TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Erik Illaste 192853IADB

Development of an Interactive Tool to Support Knowledge Management

Bachelor's thesis

Supervisor: Toomas Lepikult
PhD

Co-supervisor: Alessandro Aliakbargolkar
PhD

Tallinn 2022

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Erik Illaste 192853IADB

Teadmusjuhtimist toetava interaktiivse tööriista arendus

Bakalaureusetöö

Juhendaja: Toomas Lepikult
PhD

Kaasjuhendaja: Alessandro Aliakbargolkar
PhD

Tallinn 2022

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Erik Illaste

16.05.2022

Abstract

Mapping insights concealed in a large corpus of academic literature is a difficult and time-consuming task that often requires in-depth domain knowledge or the help of many subject matter experts. Presenting these complex data to an audience in an effective and intuitive way is challenging.

The aim of this thesis is to develop an interactive tool to support knowledge management efforts at the European Space Agency. The main goal of the resulting product is to communicate insights discovered from the scholarly data extracted from the 4S Symposium body of work from 2004–2018.

An agile iterative approach was used to reach the intended goal. Technology choices were informed by the client's requirements to the end product. A topic modeling based approach for classifying documents according to a predefined topic taxonomy is proposed and used. An interactive dashboard containing various charts was developed and evaluated based on best practices in data visualization.

This thesis is written in English and is 62 pages long, including 6 chapters, 33 figures and 16 tables.

Annotatsioon

Suures kogumis teaduskirjanduses sisalduva teabe kaardistamine on keeruline ja aeganõudev ülesanne, mis nõuab põhjalikke valdkonnaspetsiifilisi teadmisi või erialaekspertide abi. Saadud kompleksse andmestiku tõhus ja intuitiivne esitamine lugejatele on väljakutset pakkuv.

Käesoleva lõputöö eesmärk on arendada interaktiivne tööriist teadmusjuhtimise toetamiseks Euroopa Kosmoseagentuuris. Valminud tööriista põhiülesanne on 4S Sümpoosioni 2004–2018 publikatsioonidest kogutud andmetes sisalduva kasuliku teabe kommunikeerimine.

Eesmärgi saavutamiseks kasutati agiilset iteratiivset lähenemist. Tehnoloogiate valikul lähtuti kliendi nõudmistest lõpptootele. Pakutakse välja ja rakendatakse teemade modelleerimisel põhinev metoodika dokumentide klassifitseerimiseks eelnevalt määratletud teemade taksonoomia alusel. Töötati välja erinevaid andmete visualiseeringuid koondav interaktiivne armatuurlaud, mille arendamisel ja sobivuse hindamisel lähtuti andmete visualiseerimise parimatest tavadest.

See lõputöö on kirjutatud inglise keeles ja on 62 lehekülge pikk, sisaldab 6 peatükki, 33 joonist ja 16 tabelit.

List of abbreviations and terms

Bag of words	A representation of text that describes the occurrence of words within a document.
D3	Data-Driven Documents.
Intertopic distance map	A visualization of topics in two-dimensional space.
JSON	JavaScript Object Notation.
LDA	Latent Dirichlet Allocation.
LSA	Latent Semantic Analysis.
NLP	Natural language processing.
NMF	Non-Negative Matrix Factorization.
PAC network	Information about persons, papers, affiliations, and countries.
PLSA	Probabilistic Latent Semantic Analysis.
REST API	An application programming interface that conforms to the constraints of REST architecture.
Stop words	A set of commonly used words in a language.
Token	An instance of a sequence of characters in some document that are grouped together as a useful semantic unit for processing.
Topic model	A type of statistical model in machine learning to uncover abstract themes in a collection of texts.
WebGL	Web Graphics Library.

Table of contents

List of figures

List of tables

1 Introduction

The work described in this paper follows the development of parts of a software to support knowledge management. The system is being developed by Ennovatic OÜ where the author was performing work duties as part of an enterprise internship. It will be used to analyse scholarly data, in particular, the publications from the 4S Symposium from 2004–2018, and to display the results on an interactive dashboard. While pre-existing solutions for scholarly data visualization exist, they are generally focused on only one or two types of insights. The system being developed is composed of various components, dedicated to the extraction, analysis, and presentation of the data. This thesis focuses primarily on the analysis and presentation aspects. The analysis should enable classification of texts according to a specific topic taxonomy. The intended end user of the client facing interactive dashboard is the Corporate Knowledge Management team of the European Space Agency. The dashboard should be user friendly, intuitive, and clearly communicate key insights to stakeholders.

1.1 Background

Knowledge management is the process of creating, sharing, using, and managing the knowledge and information of an organization [1]. The successful employment of knowledge management methods enables an organization to improve its processes and stay competitive. The generation of useful knowledge requires transformation of an organisation's data into actionable insights. Challenges arise when attempting to distill these insights from a large volume of unstructured data. The first step in tackling this challenge is the extraction of relevant parts of the data. Natural language processing (NLP), and in particular, topic modeling, can be leveraged to extract topical information from text portions of the data. Effective data visualization supports the communication of discovered insights to end users in an intuitive and easily interpretable way.

1.2 Problem statement

Getting an overview of a large body of scientific literature is time-consuming and requires in-depth domain knowledge. An understanding of the interconnections between authors, topics and institutions is crucial when attempting to understand a field of study and its evolution through time—something that is not easily decipherable from reading the papers alone. The use of automated software to extract, analyse, and visualize information from a scholarly dataset supports getting a thorough overview of the resulting data.

Popular topic modeling approaches such as Latent Dirichlet Allocation (LDA) [2] can be leveraged to discover latent topics in a corpus of text documents and to classify the documents based on the discovered document-topic distributions. However, the topics generated with LDA can often be difficult to interpret and require manual labeling by domain experts. When required to classify papers based on a predefined set of topics, the traditional LDA approach is not optimal.

Effective data visualization is key to communicating complex data. Many techniques exist for data visualization, but it can be challenging to evaluate the appropriateness of any particular method. A general approach to validate the appropriateness of an interactive data visualization is needed. Tools exist for visualizing scholarly data, but most of them are focused on only a few key insights and require separate installation, making them cumbersome for the user. An interactive dashboard with exploratory and explanatory visualizations that communicate a combination of insights is desirable.

1.3 Purpose

This thesis focuses on the development of an interactive tool for visualizing scholarly data derived from scientific publications presented at the 4S Symposium in 2004–2018 to support European Space Agency's Corporate Knowledge Management initiative. The resulting dashboard is the user facing component in a more general-purpose system. The secondary purpose of this thesis is to describe a topic modeling based approach that enables multi-label classification of publications according to a specific topic taxonomy in situations where labeled training data is virtually unavailable. The output of the topic modeling results, among other insights, will be displayed on the interactive dashboard.

1.4 Overview of the thesis

This thesis is made up of six parts and is focused on two different topics: data visualization and topic modeling. The first part gives a general introduction. The second part provides a brief literature overview of the best practices in data visualization and topic modeling. The third part describes the methods that were used to arrive at the results: the gathering of requirements; the comparative analysis of technologies and techniques with respect to the requirements of the client; how iterative development was leveraged to arrive at the result; an approach for evaluating data visualizations; a topic modeling based technique for classifying documents to a predefined set of topics; and finally, the technologies used. The fourth part describes the interactive dashboard and the topic models that were developed. The fifth part analyses the results, details the scope and limitations, and discusses probable future directions. Part six contains the summary.

2 Literature review

This chapter provides a brief overview of the literature on two topics: best practices in data visualization and topic modeling.

2.1 A brief review of best practices in data visualization

Data visualization is an interdisciplinary field in the intersection of art and science which deals with the graphic representation of data. Effective data visualization intuitively communicates insights from data that are otherwise difficult and time-consuming to interpret. Although graphic representation is not a precise science, there are useful guidelines which govern the effective communication of insights in visual form.

Edward Tufte's classic, *The Visual Display of Quantitative Information* [3], lays out the established principles for communicating information through the simultaneous presentation of words, numbers, and pictures. "Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficacy" [3, p. 13]. He claims that data should be presented in a coherent way that does not distort the information it carries. When creating data visualizations, the emphasis should be on communicating the substance rather than anything else—minimizing "chartjunk" and maximizing the data-ink ratio ensures that only the relevant information is included in a visualization. Revealing data at several levels of detail enables the audience to understand a broad overview as well as the finer structure of the data.

Aspects of the modernist, functional and minimalist approach of Tufte, which focuses primarily on print medium, are applicable to digital media. However, it is argued, that interactive data visualization benefits from the integration of emotional (pathos) appeals into data design and this is becoming increasingly prevalent online [4]. User engagement can be enhanced by deploying color and multimodal features that elicit a psychological response, guide attention, and support cognitive processes required for interpreting the data. Offering users opportunities for commentary magnifies the level at which they are emotionally invested with the data, fostering a culture of feedback. Adding interactive features draws the user closer to the data visualization and encourages data exploration. Drawing on previous works, Hiippala further substantiates the claim of Kostelnick that the employment of multimodality and various interconnected canvases lend

visualizations a higher degree of interactivity [5], which in turn extends the audience an invitation to explore the data and influence its representation. However, it should be noted, that exploratory visualization is not objectively superior to an explanatory one. The former is intended to help the audience discover interesting aspect of the data, while the latter is meant to present the most important insights. Effective visualization should aim at combining both to form an overarching narrative frame.

Knaflic's *Storytelling with Data*, a modern complement to the work of Tufte, provides actionable insights for guiding the attention of the audience [6]. The employment of preattentive attributes like color, size, and position help guide the audience through the visualization. Attributes like labeling, text, and annotation are useful for emphasizing and deemphasizing components in the visual. The use of gestalt principles of visual perception [7] help identify unnecessary elements and ease the processing of our visual communications, thereby greatly improving the functionality, user-friendliness, and the general aesthetics of a visualization.

In general, the more intricate the visualization being viewed is, the more time it takes for the audience to understand it [8]. Therefore, it is advantageous to make use of visuals that are familiar to most and work for the majority of needs. The choice of an appropriate data visualization for a situation relies on understanding different types of data variables. The most fundamental distinction to make is between quantitative (continuous or discrete) and qualitative (categorical) data. Qualitative data are difficult to quantify, but are separable into discrete categories, which can be expressed in terms of language. Quantitative data can be measured and given numeric values. The following is a list of some of the most common types of visual displays [6].

- Plain text: useful, when only a single number or two need to be displayed.
- Table: useful for comparing pairs of related values.
- Heat map: a variation on the table, which uses color to convey the relative magnitude of the values within each cell.
- Scatterplot: an effective way of showing the relationship between two parameters.
- Column chart: useful for comparing discrete categories.

- Stacked column chart: useful for displaying cumulative totals across categories while also showing the subcomponent pieces of each category.

- Bar chart: a horizontal variation on the column chart, which is appropriate for certain layouts, especially when the number of categories is large and the names of categories are relatively long.

- Stacked bar chart: useful for displaying cumulative totals across distinct categories but also their subcomponent pieces.

- Line chart: useful for plotting continuous data, in particular, to track change over a period of time.

- Pie chart: useful, when displaying the distribution of a single categorical variable that adds up to 100%.

- Graph: useful for visualizing entities in a network and the interconnections between them.

- Word cloud: useful for displaying a set of words and/or highlighting the relative importance of words in a collection.

Effective data visualization in the digital medium should communicate data in a clear, intuitive, and engaging way and should be supported by the use of gestalt principles of design, preattentive attributes, interactive elements, and common charts that are appropriate to the data being displayed. Scholarly data are heterogenous and can be represented with a number of basic entity types. The constituents of scholarly data extractable from most publications include the title and abstract of the paper, author metadata, and citation metadata. Author metadata includes information about the institution or organization that the author is affiliated with, as well as its country of location. There are diverse relationships among these entities. While a number of specialized tools are available for visualizing the various relationships [9], most of them require separate manual installation and the results cannot be viewed on a single dashboard. “How to visualize different relationships in a single task is meaningful and challenging” [9, p. 19219]. It is desirable to have a simple to use tool that integrates various views of the scholarly data in a single interactive and intuitive dashboard.

2.2 A brief review of topic modeling

In natural language processing, topic models are unsupervised learning methods based on hierarchical probabilistic and non-probabilistic models used for revealing the underlying semantic structure of documents [10]. They can be used to achieve an understanding of the latent topics present in a body of documents without the express need to read the documents themselves. Topic modeling is especially useful and often used when there is a need to identify topics in a large collection of documents that cannot be annotated by hand [11]. They are also useful for text classification [12].

There are a number of different methods of topic modeling—the most popular non-probabilistic models include Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF) and the main probabilistic models are LDA and Probabilistic Latent Semantic Analysis (PLSA) [13]. In general, for short text classification LDA and NMF generate the most valuable outputs [14]. For texts where the average number of words per document is more than or equal to 50, and discovery of complex topic relationships is not the primary focus of the analysis, LDA is the preferred method [15]. Moreover, LDA is considered one of the most popular topic models overall [2], [14], [16], since it provides accurate results and can also be extended to infer the topic distribution in unseen documents.

The LDA model makes the assumption that each document is composed of a predefined number of topics in different proportions, and each topic is defined as a distribution over a vocabulary [13]. The aim is to learn the topics present in documents given a number of topics *k*, whereupon the percentage of the *k* topics present in each document can be found. One of the main drawback of this method is that it can be difficult to estimate the optimal number of topics present in a corpus of text—if the selected number of topics is too small, the topics become too general, if the number is too large, the topics start overlapping with each other [14]. Additionally, the discovered topics need to be manually labeled by domain experts.

Given the need to allocate documents to a predefined set of topics, existing knowledge about the number of topics alone is not sufficient to ensure that the model discovers them from the documents. In practice, artificial neural networks and Transformers are often used for multi-label text classification [17], however, these approaches require a large

volume of labeled training data, which was not available for this project. In addition, instead of binary vectors, the LDA model outputs the percentage contribution of each topic in a document, which was particularly useful for subsequent steps in the analysis. In Section 3.2.4 of this thesis, the author proposes a method for leveraging LDA to classify documents to a predefined set of topics.

3 Methodology

This chapter describes the methods and technologies used to arrive at the desired results. First, a general overview of the system being developed is provided. Second, a description of the main processes is given. This part includes subsections on requirements, technology choices informed by the latter, the approach taken for evaluating the appropriateness of visualizations and adherence to best practices, the description of topic modeling experiments, and a general description of the development technique. Finally, the technologies chosen for the project are listed.

3.1 Overview of the object

The aim of the software system being developed by Ennovatic is to generate insights from a collection of scientific papers. In particular, their client is interested in mapping the knowledge discovered in the publications presented at the 4S Symposium in the years 2004–2018.

The full process can be described in a simplified form in the following steps (Figure 1).

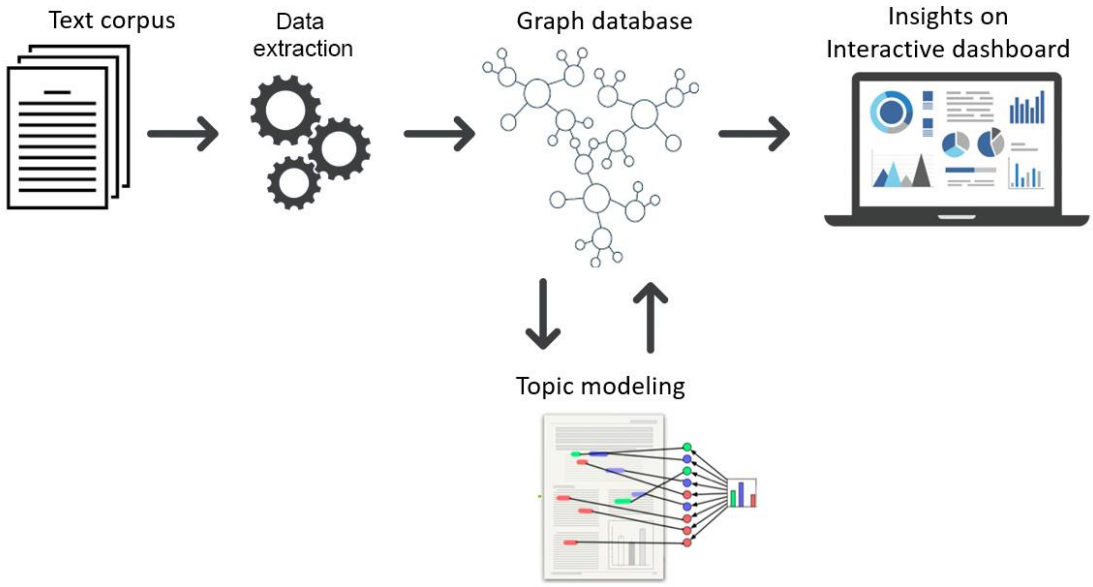

Figure 1. Simplified description of the main processes.

1) A collection of papers is gathered.

2) Metadata are extracted from the papers.

3) The data are ingested into a graph database. The resulting nodes include *author*, *conference_paper*, *entity*, *conference*, *cited_paper*. The resulting edges include *author_of*, *affiliated_with*, *citing*, and *presented_at*.

4) Topic modeling is performed on the abstracts of the papers [18], [19].

5) The results of topic modeling are imported to the database where they are shaped into *topic* nodes and *topic_of* edges. The weighted edges between paper nodes and topic nodes describe the proportion of topic present in each paper.

6) Insights generated from the scholarly data are visualized on an interactive dashboard.

This thesis focuses mainly on steps 4 and 6 in this process: topic modeling, and the development of the interactive dashboard.

3.2 Overview of processes

This section details the processes involved in the development of the interactive visual dashboard and the topic models.

3.2.1 Requirements gathering

All requirements for the solution were not established upfront but evolved as a result of frequent meetings with the client. They can be separated into functional and non-functional requirements.

The following is the list of functional requirements.

- The dashboard should include a visualization that represents how many papers were produced by different countries in any given year.
- The dashboard should include a visualization that represents how many authors published papers at the conference in any given year.

- The dashboard should include a visualization that represents the number of affiliations for each institution or organization in any given year.
- The dashboard should give an overview of how the number of papers, persons, affiliations, collaborations, and countries have evolved in the period 2004–2018.
- The dashboard should give an overview of the evolution of organic topics arising from papers presented at the conference in the period 2004–2018.
- The dashboard should give an overview of the evolution of ESA Technology Tree topics [20] discovered from papers presented at the conference in the period 2004–2018.
- It should be possible to display graph projections of the scholarly network, including relationships between organisations, countries, and co-authorship relationships with the removal of intermediary nodes.
- There should be an option to change the layout of the graph visualization.

The following is the list of non-functional requirements.

- Data should be presented in an intuitive way.
- The language of the dashboard is English.
- Free open-source software should be preferred.
- The dashboard must connect to the OrientDB graph database [21].
- The dashboard shall not be publicly accessible on the Internet, shall not collect user data, and shall not require authentication or authorization.

3.2.2 The choice of technologies

The requirements described in Section 3.2.1 informed the choice of appropriate technologies and techniques for the project.

Many tools for creating data visualizations are available. They can be broadly categorized as those which require programming knowledge and those that do not [9]. The most

popular choices in the first category include tools like PowerBI [22] and Tableau [23]. While both tools offer a range of different chart types and can be connected to an external database, the level of customizability is relatively modest, and the tools are not free to use. Additionally, neither of these tools come with a robust enough graph visualization option. The creation of a dedicated web-based dashboard using JavaScript, CSS and HTML was preferred, due to the high level of customizability it offers and the general ease with which it can be accessed by the client—aside from having access to internet connection, an operating system and a modern browser, no extra software installation is required.

Front-end frameworks encourage modular and maintainable architecture, offering the advantage that the web application can be broken up to reusable standalone components. The codebase of popular frameworks is being actively maintained and improved, and the existence of large communities of developers using these technologies ensures that it is easy to find answers to common questions. Addition of new developers to the team can be easier when a popular framework is used for the project.

According to NPM Trends (Figure 2), the current most popular framework is ReactJS [24]. The choice of this framework was further supported by its easy integration with TypeScript, availability of necessary libraries for the development of the dashboard, and the author's familiarity with it.

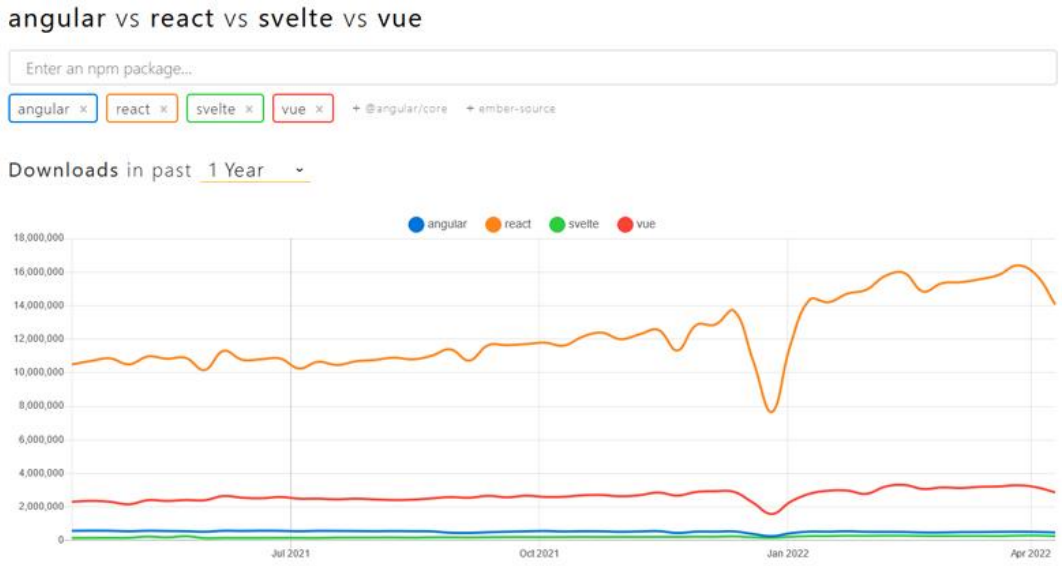

Figure 2. Trends for front-end frameworks (NPM trends).

There are a number of popular charting libraries for React [25]. The final choice was influenced by two factors. First, it was desirable for the library to offer the most common chart types. Secondly, high customizability of the basic chart types was desired. Recharts [26] is a charting library built with D3 (Data-Driven Documents) and React that offers many standard chart types which can be extended and customized as needed. In addition, Recharts has excellent documentation, and with over 17k stars on GitHub, has a considerably large userbase.

None of the popular charting libraries, including Recharts, include a standard chart type for visualizing network data. ReGraph, developed by Cambridge Intelligence is a popular option for graph visualizations among React users [27]. ReGraph comes with features like automatic layouts, network filtering, node combining, time-based analysis and its WebGL (Web Graphics Library) renderer can manage graphs with up to 100,000 items. Although the out-of-the-box capabilities are impressive, it comes with a lofty price tag. CytoScape.js is an open-source graph theory library used for graph analysis and visualization [28]. It renders in HTML Canvas (a part of HTML5 that allows for dynamic, scriptable rendering of 2D shapes and bitmap images) only but offers many graph theory features and layout algorithms. Vis.js is a dynamic, browser-based visualization library that can handle large amounts of dynamic data and interaction with the data [29]. Vis.js has comprehensive documentation, a large userbase, highly customizable options, and is designed to be easy to use. Like CytoScape.js, Vis.js does not offer WebGL based rendering. For this project, there was no explicit requirement to display an exceptionally high number of items simultaneously that would require WebGL rendering. Vis.js was chosen, mainly due to its high customizability and coherent API documentation.

Python is a general-purpose language and is very popular among data analysts and data scientists because of its simple syntax and the availability of many great libraries for data cleaning, analysis, visualization, and machine learning [30]. Jupyter notebook is a free, open source, interactive programming environment which can be used to combine live code, computational output, equations, explanatory text, visualizations, and multimedia resources in a single document [31]. Gensim is a popular library used for topic modeling implemented in Python and Cython for performance, and was the most popular tool used for LDA in many recent studies [32], [14].

3.2.3 Data visualization development

The development of the necessary data visualizations described in Section 3.2.1 were guided by the literature review on the best practices in data visualization discussed in Section 2.1. The author evaluated how aspects of each visualization adhere to these best practices. Table 1 provides a template for how each aspect was evaluated.

Table 1. Template for the evaluation of visualizations.

Aspect	**Evaluation**
Type of visualization	Is the visualization explanatory or exploratory?
Appropriateness of the chart	How does the chosen chart type make sense in the context of the data being displayed?
Interactivity	Does the visualization include interactive capabilities? If so, which ones?
Data filtering	Can the user choose which data to show? If so, what are the filtering capabilities?
Use of preattentive attributes	Are preattentive attributes used in the chart? If so, what are they?
Use of descriptive labels	Is the data interpretable, does it have descriptive labels?
Interactions with other charts	Does the chart interact with other charts?
Alternatives	What could have been done differently?

In addition, feedback from meetings with the client was used to improve the quality of the visualizations and ensure that they are interpretable and useful. To encourage further feedback, a suggestion box was added to the footer of the dashboard, which enables users to specify any chart and send their concerns or suggestions for its future improvement.

3.2.4 Topic modeling experiments

Two sets of distinct types of experiments were performed to find topics in the 4S Symposium document corpus. The client was interested in understanding both the latent topics found in the texts, which were found by applying the traditional LDA analysis approach to the corpus, as well as how the texts related to a predefined topic taxonomy—the ESA Technology Tree domains [20]. The first set of topics will be referred to as organic topics, and the second set as technology tree topics.

To find the organic topics the following steps were taken. First, the title, year, id, and abstract portions of the texts were imported from the database into a Jupyter notebook. Then, a set of stop words were defined. Stop words refer to common words that carry low-level information and are not conductive to finding meaningful topics. The stop words were extended by common names present in the dataset. The abstracts were tokenized—split into single word units. This was followed by the removal of all punctuation, digits, words shorter than three letters, and finally stop words. The resulting set of words was converted to lowercase. The last step in the preprocessing involved reducing each token to its root using the Porter stemming algorithm [33].

Next, tokens that occurred in more than 30% of the papers were removed to avoid overly general topics. The result of the preprocessing was a list with the length of the document corpus, consisting of lists of tokens for each paper. The next step involved creating the two main inputs to the Gensim LDA topic model: the dictionary and the corpus. The dictionary includes an entry for each unique token from the result of preprocessing, where the key is its index and the value the token. The corpus is a list of bag of words representations of words in each document, consisting of tuples, where the first item is the token id and the second the token count in the document.

To find the optimal number of topics, a number of LDA models with the same corpus and dictionary were created, varying the number of topics from 1–50. The coherence score for each resulting model was computed and plotted. Coherence score refers to the human interpretability of the topics [19]. The models that yielded the highest coherence scores were considered for further analysis.

Based on the keywords and the weight (relative importance) of each keyword in a topic, word clouds were generated to represent the topics visually. PyLDAvis [34] was used for

visualizing the topics in an intertopic distance map (the visualization of the topics in a two-dimensional space). Large non-overlapping circles on this visualization are generally desirable, as they represent good topics.

A heatmap was used to visualize the topic distribution with respect to each document. The results were analysed internally by the author and two experienced aerospace engineers with extensive domain knowledge. The analysis consisted of the following steps. First, an attempt was made to label the word clouds. Then, the labels were compared to the actual title and abstract of the papers that received the highest score for a given topic to validate that the labeling was accurate. Problematic word clouds were noted, and the appropriate changes to the model inputs were committed in the next iteration. The general procedure for generating organic topics is illustrated in Figure 3.

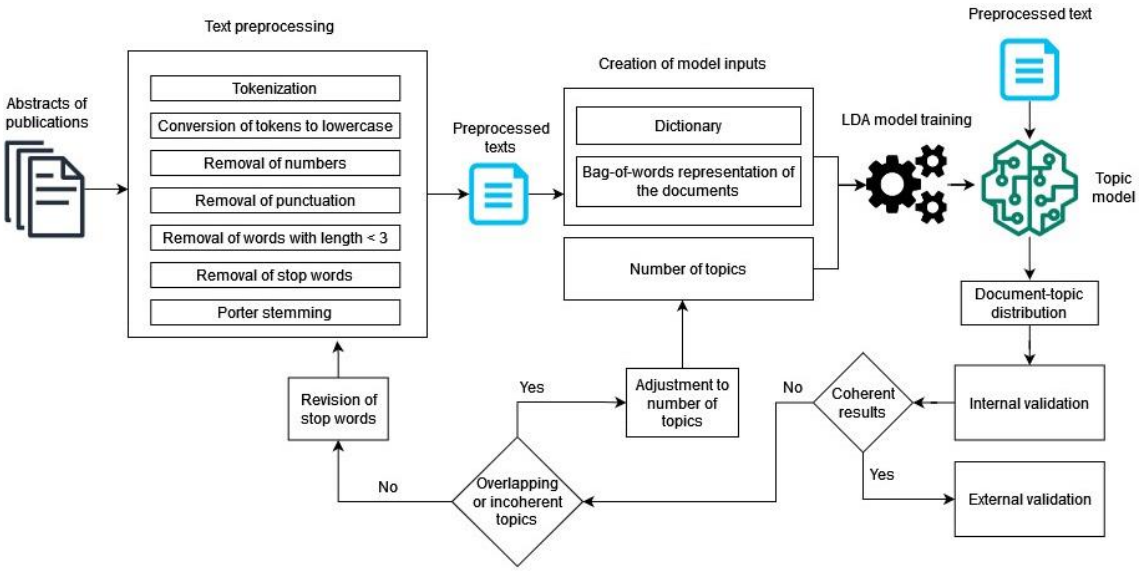

Figure 3. Procedure for generating organic topics.

To classify the document corpus according to the ESA Technology Tree, a different approach was taken. The general procedure is illustrated in Figure 4. First, a document containing text that describes each of the ESA Technology Tree domains was created. Then, each one was supplemented with relevant text from the book *Space Mission Engineering: The New SMAD* [35]. Each section was given the same preprocessing treatment as the abstracts in the analysis of organic topics. The resulting list included a list of tokens for each technology domain.

Next, for each vocabulary, ten abstracts were generated. The length of the abstracts ranged between the lowest and highest word count in the abstracts of the actual text

corpus. Each token was randomly sampled from a vocabulary without removal. The LDA model was trained on the resulting abstracts, yielding topics identifiable as the respective technology domains. Word clouds were generated for each topic and the intertopic distance map visualized using pyLDAvis. The process for reviewing the word clouds was the same as with the organic topics; vocabularies were adjusted and extended according to need. Finally, the best model was used to find the topic distribution of the actual abstracts of the 4S conference papers.

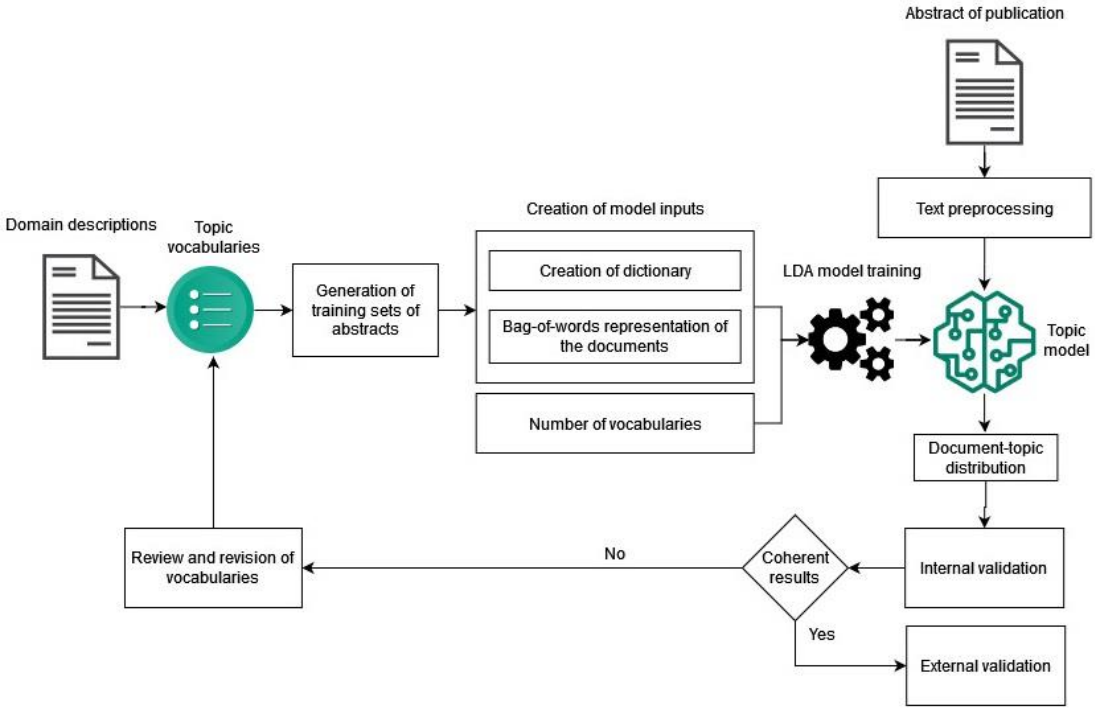

Figure 4. Procedure for generating technology tree topics.

3.2.5 Iterative development

The prototype of the solution was created in an agile work environment. Daily stand-up meetings with the team and periodic meetings with the client created a workflow based on rapid iteration, feedback, internal and external validation. A feature improvement system was integrated into the solution, consisting of a form on the dashboard, which enables users to suggest changes to the various visualizations. The suggestions are collected to a database and the team can prioritize them using a simple custom-made ticketing system.

3.3 Technologies used

Pursuant to Section 3.2.2, the solution was realized with the use of the following technologies: ReactJS, TypeScript, CSS, ReCharts, React Simple Maps, Vis.js were used for the development of the dashboard. Python, Jupyter notebook, Gensim were used for topic modeling. The simplistic ticketing system collecting user feedback was built using Flask [36] and Sqlite [37]. Visual Studio Code [38] was used as the code editor of choice by the author. Popular browsers were used to test the functionality of the dashboard.

4 Results

This chapter describes the individual data visualizations and topic models that were developed. In addition, a description of the main views and general layout of the dashboard is provided. A demo of the interactive dashboard can be viewed online [39].

4.1 Interactive data dashboard

The basic wireframe for the layout of the dashboard includes two main section (Figure 5). Each section takes up the full screen when viewed in a browser window. The user can scroll up or down to navigate between the sections. The upper portion of the dashboard holds an interactive graph visualization. A tab bar is used to navigate between three different tabs. The lower section includes visualizations specific to the selected tab.

Figure 5. Dashboard layout wireframe.

The solution used React Hooks API [40] which utilizes the functional programming paradigm. The mechanism for fetching data for the visualizations is analogous for all charts. When the user changes a selectable value for a given visualization, e.g., the year number, a function is triggered that updates the portion of the component state which holds the year value. A hook listening to changes in this value is triggered. A loading message is displayed to the user. A request is sent to the OrientDB REST API which triggers a server-side function that fetches the required data from the graph database. Upon return of the response with the data in JSON format, the portion of the component state which holds the data for the visualization is updated. The loading message will be hidden, and the chart updates to display the retrieved data.

4.1.1 Overview tab visualizations

The aim of the overview tab (Figure 6) is to provide the viewer insights of the PAC network. The PAC network includes information about persons, papers, affiliations, countries, the connections between them, and how they have evolved over time. The overview tab includes six distinct visualizations. The user can select a year for which data are displayed. The user can also specify a multiyear display style, in which case the charts on the left will be replaced with a multiline chart.

Figure 6. Overview tab.

The *authors per country chart* (Figure 7) shows the top six countries with the highest number of authors in a given year. The seventh sector of the pie chart represents the total number of authors from other countries in that year.

The *summary statistics* on the right of the pie chart provide a succinct overview of the basic numbers for the selected year: the number of papers, authors, affiliations, countries, and relations.

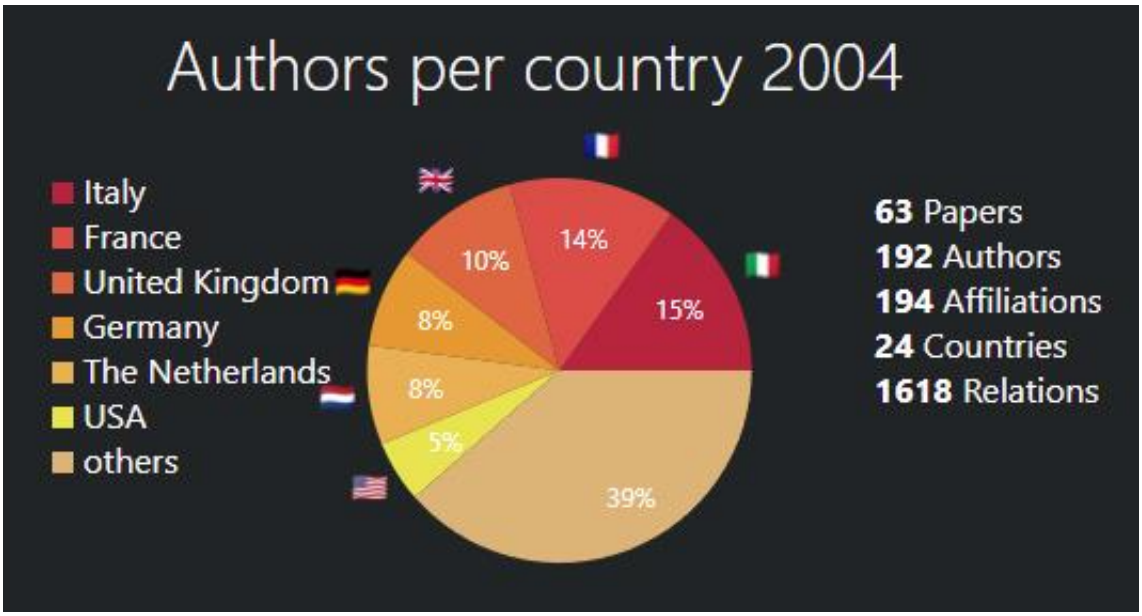

Figure 7. Authors per country (single year) and summary statistics.

The *papers per country chart* (Figure 8) displays the number of papers contributed by each country in a given year. The bars of the chart are sorted hierarchically from top to bottom. The country with the highest number of papers published is displayed at the top.

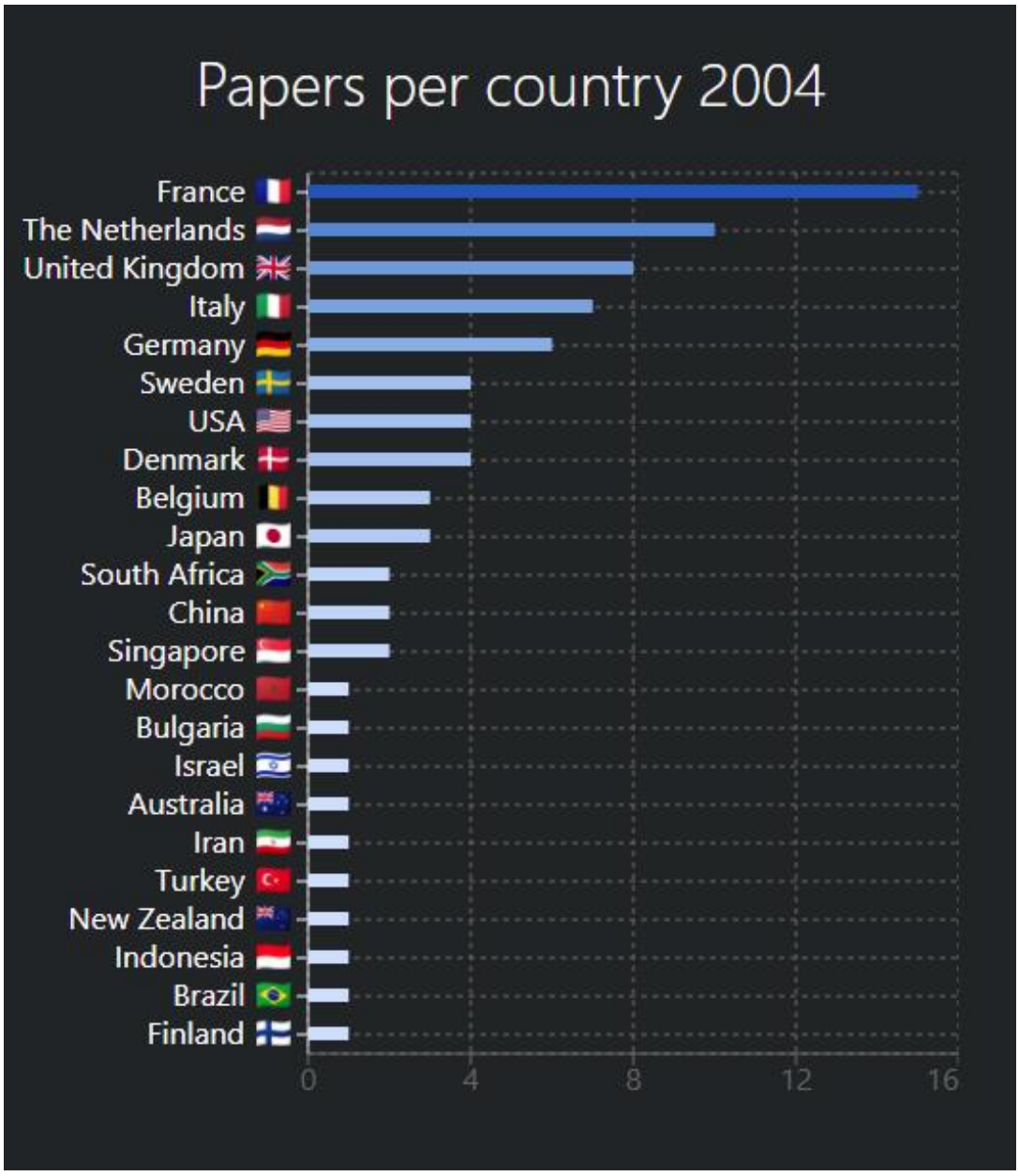

Figure 8. Papers per country (single year).

The *authors per affiliation chart* (Figure 9) shows how many authors were affiliated with an organization or an institution in a given year. The bars are sorted hierarchically, and the brush component can be used to alter the number of bars being displayed.

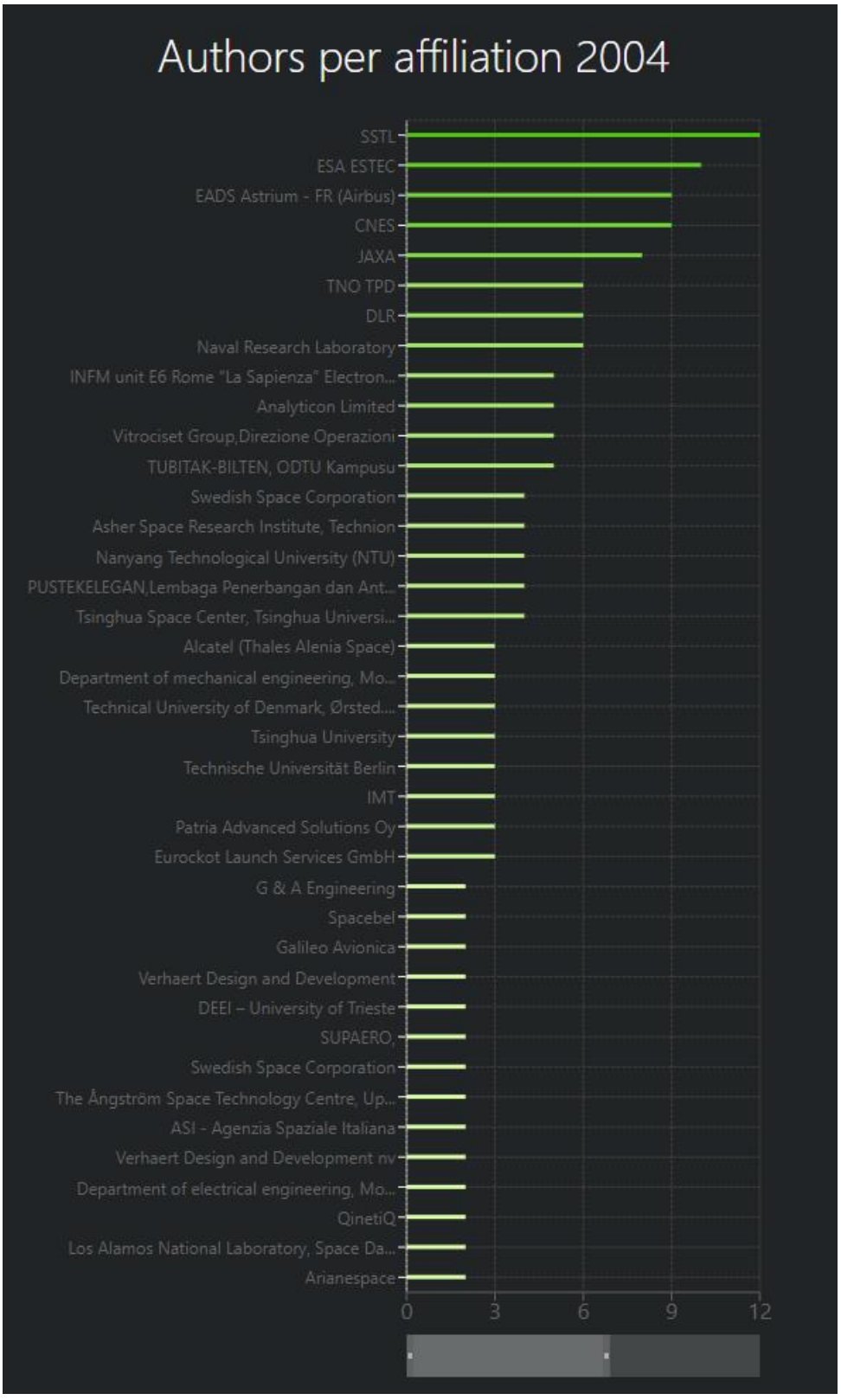

Figure 9. Authors per affiliation (single year).

The *cross-country collaborations chart* (Figure 10) can be used to compare the number of collaborations per country over a specified time period. If a publication has two authors from different countries, then this counts as a single collaboration for both countries. The visualization includes highly customizable data filtering options. The user can compare all listed countries across all years, or just a few over a shorter time period.

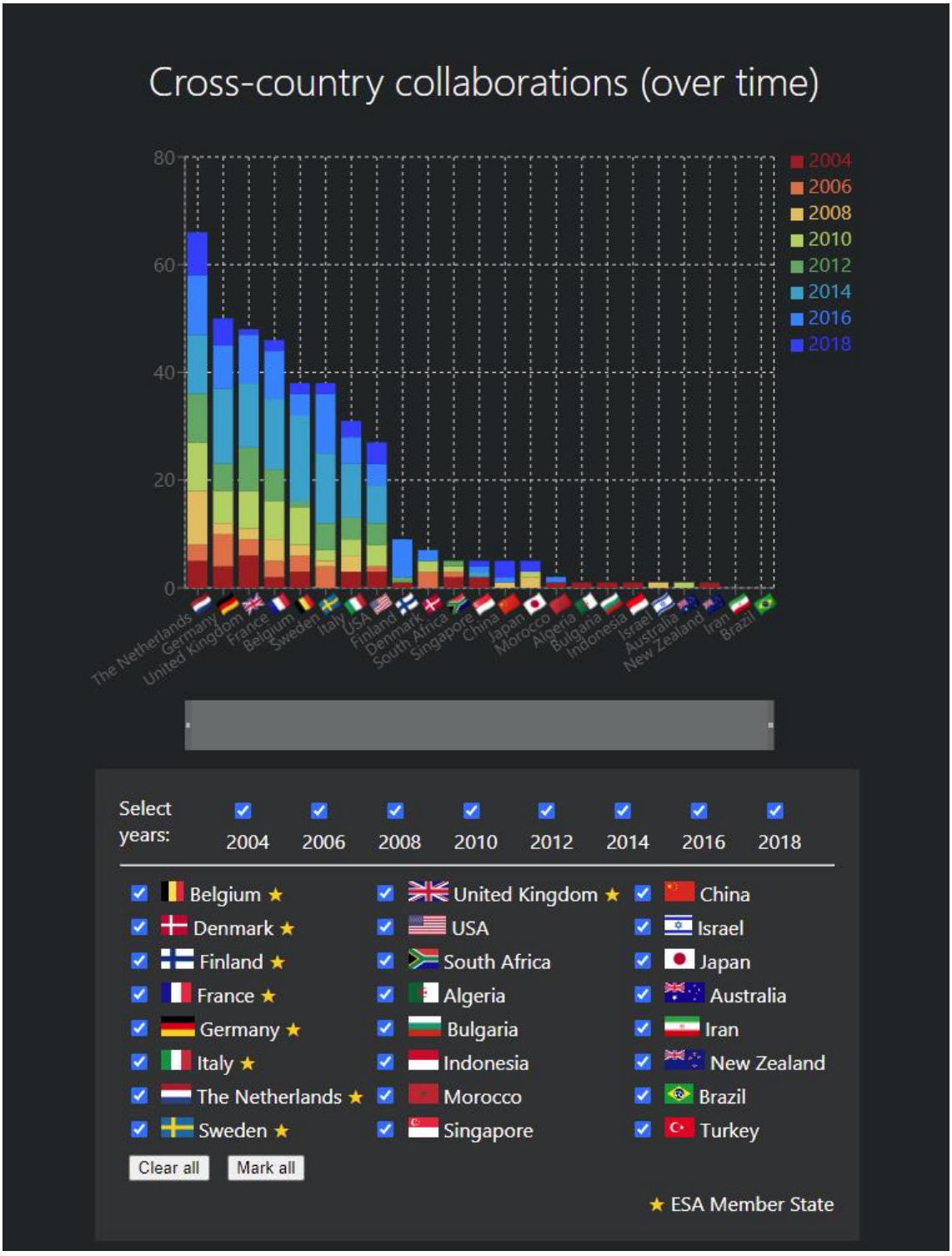

Figure 10. Cross-country collaborations (multiyear).

This multiyear chart (Figure 11) becomes visible when the user selects the multiyear display style from the select box in the first tab. In contrast with some of the charts in the single year view, this visualization shows the time evolution of the PAC item of interest. The user can specify up to five countries to include into the comparison. In addition to showing the number of authors per country in the 2004–2018 period, the chart can be used to display other PAC items. The available items are authors per country, papers per country, cross-country collaborations, affiliations per country, papers per affiliation. In the case of the last item, country names are replaced with entity names.

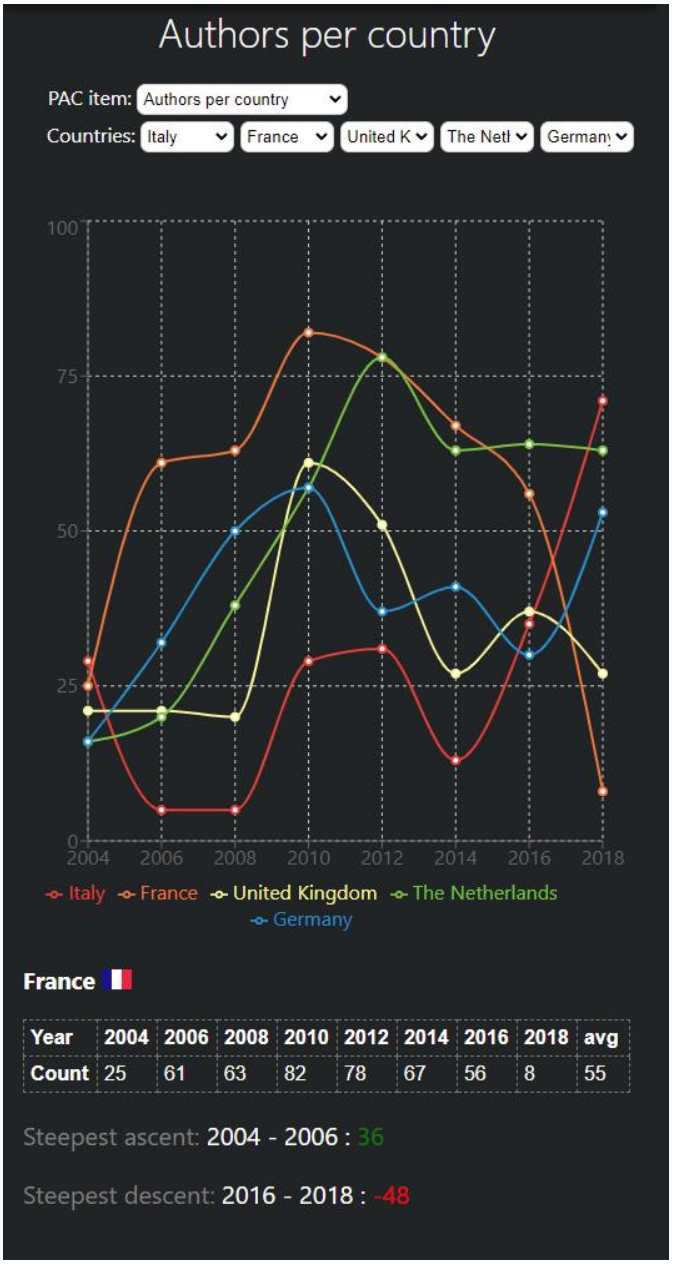

Figure 11. Authors per country (multiyear).

4.1.2 Graph visualization

The fully interactive graph visualization (Figure 12) can be found in the top portion of the dashboard. It combines all the scholarly data in the database into a network visualization composed of nodes and edges. The user can specify a year, select between predefined queries to reveal insights, change the layout, toggle physics and labels.

Figure 12. Graph visualization.

An example of a bipartite projection [41] of the results of the top ten most affiliated entities query is displayed in Figure 13.

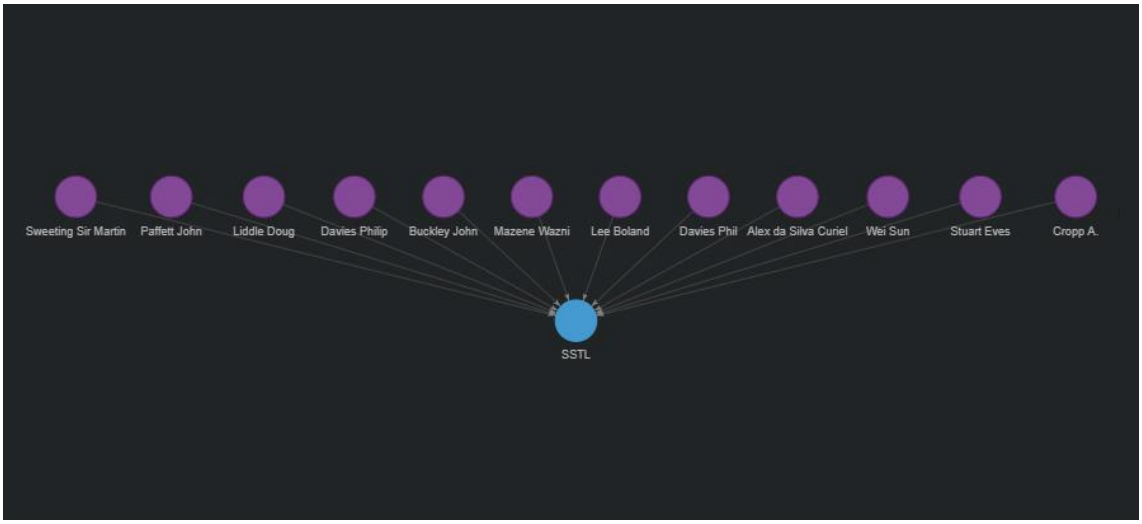

Figure 13. Bipartite projection.

The graph visualization can display a network of collaborations between countries (Figure 14). The relative size of the country nodes represents the number of collaborations for each country. The database itself does not include a country entity, but because each author is affiliated with an institution or an organization entity that contains a country field, a query can retrieve the necessary data to visualize this information.

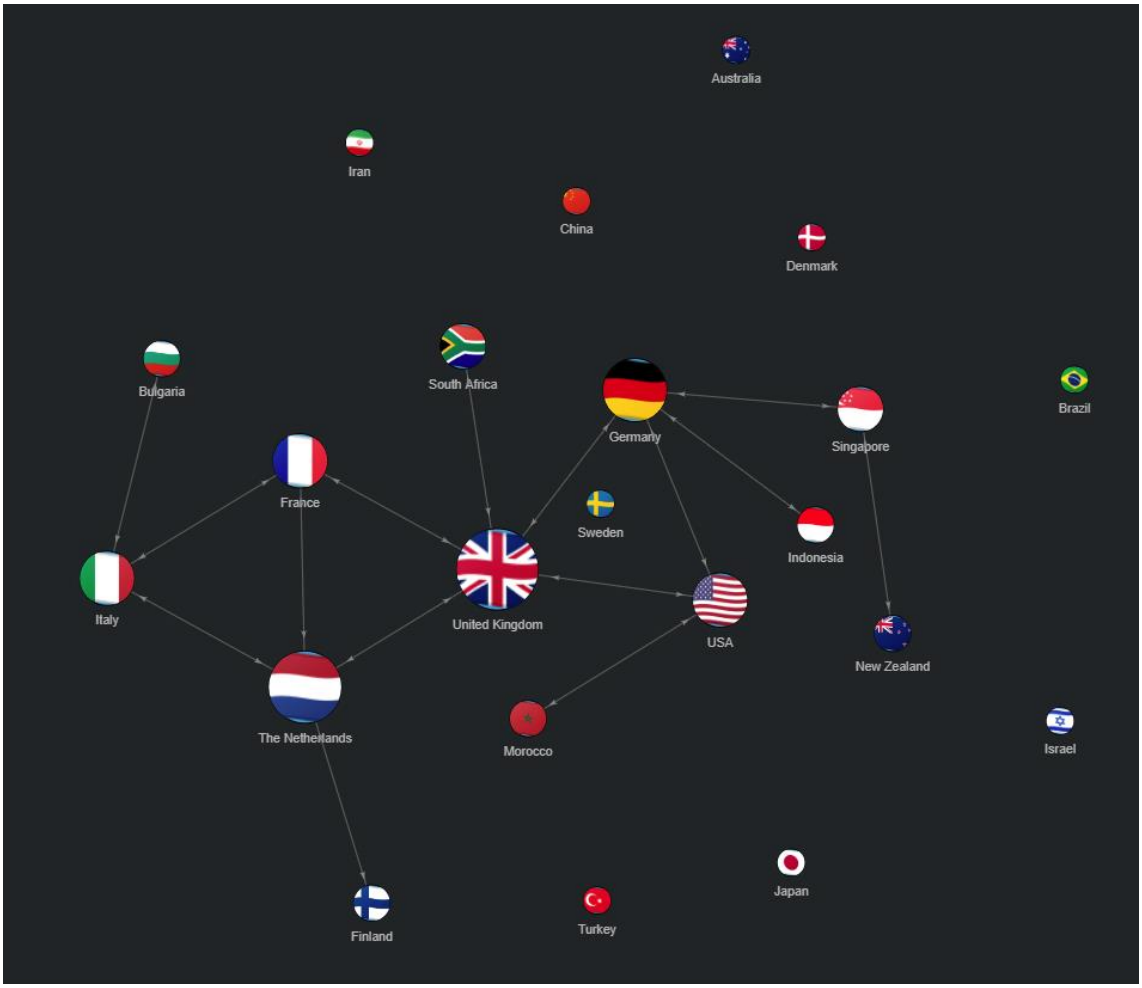

Figure 14. Country collaboration projection.

The *entity collaboration projection* (Figure 15) is a way to visualize the connections between different organizations and institutions. In the database, intermediary edges and nodes exist between these entities, but it is desirable to remove them from view when all that interests the user are the relationships between these specific entities.

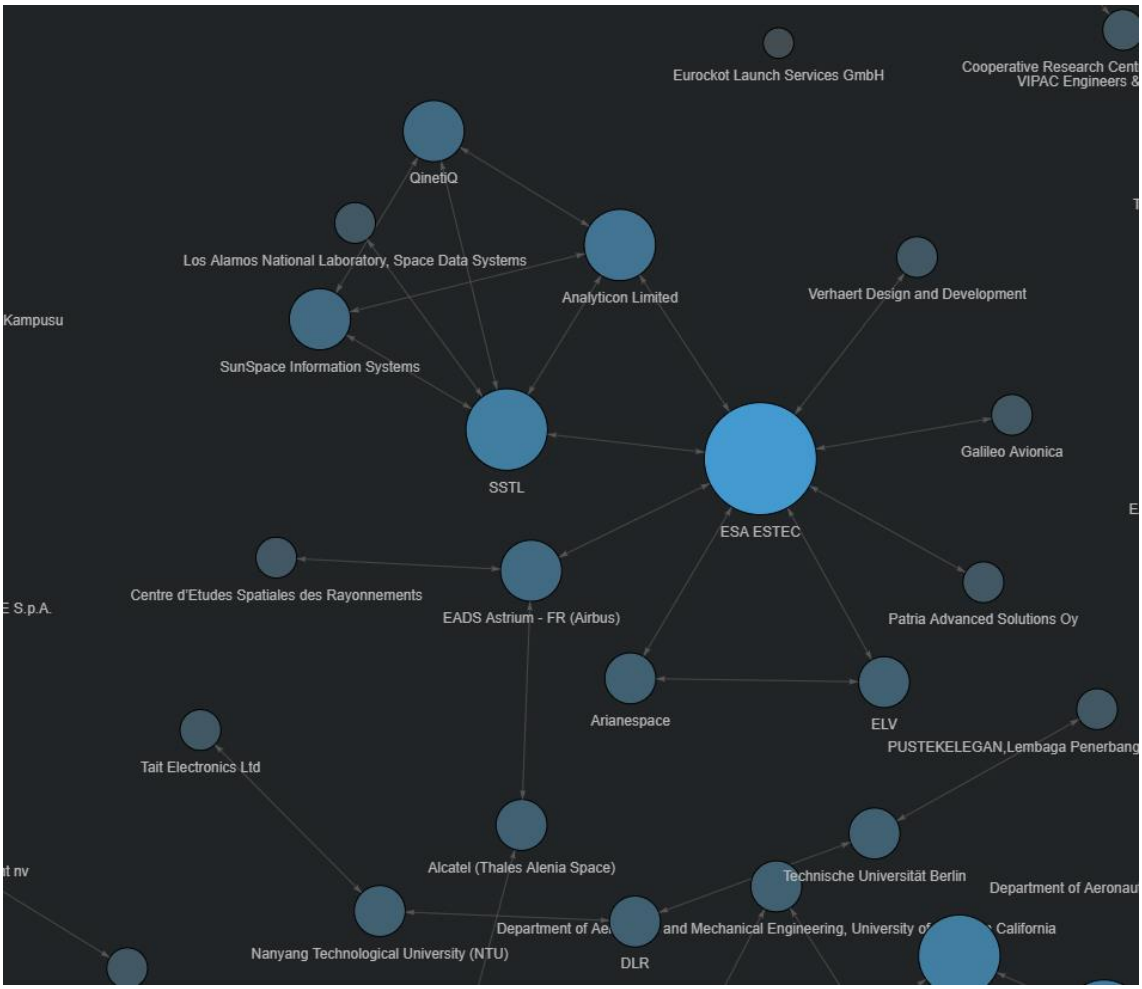

Figure 15. Entity collaboration projection.

It is also useful to remove redundant nodes and edges when trying to understand the author collaboration network (Figure 16). The nodes that appear most vibrant are the authors who have co-authored most papers with other authors.

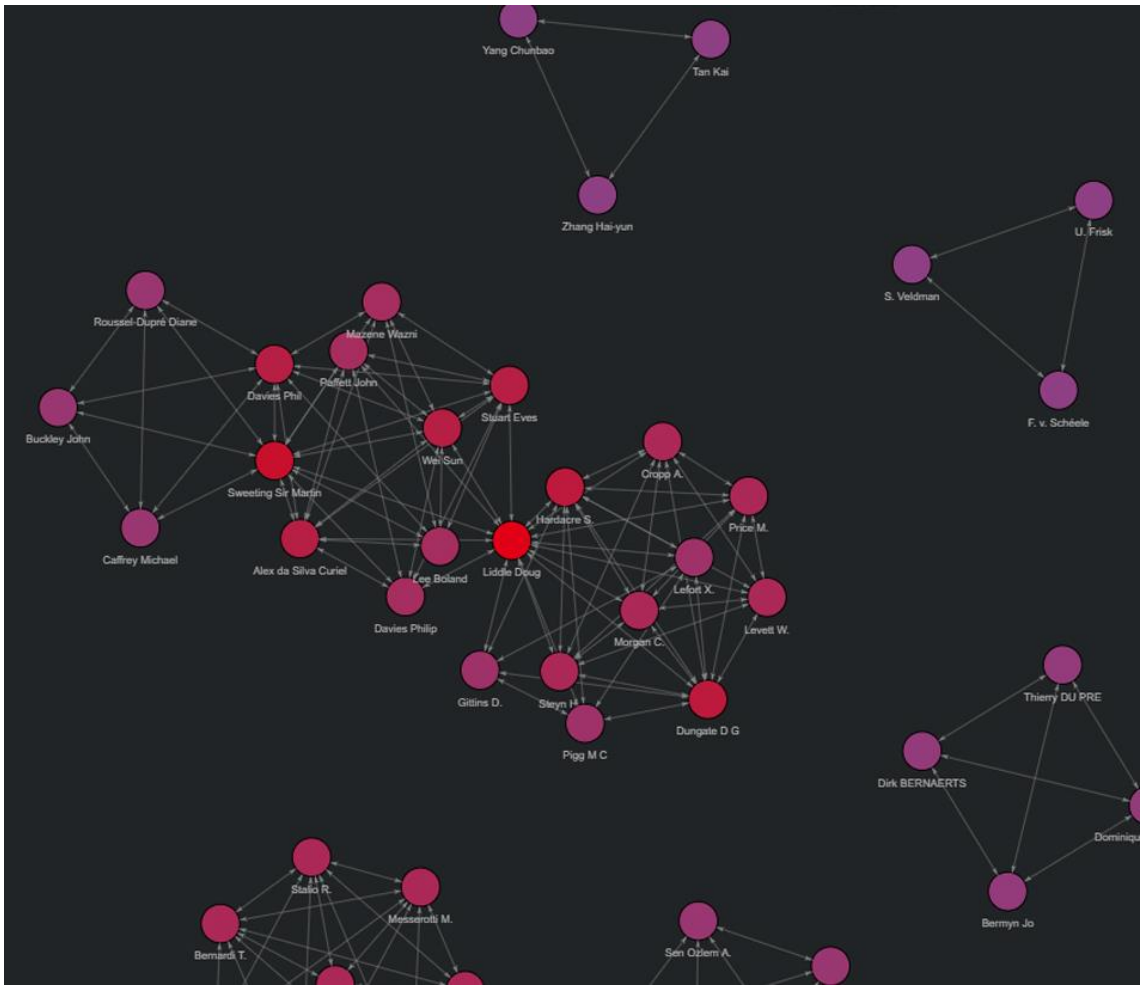

Figure 16. Collaborating authors projection.

4.1.3 Topics by country tab visualizations

The second dashboard tab (Figure 17) includes visualizations related to countries and topics and comprises of six different visualizations. In the center, a choropleth map displays overall regional patterns as well as specific data rates with respect to country contribution to the selected topic. A brief summary of the topic can be seen below the map. The charts on the left and right of the map are connected to it. When the user clicks on a country on the map the other charts update to display data about the clicked country.

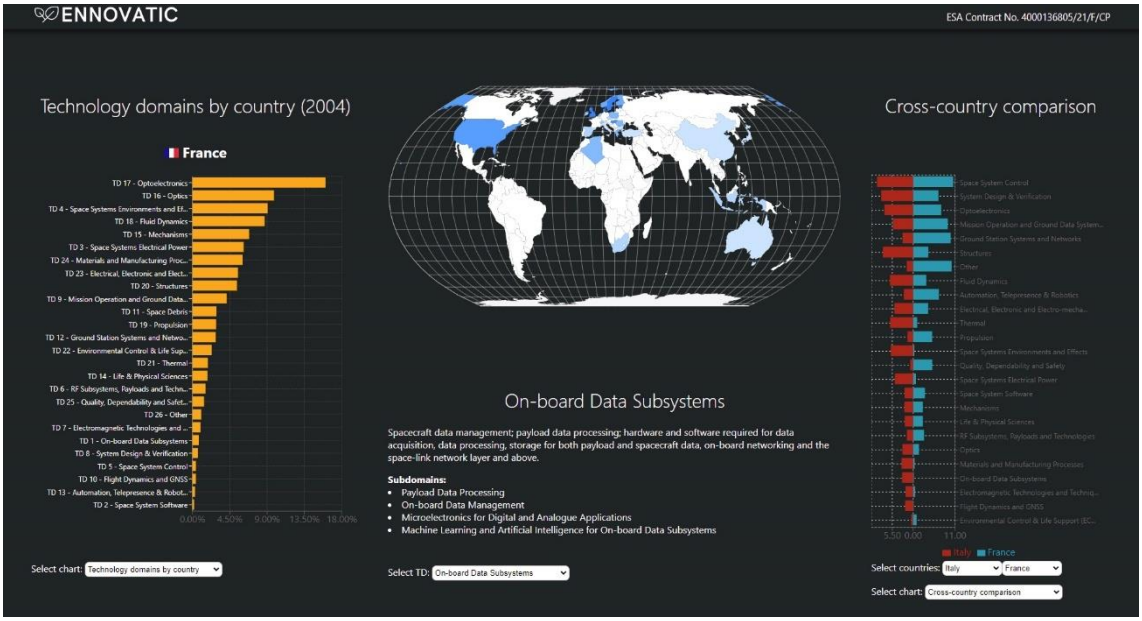

Figure 17. Topics per country tab.

The choropleth map (Figure 18) can be used to focus on a region of interest. Each publication contains a certain percentage of each topic ranging from 0% to 100%. When the user selects a topic, the colors of the countries are updated to represent the contribution percentage of each country to the selected topic, adding up to 100%. The countries that have published most on a given topic are displayed in the darkest tone of blue.

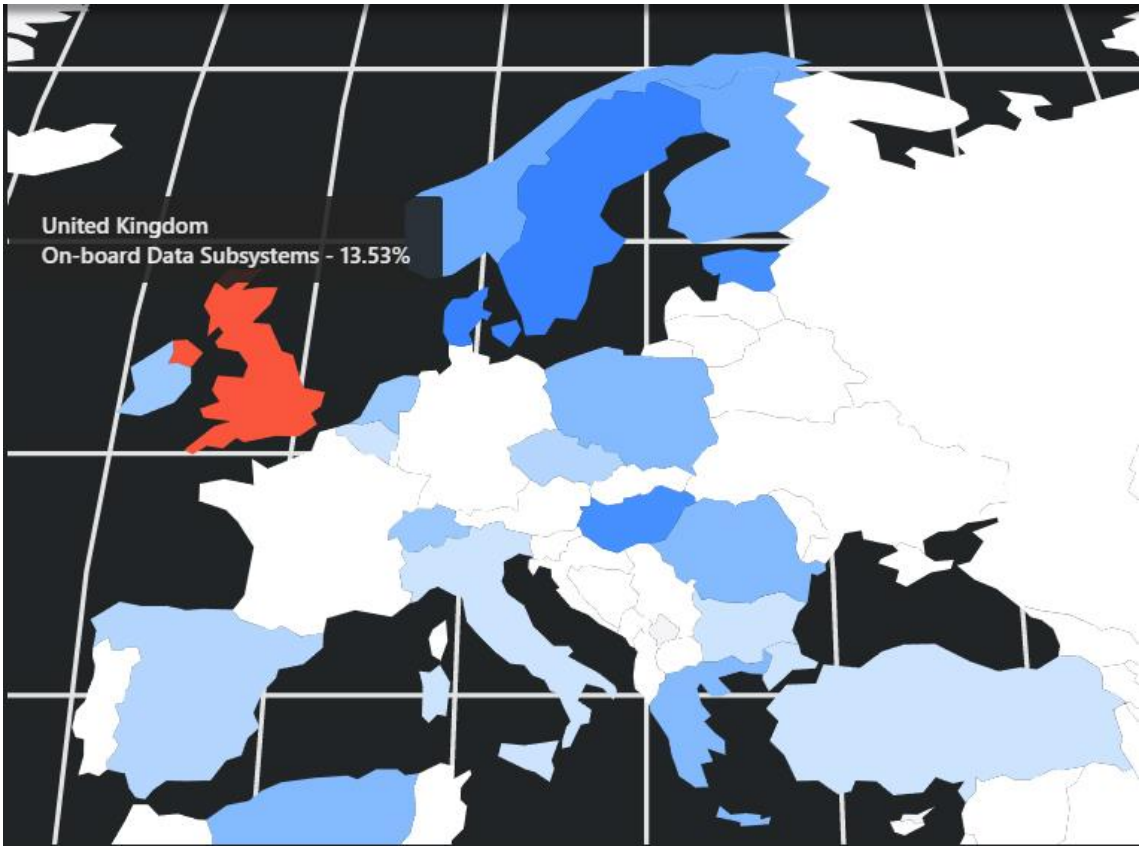

Figure 18. Choropleth map.

The *topics by country chart* (Figure 19) shows which topics the selected country has published on in a given year. The values add up to 100%.

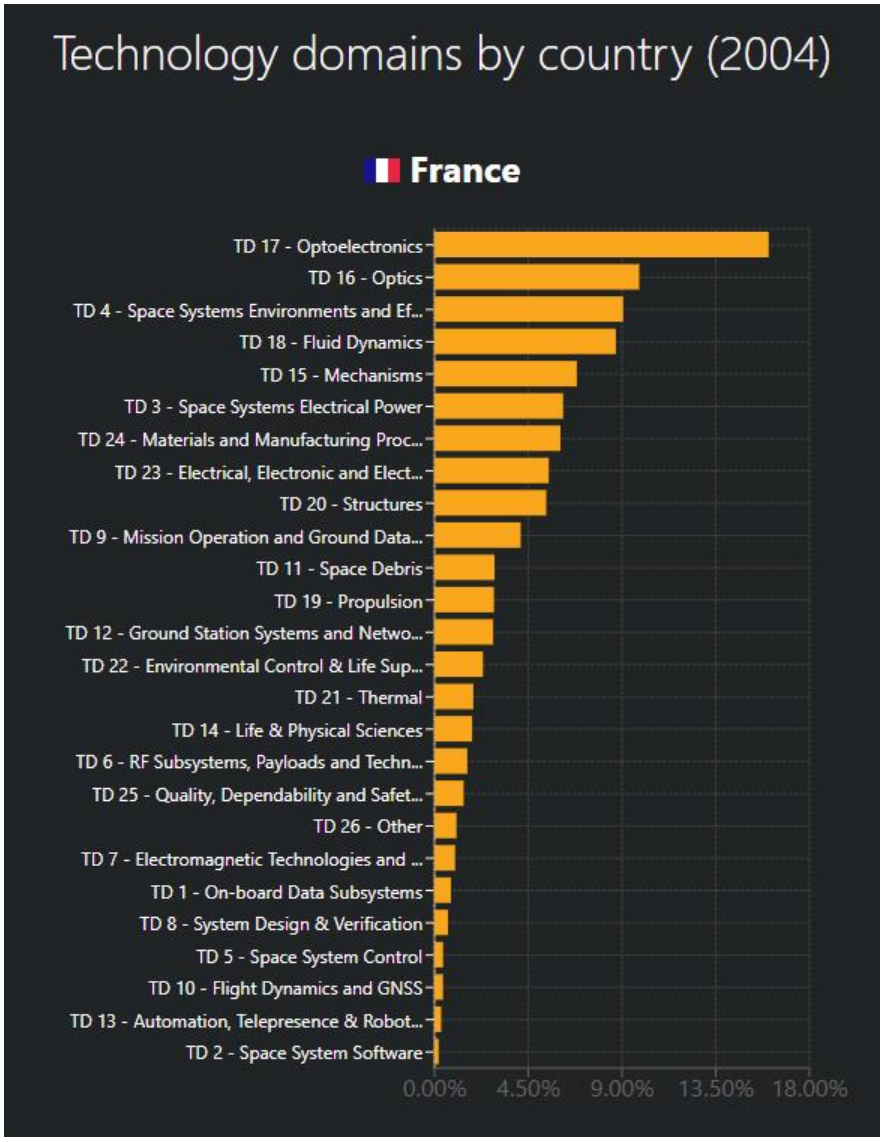

Figure 19. Topics by country (single year).

The *multiyear topics by country chart* (Figure 20) shows how countries have contributed to a specific topic over a period of time. The countries are sorted hierarchically, and the users can select the years they are interested in.

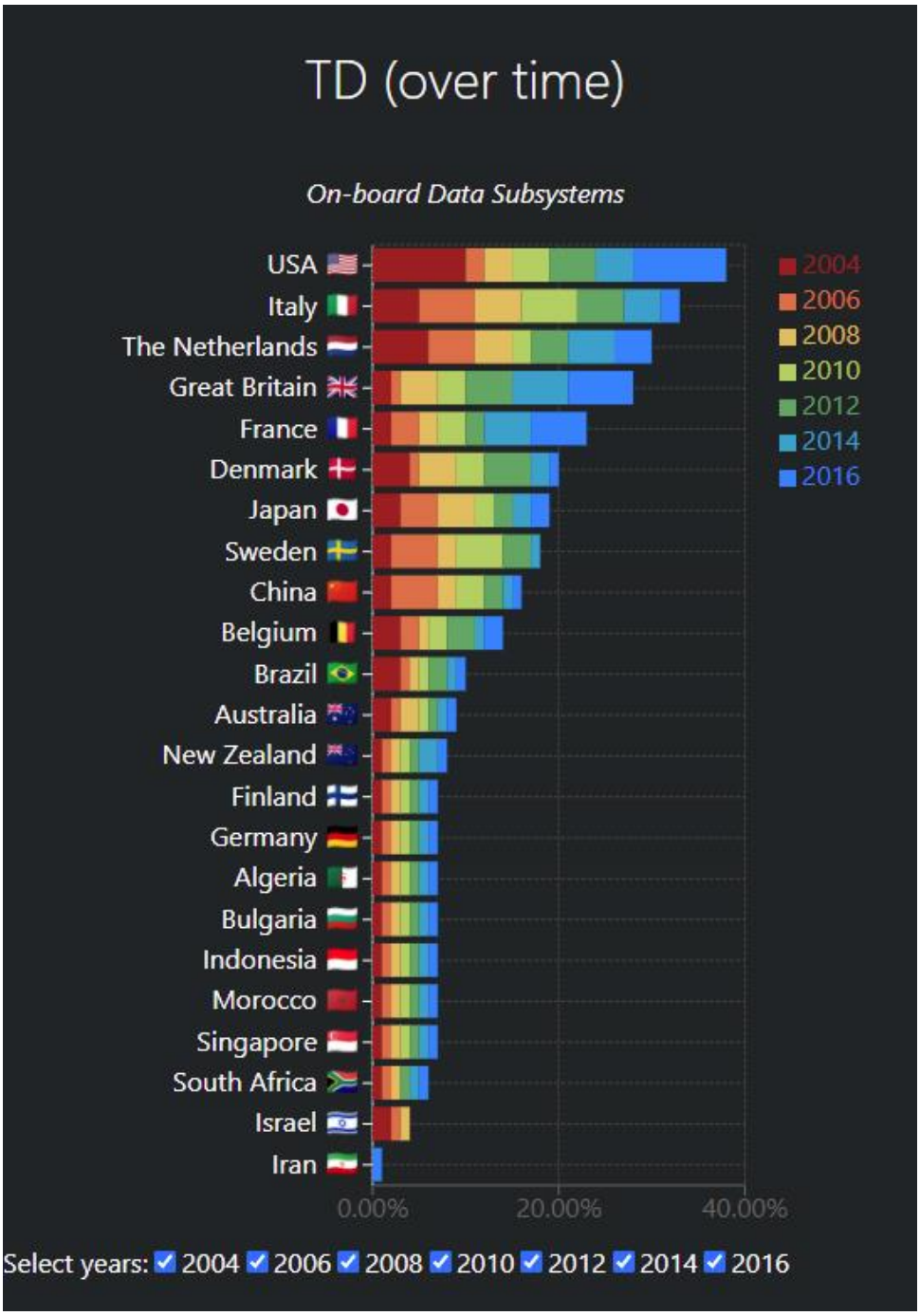

Figure 20. Topic by country (multiyear).

The previous chart can also be represented as a line chart to get a better sense of how a country's contribution to a selected topic has evolved over time (Figure 21). The small sparklines [42] below the main line chart can be reordered to focus only on the countries of interest. When the user clicks a country on the choropleth map or the sparkline for a country, the main line chart updates to display the corresponding data. A trendline shows the prevailing direction of the country's contribution to a topic over time.

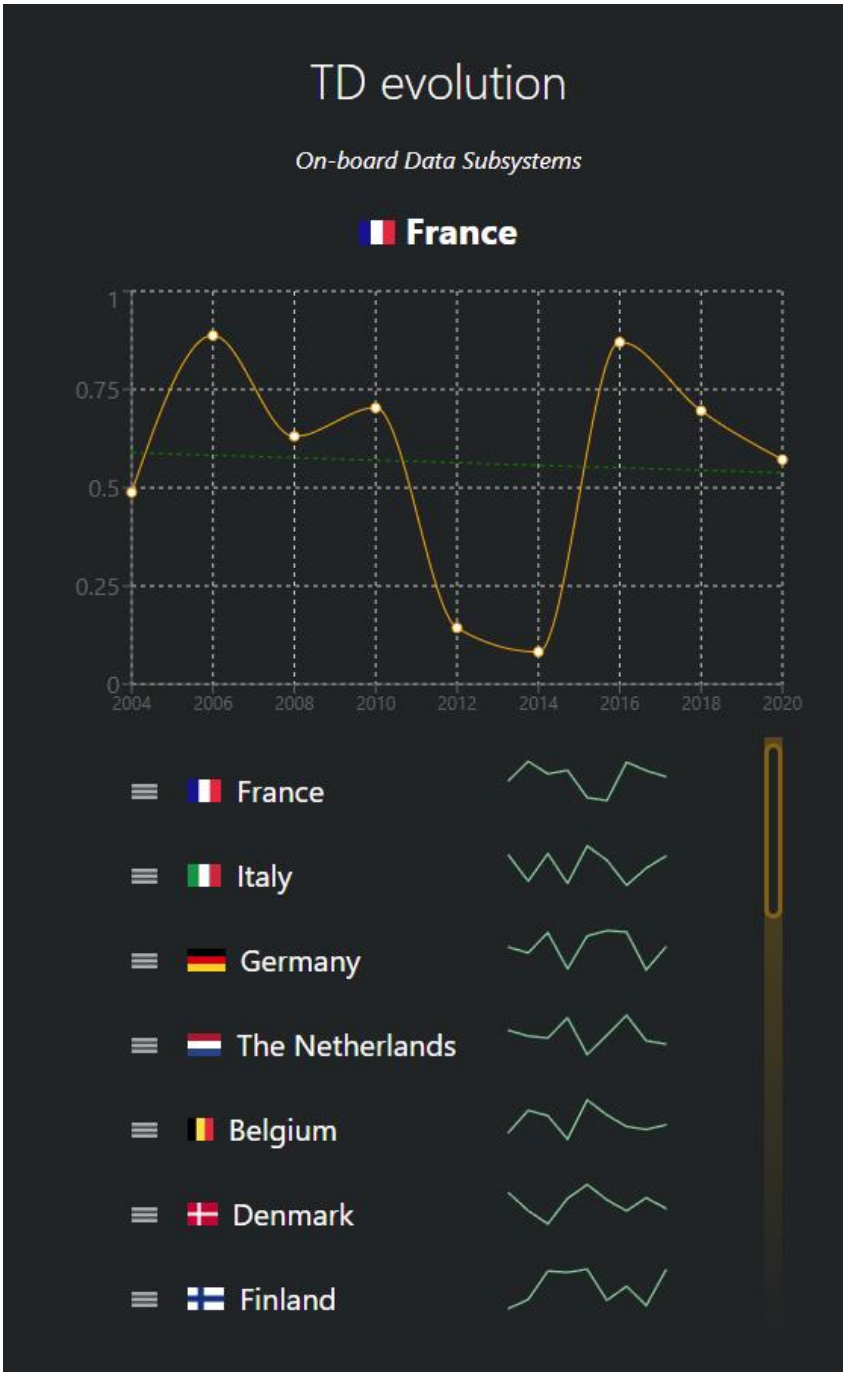

Figure 21. Topic evolution by country.

The *cross-country comparison tornado chart* (Figure 22) allows the user to compare two countries' contribution to each topic side by side.

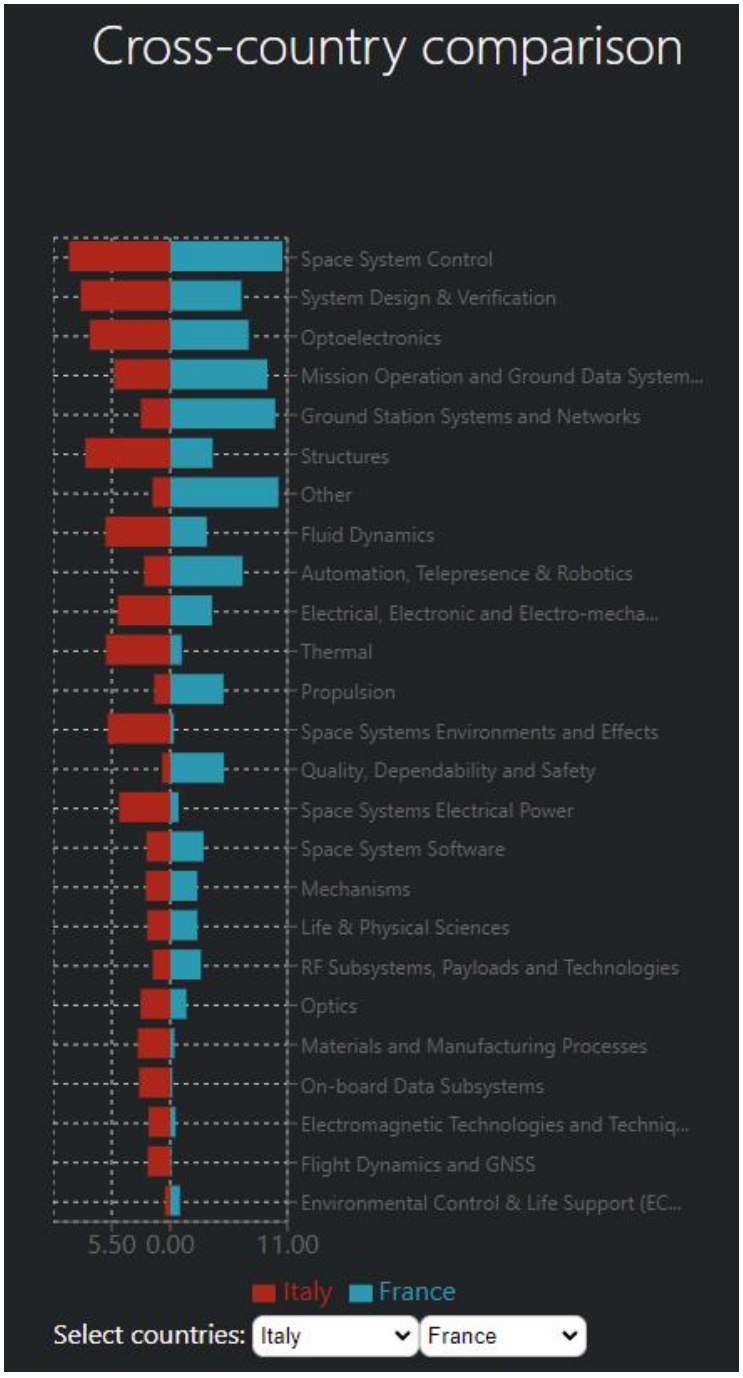

Figure 22. Comparison of country contributions to topic.

Some institutions and organizations of a selected country may have published more on a topic than others. This table gives a succinct overview of the key players in a technology domain, including entities and the authors affiliated with them (Figure 23). When the user clicks on the name of an entity or a person, the page scrolls up, the graph visualization is emptied, and only the node representing the clicked entity will be displayed in focus. The user can then explore the data further by revealing nodes it is connected to by clicking on the focused node.

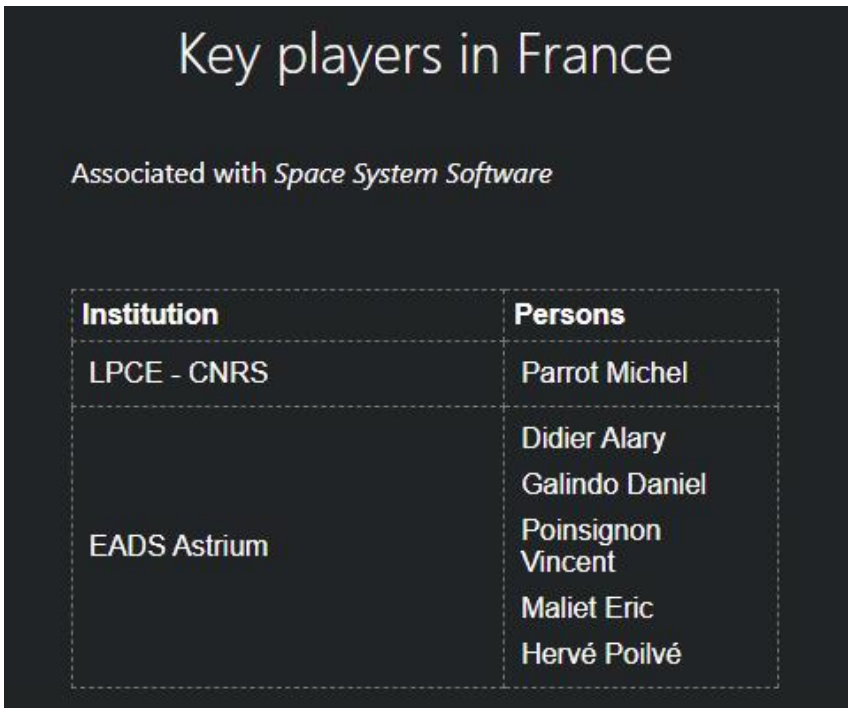

Figure 23. Key players by country.

4.1.4 Topic evolution tab visualizations

The third tab (Figure 24) is split into half, with the left side displaying a multiline chart representing the popularity trends for each topic. The user can display either organic or technology tree topics. The chart on the right displays the corresponding word cloud for the line hovered on the line chart, and the tables below provide statistics for topic popularity through time.

Figure 24. Topic evolution tab.

When the user hovers on a line or legend item below the line graph (Figure 25), the opacity of other lines is decreased to 5%, making it easy to focus solely on the topic of interest.

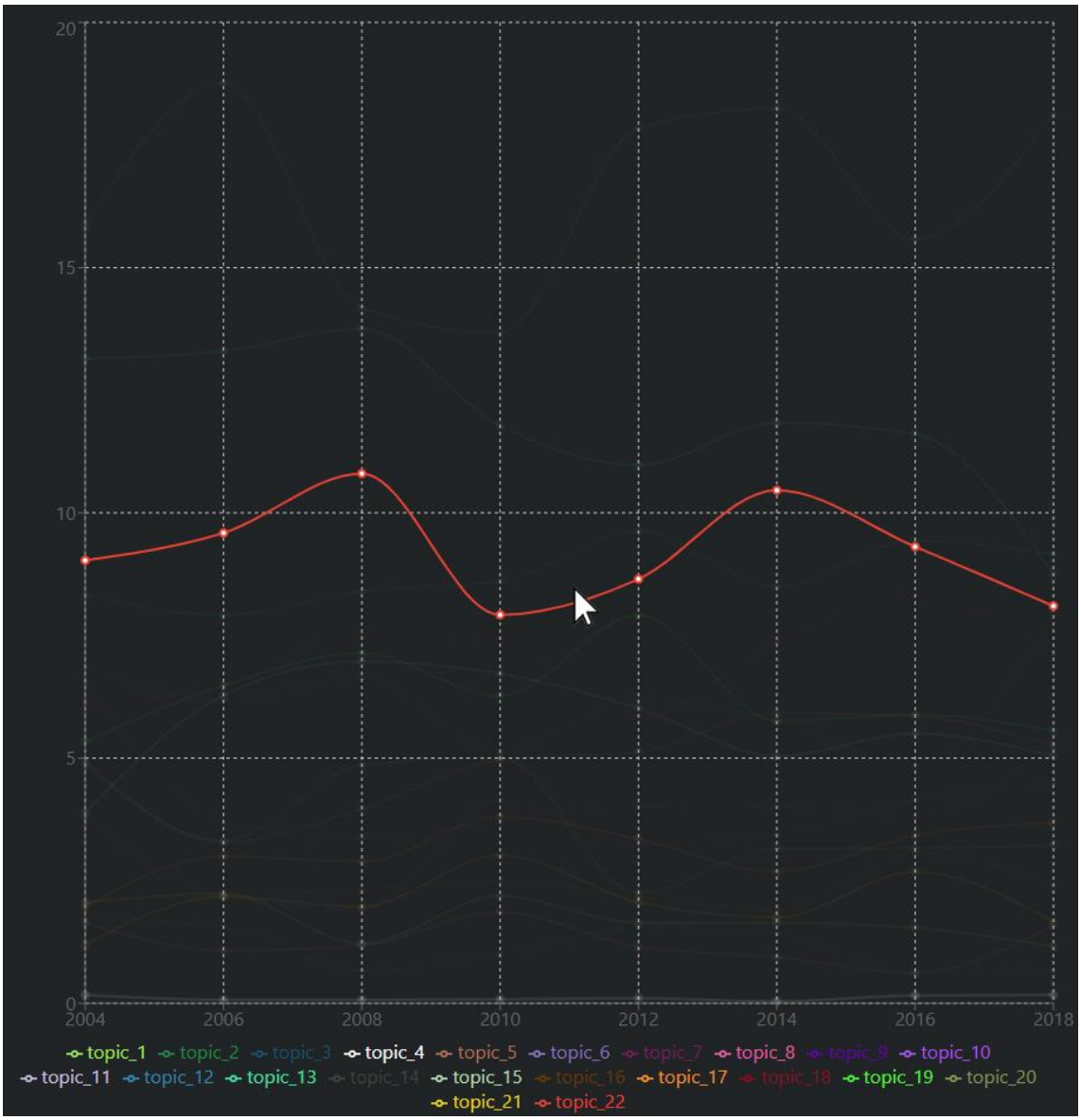

Figure 25. Topic evolution for selected topic.

The relative size of the words in the word cloud represents their importance in the selected topic (Figure 26). The table below provides an overview of the popularity of the topic in each year. The value in the first cell means that of all the papers published in 2004 2.07% were written about the selected topic, in this case topic “T16: Space debris”.

The second table can be used to understand how topic popularity has evolved across all topics in the period specified by the user. The bottom row shows the increase or decrease in popularity for each topic in the specified time frame.

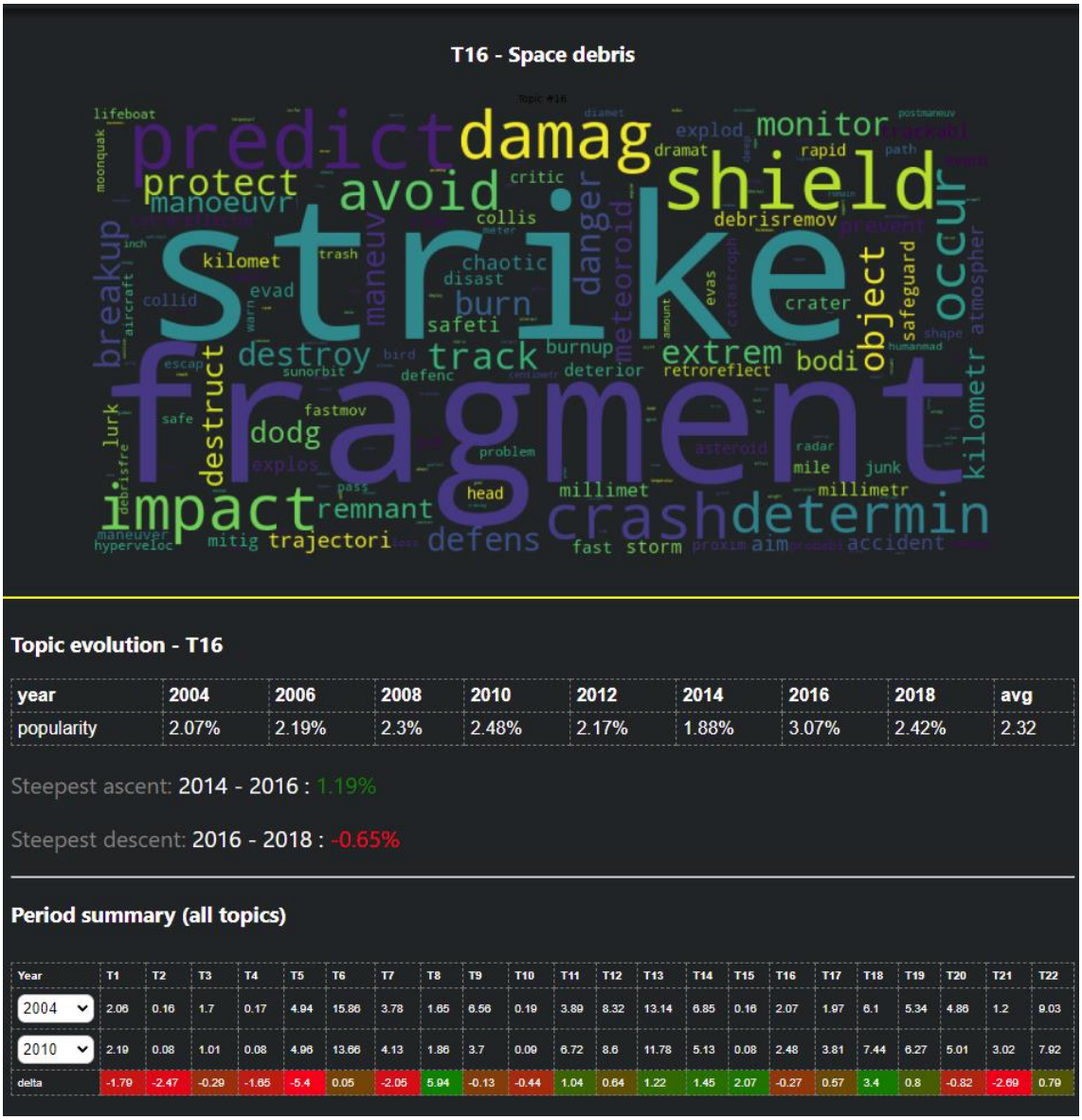

Figure 26. Topic word cloud and corresponding topic evolution summary.

4.2 Topic models

Two different approaches were taken to classify the 4S Symposium publications using the methodologies described in Section 3.2.4. Organic topics refer to topics discovered using the common approach to LDA topic modeling. Technology tree topics refer to topics specifically tailored to match ESA Technology Tree taxonomy.

4.2.1 Organic topics

The organic topics discovered from the 4S Symposium publications were the result of numerous LDA analyses performed on the dataset. The model with nineteen topics resulted in a coherence score of 0.42. Among the many experiments ran, the results of this model were considered the most interpretable by domain experts at the time of writing.

The circles represent the different topics uncovered (Figure 27). A successful analysis with interpretable results generally yields moderately large non-overlapping circles. The overlap between the topics is not substantial, but noticeable.

Figure 27. Intertopic distance map for organic topics.

An example word cloud representing an organic topic is shown in Figure 28. This word cloud was identified to represent the topic “earth observation”. Similarly, labels were given to each of the nineteen word clouds representing the topics discovered by the analysis.

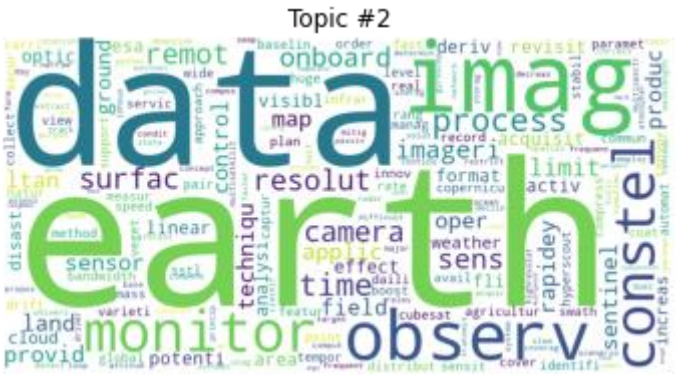

Figure 28. Example organic topic word cloud.

The heat map (Figure 29) shows the topic composition of a selection of publications.

	title	main topic	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11
0	MECHANICS OF DTUsat-2	17	0.002636	0.001432	0.002978	0.252526	0.001753	0.001743	0.002017	0.003121	0.001629	0.001175	0.001244
1	Nano-Satellite Constellation for African-centric Earth Monitoring	5	0.054852	0.020965	0.059643	0.124534	0.423465	0.006416	0.000929	0.008885	0.032417	0.000539	0.000572
2	CHEOPS: SCIENTIFIC OBJECTIVES, MISSION CONCEPT, and CHALLENGES for the SCIENTIFIC COMMUNITY	11	0.003837	0.002081	0.004343	0.010313	0.002533	0.002535	0.002950	0.004419	0.002369	0.001709	0.880409
3	THE EADS MICROPACK PROJECT A HIGHLY INTEGRATED MICROSYSTEM FOR FUTURE SMALL SPACERAFT Steven Eckersley(1), Josef Schalk(2), Martin Kluge(2), Wilhelm Ficker(2)	7	0.004122	0.002238	0.039999	0.133087	0.002722	0.002725	0.508648	0.005157	0.002547	0.001837	0.001946
4	DEFENSE AND CIVIL COLLABORATION FOR THE DEVELOPMENT OF SMALL SATELLITES FOR THE SMALL SATELLITE SYSTEMS AND SERVICES SYMPOSIUM	4	0.087913	0.001200	0.022227	0.680552	0.001460	0.001473	0.001690	0.002552	0.001366	0.000986	0.001044
5	Orbit/Attitude Determination and Control for the UMR SAT Mission	19	0.002666	0.001432	0.002958	0.094710	0.001741	0.001746	0.002016	0.003031	0.001640	0.001175	0.001245
6	IN-FLIGHT EXPERIENCE OF ANUSAT ATTITUDE BEHAVIOR AND DETERMINATION	15	0.013049	0.000601	0.001244	0.024986	0.000731	0.000732	0.000848	0.001281	0.008483	0.000493	0.000522
7	Compact Hyperspectral Imager Breadboard	9	0.001360	0.000739	0.001527	0.009320	0.000899	0.000900	0.001041	0.014382	0.793109	0.000606	0.000642
8	LAUNCH INFRASTRUCTURE FOR SMALL SATELLITE RIDESHARE	4	0.001532	0.000826	0.001726	0.886021	0.001005	0.001006	0.008918	0.019741	0.000941	0.000678	0.008298

Figure 29. Detail of heatmap for organic topics.

4.2.2 Topics tailored to ESA Technology Tree

The technology tree topics were produced as a result of training the LDA model on sets of generic abstracts generated from special vocabularies. Each vocabulary comprised of

words specific to a technology domain. The model with twenty-two topics (corresponding to the number of vocabularies) resulted in a coherence score of 0.82. The topics imaged on the intertopic distance map are mostly homogenous in size with a relatively low level of overlap (Figure 30).

Figure 30. Intertopic distance map for technology tree topics.

An example word cloud representing a technology tree topic is shown in Figure 31. This word cloud represents the technology tree topic “propulsion”.

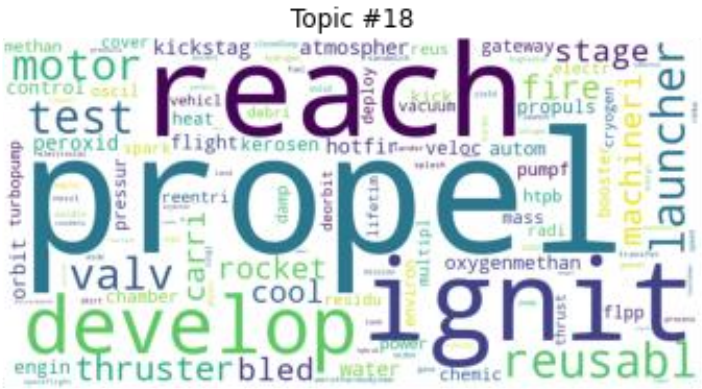

Figure 31. Example technology tree topic word cloud.

Table 2 provides some examples of papers that scored high in this topic.

Table 2. Papers that were assigned “propulsion” as main topic.

Paper title	**Main topic**	**Percentage (%)**
Setting Up a Cold Gas Propulsion System on the Microscope Satellite	#18 – Propulsion	56.17
On the Development of High Specific Impulse Electric Propulsion Thrusters for Small Satellites	#18 – Propulsion	55.86
ALMASAT-1 Cold Gas Micropropulsion System: Final Layout, Qualification and Functional Tests	#18 – Propulsion	47.97
Pulsed Plasma Thruster Development for Cubesats and Nanosatellites	#18 – Propulsion	40.08
Low-cost Launch Becomes Reality: The Spacex Falcon Family of Launch Vehicles	#18 – Propulsion	40.02

5 Analysis

This first section of this chapter provides an analysis of the data visualizations that were developed for the interactive dashboard. The second section describes the outcomes, shortcomings, and potential ways to refine the topic models. The third section provides a description of the scope and limits. The concluding section discusses future directions as well as the potential for reuse.

5.1 Data visualizations

This section evaluates each data visualization with respect to their adherence to data visualization best practices. Alternatives and potential improvements are described where applicable.

5.1.1 Overview tab visualizations

Table 3 presents the analysis of the authors per country chart (Figure 7).

Table 3. Authors per country (single year).

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	Pie chart is useful when showing parts of a whole that add up to 100%. For readability, the number of sectors should not exceed seven. Brief statistics on the number of papers, authors, affiliations, countries, and relations for the chosen year are displayed as plain text and numbers since it does not require a more elaborate chart.
Interactivity	Hovering the sectors displays a tooltip, the contents of which includes the name of the country and the number of authors for this country.
Data filtering	The user can specify a year for which data will be displayed.
Use of preattentive attributes	The yellow-red palette is used to differentiate between sectors, with red tones representing higher percentage values. In the statistics summary, numbers are represented in bold to differentiate them from the category title they represent, and to ease readability.

Aspect	Evaluation
Use of descriptive labels	A legend specifies the country names corresponding to each color. Percentages are displayed on the sectors. Country flags make the chart sectors easier to identify.
Interactions with other charts	N/A
Alternatives	Pie charts are a popular target of criticism. However, as long as the chart is not presented in 3D (skewing the perspective), and displays less than 7 categories, there are situations that warrant its use. A percentage bar chart or a tree map could have been used instead, but most audience are more familiar with the pie chart.

Table 4 presents the analysis of the papers per country chart (Figure 8).

Table 4. Papers per country (single year).

Aspect	Evaluation
Type of visualization	Explanatory
Appropriateness of the chart	A bar chart was used to fit a relatively high number of categories (countries), while maintaining the balance in page layout.
Interactivity	Hovering the bars displays a tooltip, the contents of which include the name of the country and the number of papers contributed by this country.
Data filtering	The user can specify a year for which data will be displayed.
Use of preattentive attributes	The different tones of blue used to color the bars correspond to the numeric value being represented, with darker tones representing higher values. The chart is hierarchically sorted.
Use of descriptive labels	The names of the countries are displayed on the y-axis. Country flags make the axis tick labels easier to identify.
Interactions with other charts	N/A
Alternatives	In case of a different page layout, a column chart could have been used instead.

Table 5 presents the analysis of the authors per affiliation chart (Figure 9).

Table 5. Authors per affiliation (single year).

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A bar chart was used here to fit a relatively high number of categories with long label names while maintaining the balance in page layout.
Interactivity	Hovering the bars displays a tooltip, the contents of which include the name of the entity and the number of affiliations for this entity. The tooltip contains the full expanded name of the entity.
Data filtering	The user can specify a year for which data will be displayed. A brush component below the chart enables the user to increase or decrease the number of datapoints being displayed.
Use of preattentive attributes	The different tones of green used to color the bars correspond to the numeric value being represented, with darker shades representing higher values. The chart is hierarchically sorted.
Use of descriptive labels	The names of the entities are displayed on the y-axis. If the name length exceeds the number of characters that would cause imbalance in the visual composition of the page, a trimmed version of it will be displayed.
Interactions with other charts	N/A
Alternatives	In case of a different page layout, a column chart could have been used instead.

Table 6 presents the analysis of the cross-country collaborations chart (Figure 10).

Table 6. Cross-country collaborations (multiyear).

Aspect	**Evaluation**
Type of visualization	Explanatory and exploratory. Extensive filtering options allow the user to find answers to specific questions.
Appropriateness of the chart	A stacked column chart was used to represent multiyear data for each category with each subcomponent piece representing the value for a single year.
Interactivity	Hovering the columns reveals a tooltip, the contents of which include the name of the country and a list, where each list item includes the year number and a numeric value corresponding to the number of collaborations for the country.
Data filtering	The user can specify a set of years and countries for which data are displayed. A brush component below the chart enables the user to increase or decrease the number of datapoints being displayed. Shortcut buttons for toggling all datapoints have been added.
Use of preattentive attributes	The hues of the subcomponent pieces of the columns represent different years, with more recent years displayed in tones from the blue end of the spectrum. The columns are hierarchically sorted based on the cumulative sum of the subcomponent values of the bars from left to right.
Use of descriptive labels	The names of the categories are displayed on the x-axis at an angle to improve readability. Country flags make the tick labels and selectable options easier to identify. The star symbol following the names of ESA member states make them easier to identify.
Interactions with other charts	N/A
Alternatives	Line chart can be used to represent the same data. However, in that case the user cannot see the cumulative sums of the subcomponent pieces. If the number of categories (years) is increased considerably, the rainbow color scheme will no longer be appropriate.

Table 7 presents the analysis of the authors per country chart (Figure 11).

Table 7. Authors per country (multiyear).

Aspect	**Evaluation**
Type of visualization	Explanatory and exploratory. Extensive filtering options enable the user to find answers to specific questions.
Appropriateness of the chart	A multi-line chart was used to show the time evolution of the value the specified parameter takes over a period of time.
Interactivity	Hovering the gridlines reveals a tooltip, the contents of which include the year and a list, where each list item represents a country name and the corresponding numeric value. Hovering a single line or an item in the legend decreases the opacity of other lines to 5%, effectively displaying only the hovered line.
Data filtering	The chart can be used to view time evolution of five different parameters, these include: authors per country, papers per country, cross-country collaborations, affiliations per country, papers per affiliation. Up to five different countries or entities can be entered into the comparison.
Use of preattentive attributes	Use of distinct colors for the lines improves readability.
Use of descriptive labels	The legend displays the names of the selected items.
Interactions with other charts	Below the main line chart is a simple table that displays the year-by-year values for the line being hovered, including the average. Below the table the steepest ascent and steepest descent of a line is displayed.
Alternatives	In theory a stacked bar chart could be used instead, but the deciphering of trends would become very cumbersome.

5.1.2 Graph visualization

Table 8 presents the analysis of the graph visualization (Figure 12–16).

Table 8. Graph visualization.

Aspect	**Evaluation**
Type of visualization	Exploratory and explanatory
Appropriateness of the chart	The graph visualization was used to represent the scholarly network of the 4S Symposium dataset. Interconnections between different entity types can easily be seen.
Interactivity	The nodes can be moved around. The graph is zoomable and pannable. Hovering a node or an edge reveals a tooltip containing the information about the node, e.g., the name of the author, the abstract of the paper, etc. Holding the left mouse button down on a node will remove it from view. Clicking on a node will cause a database query to be performed which returns the closest neighbors of the clicked node to the graph (if they are not present). When the directed sort method is selected with a hierarchical layout, a multipartite projection of the graph is displayed. The user can toggle the physics and labels of the nodes.
Data filtering	The user can specify the year for which data are displayed. The user can specify a query, the results of which will be displayed on the graph. The queries include top 10 most affiliated entities, top 10 most published authors, top 3 most collaborative countries, top 3 least collaborative countries, country collaboration projection, entity collaboration projection, author collaboration projection.
Use of preattentive attributes	The nodes have distinctive colors for easy identification. In country and entity collaboration projection, the size of each node represents its number of collaborations. In author collaboration projection, the color intensity of each node represents its degree (number of collaborations).
Use of descriptive labels	Labels are displayed below the nodes. The labels can be toggled on or off.
Interactions with other charts	The graph visualization is connected to key players by country chart (see Figure 23 and Table 13).
Alternatives	In theory, a very large matrix could be used to show some of the same insights. However, it would be exceedingly difficult for the user to interpret.

5.1.3 Topics by country tab visualizations

Table 9 presents the analysis of the choropleth map chart (Figure 18).

Table 9. Choropleth map.

Aspect	**Evaluation**
Type of visualization	Exploratory
Appropriateness of the chart	A choropleth map was used because it enables the user to see overall regional patterns as well as specific data rates.
Interactivity	The map is zoomable and pannable. Hovering on a country reveals a tooltip, the contents of which include the name of the country and the percentage value of the country's contribution to the selected topic.
Data filtering	The user can select a specific topic and the map is updated to display the values corresponding to each country with respect to the selected topic.
Use of preattentive attributes	The tone of blue used to color a country represents the percentage value of the country's contribution to the selected topic, with darker blue tones representing higher values.
Use of descriptive labels	N/A
Interactions with other charts	When the user clicks on a country, various charts in the same tab will be updated to display data of the clicked country.
Alternatives	A simple bar chart could be used instead, but this would remove the interesting spatial insights that a map provides.

Table 10 presents the analysis of the topics by country chart (Figure 19).

Table 10. Topics by country (single year).

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A bar chart was used to fit the relatively high number of categories being displayed while maintaining the overall visual hierarchy of the dashboard layout.
Interactivity	Hovering the bars displays a tooltip, the contents of which include the title of the topic and the corresponding percentage value. The tooltip contains the fully expanded name of the topic.
Data filtering	N/A
Use of preattentive attributes	The bars are sorted hierarchically based on the percentage value they represent from highest to lowest.
Use of descriptive labels	The name and flag of the country is displayed above the chart. The y-axis tick labels include an identification code as well as the title of the topic.
Interactions with other charts	When the user clicks on a country on the choropleth map, data about the clicked country will be displayed in this chart.
Alternatives	In case of a different page layout, a column chart could have been used instead.

Table 11 presents the analysis of the topic by country chart (Figure 20).

Table 11. Topic by country (multiyear).

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A stacked bar chart was used to represent multiyear data for each country, with each subcomponent piece representing the percentage value for a single year.
Interactivity	Hovering the bars reveals a tooltip, the contents of which include the name of the country and a list, where each item includes the year and a corresponding percentage value.
Data filtering	The user can specify a set of years for which data will be displayed.
Use of preattentive attributes	The hues of the subcomponent pieces of the columns represent different years, with more recent years displayed in tones from the blue end of the spectrum. The columns are hierarchically sorted based on the cumulative sum of the subcomponent values of the bars from top to bottom.
Use of descriptive labels	Country flags make axis tick values easier to identify.
Interactions with other charts	N/A
Alternatives	Given additional filtering options a line chart can be used to represent the same data. However, in that case the user cannot see the cumulative sums of the subcomponent pieces.

Table 12 presents the analysis of the topic evolution by country chart (Figure 21).

Table 12. Topic evolution by country.

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A line chart was used to display the time evolution of the percentage value. A trendline specifies the overall trend of the data.
Interactivity	Hovering the gridlines reveals a tooltip, the contents of which include the year, the title of the topic and the corresponding percentage value. When the user clicks on the small sparklines below, the main chart will update to display data for the corresponding country.
Data filtering	The user can drag and drop countries to form a custom order, making it easy to compare countries of interest. The user can specify the topic for which data are displayed.
Use of preattentive attributes	Points on the line highlight x-axis tick locations.
Use of descriptive labels	Country flags make the sparklines easier to identify.
Interactions with other charts	When the user clicks on a country on the choropleth map, data about the clicked country will be displayed in the main line chart.
Alternatives	N/A

Table 13 presents the analysis of the country contributions to topic chart (Figure 22).

Table 13. Comparison of country contributions to topic.

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A tornado chart was used to compare two countries across multiple distinct categories.
Interactivity	Hovering the bars reveals a tooltip, the contents of which include the full title of the topic, country names, and the corresponding percentage values for both countries.
Data filtering	The user can select which countries to compare.
Use of preattentive attributes	The chart uses complementary colors to easily differentiate between the two countries being compared.
Use of descriptive labels	The titles of the topics are displayed on the y-axis. If the name length exceeds the number of characters that would cause imbalance in the visual composition of the page, a trimmed version of it will be displayed.
Interactions with other charts	N/A
Alternatives	A table with numbers can be used instead, but the use of bars provides a faster and more intuitive overview of the data.

Table 14 presents the analysis of the key players by country chart (Figure 23).

Table 14. Key players by country.

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A table was used to succinctly summarize key players in the selected technology domain for the selected country.
Interactivity	The user can click on the values in the table to send them to the graph visualization for further data exploration.
Data filtering	N/A
Use of preattentive attributes	N/A
Use of descriptive labels	N/A
Interactions with other charts	When the user clicks on a country on the choropleth map, the key players for that country with respect to the selected topic will be displayed in the table. Clicking on a name in the table updates the graph visualization.
Alternatives	N/A

5.1.4 Topic evolution tab visualizations

Table 15 presents the analysis of the topic evolution chart (Figure 25).

Table 15. Topic evolution for selected topic.

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A multi-line chart was used to convey the time evolution of topic popularity over a period of time.
Interactivity	Hovering a single line or an item in the legend decreases the opacity of other lines to 5%, effectively highlighting only the line of interest.
Data filtering	The user can display either organic topics or technology tree topics.
Use of preattentive attributes	Easily distinguishable colors are used for the lines.
Use of descriptive labels	N/A
Interactions with other charts	When the user hovers a line or a legend item, the word cloud and topic evolution summary tables on the right update to display data about the hovered line.
Alternatives	A line chart with too many lines is sometimes referred to as a “spaghetti graph” [6, p 227]. However, the filtering options makes it possible to focus on only one line at a time. Extending the filtering options, as in the authors per country multiyear chart (see Figure 11.), could improve this visualization.

Table 16 presents the analysis of the topic word cloud and topic evolution summary chart (Figure 26).

Table 16. Topic word cloud and corresponding topic evolution summary.

Aspect	**Evaluation**
Type of visualization	Explanatory
Appropriateness of the chart	A word cloud was used to represent the keywords that make up the topic. The size of the words represents their relative importance in the topic. Tables are used to communicate succinct summaries.
Interactivity	N/A
Data filtering	In the period summary table, the user can specify a time period of interest.
Use of preattentive attributes	The cells in the period summary table are colored in tones of red or green depending on whether the popularity of the selected topic increased or decreased in the specified period. Color intensity represents the magnitude of increase or decrease.
Use of descriptive labels	The label of the topic is displayed above the word cloud. Both tables have descriptive titles and table headers.
Interactions with other charts	The chart is connected to the line chart on the left side of the tab. When the user hovers a line or legend item, the chart will update to display the data for the line being hovered.
Alternatives	N/A

The visualizations received a positive evaluation, and their appropriateness was verified by the client. The author finds that some data could be effectively visualized in more ways than one. For example, data that are related to countries, can be presented in a bar chart or a choropleth map, with the latter offering additional geospatial insights. In addition, charts showing multiyear data can be presented using either a line chart or a stacked column chart. Line charts are especially useful for understanding trends. The stacked column chart, in contrast, is better suited when the cumulative value of a parameter across a number of years is of interest—something that a line chart does not communicate well. In future iterations of the dashboard, additional functionality allowing the user to choose between alternative chart types to get insights on different aspects of the data may be included. An option to toggle between light and dark mode should be added.

To encourage future feedback from users, a suggestion box was added to the bottom of the dashboard, which enables the audience to share their ideas on how parts of the dashboard could be improved (Figure 32).

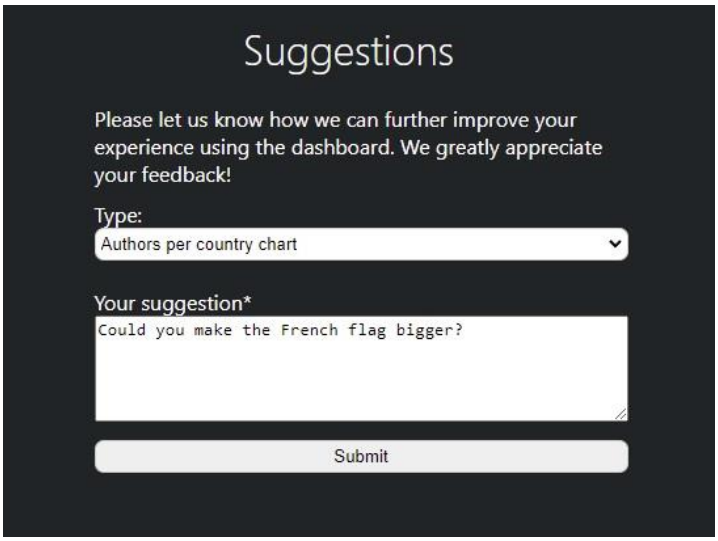

Figure 32. Suggestion box.

The suggestions made by the users can be viewed and prioritized by the team on a separate dedicated client developed for the solution (Figure 33).

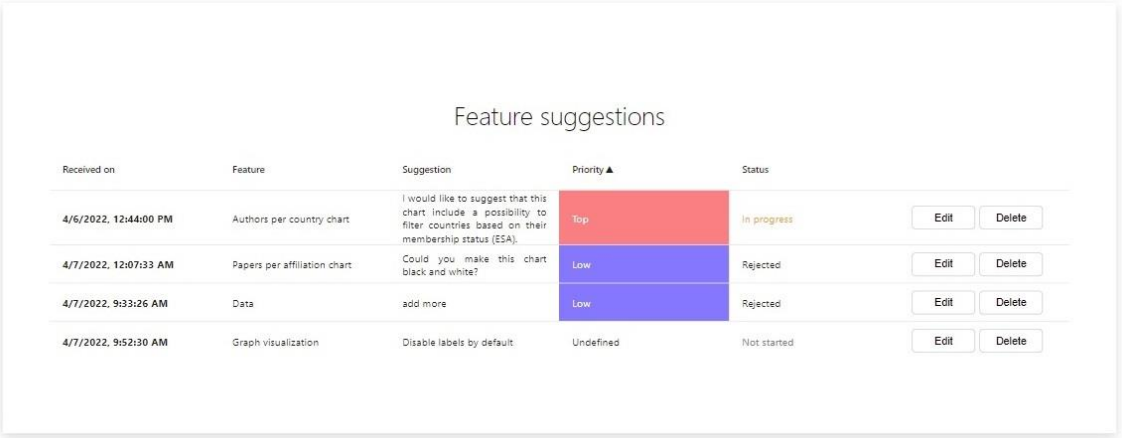

Figure 33. Custom ticketing system UI.

5.2 Topic modeling outcomes

Organic topics were discovered as a result of applying the common approach to LDA analysis to the text corpus. Multiple analyses were run, varying the number of topics, model hyperparameters, as well as the set of stop words.

One of the disadvantages of the traditional approach is that for each model, a fixed number of topics must be defined beforehand [43]. It can be difficult to determine whether a large corpus of texts is best described in terms of one hundred or just ten topics, without looking at the contents of the documents. A high coherence score alone cannot be relied on to establish the optimal number of topics for a highly interpretable model. The topics discovered were occasionally vague and merged together seemingly unrelated themes.

In some cases, the topic assigned to a document reflected the subject matter (described in the title or the abstract of a document) exceptionally well. In other cases, the connection was fairly difficult to establish. Given that the dataset was not very large, and only the abstract portion of the papers was utilized for the analysis, two enhancements should be considered to improve the performance of similar analyses—increasing the volume of documents and analysing the full texts.

The extraction of highly interpretable organic topics from the 4S Symposium publications is still a work in process and the results will improve in time.

Training the LDA model to distinguish technology tree topics in documents was achieved by extracting specialized vocabularies from texts written on the corresponding technology domains. The model performance was enhanced iteratively by improving the vocabularies.

The model yielded a relatively high coherence score. Overall, the majority (91%) of technology tree topic were easily identifiable by looking at their respective word clouds. The circles on the topic interdistance map on Figure 28, representing the topics, are in general very homogenous in size. However, there is some overlap, and one of the circles is noticeably larger than the others. Multiple technology domains may use similar terms, which causes overlap in the vocabularies. This can be corrected by making the vocabularies more specific.

It was noted by experts reviewing the results, that the predefined topic taxonomy does not account for all important topics present in the publications—papers describing mission design and other broad topics require additional vocabularies. Another interesting problem that arose during the review process of the technology tree topics, was that publications that were assigned an accurate topic in general, could be more accurately classified to distinguish the general subject area from the specific object of research. For example, a paper may be assigned the topic "space environments and effects", but this general label does not carry enough information for deciding whether the paper is about space environment per se, or technological applications specific to the space environment. Some vocabularies could be split into two separate ones, with the first including more general terms and the second the related technological terminology. Improving and extending the vocabularies to enable even more accurate classification is the subject of ongoing work.

5.3 Scope and limitations

The interactive dashboard is accessible solely to the client and not publicly available on the internet. With the exception of a minimalistic service developed to enable users of the interactive dashboard to send feedback, and the OrientDB server instance, no additional dedicated backend infrastructure was planned or developed. For these reasons, authentication and authorization mechanisms were not among the requested functionalities described by the client. The dashboard communicates with the OrientDB server instance via the HTTP REST protocol, utilising server-side functions written to query the database and retrieve data in JSON format. The dedicated user created for interfacing with the database was granted read-only permissions.

The intended date of completion for the project is in the second half of 2022, and the features and functionality described in this thesis are subject to change. The described dashboard is a prototype of the final product. Due to the database contents being incomplete at the time of writing, the visualizations presented in this work may not reflect the state of final data. The topic models described are not finalized, and the classification results they yield are still subject to rigorous internal and external validation. All source code produced for the project is the intellectual property of Ennovatic OÜ and will not be shared.

5.4 Future directions

The work described in this thesis focused on the development of only parts of a more general system for scholarly data analytics. As described in Section 3.1, the performance of topic analysis and the output of the interactive dashboard rely on some preceding steps, namely, extraction of metadata from raw documents and ingestion of these data in a graph database. The development of effective methods for extracting data from raw documents is an area of ongoing research [44]. Dealing with non-uniform formatting is a big open challenge in scholarly data extraction from scientific publications.

Given a robust enough data extraction system and increased automation between the various parts of the system, the software being developed could in theory be used to conduct similar analyses on any corpora of academic and scientific literature. The described multi-label classification approach based on LDA topic modeling can be used to classify documents to any set of topics, provided that the topics are known, and vocabularies for them can be generated.

The interactive dashboard and the OrientDB server-side functions were designed to be flexible enough to allow for the querying and displaying of analogous scholarly data insights for any dataset with a similar structure and size, irrespective of the subject matter. To the best knowledge of the author, no end-to-end software enabling the transformation of a corpus of papers into actionable insights displayable on an interactive dashboard exist at the time of writing of this thesis.

6 Summary

Effective utilization of data can support an organization's knowledge management efforts. However, large and complex datasets are often difficult to interpret. This thesis was focused on the development of parts of a software enabling the discovery and communication of actionable insights from the scholarly data extracted from the 4S Symposium body of work from 2004–2018. The topic distribution of the documents, with reference to the ESA Technology Tree, was discovered by training an LDA model on specialized vocabularies. The data visualizations were developed and evaluated based on accepted best practices. The resulting web-based interactive dashboard includes explanatory and exploratory visualizations that provide insights on the evolution of the 4S Symposium. Topic modeling was successfully used for multi-label classification of a corpus of texts to a predefined set of topics. The prototype of the dashboard was approved by the client. The proposed method for document classification could be extended to other domains. The visualization tool is flexible enough to enable displaying of analogous scholarly data insights for any dataset with a similar structure and size, irrespective of the subject matter.

References

[1] J. P. Girard and J. L. Girard, "Defining knowledge management: Toward an applied compendium," *Online Journal of Applied Knowledge Management,* vol. 3, no. 14, pp. 1–20, 2015.

[2] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," *Journal of machine Learning research,* vol. 3, pp. 993–1022, 2003.

[3] R. T. Edward, The Visual Display of Quantitative Information, Chesire, Connecticut: Graphics Press, 2001.

[4] K. Kostelnick, "The Re-Emergence of Emotional Appeals in Interactive Data Visualization," *Technical Communication,* vol. 63, no. 2, pp. 116–135, 2016.

[5] T. Hiippala, "A multimodal perspective on data visualization," in *Data Visualization in Society*, M. Engebretsen and H. Kennedy, Eds., Amsterdam, Amsterdam University Press, 2020, pp. 276–293.

[6] C. Nussbaumer Knaflic, Storytelling with data, John Wiley & Sons, 2015.

[7] "Gestalt Principles," Interaction Design Foundation, [Online]. Available: https://www.interaction-design.org/literature/topics/gestalt-principles. [Accessed May 15, 2022].

[8] R. Reitsma and A. Marks, "The Future of Data: Too Much Visualization, Too Little Understanding?," *Dialectic,* vol. 2, no. 2, pp. 109–130, 2019.

[9] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong and F. Xia, "A Survey of Scholarly Data Visualization," *IEEE Access,* vol. 6, pp. 19205–19221, 2018.

[10] D. M. Blei, "Probabilistic topic models," *Communications of the ACM,* vol. 55, no. 4, p. 77–84, 2012.

[11] M. Hoffman, D. Blei and F. Bach, "Online Learning for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems,* vol. 23, p. 856–864, 2010.

[12] V. B. Sowmya, B. Majumder, A. Gupta and H. Surana, Practical Natural Language Processing, O'Reilly Media, Inc., 2020, p. 250.

[13] P. Kherwa and B. Poonam, "Topic Modeling: A Comprehensive Review," *EAI Endorsed Transactions on Scalable Information Systems,* vol. 7, no. 24, 2020.

[14] R. Albalawi, T. H. Yeap and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Frontiers in Artificial Intelligence,* vol. 3, no. 42, 2020.

[15] I. Vayansky and S. Kumar, "A review of topic modeling methods, Information Systems," 2020.

[16] Z. Liu, M. Li, Y. Liu and M. Ponraj, "Performance evaluation of latent dirichlet allocation in text mining," in *Proceedings - 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2011*, Shanghai, 2011.

[17] M. Maia, J. E. Sales, A. Freitas, S. Handschuh and M. Endres, “A Comparative Study of Deep Neural Network Models on Multi-Label Text Classification in Finance,” in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, 2021.

[18] N. Garzaniti, Z. Tekic, D. Kukolj and A. Golkar, “Review of technology trends in new space missions using a patent analytics approach,” *Progress in Aerospace Sciences,* vol. 125, 2021.

[19] S. Syed and M. Spruit, “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation,” in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Tokyo, 2017.

[20] E. Communications, “ESA Technology Tree, version 3.0,” April 2020. [Online]. Available: https://esamultimedia.esa.int/multimedia/publications/STM-277/STM-277.pdf. [Accessed May 15, 2022].

[21] “OrientDB Community Edition,” OrientDB Ltd, [Online]. Available: https://orientdb.org/. [Accessed May 15, 2022].

[22] “Microsoft Power BI: Data Visualisation,” Microsoft, [Online]. Available: https://powerbi.microsoft.com/. [Accessed May 15, 2022].

[23] “Tableau Online,” Tableau, [Online]. Available: https://www.tableau.com/products/cloud-bi. [Accessed May 15, 2022].

[24] “React - A JavaScript library for building user interfaces,” Facebook, [Online]. Available: https://reactjs.org/. [Accessed May 15, 2022].

[25] E. Ercan, “A Comparison of Data Visualization Libraries for React,” Capital One, 23 9 2020. [Online]. Available: https://www.capitalone.com/tech/software-engineering/comparison-data-visualization-libraries-for-react/. [Accessed May 15, 2022].

[26] “Redefined chart library built with React and D3,” Recharts, [Online]. Available: https://github.com/recharts/recharts. [Accessed May 15, 2022].

[27] “ReGraph - Graph Visualization Software for React Developers,” Cambridge Intelligence, [Online]. Available: https://cambridge-intelligence.com/regraph/. [Accessed May 15, 2022].

[28] “Cytoscape.js,” Cytoscape, [Online]. Available: https://js.cytoscape.org/. [Accessed May 15, 2022].

[29] “Vis.js - A dynamic, browser based visualization library,” Vis.js, [Online]. Available: https://visjs.org/. [Accessed May 15, 2022].

[30] “9 Top Programming Languages for Data Science,” edX, [Online]. Available: https://blog.edx.org/9-top-programming-languages-for-data-science. [Accessed May 15, 2022].

[31] J. Perkel, “Why Jupyter is data scientists’ computational notebook of choice,” *Nature,* vol. 563, p. 145–146, 2018.

[32] “Gensim: Topic modelling for humans - Radim Řehůřek,” RARE Technologies Ltd., [Online]. Available: https://radimrehurek.com/gensim/. [Accessed May 15, 2022].

[33] P. Willett, “The Porter stemming algorithm: then and now,” *Program Electronic Library and Information Systems,* vol. 40, 2006.

[34] “pyLDAvis 2.1.2 documentation,” pyLDAvis, [Online]. Available: https://pyldavis.readthedocs.io/en/latest/readme.html. [Accessed May 15, 2022].

[35] J. R. Wertz, D. . F. Everett and J. J. Puschell, Space Mission Engineering: The New SMAD, Hawthorne, CA: Microcosm Press, 2011.

[36] A. Ronacher, “Welcome to Flask — Flask Documentation (2.1.x),” [Online]. Available: https://flask.palletsprojects.com/en/2.1.x/. [Accessed May 15, 2022].

[37] “Appropriate Uses For SQLite,” SQLite, [Online]. Available: https://www.sqlite.org/whentouse.html. [Accessed May 15, 2022].

[38] “Visual Studio Code - Code Editing. Redefined,” Microsoft, [Online]. Available: https://code.visualstudio.com/. [Accessed May 15, 2022].

[39] E. Illaste, “Demo of Interactive Dashboard,” Youtube, [Online]. Available: https://youtu.be/c-c83AusQ-8. [Accessed May 15, 2022].

[40] “Hooks API Reference – React,” Facebook, [Online]. Available: https://reactjs.org/docs/hooks-reference.html. [Accessed May 15, 2022].

[41] E. Ma, “Chapter 9: Bipartite Graphs - Network Analysis Made Simple,” [Online]. Available: https://ericmjl.github.io/Network-Analysis-Made-Simple/04-advanced/01-bipartite/. [Accessed May 15, 2022].

[42] “Sparklines: Another masterpiece of Edward Tufte,” Bissantz & Company GmbH., [Online]. Available: https://web.archive.org/web/20070311173343/http://www.bissantz.com/sparklines/. [Accessed May 15. 2022].

[43] T. Griffiths and M. Steyvers, “Finding Scientific Topics,” *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101 Suppl 1, p. 5228–35, 2004.

[44] F. Peng and A. McCallum, “Information extraction from research papers using conditional random fields,” *Information Processing & Management,* vol. 42, no. 4, p. 963–979.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I, Erik Illaste

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Development of an Interactive Tool to Support Knowledge Management", supervised by Toomas Lepikult, and co-supervised by Alessandro Aliakbargolkar
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

16.05.2022

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.