVE). This technique helped identify models used in the texts without any prior knowledge of their existence.

— A specific task was the utilisation of the neural word embedding to discover typos in the text of the IA reports, (e.g. for GEM-E3 we got GEME3, GEM E3 and GEME 3, which are distinguishable only thanks to the same word context).

— Understanding what the model does was critical for this analysis. In this case, the developed Subject Verb-Object Analysis was a game-changer, see box 1 on **Subject-verb-object analysis** below.

— Statistics and graph analysis of occurrences (frequency of model use in IA) and co-occurrences (how often model names are mentioned together in IA) was provided, which helped the post-processing follow-up of text mining results.

Box 1: Subject-verb-object analysis

Approach: Subject-verb-object analysis

A typical example was the need to understand what a computational model, which has just been identified, can be used for. First, we ran a search through the ES database for the word FIDELIO. We got circa 500 results out of 75.000.000 paragraphs.

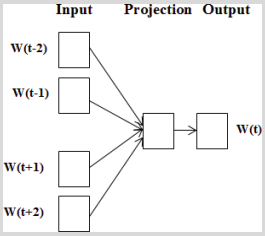
An analysis of dependency parsing where the word FIDELIO was in the subject and the lemma of the root verb was BE yielded six results (three shown):

```
PUBSY:JRC81864:1 FIDELIO is appropriate for the impact assessment purposes
of diverse ( economic and/or environmental ) policy questions of our times


cellar:6bba687f-ffe9-11e6-8a35-01aa75ed71a1:8 FIDELIO is a more powerful
and flexible ( hence , realistic ) model for policy impact assessment
purposes .


cellar:a20ded17-9ef9-4a90-a9d4-2d6302a92eca:191 FIDELIO is a demand-driven
model
```

Box 2: Disambiguation

Approach: Disambiguation

Computational models that are so important for policy anticipation and planning are often identified by acronyms that are the same as common words (e.g. EVE, LUISA, GAINS, RAINS, G2, MAGNET, …). The ability to understand which terms refer to actual models and which bear a different meaning was tested using the CBOW topology.



The advantage of CBOW is that it encodes context terms to help predict the meaning of each term. Therefore, by considering the words surrounding e.g. MAGNET, we were able to tell, by using a very simple Bayesian classifier which occurrence of MAGNET points to a model. When the CBOW network returned terms like 'impact assessment' and 'cost benefit analysis', we knew we were probably dealing with a model, while words like 'neodymium', 'magnetic', 'attractive force', etc. identified a ferromagnetic material.

To our surprise we often found other categories, such as 'touristic magnet'.

5.1.3 Results

The **results of text mining** gave us a good overview of all IAs where models were mentioned, the frequency of these mentions in the IAs and the co-occurrences of different models mentioned in the same IA. These results underwent further **post-processing** to ensure that the list of identified IAs and the models they referred to was complete, to eliminate false positives (i.e. to cross-check whether the model name mentioned in the IA was really a model), and to understand the role of the model in the IA.

Detailed results are covered in a separate publication (Acs et al., 2019).

5.1.4 Conclusion

Using the knowledge base was pivotal for the work in MIDAS, and allowed to complete an inventory and an analysis of models used for Impact Assessments since their first introduction in the EC policy formulation in 2002.

The conclusions are covered in a separate publication (Acs et al., 2019).

5.2 Understanding policy processes for EC Impact Assessments

5.2.1 Context and question

As part of the support to the Better Regulation Agenda, the Competence Centre on Modelling is supporting in particular the Impact Assessment process (see chapter 2). In this context, the competence centre performed an analysis of the evolution of the term Impact Assessments over some decades. The aim was to pick up signals that would allow us to analyse the existing situation and propose improvements for processes related to impact assessments.

5.2.2 Method

The neural network was used to investigate the term 'impact assessment', using an algorithm that is capable of providing approximate components. In other words, it describes what an impact assessment, in the perspective of the Commission, consists of. The method itself has already been discussed in chapter 4. In order to see the evolution of the term, we split the analysis into decades, and later on into 5-year periods, when enough documents became available.

5.2.3 Results

The term first appeared sporadically in the 1960s, linked to impacts of natural disasters. In 1980 it started to be correlated with assessments of impacts of political decisions, which however were not yet formalised. In the 1990s, we see the emergence (and dominance) of environment impact assessments as the most frequent form of assessment. The term as we know it today emerged in the years following 2000, coinciding with the decision to make impact assessments mandatory for all policy decisions. The higher the frequency of use of the term in the knowledge base, the better the results of the neural network. Occurrence was low until the early 1990s, but has been rising ever since. This should be kept in mind when viewing the results.

Below we compare two graphs illustrating the viewpoints of the components of Impact Assessments for two different half-decades. We projected into a graphical representation of a network for two levels of word similarities, resulting in two levels of nodes showing how the meaning is gravitating towards the term in question.

Figure 11 shows the results of the neural network for the years 1990 to 1994. The frequency of the term was low, however it already provides some meaningful results The various nodes that are returned show three different types of components that are directly linked to the main term: 1) thematic components, such as environmental impacts, water resources, and environmental planning; 2) methodological components such as risk assessment; and 3) what might be described as procedural component,s such as scenario development and monitoring. Further components, e.g. methods such as sensitivity analysis and risk analysis, become visible in the second set of nodes.

Figure 11: Impact Assessment 1990 – 1995, frequency 206

For the most recent years 2015 to 2018 (Figure 12), the frequency of the term was much higher. If we compare the two results, they differ widely. The terms now reflect almost exclusively procedural aspects as well as tools and bodies related to these procedures, such as regulatory scrutiny board, staff working document and impact assessment guidelines. Any relation to thematic components, such as environmental, social or economic impacts, is not visible in the nodes. Even second order nodes still mostly show procedural aspects. The only node that remains visible that is not directly related to procedural aspects is the one of external studies and contractors, which implies that a majority of IAs is carried out by external capacities.

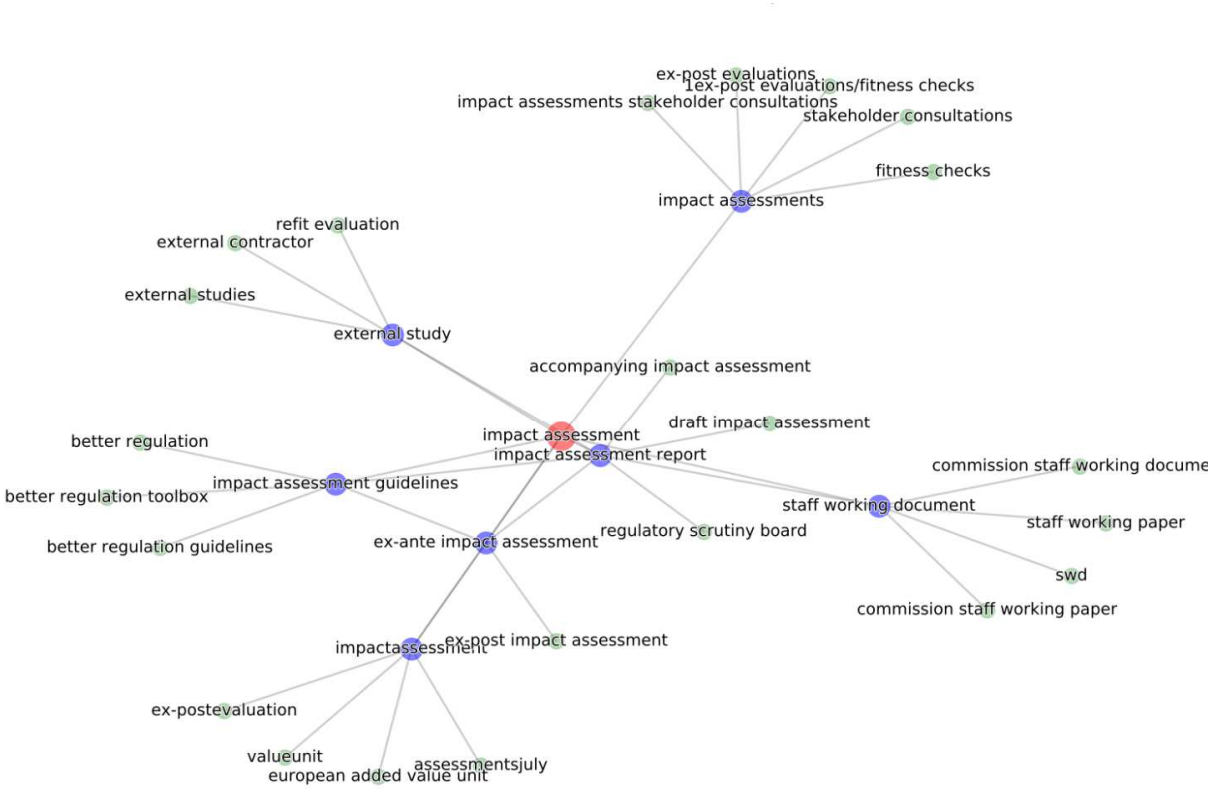


Figure 12: Impact Assessment 2015 – 2018, frequency 21890

5.2.4 Conclusion

A possible interpretation of these results is that the focus of the Commission moved away from impact assessment themes and methodology towards procedural components, which can also be interpreted as a more harmonised approach. As the majority of mentions in the knowledge base have this focus, a large amount of work has been put into the procedures, with direct implications for their application by experts and the interpretation by lay people. Developing a team of impact assessment specialists that is able to guide, harmonise, and apply impact assessment procedures and methodologies seems like a great added value that is certainly called for. The Commission also needs to be mindful that it is outsourcing a large part of the knowledge linked to impact assessments, and that it needs to reflect on its capacity to do impact assessments in-house, to ensure aspects of transparency, traceability and reproducibility.

5.3 Enhancing the policy relevance of INSPIRE Directive data resources for the European Commission and the Member States

The new INSPIRE Geoportal, officially launched on 18.9. 2018 during the INSPIRE Annual Conference in Antwerp, serves as a one-stop shop for public authorities, business and citizens to discover access and use of ca. 160.000 national geospatial data sets related to the environment in Europe. The Geoportal represents a key objective of the INSPIRE Directive[27] that requires the Commission to establish a community geoportal as an access point to the Member States infrastructures through network services.

One of the important components of the system is metadata provided for MS data sets, as well as network services. The Implementing rules together with the Technical Guidelines that provide the necessary set of legal / technical requirements and recommendations to make the whole distributed EU system interoperable have been in force since 2014.

Despite the efforts made by hundreds of national data providers to describe their relevant data sources by metadata, the full potential has not yet been reached. For instance, the ability to search for a thematic domain or policy-relevant aspects e.g. "show me all the data sets related to waste management" is not directly possible. There are three major reasons for that:

1) Data providers are understandably reluctant to create metadata unless a clear use case has been identified, and revisiting collections which have already been classified is time-consuming and expensive. This limits the completeness of manual metadata, particularly as new use cases and requirements appear over time.
2) The INSPIRE Metadata profile is rather a set of technical and administrative elements: with the exception of INSPIRE data themes, it contains thematic /domain classifiers. The INSPIRE data scope is represented by 34 data themes and metadata both for data sets and the network services have to be associated to at least one theme.
3) Multilingualism of INSPIRE Metadata – the metadata have been provided in the national language; only for some we have collected also the English translations authorised by the data provider.

5.3.1 Context and question

The scope was to investigate the possibility to apply text mining and machine learning methods for increasing the policy relevance of MS data sources by adding a new analytical layer facilitating thematic / policy relevant data categorization.

As a concrete example, DG Environment (INSPIRE lead DG) selected the newly created list of priority data sets[28] for environmental reporting. In our case, we wanted to identify in the EU Geoportal all the data sets related to Natura 2000.

5.3.2 Method

Large-scale data infrastructure semantic analysis (2 million metadata records) has already been covered under the example of GEOSS (Craglia, Hradec, 2018), and subject-

[27] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

[28] https://webgate.ec.europa.eu/fpfis/wikis/display/InspireMIG/Action+2016.5%3A+Priority+list+of+datasets+for+e-Reporting

origin-time policy analysis was used for the 200 000 metadata records of INSPIRE Infrastructure (Hradec, Lima, Tomas, 2017). Here we demonstrate the deep extraction of data sets relevant to a specific policy, a task that is most often requested by policy analysts.

A. Finding the relevant terms from the neural network pre-trained on legal documents and technical reports

1. We have started by identifying the related keywords visually using the SeTA web application. The first term used was 'NATURA 2000 sites'.
2. Later we kept browsing the neural ontology to identify all terms relevant to our search.

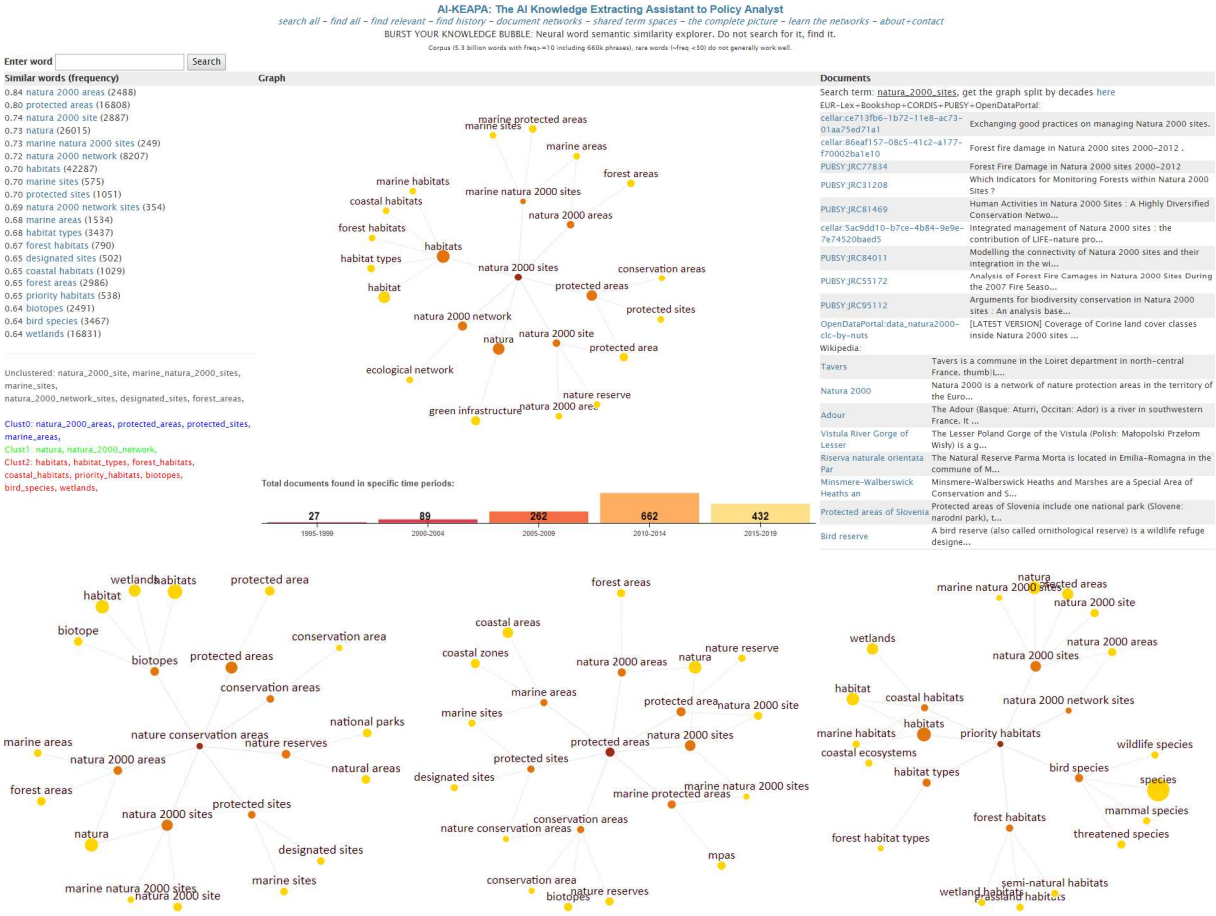


Figure 13: Ontology extracted from the neural embedding used to find all terms relevant to Natura

B. Term extraction from metadata

1. We translated titles and abstracts of all harvested metadata records (i.e. 160 000 records) into English to create a uniform text base ready for further processing using the EC Machine Translation.
2. We used the newly created neural network trained on the EU legislation and all technical reports present in EUR-Lex and EU bookshop to search for different spelling of the keyword 'NATURA 2000' or for the use of similar keywords. As a result, we obtained a data scope of the possibilities of the use of the NATURA 2000 keyword.
3. A specialised small neural network (Word2Vec) was trained just on the metadata records. As a result, we identified easily by word similarity different names for the

same concept, e.g. "Natura 2000" is the same as n2k, n2000, natura, natura2000, but also "natura 200" etc.

4. The results from the previous step were compared with the simultaneous direct manual search of the EU Geoportal for Natura 2000 related data sets using the combination of the term Natura 2000 and its lingual equivalents translations. The search was also extended by title, ID, colloquial naming of two relevant Directives (Habitats and Birds) that also regulate the Natura 2000 sites.
5. Concrete data mining algorithms linking to the ID of individual data records (SGRank, n-grams=1-6, idf=on, window=7) were applied to extract the most relevant key phrases.
6. The whole process was repeated several times (iterations) in order to obtain the best available results, i.e. the list of data sets related to Natura 2000. These results were also compared with free additional keywords used by some data providers to identify the scope of their data sets.

5.3.3 Results and preliminary conclusions

— It is possible to semi-automatically (several round of iterations) create a new thematic layer, e.g. policy domain classification, of huge EU-wide data source (160 000 data sets records) based on the actual data content.

— It is planned to enhance the analytical, matching capabilities of neural network by adding the documents from the INSPIRE knowledge base not published in EUR-Lex.

— The test provided valuable feedback to the data providers regarding the quality and consistency of their metadata descriptions.

— It was proven that the primary source of information about each of the data set was the abstract that can be mined with ML techniques to often achieve more consistent results compared to human classification, e.g INSPIRE themes keyword.

— When creating the base corpus it is necessary to add other available relevant sources e.g. documents from the INPSIRE knowledge base[29] not published in EUR-Lex.

Box 3: Analogy and cross-domain inference of domain knowledge

Approach: Analogy and cross-domain inference of domain knowledge

Vector space where we store the knowledge learnt by the neural network can be used to extract knowledge from other domains. Below are a few examples where this technique is actually rather useful, especially for domain experts who are to collaborate with experts from other domains.

While the distance "Concept – Scope 1" defines the vector of analyst's knowledge ("INSPIRE Directive" actually IS a "directive"), the shift "Scope 1 – Scope 2" defines the question (what "data" are most relevant to the "INSPIRE Directive"?).

These analytical results confirm that vector space analytical results are useful for knowledge extraction. They can directly serve as input for the complex analysis of documents

[29] https://inspire.ec.europa.eu/

Table 4: Vector space analogy to extrapolate analyst domain knowledge to other domains

Concept	Scope 1	Scope 2	Result
natura_2000_sites	**natura**	**birds**	**breeding_sites** **breeding_bird_species** **foraging_habitats**
european_flood_alert_system	**flood**	**forest_fire**	**european_forest_fire_information_system** **european_forest_fires_information_system** **forest_fire_information_system**
european_flood_alert_system	**flood**	**forest_fire**	**european_forest_fire_information_system** **european_forest_fires_information_system** **forest_fire_information_system**
gem-e3_model	**model**	**directive**	**2012/27/eu_directive** **f-gas_directive** **energy_taxation_directive**
inspire_directive	**directive**	**data**	**resource_data** **spatial_data** **datasets**
agri-environmental_measures	**agriculture**	**air_pollution**	**air_pollution_abatement** **air_pollution_policies** **air_pollution_measurements**

5.4 Mapping Dual-use goods and technologies

The history of nuclear, chemical, biological, missile-related proliferation has shown how the piecemeal illicit procurement of dual-use[30] strategic goods and technologies, rather than of turn-key facilities, has become a key concern in the development of competences and capabilities for the development of weapons of mass destruction, particularly since the 80's and 90's.

In a broader context, strategic, or "dual-use", are items with both civil and military applications, including systems, equipment, components, materials, software and technologies for manufacturing, aerospace, electronics, chemical, biological and medical, nuclear, telecommunications, cyber-security, marine, navigation, avionics, laser applications, energy production, human rights protection and many other applications.

The evolution of technologies and processes allows discovering new areas and opportunities (New Evolving and Emerging Technologies[31]), but may also bring about more challenges to the non-proliferation framework, which therefore needs to adjust and evolve to address them, while safeguarding research and legally authorised trade and exchanges.

In 2009 the EU adopted a new regulation setting the Community regime for the control of exports, transfer, brokering and transit of dual-use items[32]. Annex I to this Regulation lists all the export controls agreed upon in key multilateral export control regimes (such as the Australia Group, the Chemical Weapons Convention, the Wassenaar Arrangement, the Nuclear Suppliers Group and the Missile Technology Control Regime) and it is annually amended by means of a Commission Delegated Act[33].

5.4.1 Context and purpose

We investigate the possibility to apply text mining and machine learning methods to find connections and interrelations between dual-use items and technologies, adding a new analytical layer facilitating thematic / policy relevant data categorization. The links would gain even more relevance when the dual-use items are linked to the potential proliferation of Weapons of Mass Destruction.

The purpose of this research study is to assess the functionality of the SeTA tool applied on dual-use goods and technologies, for getting information about activities related to dual-use (scientific publications, conferences, patents, projects) performed by a range of actors (States, organisations, companies, universities, research centres, etc.).

A more ambitious goal would be to map the items controlled in Annex 1 of the aforementioned Regulation, also known as the "EU dual-use control list", and the new emerging or evolving technologies in the dual-use area.

Mapping the activities related to these goods will make it possible to monitor developments and trends of both controlled items and emerging technologies. It is important to make this knowledge available in the most efficient way to adapt policies to technological changes, to ensure that the right goods are being controlled and to make

[30] Art. 2.1 of the Council Regulation (EC) No 428/2009 of 5 May 2009 defines dual-use goods such as: "items, including software and technology, which can be used for both civil and military purposes, and shall include all goods which can be used for both non-explosive uses and assisting in any way in the manufacture of nuclear weapons or other nuclear explosive devices".

[31] Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. Research Policy, 44(10), pp. 1827-1843.

[32] Council Regulation (EC) No 428/2009 of 5 May 2009 setting up a Community regime for the control of exports, transfer, brokering and transit of dual-use items, OJ L 134, 29/5/2009, p. 1–269.

[33] The one currently in force is the Commission Delegated Regulation (EU) 2018/1922 of 10 October 2018 amending Council Regulation (EC) No 428/2009 setting up a Community regime for the control of exports, transfer, brokering and transit of dual-use items, OJ L 319, 14.12.2018, p. 1–252.

export control actors aware of certain intelligible technology transfers (such as academic collaborations or participation to common projects).

5.4.2 Outcome

Several searches were conducted employing specific dual-use keywords with the text mining tool.

As a first example, the conducted query using the keywords "dual-use" and "WMD" (Weapons of Mass Destruction), provides the following diagram.

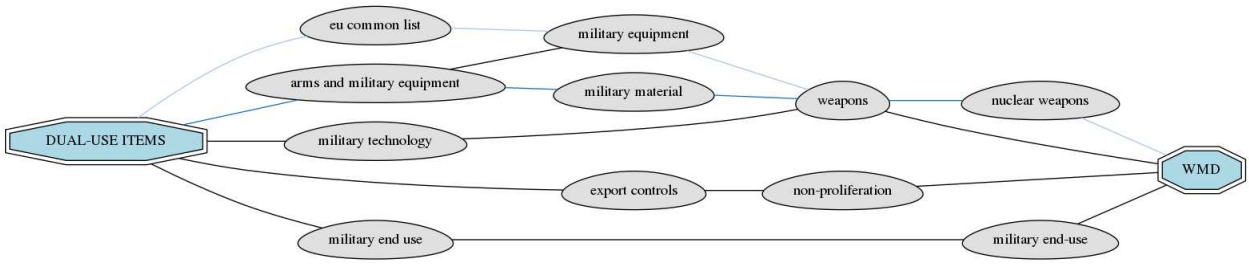


Figure 14: Ontology extracted from the neural embedding used to find all terms relevant to Dual-use items and Weapons of Mass Destruction (WMD)

The text mining tool links the two keywords, 'dual-use' and 'WMD', to the most common keywords related to them, such as non-proliferation, nuclear weapons, military equipment, EU common control list or export controls. Therefore, the usefulness of this tool is, in first instance, demonstrated, since the provided results match perfectly with the jargon used in the field of dual-use and non-proliferation.

A second example focuses on assessing how a new emerging technology is linked to dual-use.

A new emerging technology might pose a risk for the proliferation of WMD, however, in some cases this type of technology is not included in the items control list of the EU regulation for the time being. The query about 'artificial intelligence' as a new emerging technology and dual-use items provides a large amount of related keywords (see figure 15), which can be classified in four groups:

a) Already controlled technologies by EU dual-use regulation such as robotics, cyber-surveillance technology, image processing, unmanned aerial vehicles (UAV), cryptography, image processing, or intangible transfers.

b) Non currently controlled technologies: biometrics, Online Analytical Processing (OLAP), social network analysis, knowledge engineering, intelligent machines, etc.

c) Common keywords used in the field of dual-use and export control: control list, military use, brokering, missile technology control regime, etc.

d) Other keywords unrelated to dual-use: European food research, ICT implants, etc.

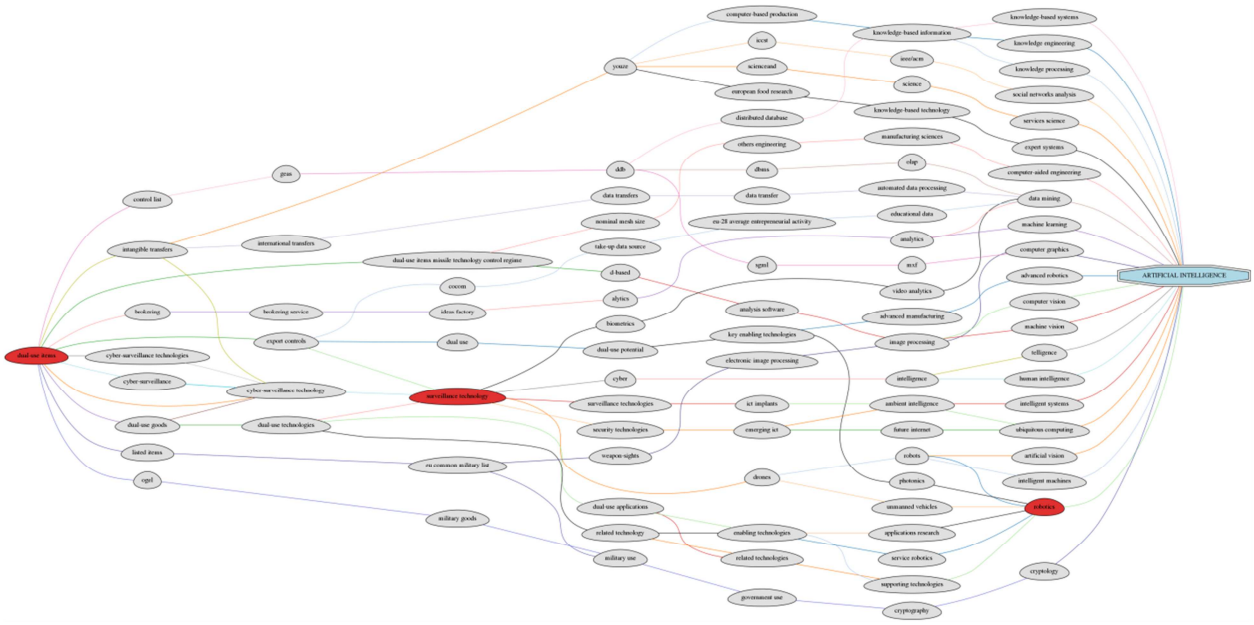


Figure 15: Ontology extracted from the neural embedding used to find all terms relevant to Artificial Intelligence and Dual-use

A third example looks at linking dual-use items with certain destination countries like Syria, where it has been proved that the import of dual-use items led to the production of WMD, which later were used against the citizens. The related keywords we found were: weapons, arms supplies, ballistic missiles, technical assistance and brokering services, export controls, dual-use exports, and also a list of countries like Iraq, Iran, and North Korea.

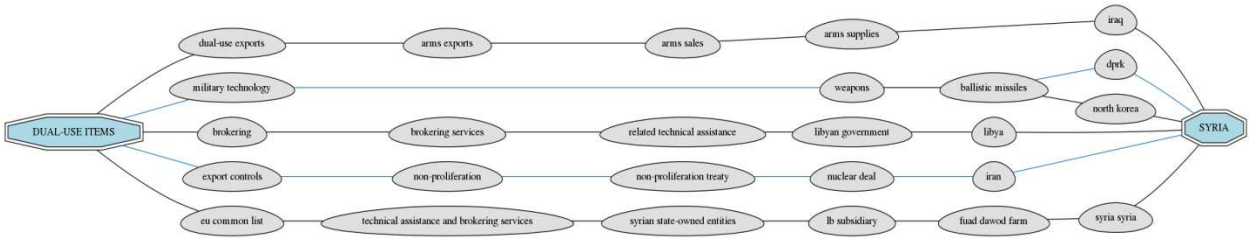


Figure 16: Ontology extracted from the neural embedding used to find all terms relevant to Syria and Dual-use

6 Future research

The current research has highlighted the practical use of word embeddings and shown that the use of phrase embeddings can significantly increase quality and allow the automatic production of high-quality ontologies covering the complete document corpus. The use of basic vector space calculations has allowed the transfer of knowledge between different domains. The key advantage of the automated analytical approach taken is the way it can easily scale to extremely large corpora, supporting human analysts to cope with documents at a scale which could not be tackled unaided.

A number of future research directions are clearly of interest, including

1. Fact checking – integration of EUROSTAT databases with semantic sentence parsing.
2. Enrichment – extracted information can be accompanied by related information from other sources. Currently the team uses Wikipedia extracts as an outer domain knowledge source, but other sources are being considered as well.
3. Bias and intent in text – analytical documents are supposed to be unbiased and identification of hints from sentences can help analyst focus on those parts of text.
4. API – fully documented provision of all the data and algorithms on the fly including access to the knowledge base.

7 Conclusions

The ability to extract information from large bodies of text is of increasing importance to policy making, impact assessment and scientific research. There are many challenges, ranging from collection of the source documents, to recognising and extracting the text, to finding ways to make sense of it and to recognise concepts, how they are linked together and how this changes over time. Recent advances in natural language processing, text mining and machine learning have enabled JRC researchers to produce and test a system, SeTA, which overcomes many of these challenges and allows researchers and analysts to explore large text corpora and to present their results in a simple, meaningful way to decision makers.

Example uses of the system include:

- From a starting concept, keywords or phrases, find documents and datasets which are similar or related at the semantic level, without the need for manual tagging;
- Interactive, visual exploration of knowledge domains through automatically identified key concepts, their links and the documents containing them;
- Understanding and linking the meanings of technical jargon from different domains;
- Exploring changes in meaning of words or phrases over time, for example under different Commissions.

The SeTA system will be provided as a service of the Competence Centre on Text Mining and Analysis, and a pilot version is already available online as a web application. Input from users of this version will help to guide future developments. The value of the system has already been demonstrated through a number of use cases, and this type of semantic text analysis is mature enough to be of practical use to policy analysts and policy makers.

References

Acs, S., Ostlaender, N., Listorti, G., Hradec, J., Hardy, M., Smits, P., Hordijk, L. (2019). Modelling for EU Policy support: Impact Assessments. Analysis of the use of models in European Commission Impact Assessments in 2003-2018, JRC Technical Report, European Commission, Ispra, Upcoming.

Bojanowski, P et al., Enriching Word Vectors with Subword Information, Facebook AI Research, eprint arXiv:1607.04606

Broughel, J. (2015) 'What the United States Can Learn from the European Commission's Better Regulation Initiative', European Journal of Risk Regulation 6(3): 380–381.

Craglia, M., Hradec, J., Nativi S., Santoro, M. (2017) Exploring the depths of the global earth observation system of systems, *Big Earth Data*, 1:1-2, 21-46, doi:10.1080/20964471.2017.1401284

Craglia M. (Ed.), Annoni A., Benczur P., Bertoldi P., Delipetrev P., De Prato G., Feijoo C., Fernandez Macias E., Gomez E., Iglesias M., Junklewitz H, López Cobo M., Martens B., Nascimento S., Nativi S., Polvora A., Sanchez I., Tolan S., Tuomi I., Vesnic Alujevic L., *Artificial Intelligence - A European Perspective*, EUR 29425 EN, Publications Office, Luxembourg, 2018, ISBN 978-92-79-97217-1, doi:10.2760/11251, JRC113826

Danesh, S., Sumner, T., SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction, ACL, 2015

European Commission (2015) 'Better Regulation for Better results - An EU agenda', COM (2015) 215 final, Brussels, 19 May.

European Commission (2017) 'Better Regulation Guidelines', SWD(2017) 350 final, Brussels, 7 July.

EPRS (2015) Ex-ante impact assessment in the European Commission's new Better Regulation Guidelines, European Parliamentary Research Service.

Golberg, E. (2018) '"Better Regulation": European Union Style', Harvard Kennedy School, Mossavar-Rahmani Center for Business and Government, M-RCBG Associate Working Paper Series (98).

Honnibal, M., Montani, I., 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (to appear).

Hradec, J., Lima, V., Tomas, R., & Fullerton K. (2017). INSPIRE Visual Analytics - dissecting the infrastructure. *INSPIRE Conference 2017 proceedings*. Strasbourg, France, 2017

JRC Strategy 2030, https://ec.europa.eu/jrc/sites/jrcsh/files/jrc-strategy-2030_en.pdf

Le, Quoc and Mikolov, T, (2014) Distributed Representations of Sentences and Documents, arXiv:1405.4053v2

Impact Assessment Institute (2017) 'A year and a half of the Better Regulation Agenda : what happened ?', Final study. Řehůřek, R., Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. p. 46--50, 5 pp. ISBN 2-9517408-6-7.

Mikolov, T, et al., Distributed representations of words and phrases and their compositionality. Proceedings NIPS'13 *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2: 3111-9. Lake Tahoe, Nevada. December 2013.

Mikolov, T, et al., Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

Peters, Matthew E. et al, (2018), Deep contextualized word representations, arXiv:1802.05365

Radaelli, C. M. (2018) 'Halfway Through the Better Regulation Strategy of the Juncker Commission: What Does the Evidence Say?', *JCMS: Journal of Common Market Studies* 56: 85–95.

Renda, A. (2015) 'Too good to be true? A quick assessment of the European Commission's new Better Regulation Package', CEPS Special Report (108).

List of abbreviations and definitions

AI	Artificial Intelligence
API	Application Program Interface
CBOW	Continuous Bag of Words
CC-MOD	JRC Competence Centre of Modelling
ES	ElasticSearch
IA	Impact Assessment
JRC	Joint Research Centre
JEODPP	JRC Earth Observation Data Processing Platform
MIDAS	Modelling Inventory and Knowledge Management System of the European Commission
NLP	Natural Language Processing
POS	Part of Speech
SeTA	Semantic Text Analyser
TMA-CC	JRC Competence Centre on Text Mining and Analysis

List of boxes

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-29708-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi:10.2760/577814

ISBN 978-92-76-01518-5