DOI: 10.1111/cgf.70158

COMPUTER GRAPHICS *forum*
Volume 44 (2025), number 6, e70158

Exploratory Analysis of Scientific Publications for University Governance

A. Gràcia,[1] L. Padró,[1,2] E. Alarcon[3] and P. Vázquez[1,4]

[1]Dep. Computer Science - UPC, Barcelona, Spain
{alexandre.gracia, lluis.padro, pere.pau.vazquez}@upc.edu
[2]TALP Research Center - UPC, Barcelona, Spain
[3]Dep. Electrical Engineering - UPC, Barcelona, Spain
eduard.alarcon@upc.edu
[4]ViRVIG Group - UPC, Barcelona, Spain

Abstract

Research-oriented universities often comprise numerous researchers of various types and possess complex research structures that encompass research groups, departments, laboratories, and research institutes. In this situation, understanding the university's strengths and areas of excellence requires careful examination. Additionally, individuals at different levels of governance (e.g., department heads, directors of research institutes, rectors) may seek to establish synergies among researchers to tackle issues such as international project applications or industry technology transfer. University officials and faculty members frequently require the expertise of specific research groups or individuals, but struggle to obtain this information beyond their personal networks. This limits their ability to locate necessary resources effectively. Fortunately, most institutions have databases containing publications that could provide valuable insights into areas of strength within the university. In this article, we present a visual analysis application capable of addressing these questions and assisting management in making informed decisions regarding governance measures such as creating new research institutes. Our system has been evaluated by domain experts, who found it highly beneficial and expressed interest in utilising it regularly.

Keywords: visualisation, visualisation; information visualisation, visualisation; visual analytics

CCS Concepts: • Human-centred computing → Information visualisation; • Computing methodologies → Natural language processing

1. Introduction

Research-based universities face growing complexity in their strategic governance, in a context in which, conversely, there is a growing trend and need to foster interdisciplinary science for higher impact. To address such complexity, universities face a growing need to articulate multidisciplinary activities. However, their large size (e.g. the University of Vienna has 7.5K researchers, Paris Saclay has 9K professors and researchers, and Harvard has around 4.8K) makes it difficult to get a holistic understanding of their research areas. To overcome such complexity, we introduce Atlas, a tool that enables the exploratory analysis of the landscape of the publications by the researchers in our university. Its goal is to help university officials understand the strong areas and make informed decisions on strategic issues. Some examples include the creation of new multidisciplinary research structures, supporting emerging research areas, or promoting the internal building of teams for funding applications. Our approach has several advantages: unlike other approaches, by integrating both standard classification information given by publishers (Scopus classifications) and content-based created clusters and terms, we can illustrate the research areas of the university is strong at. We also provide tools to analyse the *evolution* of the research over time. Our exploratory analysis tool is designed to tackle concrete governance challenges:

- Understanding the areas of research the university is working on: useful for deciding which new research areas to promote, or the creation of new degrees, masters.
- Ability to search for groups or individuals with expertise in a certain problem or data: necessary to help solve technology transfer needs from industry, or to build teams for international projects.

- Finding interdisciplinary and multidisciplinary research at the intersection of several disciplines: useful for the creation of new research units.
- Identifying potential emerging areas of research for the development of cutting-edge projects and technologies.

The entire process is made possible by a multiple-view application that uses a robust automated data processing pipeline and combines the extracted information from the text analysis with external sources such as Scopus classifications to automatically determine names for the research clusters that properly identify the research within. Moreover, both our layout and exploratory paths differ from previous approaches (e.g. [CRF*21]). For example, we incorporate links to demonstrate relationships between documents from different disciplines that may be related (e.g. both deal with the same kind of datasets), created through text analysis. Furthermore, we create visualisation methods and interaction tools that highlight and provide contextual-guided details on the elements of interest (clusters, papers, researchers, to name a few), different from images or opening names (like in [EHA*23]). By integrating both data extracted from the documents and metadata from the journals, such as the Scopus categories, we enable both top-down and bottom-up exploration of the data. Our top-down exploration begins with the knowledge areas as defined by Scopus, and then allows users to drill down to the concrete research produced within represented as structures (clusters) that emerge from the data itself. This is useful for undertaking problems such as building multidisciplinary groups in a certain area. Other goals, such as the creation of research structures (i.e. research institutes), or finding potential collaborators, can be achieved adequately by finding groups that work with similar issues or data. This greatly benefits from a bottom-up exploration that can be achieved through the analysis of the contents of the papers themselves.

The rest of the paper is organised as follows: Section 2 describes the related work. The system is explained in Section 3 and the methods and algorithms needed to achieve the layout are described in Section 4. Section 5 describes our application and the main views, together with the implemented interaction techniques. Section 6 describes some use cases and in Section 7 we present the evaluations carried out and their results. We discuss our results and summarise the lessons learned in 8, and we analyse limitations and extensions in Section 9 which also concludes our work.

2. Related Work

The visualisation of large corpora of documents has many potential applications, such as searching for relational patterns between documents [HT04, CSL*10, FHKM17, LTW*18], understanding the contents of a corpus through topics [CB12, GOB*12, DN18] or even the exploration of document embedding techniques [JSR*19].

2.1. Text corpus exploration

Our research falls within the category of text corpus exploration (TCE) [GLB24]. The objective of TCE tools is to facilitate the exploration of substantial volumes of documents by users, enabling them to acquire an understanding of the corpus. This encompasses uncovering novel insights, comprehending the distribution of documents within the corpus, including their quantity, clusters, and potential relationships, and identifying novel documents. TCE is a form of exploratory search [SRA22, Mar06], where the users perform three activities: lookup, learn, and investigate. Among those, the final two constitute the exploratory search, as the former involves the exploration of known information.

While the focus of this paper is on scientific literature, related work has also addressed the visualisation of diverse non-scientific text corpora. These systems often target distinct analytical goals tailored to the specific characteristics and uses of the data source. For instance, the analysis of social media content is frequently explored for relevant data, frequent terms, or detecting sentiment polarity [LLZ*16, HAAE17, HKH*14]. Karduni et al. analyse the text from social media publications to extract features that may identify them as misinformation [KCW*19]. For broader sensemaking tasks, visualisation tools have been developed to support the exploration of massive, potentially unstructured document sets, such as Wikipedia articles [BLB*14, PCE*19], or aiding fields like investigative journalism [BISM14].

2.2. Visual analysis of scientific documents corpora

Scientific document corpora analysis often encompasses different and more general objectives that differ from these specific, likely immediate analytical needs. Moreover, the exploration of scientific documents collections, such as research articles, theses, or patents, is highly challenging due to the structured nature of the data (including metadata like authors, affiliations, publication venues, and citations) and the complex semantic relationships between documents. Dedicated visual analytics systems aim to provide deep exploratory capabilities, which can be realised through different representation strategies. 2D spatial layouts, obtained through 2D projections of embeddings, or node-link layouts of different types, are highly popular, but other visual representations that do not make those spatializations the central part of the exploration, have also been used. We start our analysis with the latter.

2.2.1. *Systems not primarily using 2D spatial layouts*

A significant body of work explores scientific corpora using various types of visualisation techniques. These systems generally prioritise understanding the analytical model itself rather than directly visualising the document space.

List-based interfaces provide one alternative, emphasising the exploration of entities and their attributes extracted from documents. The Jigsaw system, for example, relies heavily on coordinated list views to help users make sense of document collections by examining connections between extracted entities like people, places, and organizations [GLK*13]. More recently, the VITALITY system uses enhanced interactive list widgets as the primary means for exploring research articles, employing a 2D layout only as an auxiliary view to support serendipitous discovery [NKWW22]. El-Assady et al. [ESS*18] develop a system of interactive connected lists explicitly designed for comparing the outputs of different topic modelling algorithms applied to the same corpus. Termite [CMH12] uses topic-word matrices to help users assess the quality and interpret the meaning of textual topic models.

Temporal dynamics within scientific literature are often visualised using techniques adapted for time-series data. Streamgraphs, such as ThemeRiver [HHN00], and related techniques like stacked graphs are employed. TIARA [LZP*12], for instance, uses latent Dirichlet allocation (LDA) to model topics in a corpus over time, improves results by ranking topics, and then uses extended stack graph visualisations to allow exploration of topic evolution.

Other methods create abstractions of documents, like document cards. These are synthesised document thumbnails [SOR*09] that offer compact representations of articles intended to represent their key semantic features as a mixture of images and relevant key terms.

Rosenthal et al. [RMB19] also created a set of tools for the analysis of research production. Their tools are focused on communicating the growth, and focus specifically on university structures, such as departments, institutes, and research groups. Therefore, they do not create a map of the fields present in the institution. *In our case, the university has a clear list of research groups, departments, and institutes. And the governance officials also have access to indicators broken down by those units. What the officials lack is a map of the overall disciplines that are focus of research, and the researchers that play a role in those areas*.

Collectively, these non-spatialisation approaches demonstrate a focus on specific facets of scientific corpora: relational structures, entity details, model diagnostics, temporal patterns, or other specialised analytical goals. But our goals are more effectively satisfied with an approach where the exploration starts displaying the whole landscape of documents.

2.2.2. *2D spatialisation: Graphs and node-link diagrams*

Graphs and node-link diagrams are prevalent representations that have been used widely to emphasise relationships between documents. These relationships depend on the features considered relevant by the tasks supported by the system. For example, the StarSPIRE technique represents documents as dots that are transformed to textual windows, and lays them in a force-directed layout for semantic interactions [BNHL14], where links are created when document share entities, such as words.

In bibliometrics, node-link diagrams are used for visualising network structures derived from scientific literature in a 2D space [BWOW20]. Common applications include displaying the co-occurrence of keywords or languages, and mapping co-citation networks [CNS23]. While useful for showing connections, many standard applications of node-link diagrams in this area may lack sophisticated interactive exploration features to facilitate deeper data comprehension [RB21, WRLW21]. Some approaches enhance these diagrams with labels [WRLW21] or integrate them with widely used bibliometric analysis tools like VosViewer [Won18] or Citespace [Che06] for more detailed subfield analysis. Dunne et al. [DSG*12] integrated statistical information about citations directly into force-directed node-link diagrams, coupling them with ranked lists to support literature exploration by highlighting influential papers and connections.

Beyond the bibliometrics field, Dang and Nguyen [DN18] used a force-feedback algorithm guided by term co-occurrences to spatialise documents. Lee et al. [LKC*12] also employ a force-directed layout as part of a multi-view system focused on cluster communication and exploration. The forces are encoded by the similarity between documents, calculated as cosine similarity between the bag-of-words representation of each node. TopicNets [GOB*12] created a web-based system for the visual analysis of large sets of documents using topic modelling. In addition to scientific publications, their system also deals with grant proposals. The 2D layout is created using a force-feedback approach that uses both topic intersection and topic similarity to place the elements. The system also includes operations to analyse document sections, select multiple elements, and perform real-time topic modelling on selected subsets. FacetAtlas [CSL*10] aims to display both global and local patterns simultaneously on a force-directed layout to reveal relational patterns under different facets, that includes unstructured search. Global relations are displayed through the use of a density map; and local relations are conveyed through the composition of nodes and edge bundling techniques. Wang et al. also use a 2D graph to illustrate associations between scientific papers and patented inventions [WQQ*24]. Argo Scholar is a search tool that let's the user search for articles and references, and uses a 2D layout that can be edited by the users to display the different elements [LYM*22].

These graph-based approaches explicitly represent relationships (citations, co-occurrences, conceptual links, similarity) as edges, using layout algorithms to position nodes (documents, concepts) in 2D based on these connections, often aiming to reveal network structure and support exploration through interaction with the graph. A broader overview of text visualisation techniques developed up to 2019 can be found in the survey by Kucher and Kerren [KK15].

2.2.3. *2D spatialisation: Projections*

Creating a 2D spatial layout by projecting high-dimensional document representations in some latent space (e.g. [LLD04]) is another frequently employed strategy for providing a holistic overview of a scientific document corpus. These systems vary significantly in the data sources used, the computational processes for projection, and the specific analytical tasks they support.

Some systems focus on specific goals like promoting serendipitous discovery (Serendip [AKV*14], VITALITY [NKWW22] – though VITALITY uses the 2D layout auxiliarly) or enabling the exploration of the embeddings themselves [JSR*19]. Recent work by Gleicher et al. [GLB24] highlights the need for explanations within these exploratory tools, helping users understand why documents appear close in the projection.

Other approaches aim to visualise themes and clusters. For example, Wise et al.'s Galaxies [WTP*95]. Their system for document exploration has two different views, one completely 2D, named Galaxies, and another one with heightfields (where the elevation depicts the theme strength), named Themescape. They extracted terms and frequencies from around 20K documents in their INSPIRE system. For the layout, they propose to build clusters in high dimensional space, and project the centroids using a non-linear projection algorithm using Multi-Dimensional Scaling (MDS) [CA98]. Fried and Kobourov [FK14] focused on mapping major topics and relationships in Computer Science, using as input words and phrases

extracted from DBLP article titles, that then are converted into cities of the 2D map.

Topic modelling outputs are frequently used as the basis for projection. Choo et al. [CLRP13] created the UTOPIAN system, where documents from the Information Visualisation and VAST dataset were used. They produce a 2D layout based on a non-negative matrix factorisation [PT94]. The UTOPIAN system uses the 2D layout as a starting point for navigation through the topics. It provides certain interaction capabilities for topic manipulation, including merging, creation, and splitting. Chuang et al. [CRMH12] developed the Stanford Dissertation Browser to analyse PhD theses from Stanford University and their evolution over time based on a LDA [BNJ03] approach to model the relevant topics. Their input is the abstract of the thesis, from which they extract the words, and use either word similarity or topic similarity to relate the theses of the different departments and use PCA for the projection.

Heimerl et al.'s DocuCompass [HJH*16] enables a comprehensive exploration through lenses that provide up to ten highly ranked terms that characterise the region, extracted using term frequency-inverse document frequency (TF-IDF) or G^2. It can utilise any 2D spatialisation, and propose different ways to characterise the documents to be analysed. Extra meta data can also be included in the visualisations, such as citations over time.

More recent systems often leverage embeddings and advanced dimensionality reduction (DR) techniques like t-SNE [VdMH08] and uniform manifold approximation and projection (UMAP) [MHSG18]. Kim et al. [KKP*17] presented Topiclens, an interactive system for topic exploration that efficiently recomputes topics for subsets of documents using a semi-supervised t-SNE [VdMH08] approach. They build the clusters using a fast rank-2 nonnegative matrix factorisation [KP13], and enable detailed exploration using a lens-based exploration. Raval et al. create 2D scatterplots of sentence embeddings of scientific papers, to provide interactive explanations for those embeddings [RWVW23].

Lafia et al. [LKCH21] bears some similarity to our system, since they also analyse data from a research institution beyond common bibliometrics. In their case, they analyse the production from the Earth Research Institute, which is difficult to compare for articles from different areas of research. They use a combination of the title, abstract, and authors of papers, and analyse them using TF-IDF ([Jon04], a statistical concept that has proven to be useful for document summarisation [CAS16]). They subsequently construct 2D maps using either t-SNE or UMAP. *However, unlike ours, their approach does not build the embeddings from the whole text of the documents (which appear to work slightly better* [AB17]*), and we also include richer metadata such as the journals' categories. Furthermore, our system identifies research areas directly from the data, rather than classifying papers based only on predefined topics, making it adaptable to emerging fields*.

Li et al.'s Galex system [LZJZ20] supports hierarchical exploration of Computer Science disciplines, starting from predefined areas. They work an embedding by using titles, keywords, and abstracts, and key phrase extraction by ToPMine [EKSW*14]. Then, document vectors are obtained via doc2vec, and the projection is achieved through t-SNE. Their system has a set of predefined areas of interest, obtained from CSRankings. *In contrast, our approach is more general, since it adapts to the contents of the documents, by extracting the areas from the data itself. This is especially useful for emerging areas of research, that may have not a proper classification already defined. Moreover, their system focuses in Computer Science, while our dataset covers all the disciplines present in our university, which go from highly technical, to others such as Architecture or Medicine. Furthermore, interdisciplinary connections go beyond papers that are similar in title or abstract. We also use relevant terms and higher level Scopus classifications to guide exploration and make sense of the whole landscape*.

Caillou et al.'s Cartolabe [CRF*21] focuses on the exploration of massive datasets (like arXiv, although they include other datasets from political debates or Wikipedia). They compute doc2vec embeddings [LM14] from abstracts, and used UMAP to create the 2D layout, primarily supporting theme organisation via zoom/pan and search. They also add metadata from laboratories or researchers. For the exploration, they provide zoom and pan features, and allow users to input search for individual components. *Like them, we handle research documents and organise them using a UMAP projection based on doc2vec embeddings. However, the similarities end from this point forward. Cartolabe is designed for the exploration of large general datasets, while our system is explicitly designed for institutional governance tasks, offering deeper analytical features. Although they also organise the documents by theme, they do not provide an analysis of relevant terms or enable connections between articles from different areas. And they do not provide direct access to the publications themselves. We also incorporate other meta data, such as the year, the Scopus classification, or the growth, and we explore both through the creation of article and cluster detail views. We provide tools for cluster analysis, including but not limited to size, authors, and related clusters. In addition, our exploratory tool is time-aware. This enables us to gain insights such as the evolution (whether it is growing, steady, or decreasing) of a specific area (cluster) over time. Moreover, we also offer top-down exploration based on Scopus classification. Furthermore, unlike them, we consider the entire document, not just the abstract*.

2.3. Conclusion

The visualisation of text corpora encompasses a wide array of techniques and objectives, highly dependent to the nature of the documents and the specific analytical goals. For example, unlike Rosenthal et al. [RMB19], whose tools focus on communicating metrics within predefined organisational units (departments, institutes), *our system generates a content-based map of the research landscape itself. This addresses the need identified in our context for understanding the thematic structure and identifying key players across the institution, independent of existing administrative boundaries*.

For scientific literature, while methods like list-based views and temporal visualisations effectively address specific aspects, 2D spatialisations (both graph-based layouts and DR projections) remain a common and powerful approach for providing a holistic overview of the document landscape. Graph layouts excel at showing explicit relationships, while projections are increasingly based on semantic representations derived from topic models or document

embeddings, rather than surface-level features. Nevertheless, while the 2D projection provides a valuable overview, and serves as a crucial starting point, the analytical power frequently emerges from the interactions and integrated analyses built upon it. Tools like interactive querying, filtering, and analytical capabilities are required to allow users moving beyond the overview and serendipitous discovery, which, despite highly useful, can be insufficient in certain scenarios.

Atlas builds upon existing work using 2D spatialisations (specifically projections). The novelty of the system lies in its integrative approach, which couples full-text embedding-based spatialisation with the incorporation of heterogeneous metadata and derived analytical indicators (including topic clusters, interconnections, growth ...). This synergistic framework allows data-driven exploration and discovery of research domains across diverse academic disciplines, while also offering interactive, decision-support tools tailored to the strategic and operational needs of university leadership. It also distinguishes itself through key aspects: the data gathering and derivation process (i.e. data-based clusters, cluster naming, connections between articles) and the design elements (cluster views, growth map, time-based exploration, etc.). These elements are not common of other systems, and have been specifically tailored to satisfy our concrete needs.

3. Overview

Research-intensive universities have complex organisational structures, which make strategic governance a challenging enterprise. Our university, for example, has a hierarchical structure, consisting of various organisational units, such as schools, departments, research institutes, laboratories, and research groups. Each of these entities operates independently, with departments potentially delivering educational programs across multiple schools, and research groups originating from diverse departments. Consequently, there is limited familiarity among the extensive cohort of researchers (3.7K), and different research groups may unknowingly engage in closely related research themes. Given this complexity, coupled with the increasing demand to promote interdisciplinary scientific endeavours, it presents significant challenges in effectively coordinating multidisciplinary groups and cultivating emerging research fields or teams. These go beyond the capabilities of existing solutions. Other tools are insufficient to tackle certain problems that are specific for large research institutions. University administrators frequently encounter tasks such as:

- Characterising the university's research landscape: The need to articulate institutional research strengths (e.g. in the context of university alliances like Unite!) cannot be achieved with lists of research groups, that lack detailed exploration of research areas, including publication volume and contributing researchers.
- Evaluating critical mass in research domains: Assessing the density of expertise in specific fields is crucial for strategic decisions regarding the formation of research structures (e.g. institutes) intended to foster collaboration and facilitate large-scale project applications. Current methodologies often lack the granularity required for such evaluations.
- Identifying researchers with specific expertise: Facilitating connections between external entities (e.g. companies seeking technology transfer partners) and relevant university researchers requires sophisticated search functionalities capable of navigating the dynamic evolution of research areas. Similarly, media inquiries for expert commentary require efficient identification of faculty with specialised knowledge.
- Assessing the internal status and evolution of research fields: Determining the current and past trajectory of specific research topics within the university is essential for strategic planning, including decisions regarding resource allocation and the identification of potential collaborators for research proposals.

The detection of specific areas of investigation within the university itself is a daunting task. Atlas was conceived in discussions between some of the authors and university officials. Its goal was to address two complementary purposes: fostering enhanced interdisciplinarity and facilitating multidisciplinary collaboration. The ultimate objective was to leverage existing databases to develop an interactive knowledge map specifically designed for open and strategic management, with the goal of generating a more impactful outcome for society.

3.1. General goals

At the start of the project, together with the university officials, we set four high-level governance goals that the system had to meet:

R1: Getting a list of areas the institution is researching into.
R2: Facilitate finding people and groups who are strong in certain areas.
R3: The detection of emerging research areas.
R4: Finding areas where multidisciplinary teams could be built.

The first objective is rooted in the necessity to gain enhanced insights into the strengths and limitations of research endeavours. The second objective arises from our university's aspiration to participate in international funding opportunities, which often require interdisciplinary collaborations. Hence, it was crucial to identify individuals or research groups with expertise in specific areas. The third component focuses on fostering and advancing research in emerging fields, which are frequently difficult to identify at an early stage. Consequently, one of the aims was to explore the feasibility of identifying novel research areas. Furthermore, the university expressed interest in identifying areas of interest that are investigated from different perspectives by distinct research teams. This knowledge could facilitate the establishment of research institutes capable of approaching multifaceted research topics (e.g. Biomedicine) from different angles.

It may be infeasible to tackle an entire research field for many such goals. Hence, we have chosen to pursue more specific divisions referred to as "clusters" to represent distinct subfields. As will be described later, unlike other approaches, like Li et al.'s Galex [LZJZ20], which have predetermined areas of research, we are completely agnostic to the data distribution. Therefore, in our case, clusters are extracted from the data themselves.

3.2. Specific requirements

The project development spanned several years and required the collaboration of several divisions of the university. Initially, university

officials and library managers were involved. The former provided information on the general goals of classification systems and enabled access to database data and metadata in the needed format. As we proceeded with the development of initial tool versions for beta testing, we incorporated additional potential stakeholders into our discussions and initial usability evaluations. We included former university officials and other faculty with diverse current or past responsibilities, such as directors of departments, heads of research groups, or school directors (in our university, a school is responsible for multiple degrees). As a result, certain stakeholders expressed interest in incorporating additional functionalities that would be beneficial for their work. Some of these requested features include:

SR1: Understanding the behaviour of a certain cluster over time.
SR2: Finding relevant authors in an area of research or cluster.
SR3: Finding concrete keywords that may be related to a cluster.

As a consequence, we added some additional information to our views and introduced new tools that were not originally planned.

4. Data processing and derivation

The data used in the application consists of all papers published in scientific journals and conferences authored by at least one member of our university in the 2010–2020 time span (to avoid embargo periods). The total number of papers is around 20K. After discarding the papers written in a language other than English (see below) or that were too short, the total number of papers included is 15 473.

4.1. Data analysis and layout construction

Articles are automatically downloaded from the university publication database via internal API. Besides the PDF files, additional metadata were obtained (authors, journal, publication date, Scopus categories, etc.). Documents were then processed as follows:

- *Conversion to plain text*. PDF files are transformed to UTF-8 encoded plain text using Apache Tika (https://tika.apache.org), a powerful toolkit capable of extracting metadata and text from a variety of file formats.
- *Language identification and filtering*. The central 50% of each document is extracted, stripping out the first 25% (likely to contain a title or abstract, sometimes in several languages) and the last 25% (likely to be a reference section). This central part is fed to a language detector, and only documents identified as English are kept. This step also filters out ill-formed documents –due to original PDF format (e.g. scanned PDF) or problems in PDF-to-text conversion– as well as documents consisting mainly of formulas, images, tables, or other non-text content. Because those are not recognised as English text.
- *NLP processing*. Selected documents are processed (all text is used here, not just the central part extracted for language detection) with FreeLing NLP suite [PS12], a powerful open-source library, which offers a large variety of language analysis functionalities. In this case, tokenisation, sentence splitting, part-of-speech (PoS) tagging, and lemmatisation were applied, obtaining a normalised version of the text, where verb forms are mapped to their infinitive form, nouns are mapped to their singular form, and the use of a word is distinguished depending on its PoS tag. For instance, the sentence *Results produced by the experiment resulted in larger uncertainty* would be converted to `result_N produce_V by_IN the_DT experiment_N result_V in_IN large_J uncertainty_N`. The goal of this step is to obtain a version of the document in a 'normalised' vocabulary, so that similarities can be found in later steps, regardless of the used verb tense or other morphological variations.
- *Document embeddings*. The next step consists of creating a latent vector space where documents and words in them are represented as points, and distances between them can be computed. For this, we used gensim (https://radimrehurek.com/gensim/) implementation of doc2vec [LM14]. Word and document embeddings are created by training a neural network (NN) to relate a word to its context (or a document ID to the words it contains). The hidden layer of this NN encodes these relations in a compressed way, and that vector is used to represent a word or a document in the latent space. Embeddings have been proven to have interesting semantic properties: vectors close in the latent space correspond to words or documents semantically similar, and vector subtraction/addition operations can be used to perform analogies (e.g. the answer to *man is to king as woman is to X* can be obtained subtracting the vector for `man` from the vector for *king* and adding the result to the *woman* vector). We fed doc2vec with the 15 473 documents and obtained a model assigning each of them a vector in a latent space of 400 dimensions (the final value for the embedding dimension was selected empirically, together with other parameters for the clustering algorithm, as described below).
- *Clustering*. Once the document vector space is created, it is projected from 400 to 10 dimensions using UMAP [MHSG18] and then we run the HDBSCAN [CMS13] clustering algorithm, to group similar documents in several thematic groups. UMAP is one of the most effective DR algorithms [EMK*21]. Using HDBSCAN in the latent space was too slow for the number of papers (and does not scale for larger sets) and reducing to two dimensions prior to the classification loses too much information for the clustering. The parameters of the algorithm were explored in a grid search, and the parameter combination producing the best-balanced results was chosen (see 4.2).
- *Graph construction*. Once the documents are grouped into clusters, a graph is built as a final step before visualisation: TF-IDF is computed in the usual Information Retrieval way, obtaining a rank of relevant words for each document (i.e. words that are frequent in a document but not frequent over all –or many– of them). A graph is built where each document is a node, and an edge is added between two documents if they share a minimum number m of their top-k TF-IDF terms. Note that an edge is added regardless of whether both documents belong to the same cluster. Parameters m and k are also selected empirically to produce a reasonable number of edges. We calculated up to 20 most frequent terms per documents. For each pair of documents, we rank their similarity, based on the number of terms they share. Then, we show at most, 11 connections for each document. Each connection requires two documents to share at least three common terms. Words with high TF-IDF for each cluster (compared to other clusters) are also computed, so that relevant terms for a cluster can be visualised later.
- *2D Projection*. Finally, the document coordinates are projected to 2D using UMAP again. The final visualisation will depict the

documents in different colours (using a palette of 64 colours that maximise visual discriminability, calculated using CIE Lab, a perception-designed colour space), depending on the cluster they belong to, and edges among documents are also visualised.

To provide further insights into the processed data, several extra fields are derived:

- Document details: We generate a list of document terms, from a TF-IDF, a list of Scopus tags (top-level Scopus categories of the paper), and the link to the public university database.
- Cluster names and frequent terms: Names are obtained by using the most common level 2 Scopus classification and the three more relevant terms as determined by the TF-IDF analysis. The other frequent terms are also stored.
- Cluster stats and data: we calculate the total number of elements, publications per year, one-year growth, and the university average publications per year per cluster. In addition, we also extract a histogram of the Scopus categories of the cluster.
- Related clusters: We generate a list of clusters with at least 4 connecting nodes with the current cluster.
- Publications list and relevant authors: Computed per cluster or region of interest, together with the contribution percentage of the authors to the papers in the cluster. These are available in the cluster view or region of interest view.

Finally, the calculated data (node positions, links, cluster information, and all derived information) is saved as a json file that is the input of the visualisation tool.

4.2. Model validation

Despite its age, doc2vec remains superior to other newly developed systems such as BERT and BERT-derived approaches. For example, Kohlmeyer et al. found that it works better than BERT, ROBERTA, or XLM for book classification [KRK21], while it also outperformed GloVe and FastText in scientific paper classification [GV22]. It also works better than other technologies developed for large documents, as it been shown with legal documents or novels [PMP*21, KRK21], and with scientific documents, in a recent publication [RV24]. Additionally, doc2vec is orders of magnitude faster than other, more complex techniques, and it is also faster to train. Large Language Models may compete with doc2vec, however, they are much slower at creating embeddings, and their training is very environmentally impactful. Notwithstanding, we envision that applying Low-Rank Adaptation [HysW*22] to a pretrained large language model may be a good balance between the computational effort and the benefits.

To determine the optimal number of dimensions to use with doc2vec, we conducted an analysis using a test case consisting of 1131 documents from various departments manually classified into 16 classes. We applied K-Means and HDBSCAN algorithms to automatically classify the documents, utilising doc2vec embeddings with dimensions 50, 100, 200, 300, 400, and 500. Confusion matrices were examined, revealing that the 400-dimensional embeddings performed nearly as well (goodness of fit of 85.36%) as the 500-dimensional embeddings (1% difference) in the classification tasks, while the 500-dimensional embeddings requiring roughly double the time. Smaller dimensions resulted in significant classification errors, indicating less consistent projections. This observation was further supported when considering the clustering and DR algorithms, as explained below.

UMAP is a DR technique renowned for its cluster identification capabilities [XZS*21]. We validated the UMAP parameters using a grid search with: *n_neighbours* between 30 and 55 (with increments of 5), *min_dist* in the range $[0-0.3]$ (with increments of 0.05), *min_cluster* size in $[30-100]$ (with increments of 5), and *min_samples* was evaluated between 1 and 91 with increments of 10. We measured the *validity index*, *relative validity*, number of clusters, and unclassified documents. We then created an empirical function that weights the number of clusters (prioritising around 50), a high *validity index* and reduction of unclassified documents number. The optimal parameters found were $n\_neighbours = 30$, $min\_dist = 0$, and minimum cluster size of 10. The metric use was cosine similarity, the standard for document embeddings. Solely relying on the validity index as a measure generated an excessive number of clusters and unclassified nodes. The resulting UMAP projection was independently reviewed and approved by five individuals from the production and supervision teams who were not involved in authoring this paper. Additionally, the projection was reevaluated after incorporating an additional year of papers in the later stages of development. While some tools exist for visually exploring latent spaces [LJLH19, FKM20], certain types of data, such as images, may be more easily assessed compared to other types, such as documents.

Since our visual classification involves the clusters' colours (computed in the 10-dimensional space), clusters are formed by items that are not always the closest neighbours in 2D. However, we found that, for our data, the number of documents from one cluster that, after projection, have a 2D position within the boundaries of a different cluster, is very small. To ensure all the elements can be properly identified, our cluster selection de-emphasizes the non-selected clusters. Alternative highlighting techniques could be used, such as adding an enclosing semi-transparent layer, similar to the growth map, or increasing the size of the items of the selected cluster. Even a combination of those could also work, depending on the number of nodes. We could, however, evaluate the 2D layout using methods specifically designed for this, such as the approach by Stahnke et al. [SDMT16]. This could help, in the future, for comparing different projection techniques, should we include more data in the dataset, or if we change the algorithm for embedding construction.

5. Visualisation

5.1. Overview of the tool

To achieve the aforementioned goals, we have designed a multiview application, as depicted in Figure 1. Certain views are always present, such as the navigation view (A) or the list of clusters (B), whereas others are generated on demand.

The central view (A) provides a 2D spatialisation of the publications, represented as nodes. As outlined in Section 4, the 2D positions of the nodes are determined by the document contents employing a DR projection with UMAP. Hence, closer papers are related by their contents. Additionally, we also connect papers with

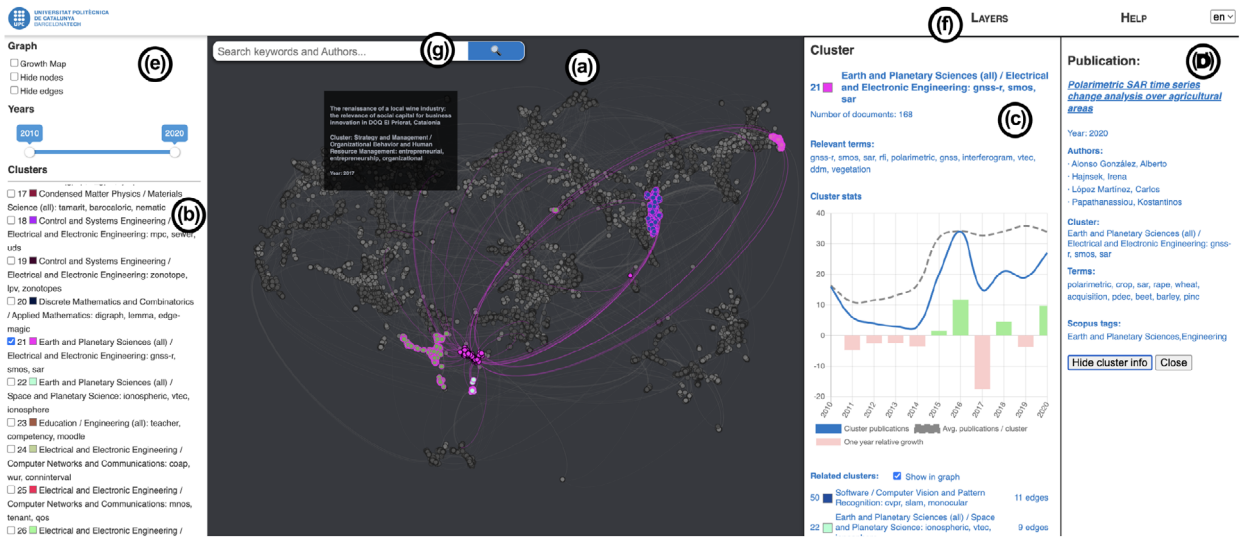

Figure 1: *Exploration of the universities' research space for the period 2010–2020. The central view (a) shows one node per publication. The system automatically finds similar papers and close areas of research by extracting the text of documents, and building embeddings using doc2vec and UMAP projections. Clusters are automatically built and labelled by analysing frequent terms. Their details can be explored interactively (c) by showing the cluster information in the paper view (d), or by direct selection over the cluster list (b). Other explorations are also allowed through the control panel (e), or by opening the layers menu (f), or through direct search of authors or terms (g).*

common, highly-frequent terms. The rationale behind that is that papers addressing very different problems (and thus, including domain-specific wording) may be related (for example, they might deal with similar data). And this likely points to different areas that may be connected, and where interdisciplinary groups might be built. One of such examples is shown in Section 6. The interactive exploration of clusters and connected articles solves requirement **R2**.

The user can navigate the dataset using pan and zoom in the central view, or they can filter the data using the top-left control panel or the left clusters list. They can also obtain details on demand. The top-left view supports navigating over the years, displaying the growth map, or filtering out the nodes (to facilitate the growth map exploration). The left view facilitates cluster selection, resulting in a view that offers additional tools for the interactive exploration of cluster details (as in Figure 1-right). The detail views, including cluster and paper details, always appear to the right of the screen. Scopus layers can be revealed using the *Layers* menu. Furthermore, an additional Help menu provides guidance on the use of the application.

5.2. Understanding the fields of research

One of the primary objectives of the tool, as defined by requirement **R1**, is to provide an understanding of the areas of research that our university is publishing. This is accomplished through a combination of views, including the 2D layout, cluster list, and various detail views. The 2D spatialisation provides an overview by displaying all papers in a colour-coded manner based on the area that has been calculated by them. The nodes can be filtered by year by accessing the top-left panel. However, the layout is calculated using a UMAP projection of the entire set of papers. That is, we do not project independently per year (or per year range). Doing so would yield greatly changing layouts due to the inclusion of different papers. Moreover, the randomness of the UMAP algorithm would even result in slightly different projections for the same year if executed multiple times. This presents a challenge as users anticipate consistent paper and cluster positions when filtering based on the year range. Maintaining the relative distribution of papers while filtering out specific years ensures the preservation of spatial continuity among elements. Consequently, this facilitates the exploration of the underlying structure. Initially, a force-feedback algorithm was used for the layout, but the resulting structure was loosely related to the papers contents. Our outcome is a 2D layout that effectively portrays the research field landscape, offering valuable insights into areas of concentrated activity.

The **Cluster List** (View B in Figure 1) lists the areas of research. This list serves a dual purpose: First, it gives an idea of the research areas represented in the corpus, their high-level distribution, along with insights on relevant words within the area, as explained later. Second, it functions as an interactive filter. This overview plays a crucial role in enabling exploration, zoom, several filters, and extra superposed layers can be applied. Further analysis is achieved through a set of accompanying detail views, that are interactive. These enable a deeper exploration of the selected cluster or document upon demand, as explained below. It is built by taking the identified clusters, and giving them a meaningful name. This is not straightforward: using the extracted terms is only useful for people who are familiar with each area, and Scopus classification is too general (e.g. "Computer Science"). Nonetheless, it should be noted that classification systems are not perfect [WW16], and that often

journals are not properly classified. We also considered other systems, like ACM's. But the ACM classification is too narrow for our university. The other alternative source of metadata for publications is the Web of Science. Despite both of them have their strengths, Scopus coverage is slightly larger [Pra21], and is the one commonly used by our librarians. After evaluating both top-down and bottom-up naming alternatives, we defined the following strategy: we analyse the level two Scopus categories of all the publications in the cluster and select the two more frequent categories. Then, we add the three most frequent terms in the cluster that are not frequent in the rest of clusters. This gives names that are understandable at a high level, and insights for people familiar with the area of research, thus solving **R1**.

Individual nodes correspond to papers, and the assigned colours give the user insights on the contents of the cluster using the list of clusters/areas of research on the left. We explicitly encode the list of clusters, instead of simply displaying them in the main window, to enable users to scan all of them, thereby aiding in identifying the major areas of research of the institution. Since these disciplines are one of the entry points for exploration, we want the list to always be present.

To locate related articles, the users have three primary options: exploring the papers in a cluster, exploring the neighbourhood of an interesting paper, or following the connections. It may not be sufficient to explore data solely in a reduced space to find relationships between documents [NA18, EHA*23]. The distortions generated for the projection may adversely impact the users' ability to locate related items. To overcome these limitations, we employ two strategies. Initially, we detect the clusters in a 10-dimensional space, thereby ensuring that the information loss is not as severe as with the final reduced space. Secondly, we also analyse frequent words in the original texts and analyse them using TF-IDF. Then, documents containing common frequent words are connected, and those connections are depicted in the graph.

5.3. Getting insights from the data

Several goals require obtaining more fine-grained details of the data. For example, requirement **R2** requires that researchers have a strong presence in a certain area. We solved those through the inclusion of a set of detail views. These will integrate the metadata we have on the different items, as explained below.

The first detail view added was the **Document view**. It provides details on the selected publication: title, authors, year, cluster, common terms, and Scopus tags (see Figure 2-right). Since the floating label appears only with the title, cluster name, and year, this view is used to provide additional details. Furthermore, we use the displayed information to further accelerate other searches. Authors and terms can be clicked, and they are copied to the Search button. This is needed for Use cases 3 and 4. Besides, the article title is also a hyperlink to the university database entry, that contains extra publication details and a preprint version if available.

To enable the users to understand magnitude of a discipline, and researchers who work in that area, we have implemented the **Cluster details view**. It is a rich-content view. It provides information on the name of the cluster, its relevant terms, the number of documents, etc.

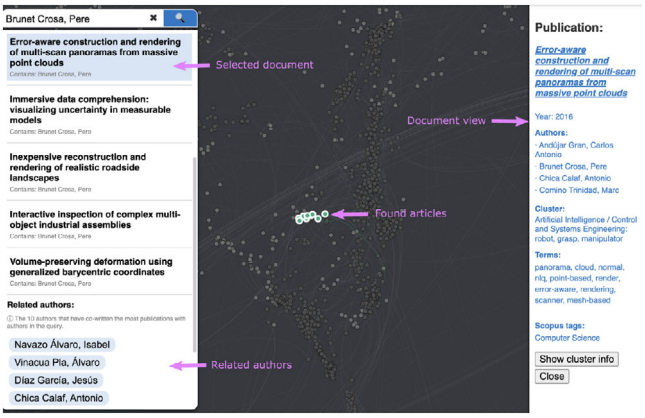

Figure 2: *The document view (right) shows the details of the document and provides some interaction elements. Authors or terms can be added to the search box (left) by clicking onto them. Upon search, the found articles are listed (left) and highlighted in the main view (centre). The list also provides a list of authors that can also be added to the search by clicking.*

These are shown in the top part of Figure 3-left (a–b). In addition, it also plots a set of stats (relative growing rate, number of publications per year, and average publications of all the clusters) intended to provide a better understanding of the evolution of the cluster along the years (as shown in Figure 3-c). These charts are interactive, the user can hover to get the details. Scrolling down, other contextual information is also shown, such as the names of the related clusters together with the number of the connections (D). This helps users gain insights into the relevance of the relationship between clusters. Moreover, the connected clusters can also be highlighted on the map by activating the *Show in graph* toggle. This colourizes those with their cluster colour, as seen in Figure 4. Finally, we also provide two dropdown lists that appear upon request with the authors of the cluster (sorted alphabetically or by the % of publications in the cluster), and the (selectable) list of publications in the cluster (3-right f and g). By adding the list of authors, which can be sorted both alphabetically and by their relative contribution to the cluster, we can solve the requirement **SR2**, to find researchers that are strong in a certain area. The name of the cluster is generated based on a high-level classification by Scopus and the three most frequent terms of the cluster. This name, combined with a comprehensive list of the most frequently employed terms of the cluster, helps the user understand what the cluster represents as a research discipline. This is the specific requirement **SR3**, and cannot be resolved solely through Scopus classification, as it provides broad terms. Both the list of authors and the list of articles, are designed as dropdown options to save space, because some of the use cases will not need them.

The previous view does not provide a direct search of individual researchers or articles that are associated with a specific topic. This is required by the goal **R2**. To solve this, the exploration is complemented with the **Search tool**. Although the inclusion of a search function was not initially considered during the initial iterations of the system, it emerged as a desired feature during beta testing. We have implemented it as a two-level process: Initially, users can add terms (see Figure 5-left) or authors (Figure 5-right).

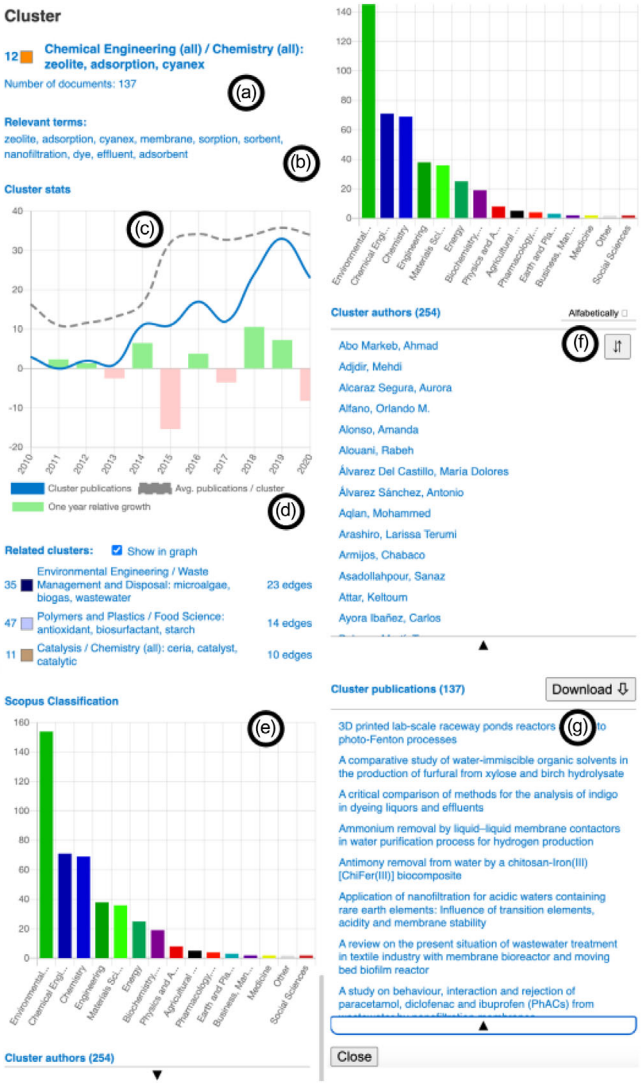

Figure 3: *Available information in the cluster view when user scrolls down, besides the cluster name: (a) number of documents, (b) relevant terms, (c) cluster stats, (d) related clusters, (e) Scopus categories in the cluster, (f) authors' list and (g) cluster publications. The last two, are dropdowns that only unroll upon user click.*

If the words typed start with uppercase, the system automatically searches for authors. By clicking the search button, a search is triggered inside the database (both for exact and partial matches). A results box will then appear with both search suggestions and found results (see Figure 6). The papers in the list are highlighted and can be selected directly from the list. Search suggestions can be incorporated into the search (also manually, using a comma as a separator). The search is executed as an *or* of all the terms. When there is an author involved in the search, the list of papers they have published appears in the results box. Papers, as well as suggested terms, appear coloured in blue to communicate the idea that they can be selected through clicking. This allows for a specific bottom-up search from the publications themselves, instead of a top-down analysis when

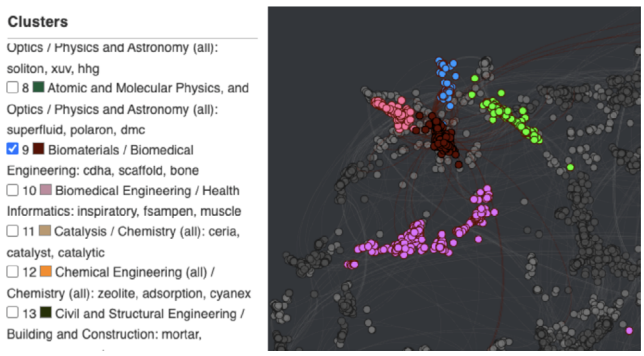

Figure 4: *When the user selects a cluster, besides highlighting it, we also highlight the related ones (with a higher number of links). The relevance is determined by the number of connections, which can be seen in the cluster details view.*

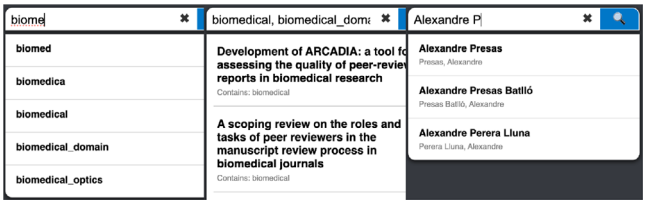

Figure 5: *When typing a word, the system suggests terms found in the publications (left). If more than one term is searched for, the results mark the words that were found in the articles (centre). If the word entered in the search box is capitalised, the auto-completion suggests names of authors if the word is capitalised (right).*

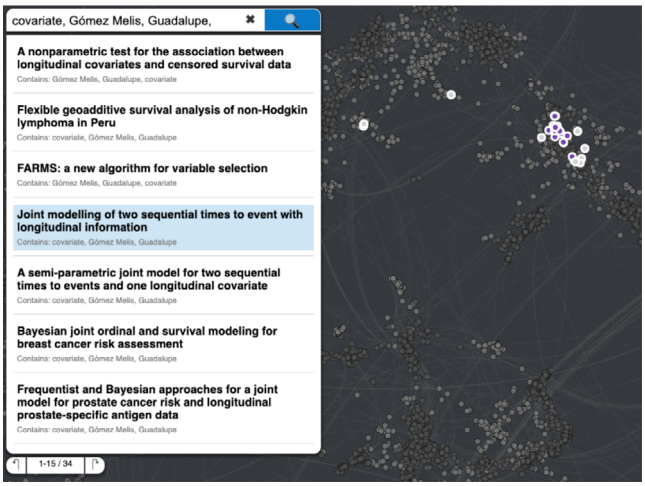

Figure 6: *Search result: the publications where the terms appear, or belonging to the author used as input, are highlighted in the main view. If the papers list is above 15, scrolling buttons appear (bottom left). The highlighted article is the one that has been selected, and is shown in the document view (clipped here to enlarge the rest of the contents).*

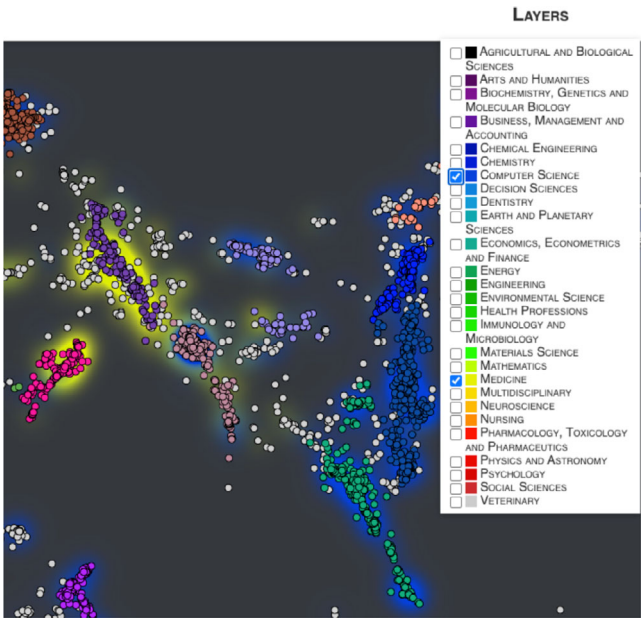

Figure 7: *Scopus classification selections: By toggling the Scopus categories on and off, we can see papers that may be related by higher-level areas. To highlight those papers, we generate a layer onto the main view with a halo coloured from the Scopus category.*

starting through the clusters list. By searching for terms, and then navigating through the papers related to these terms, users can find researchers with expertise in certain areas, for example to find teachers for concrete courses, required by **R2**.

Another aspect that is relevant to university governance is the detection of multi-disciplinary or interdisciplinary areas of research, as depicted in the requirement **R4**. To achieve this goal, we created the **Scopus layers**: The menu (F in Figure 1) shows the *Layers* button. Upon click, a floating window with Scopus categories is shown. By toggling any of the categories, a semi-transparent layer, that shows which papers are classified under the category, appears in the main view (see Figure 7). This facilitates top-down exploration of the nodes. By selecting multiple layers, we can find multi-disciplinary or interdisciplinary areas.

5.4. Extra interactions

Throughout the previous section, we already introduced some of the interactions that are enabled by the different views. In this section, we present some specific set of cross interactions that deserve further details. These are aimed at addressing the high-level problems specified in our requirements. We have included a video that demonstrates the functionality of all the features.

Map exploration. The central view is designed as the entry point for the exploration session. Besides the usual navigation tools (zoom and pan) and clicking to select a document, we have other interactions. Right click toggles off selection of a document, cluster, or region. The growth map, nodes, and links can be toggled on and off with the top left panel.

Figure 8: *To further inspect the overall production of the papers and get a sense of how the clusters are evolving, we have the* Growth Map *that shows a heatmap that indicates the clusters with higher growth in the most recent years.*

On top of that, more complex insights can be obtained through a couple of techniques: Region Of Interest, and Time filtering. The *Region Of Interest* (ROI) can be selected through right-click and dragging actions. When an ROI is chosen, all documents within that region are selected, while the remaining documents are visually de-emphasised. Furthermore, an analysis of the documents within the ROI is conducted, and a details view similar to the cluster view is dynamically generated upon request. This view presents various information about the selected papers, including their Scopus classifications displayed as a histogram, as well as lists of authors and publications. Additionally, a toggle on the top of the view enables the selection of unclassified nodes (i.e., nodes not assigned to clusters). *Time filtering* is also possible through the use of the years slider. The user can manipulate the maximum and minimum ranges to increase or decrease the number of displayed years. Consequently, documents falling outside the specified range are hidden. Moreover, both sliders can be adjusted simultaneously by selecting the highlighted range between them, thereby coordinating changes to the minimum and maximum limits. This functionality aids in the identification of emerging areas, as required by **R3**.

Growth map. Adds a layer with a heatmap that highlights the regions that have grown recently (the last years of the range), as shown in Figure 8. This layer is semitransparent (opacity 0.75) and appears in the background, to avoid occluding the other elements in the view, such as nodes or dynamic labels. It is implemented as a hierarchical set of layers with distinct levels of detail that are modified in accordance with the zoom level. The heatmaps are implemented using the deckGL's HeatmapLayer class and use ColorBrewer's YlOrRd 6-class palette for the visual encoding. The papers belonging to the first year of the range are assigned the yellow colour, and the papers from the last year are assigned the red colour. The values are calculated internally using Gaussian kernel density estimation with weights that aggregate the points within the region instead of calculating the mean, which would average out the signal. This configuration ensures that the layer is visible at different zoom levels. Regions with recent growth but no cluster assigned may indicate an emerging area of research, solving **R3**. These maps resemble the GRAM system [BEH*18] but they encode citations instead of production increases over time. Furthermore, we only draw maps around the nodes to only encode the regions with actual nodes instead of the whole plane.

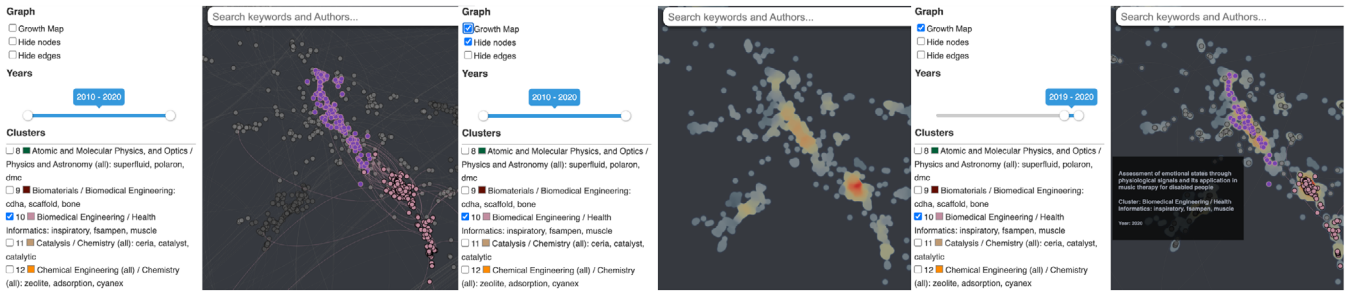

Figure 9: *The user searches for expertise in biomedical engineering using the cluster view (left), then they test whether this area is actually growing (centre), using the growth map, and validates the result by hovering over the papers to check whether these are actually in the correct area (right).*

6. Use cases

In this section, we present some useful scenarios that can be addressed using our tool. We start with some basic analyses, and then proceed to more complex ones.

Use case 1: Demonstrating prowess in a certain discipline. University officials often evaluate research progress to report to funding entities. In Catalonia, the regional government is a funder, and it is also in charge of the different degrees that are taught in public universities. Therefore, the degrees that are imparted by the universities, as well as the number of students that can enroll to a degree each course, depend on a negotiation between both. Our university has recently (between 1 and 2 years ago) applied for the creation of a degree in medicine. The university must demonstrate that, despite not currently teaching this degree, it has experience in closely related areas within the university. There exist various methods to demonstrate this. One straightforward approach would be to quantify the number of papers published per year in related fields. However, a more comprehensive analysis, such as one that focuses on areas experiencing significant growth, could provide more valuable insights. The users can simply look for the clusters that are related to this, such as cluster 10, which is named *Biomedical Engineering/Health Informatics*, as depicted in Figure 9-left. To demonstrate that the field of research is actually active, the growth map can be toggled. Nodes can be hidden, and the user will immediately notice that the cluster has a red region in the centre that indicates recent growth (centre). To obtain additional information regarding the concrete research conducted lately, the user may reactivate the nodes and use the year filter to select 2019 and 2020. This will only include papers published within the past two years of the sample. By hovering, one can see that certain papers are addressing medical issues like apnea analysis, respiratory muscle simulation, and so forth (right). Another cluster that is related to medicine is Cluster 9, with the name *Biomaterials/Biomedical Engineering*. When the cluster is selected, the user will find that it is strongly connected to four other clusters related to materials, such as polymers and plastics, materials chemistry, etc. As they will be highlighted by default, they can be hidden from the 2D map by selecting the toggle button in the cluster view. The user then can repeat the same operation: hide nodes, toggle the growth map. The result is that this cluster is healthy again and that the research in the area has increased in the previous years. This can also be confirmed by using the *Cluster stats* section of the Cluster View.

Use case 2. Participating in the EU mission: Healthy oceans, seas, coastal and inland waters. When the university is interested in participating in a concrete call, or when it is contacted by an external institution or company to explore collaboration potential, officials may need to explore the formation of multi-disciplinary groups that fulfill certain parameters. This is related to requirement **R4**. In this example case, the university is willing to participate in a certain mission. After reading the call, the university seeks a team of researchers interested in different aspects of ocean and sea health preservation. This can be achieved by highlighting all groups within the *Earth and Planetary Sciences* Scopus category and inspecting the clusters inside. After checking a couple of clusters, where we find the areas of astronomy and geology, we identify a cluster whose label is *Water Science and technology/Ocean Engineering*, circled in yellow in Figure 10. By hovering over the papers, we see that some of them deal with the analysis of waves' effects in the coastal zones. Then, we want to incorporate another group, more related to climate analysis. To do so, we select this cluster and enable the *Environmental Science* Scopus category. In this case, it turns out that one of the closely related clusters (as indicated in the cluster view) belongs to this area. It is the green one slightly below the initially selected (marked in orange in Figure 10). If we hover over the papers, we see some that deal with climate prediction and analysis, and its effects on elements such as rainfall. By analysing the most relevant authors in both clusters through the cluster view, one can get their names and contact them.

Use case 3. Creating a research institute. Around two years ago, the head of the Centre de Recerca en Enginyeria Biomèdica (Biomedical Engineering Research Centre) was looking for critical mass for the creation of a research institute in biomedicine (the goal was to make the research centre to qualify for a higher classification in the reseach system, which would allow it to apply for certain grants that are now not possible). Therefore, he wanted to find other groups that had research in medicine and similar areas to increase the size of the centre, and then apply for the higher-level denomination. With the current available tools, they had to scan all research groups' webpages in our university, then articles, and ultimately find people who were doing research related to medical data, medical devices, or other related areas. It took several people many hours of work. With Atlas, a big portion of the groups can be identified in few minutes. We can simply search a word such as *biomedical*. While typing, the search box shows similar names

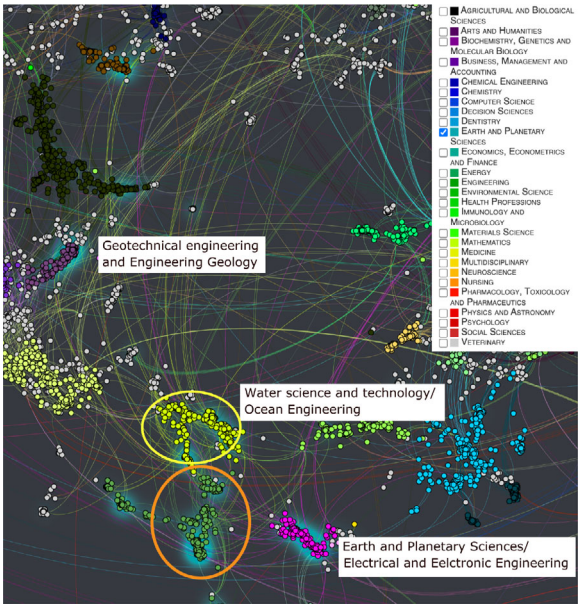

Figure 10: *Creating a group for applying to a call under the* Healthy Oceans, seas, coastal and inland waters *EU Mission. First, the* Earth and Planetary Sciences *Scopus layer is toggled. This leads us to a cluster (marked in yellow) that investigates ocean and sea waters. In this case, a closely connected cluster (as detected by the links and identified in the cluster view, here marked in orange), contains articles on climate analysis.*

and *biomedical_domain* surfaces as a term (see Figure 5-left). Most probably coming from a list of capitalised keywords (since they are interpreted as a unity, such as an author name). By selecting both, we quickly find papers related to those disciplines (see Figure 5-left). We can then navigate to the papers. For example, a paper regarding medical engineering applied to biomedicine. It has just another linked paper, but if we open the cluster it belongs to, and we check the papers of the cluster, we start finding publications related to engineering research in medical devices that were not found directly with a single word. Moreover, by checking the link of the papers, we can get to the research groups at our institution. By performing the same process iteratively (e.g. for all the papers appearing in the list), we can end up with a thorough list of authors and research groups that might be interested in the creation of such an institute. As an example, within 5 min we found the following research groups related to biomedicine: (i) industrial robotics (robots applied to biomedicine), (ii) a group from the Chemistry department (biomaterials for regenerative therapies), (iii) natural language processing (named entity recognition in the biomedical domain), (iv) an optical research (deal with flow measurement), (v) nanoengineering (polymers for detecting biomolecules), (vi) biostatistics (quality of reports in biomedical research), (vii) biomechanical engineering, (viii) biosignal analysis (rehabilitation and therapy), (ix) robotics, and intelligent systems (wearable robots for rehabilitation). Note that our university has multiple campuses in different cities, totaling 3.7K researchers, making this exploration difficult by other means. Thus, oftentimes, researchers are unaware of other groups doing related or complementary research.

Use case 4. Searching for expertise in some area or data type. Recently, a physiotherapist approached a member of the Computer Science Department with expertise in virtual reality because she wanted to do a PhD thesis in the use of virtual reality for rehabilitation. As this presents an excellent opportunity for individuals within the Computer Science Department to collaborate with colleagues across the university, we can seek additional expertise in this field. The procedure is straightforward: The user enters 'rehabilitation' into the search box and searches for it. The outcome will consist of a list of articles and a list of associated terms. Since the goal is to work in physical rehabilitation, instead of using VR for neurorehabilitation, the user selects the terms 'exoskeleton,', 'gait,' and 'movement' and updates the search. The user can then search for the authors and the contents of the papers from the resulting list by clicking the link to the document in the document view. This will enable the user to locate the research group and other relevant data from the authors.

7. Evaluation

We first carried out a semiformal usability test with a relatively large group (17) of university officials and leading researchers with diverse backgrounds representative of the research fields in the university. The study used a modified version of the SUS usability test. The participants were highly enthusiastic and proposed some extensions based on their needs, and subsequent meetings with a subset of these users served to further refine the new features. The current version was analysed through a formal user study with five domain experts with different responsibilities: (i) A vice rector of research of another university, (ii) the head of an external department, (iii) the vice dean of postgraduate studies, (iv) the head of a research institute within the university, and (v) a lecturer who had obtained a position less than three months before the study. Except for the first person, the rest belong to the same university. None of them has previous background in Visualisation. The study had four parts: Introduction and tutorial, training, tasks, and questionnaire.

Tutorial: After the goals of the tool were introduced, it was demonstrated. All the features were visited at least once. Participants were instructed to ask questions, so this process lasted from 10 to 30 min, depending on the discussion. **Training**: Participants were instructed to use the tool until they felt comfortable achieving the described tasks. This self-training was partially guided: we ensured that all the relevant features were visited by indicating the participants to do so with comments such as 'did you test the year filter?', 'can you check that cluster and see if it contains the types of papers you would expect?'. To ensure that the participants became familiar with the features, we had a written script with the features to explore. Some users were interested enough to explore all features of the application, so we only hinted at some of those features if they were necessary. This stage took the participants between 20 and 30 min, although one participant spent around 40 min testing the tool. **Tasks**: After the training, users were asked to perform five tasks (which took them around 5–7 min):

- Select two different clusters and check whether the papers inside the cluster made sense.

Table 1: *Questionnaire results: Except for the names of the clusters and the relevant papers, all results exhibit values over 6. Many participants stated that the clusters appeared correct, but that the names were not completely describing them, despite not being wrong either.*

Questions: 'I believe the application…'	Avg.
helps me understand the different research areas	6.0
lets me understand who is collaborating with anybody	6.2
can be used to find interdisciplinary research areas	6.6
lets me find groups/people for a new research institute	6.0
is suitable to understand which areas are growing	6.0
cluster names suitably identify the papers	5.2
lets me easily find relevant publications in an area	4.6
I believe I could use this application frequently	6.0
I believe the application is easy to use	6.2
I believe the application is easy to learn	6.6
I feel confident using the system	6.2

- Look for a growth region and check the cluster inside and determine if the growth stats corresponded to a cluster of recent growth as expected.
- Search a term ('photogrammetry') and determine whether the institution was doing research that included this concept.
- Look for papers at the intersection of 'Medicine' and 'Computer Science' and determine whether there was research being carried out at the university.
- Search a given author (unknown to the participant) and check what research is the author involved into.

All the participants managed to do the tasks effectively and efficiently (all of them were performed in 1–3 min, approximately).

Questionnaire: After those tasks were completed, we asked them to fill a questionnaire using a Google Form. The questionnaire examined two different aspects: the features of the application and the usability. Users were asked to answer in a 1–7 Likert scale. There were two open questions asking for comments or suggestions. The results are shown in Table 1 (questions are abbreviated here to save some space, they are commonly written as "I believe that the application …"). The participants were positively impressed by the features of the tool. They ranked all the features with good grades, only the questions regarding the names of the clusters, or the relevant papers in an area were aspects that, though not incorrect, could be improved, as some of them commented.

Comments and insights Most of them found the application very useful and that it could be of utility for their work. The vice rector of research from the external university said, 'I would like to have this application in my university', 'it lets you see what the people are researching'. When further asked about this, she noted that her university has four different locations, and, though it is easy for her to gain a global knowledge on the main research areas of the different groups, it is very difficult for her to distill and get insights on an individual levels. Thus, our tool would allow her to get insights on: (a) Who is working on a certain research area as their main focus, or only on a tangential way, (b) which are the relationships/connections between the different researchers of the institution, (c) engage in promoting certain collaborations between groups in certain areas of interest, and (d) finding researchers that can work in projects proposed by the nearby industry or by public institutions. The vice dean of Postgraduate studies found a use case we had not thought of when designing the application: looking for a lecturer in a specific area. A professor had left to a startup recently, and it had been very complex to find a substitute. And he said that he had looked for a lecturer when training with the application during the experiment and actually successfully found the actual person that now was teaching the course. The difference was that the application guided him to the person within seconds, where the actual process without the application, months ago, had cost him a lot of effort and talking to many people. The director of the research institute wished he had had the application available two years ago: to increase the multidisciplinarity of the institute, he had scanned all the research groups in the university *manually*, a process that took several days. He believes that the same task would have been a matter of hours instead of days with the tool.

8. Discussion

8.1. Exploratory analysis of document corpus

The necessity of the tool was made evident in our university due to the high-level inquiries that the governance team had to address when addressing diverse long-term objectives. The evidence of its significance is exemplified by the fact that a distinct governance board has maintained an interest in the development of the tool. Our exploration is heavily based on the cluster concept, which signifies a specific area of research. We also examined topics, such as other systems [CMH12, GOB*12, LZP*12, CLRP13, KKP*17, DN18, ESS*18]. Unfortunately, they are not suitable for our institution because topic analysis leads to a high variety of elements that are difficult to classify. Therefore, our system requires a higher level of classification (we use the Scopus second-level classification) to communicate about the disciplines.

Our visualisation tool revolves around the discipline/cluster inference from the contents of the documents. Therefore, to provide an answer to the question about the fields of expertise of our university, we designed a novel pipeline for cluster detection, that combines the analysis of the whole documents, clustering at 10 dimensions, and a naming strategy that uses a predefined hierarchy, as well as information extracted from the data themselves. As a result, we can provide a list of clusters with names that describe them both top-down (with Scopus names) and bottom-up (with frequent words within the cluster), and a 2D spatialisation with coloured clusters. This can be used to describe our university research wise. The cluster view provides numerical information on the number of elements of the cluster as well as its progression, which facilitates comparing the state of each subfield.

Nevertheless, reducing the dimensionality results in the loss of information. Hence, the outcome may be difficult to explore since it may conceal relationships or lead users to infer relationships that are, in fact, nonexistent [NA18, JPN15]. Unlike other approaches, where the main goal is to classify papers based on the discipline (e.g., [LZJZ20]), our goals encompassed discovering relationships at different levels. Thus, our 2D layout includes relationships that can be found by spatial proximity, clusters, and links. This

overcomes the problem of spurious relations that might arise solely based on 2D positions. Previous literature does not help finding these connections unless there are direct citations or co-citations between documents.

The Scopus layer, that provides a higher-level classification of the papers, helps university officials understanding potential interdisciplinary areas of research, and can also be used, in combination with the cluster view, to create a multidisciplinary team to address complex challenges such as EU Missions, or XPrize competitions. The growth map easily displays in which areas the university is growing or which ones may be stagnant. It can be used for the early detection of potential emerging research fields.

A set of interactive tools enable deeper analysis of the data, such as the searching function, that can be used to find terms or authors. The combination of these views and layers and the interaction techniques, enable solving complex tasks, as demonstrated by the different use cases. For example, the vice dean of Master Studies was able to quickly find the adequate faculty for teaching a specific advanced course. The very same task had taken him hours and multiple phone calls without the use of Atlas. The presence of the powerful search tool has several advantages that help users save time. First, if the area of interest is already known, it can be searched in the box. The same applies to the authors. Secondly, the search tool additionally provides autocomplete with suggestions, which is highly advantageous for reducing time. It can also be populated from other views, such as the document view, with just a few simple clicks. The resulting list is not constrained to a predetermined number of documents. Moreover, it generates a list of related authors that can be incorporated into the search box with a single click.

Document embeddings, in particular, in combination with sophisticated projection techniques (such as UMAP [MHSG18]) results in layouts where neighbourhoods tend to be related and are more stable than force-based layouts [Dri12]. Nevertheless, reducing the dimensionality results in the loss of information. Hence, the outcome may be difficult to explore since it may conceal relationships or lead users to infer relationships that are, in fact, nonexistent [NA18, JPN15]. Unlike other approaches, where the main goal is to classify papers based on the discipline (e.g., [LZJZ20]), our goals encompassed discovering relationships at different levels. Thus, our 2D layout includes relationships that can be found by spatial proximity, clusters, and links. This overcomes the problem of spurious relations that might arise solely based on 2D positions. Furthermore, the users can establish connections by utilising common terms or authors' names, which can be accessed through a search function, as well as a comprehensive list of authors.

Some researchers have created tools for the specific communication of the DR results. For example, Chatzimparmpas and colleagues developed a tool for easily inspecting aspects of the reduced space, such as the accuracy or the effects of different hyperparameters [CMK20]. Cutura et al. created VisCoDeR, a visual tool for exploring the effect of hyperparameters of DR algorithms, as well as directly comparing the resulting projections [CHAS18]. Our objective, in contrast, is to guide the university officials to the details (number of articles, growth, etc.) of the clusters or articles.

8.2. Generalizability

The modular architecture of our system facilitates straightforward replication and exhibits strong potential for generalisation across diverse institutional contexts and document types. The primary constraint for deploying an identical pipeline at another university lies in the accessibility of the specific data fields extracted via the current API. Given that the acquired data is structured in json format, subsequent processing stages are readily transferable to other research centres. All downstream components, including cluster generation, term extraction, and link analysis, are derived directly from this data. In scenarios involving institutions with substantially different research foci (e.g. humanities-centric), the principal adaptation would involve retraining the doc2vec model to generate domain-specific document embeddings. Alternatively, as previously noted, Low-Rank Adaptation (LoRA) [HysW*22] could be applied to a pre-trained Large Language Model using a corpus of relevant scholarly works. Consequently, we posit that the migration of our software infrastructure to other institutions presents a manageable challenge, as the foundational information required is inherently present in scholarly publications. The cluster naming convention employed in our application adheres to the Scopus classification system, a widely adopted standard across various academic disciplines. While institutions whose librarians do not utilise the Scopus classification would require a modification of these labels, such an adjustment constitutes a relatively simple undertaking.

Furthermore, the generalizability of our system extends to alternative data modalities, such as project proposals and technology transfer agreements. Although the specific analytical tasks would need adaptation, the underlying Natural Language Processing (NLP) pipeline and the majority of visualisation and interaction motifs are likely reusable across these diverse data streams.

8.3. Lessons learned

Throughout the development of the project, we encountered numerous issues regarding data acquisition and processing, cluster creation and representation, and differences in the perception of the relevant goals from university officials.

Data acquistion. First, the acquisition of the data can be non-trivial. Every institution has a different database with different metadata, and the metadata are often not publicly available through an API. For instance, our website https://upccommons.upc.edu provides a substantial amount of information in a structured fashion, but downloading the required data in bulk fashion is not allowed. It was necessary to create specific scripts and needed special permissions to access the required data. We also created a version of the system with data from the public Spanish publications aggregator Recolecta https://recolecta.fecyt.es/. However, the system lacks significant information that is essential for the exploration, such as the classification of journals. Furthermore, it is not specifically designed to facilitate bulk downloads. Therefore, we created scraper scripts. Unfortunately, aggregators cannot guarantee that the links they point to will be working and have the appropriate tag for the PDF download.

Cluster creation and naming. The creation of cluster names was very complex. First, trying to classify all the nodes in a

cluster does not work, since there are papers that deal with areas that are very different from the more general research. Adding them to other existing clusters may be confusing to researchers, and joining all outliers together may not make sense. These documents tend to be separated, commonly isolated from the other articles, in the final 2D view. Second, determining the adequate size of the clusters is also tricky: we do not want the number of clusters to be too small, which will hide the variety, or too many clusters, which would be difficult to explore. We believe that the optimal clustering algorithm must be fine-tuned by leveraging the knowledge within the institution. Creating the clusters in 10D and then using 2D layouts does not generate a large visual distortion. That is, the majority of the nodes appear in the 2D layout close to each other and only a small portion of the nodes have a location with shows them intermingled with other clusters. We manually inspected around the 30% of the clusters, and found that the number of nodes that are not in the surroundings of the cluster they belong to, and are placed inside other clusters (without counting the unclassified nodes) was less than 1.2% of the nodes of each cluster. The resulting 2D distribution, including the presence of the unclassified nodes, did not appear to represent any issue for the people using the tool in the evaluation sessions.

Different users, different goals. When we began showing initial versions to different levels of university governance, various opinions were expressed. One of the controversial issues relates to the metrics that were measured. Some officials suggested the idea of having quantitative metrics that might lead to the evaluation of the researchers. We decided against this idea for several reasons. First, the dataset is incomplete, which would be unfair. Second, the university has established other methods for research evaluation. Third, the definition of metrics has implications for what constitutes a more significant contribution, which can be challenging to objectively assess, particularly within an institution with a diverse range of research.

9. Conclusions and Future Work

9.1. Conclusions

In this paper, we have presented an exploratory tool designed to facilitate governance decisions regarding research management within a university. The application is currently available to all users inside the institution network. The multiple view system allows to get highly valuable insights that are difficult to achieve with the current tools, such as understanding globally the areas of research within the university, getting to know who is an important actor in each area, detecting emerging areas, or looking for researchers with a certain concrete expertise. Internal databases do not provide a global view of the research in the university, and combining queries to the database and other tools such as Google Scholar can result in lists of articles, but no relationships are available. Numerous applications have recently appeared in this domain, with the main goal of facilitating the research of literature, such as http://researchrabbit.ai, https://scite.ai/, https://www.connectedpapers.com/, or https://researcher.life/. These tools primarily provide a compilation of articles and other services, such as summaries or aids for authoring papers. Research Rabbit or Connected Papers, however, also provide paper graphs, mostly as node-link diagrams, that are related to citations or authorship. However, these tools do not generate disciplines nor provide us with the necessary level of exploration.

9.2. Future Work

The current version of the application includes 15K+ articles. The final goal is to include all papers available in readable PDF (older papers contain scans of the original documents). Although no scalability tests were specifically conducted, the rendering subsystem uses deckGL https://deck.gl/, which is intended to render massive datasets. Thus, we anticipate no significant issues for datasets of up to two orders of magnitude larger than the current set. A potential limitation may be the trained model. As previously mentioned, we utilise doc2vec, which remains superior to more modern techniques, as shown in a recent study [RV24]. However, finding better embeddings is also another line of research. Some alternatives could be SPECTER [CFB*20]. However, this method is also trained on the title and abstract, rather than the full document. It would be interesting to test with longer text inputs. Furthermore, we also believe that utilising a pre-trained model without the need for fine-tuning would be advantageous for the environment. In this regard, Large Language Models have the potential to provide a solution. Despite their impact on the environment, their usage once trained would not, at the very least, increase the CO2 footprint as compared to training new models.

Another potential problem that will arise when we scale up the number of documents is the clutter. The current dataset exhibits dense regions, but zooming in does reveal all individual articles without problem. For much larger datasets (e.g, one order of magnitude larger, with more articles in each discipline) this might not be enough. Therefore, some strategy should be designed. This is a well-known problem that has been previously analysed [ED07].

In our case, we used a set of documents within research areas that are representative of our university. If the system were to be deployed at a different university with a different research focus (e.g. one with the most contributions in the social sciences), and we do not change the current method of obtaining the embeddings, retraining the model would be necessary. Nonetheless, this is a one-time task. The addition of a large new set of documents will require several days of computation, since most of the derived data will need to be regenerated. Given that this operation typically occurs once or twice a year, it is not considered a significant concern. However, if a substantial amount of new data were to be added, such as doubling the size of the database, it may be necessary to readjust the cluster parameters. This adjustment process is largely automated through a set of scripts, although a domain expert validates the final decision. Additionally, an authentication mode has been implemented to facilitate the inclusion of user profile-based restrictions, as the application may potentially display sensitive or private information. This feature enables different roles within the university governance to have appropriate access restrictions.

The system is institution-agnostic, since all the pipeline requires only the PDFs of the files to analyse. We used data that is available publicly. Thus, any other university could use it, quite straightforwardly. The current system is still being extended with new features.

In the future, we want to enhance the clusters' analysis with new views that facilitate more insights, such as cluster changes over time, or visual comparison of two clusters' stats. It seems quite straightforward, since all the elements are already calculated, but a better layout to facilitate visual comparisons (and maybe some extra metrics) might be required. Although the current design is tightly linked to the university governance needs, a similar design could be suitable for many companies to find research related to their products, for example. Moreover, by changing the data, the NLP stack and many of the visualisation motifs (e.g., the 2D layout or the inspection views) can also be used to tackle other problems such as getting insights of the technology transfer projects funded by an institution, comparing the degrees' curricula of different universities, searching for patents similar to a new product, and so on.

The data that support the findings of this study are openly available in https://upccommons.upc.edu. The authors have no conflict of interest.

Acknowledgements

This project has been supported by PID2021-122136OB-C21 from the Ministerio de Ciencia e Innovación, Spain, by 839 FEDER (EU) funds.

References

[AB17] Alvarez J. E., Bast H.: *A Review of Word Embedding and Document Similarity Algorithms Applied to Academic Text*. Bachelor thesis, University of Freiburg (22nd October 2017).

[AKV*14] Alexander E., Kohlmann J., Valenza R., Witmore M., Gleicher M.: Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2014), pp. 173–182. https://doi.org/10.1109/VAST.2014.7042493

[BEH*18] Burd R., Espy K. A., Hossain M. I., Kobourov S., Merchant N., Purchase H.: GRAM: Global research activity map. In Catarci T., Norman K. L., Mecella M. (Eds.), *Proceedings of the 2018 International Conference on Advanced Visual Interfaces* (2018), pp. 1–9.

[BISM14] Brehmer M., Ingram S., Stray J., Munzner T.: Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2271–2280. https://doi.org/10.1109/TVCG.2014.2346431

[BLB*14] Burch M., Lohmann S., Beck F., Rodriguez N., Di Silvestro L., Weiskopf D.: Radcloud: Visualizing multiple texts with merged word clouds. In *2014 18th International Conference on Information Visualisation* (2014), pp. 108–113. https://doi.org/10.1109/IV.2014.72

[BNHL14] Bradel L., North C., House L., Leman S.: Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2014), pp. 163–172. https://doi.org/10.1109/VAST.2014.7042492

[BNJ03] Blei D. M., Ng A. Y., Jordan M. I.: Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

[BWOW20] Bales M. E., Wright D. N., Oxley P. R., Wheeler T. R.: Bibliometric visualization and analysis software: State of the art, workflows, and best practices, Cornell University Library, 2020. https://ecommons.cornell.edu/handle/1813/69597

[CA98] Carroll J. D., Arabie P.: Multidimensional scaling. In: Birnbaum M. H. (Ed.), *Handbook of Perception and Cognition (Second Edition), Measurement, Judgment and Decision Making*, Academic Press (1998), pp. 179–250.

[CAS16] Christian H., Agus M. P., Suhartono D.: Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications 7*, 4 (2016), 285–294.

[CB12] Chaney A., Blei D.: Visualizing topic models. In *Proceedings of the International AAAI Conference on Web and Social Media* (2012), vol. *6*, pp. 419–422.

[CFB*20] Cohan A., Feldman S., Beltagy I., Downey D., Weld D. S.: SPECTER: Document-level representation learning using citation-informed transformers. In: Jurafsky, D., Chai J., Schluter N., Tetreault J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (2020), pp. 2270–2282.

[CHAS18] Cutura R., Holzer S., Aupetit M., Sedlmair M.: Viscoder: A tool for visually comparing dimensionality reduction algorithms. In *26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2018* (2018), i6doc.com publication, pp. 105–110.

[Che06] Chen C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology 57*, 3 (2006), 359–377.

[CLRP13] Choo J., Lee C., Reddy C. K., Park H.: UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 1992–2001. https://doi.org/10.1109/TVCG.2013.212

[CMH12] Chuang J., Manning C. D., Heer J.: Termite: visualization techniques for assessing textual topic models. In *International Working Conference on Advanced Visual Interfaces, AVI 2012, Capri Island, Naples, Italy, May 22–25, 2012, Proceedings* (2012), G. Tortora, S. Levialdi, M. Tucci (Eds.), ACM, pp. 74–77. https://doi.org/10.1145/2254556.2254572

[CMK20] Chatzimparmpas A., Martins R. M., Kerren A.: t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics 26*, 8 (2020), 2696–2714. https://doi.org/10.1109/TVCG.2020.2986996

[CMS13] Campello R. J. G. B., Moulavi D., Sander J.: Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II* (2013), J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), vol. *7819* of *Lecture Notes in Computer Science*, Springer, pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

[CNS23] Crosthwaite P., Ningrum S., Schweinberger M.: Research trends in corpus linguistics: A bibliometric analysis of two decades of scopus-indexed corpus linguistics research in arts and humanities. *International Journal of Corpus Linguistics 28*, 3 (2023), 344–377. https://doi.org/10.1075/ijcl.21072.cro

[CRF*21] Caillou P., Renault J., Fekete J., Letournel A., Sebag M.: Cartolabe: A web-based scalable visualization of large document collections. *IEEE Computer Graphics and Applications 41*, 02 (March 2021), 76–88. https://doi.org/10.1109/MCG.2020.3033401

[CRMH12] Chuang J., Ramage D., Manning C., Heer J.: Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), pp. 443–452.

[CSL*10] Cao N., Sun J., Lin Y.-R., Gotz D., Liu S., Qu H.: Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1172–1181. https://doi.org/10.1109/TVCG.2010.154

[DN18] Dang T., Nguyen V. T.: ComModeler: Topic modeling using community detection. In *Proceedings of the EuroVis Workshop on Visual Analytics* (2018), C. Tominski, T. von Landesberger (Eds.), EuroVA '18, The Eurographics Association, pp. 1–5. https://doi.org/10.2312/eurova.20181104

[Dri12] Drieger P.: Visual Text Analytics Using Semantic Networks and Interactive 3D Visualization. In *EuroVA 2012: International Workshop on Visual Analytics* (2012), K. Matkovic and G. Santucci (Eds.), The Eurographics Association. https://doi.org/10.2312/PE/EuroVAST/EuroVA12/043-047

[DSG*12] Dunne C., Shneiderman B., Gove R., Klavans J., Dorr B.: Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology 63*, 12 (2012), 2351–2369. https://doi.org/10.1002/asi.22652

[ED07] Ellis G., Dix A.: A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1216–1223.

[EHA*23] Eckelt K., Hinterreiter A., Adelberger P., Walchshofer C., Dhanoa V., Humer C., Heckmann M., Steinparz C., Streit M.: Visual exploration of relationships and structure in low-dimensional embeddings. *IEEE Transactions on Visualization and Computer Graphics 29*, 7 (2023), 3312–3326. https://doi.org/10.1109/TVCG.2022.3156760

[EKSW*14] El-Kishky A., Song Y., Wang C., Voss C. R., Han J.: Scalable topical phrase mining from text corpora. In *Proceedings of the VLDB Endowment* (Novemeber 2014), vol. *8*, VLDB Endowment, pp. 305–316. https://doi.org/10.14778/2735508.2735519

[EMK*21] Espadoto M., Martins R. M., Kerren A., Hirata N. S. T., Telea A. C.: Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics 27*, 3 (2021), 2153–2173. https://doi.org/10.1109/TVCG.2019.2944182

[ESS*18] El-Assady M., Sperrle F., Sevastjanova R., Sedlmair M., Keim D. A.: LTMA: Layered topic matching for the comparative exploration, evaluation, and refinement of topic modeling results. In *2018 International Symposium on Big Data Visual and Immersive Analytics, BDVA 2018, Konstanz, Germany, October 17–19, 2018* (2018), IEEE, pp. 1–10. https://doi.org/10.1109/BDVA.2018.8534018

[FHKM17] Federico P., Heimerl F., Koch S., Miksch S.: A survey on visual approaches for analyzing scientific literature and patents. *IEEE Transactions on Visualization and Computer Graphics 23*, 9 (2017), 2179–2198. https://doi.org/10.1109/TVCG.2016.2610422

[FK14] Fried D., Kobourov S. G.: Maps of computer science. In *IEEE Pacific Visualization Symposium, PacificVis 2014, Yokohama, Japan, March 4–7, 2014* (2014), I. Fujishiro, U. Brandes, H. Hagen, S. Takahashi (Eds.), IEEE Computer Society, pp. 113–120. https://doi.org/10.1109/PACIFICVIS.2014.47

[FKM20] Fujiwara T., Kwon O.-H., Ma K.-L.: Supporting analysis of dimensionality reduction results with contrastive learning. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 45–55. https://doi.org/10.1109/TVCG.2019.2934251

[GLB24] Gleicher M., Leppenan K., Bai Y.: Enhancing text corpus exploration with post hoc explanations and comparative design. *CoRR abs/2406.09686* (2024). http://arxiv.org/abs/2406.09686

[GLK*13] Görg C., Liu Z., Kihm J., Choo J., Park H., Stasko J.: Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics 19*, 10 (2013), 1646–1663. https://doi.org/10.1109/TVCG.2012.324

[GOB*12] Gretarsson B., O'Donovan J., Bostandjiev S., Höllerer T., Asuncion A. U., Newman D., Smyth P.: TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology 3*, 2 (2012), 23:1–23:26. https://doi.org/10.1145/2089094.2089099

[GV22] Gómez J., Vázquez P.-P.: An empirical evaluation of document embeddings and similarity metrics for scientific articles. *Applied Sciences 12*, 11 (2022), 5664. https://doi.org/10.3390/app12115664

[HAAE17] Humayoun S. R., Ardalan S., AlTarawneh R., Ebert A.: TExVis: An interactive visual tool to explore twitter data. In *19th Eurographics Conference on Visualization, EuroVis 2017 - Short Papers, Barcelona, Spain, June 12–16, 2017* (2017), B. Kozlíková, T. Schreck T. Wischgoll, (Eds.) Eurographics Association, pp. 151–155. https://doi.org/10.2312/EUROVISSHORT.20171149

[HHN00] Havre S., Hetzler B., Nowell L.: Themeriver: visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* (2000), pp. 115–123. https://doi.org/10.1109/INFVIS.2000.885098

[HJH*16] Heimerl F., John M., Han Q., Koch S., Ertl T.: DocuCompass: Effective exploration of document landscapes. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2016), pp. 11–20. https://doi.org/10.1109/VAST.2016.7883507

[HKH*14] Ha H., Kim G.-n., Hwang W., Choi H., Lee K.: CosMovis: Analyzing semantic network of sentiment words in movie reviews. In *2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV)* (2014), pp. 113–114. https://doi.org/10.1109/LDAV.2014.7013215

[HT04] Hetzler E., Turner A.: Analysis experiences using information visualization. *Computer Graphics and Applications, IEEE 24*, 5 (September 2004), 22–26. https://doi.org/10.1109/MCG.2004.22

[HysW*22] Hu E. J., yelong shen, Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W.: LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (2022). https://openreview.net/forum?id=nZeVKeeFYf9.

[Jon04] Jones K. S.: A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 60*, 5 (2004), 493–502. https://doi.org/10.1108/00220410410560573

[JPN15] Joia P., Petronetto F., Nonato L. G.: Uncovering representative groups in multidimensional projections. *Computer Graphics Forum 34*, 3, 281–290.

[JSR*19] Ji X., Shen H.-W., Ritter A., Machiraju R., Yen P.-Y.: Visual exploration of neural document embedding in information retrieval: Semantics and feature selection. *IEEE Transactions on Visualization and Computer Graphics 25*, 6 (June 2019), 2181–2192. https://doi.org/10.1109/TVCG.2019.2903946

[KCW*19] Karduni A., Cho I., Wesslen R., Santhanam S., Volkova S., Arendt D. L., Shaikh S., Dou W.: Vulnerable to misinformation?: Verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2019), IUI '19, ACM, pp. 312–323. https://doi.org/10.1145/3301275.3302320

[KK15] Kucher K., Kerren A.: Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)* (2015), pp. 117–121. https://doi.org/10.1109/PACIFICVIS.2015.7156366

[KKP*17] Kim M., Kang K., Park D., Choo J., Elmqvist N.: Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 151–160. https://doi.org/10.1109/TVCG.2016.2598445

[KP13] Kuang D., Park H.: Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11–14, 2013* (2013), I. S. Dhillon, Y. Koren, R. Ghani, T. E. Senator, P. Bradley, R. Parekh, J. He, R. L. Grossman, R. Uthurusamy (Eds.), ACM, pp. 739–747. https://doi.org/10.1145/2487575.2487606

[KRK21] Kohlmeyer L., Repke T., Krestel R.: Novel views on novels: Embedding multiple facets of long texts. In *WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence, Melbourne VIC Australia, December 14–17, 2021* (2021), J. He, R. Unland, E. S. Jr., X. Tao, H. Purohit, W. van den Heuvel, J. Yearwood, J. Cao (Eds.), ACM, pp. 670–675. https://doi.org/10.1145/3486622.3494006

[LJLH19] Liu Y., Jun E., Li Q., Heer J.: Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum 38*, 3, 67–78.

[LKC*12] Lee H., Kihm J., Choo J., Stasko J., Park H.: iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum 31*, 3, 1155–1164.

[LKCH21] Lafia S., Kuhn W., Caylor K. K., Hemphill L.: Mapping research topics at multiple levels of detail. *Patterns 2*, 3 (2021), 100210. https://doi.org/10.1016/J.PATTER.2021.100210

[LLD04] Landauer T. K., Laham D., Derr M.: From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Sciences 101*, suppl 1 (2004), 5214–5219.

[LLZ*16] Liu M., Liu S., Zhu X., Liao Q., Wei F., Pan S.: An uncertainty-aware approach for exploratory microblog retrieval. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 250–259. https://doi.org/10.1109/TVCG.2015.2467554

[LM14] Le Q. V., Mikolov T.: Distributed representations of sentences and documents. *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, JMLR Workshop and Conference Proceedings (2014), vol. *32*, pp. 1188–1196.

[LTW*18] Liu J., Tang T., Wang W., Xu B., Kong X., Xia F.: A survey of scholarly data visualization. *IEEE Access 6* (2018), 19205–19221. https://doi.org/10.1109/ACCESS.2018.2815030

[LYM*22] Li K., Yang H., Montoya E., Upadhayay A., Zhou Z., Saad-Falcon J., Chau D. H.: Visual exploration of literature with argo scholar. Al Hasan M., Xiong L. (Eds.), *Proceedings of the 31st {ACM} International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022* (2022), pp. 4912–4916. https://doi.org/10.1145/3511808.3557177

[LZJZ20] Li Z., Zhang C., Jia S., Zhang J.: Galex: Exploring the evolution and intersection of disciplines. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 1182–1192. https://doi.org/10.1109/TVCG.2019.2934667

[LZP*12] Liu S., Zhou M. X., Pan S., Song Y., Qian W., Cai W., Lian X.: TIARA: interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST) 3*, 2 (2012), 25:1–25:28. https://doi.org/10.1145/2089094.2089101

[Mar06] Marchionini G.: Exploratory search: from finding to understanding. *Communications of the ACM 49*, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

[MHSG18] McInnes L., Healy J., Saul N., Großberger L.: UMAP: uniform manifold approximation and projection. *Journal of Open Source Software 3*, 29 (2018), 861. https://doi.org/10.21105/JOSS.00861

[NA18] Nonato L. G., Aupetit M.: Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics 25*, 8 (2018), 2650–2673.

[NKWW22] Narechania A., Karduni A., Wesslen R., Wall E.: VITALITY: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2022), 486–496. https://doi.org/10.1109/TVCG.2021.3114820

[PCE*19] Paul C. L., Chang J., Endert A., Cramer N., Gillen D., Hampton S. D., Burtner R., Perko R., Cook K. A.: Textonic: Interactive visualization for exploration and discovery of very large text collections. *Information Visualization 18*, 3 (2019). https://doi.org/10.1177/1473871618785390

[PMP*21] Polo F. M., Mendonça G. C. F., Parreira K. C. J., de Godoy Gianvechio L., Cordeiro P., Ferreira J. B., de Lima L. M. P., do Amaral Maia A. C., Vicente R.: Legalnlp - natural language processing methods for the brazilian legal language. *CoRR abs/2110.15709* (2021).

[Pra21] Pranckute R.: Web of science (wos) and scopus: The titans of bibliographic information in today's academic world. *Publications 9*, 1 (2021), 12. https://doi.org/10.3390/PUBLICATIONS9010012

[PS12] Padró L., Stanilovsky E.: Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC* (2012), N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), European Language Resources Association (ELRA), pp. 2473–2479.

[PT94] Paatero P., Tapper U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*, 2 (1994), 111–126. https://doi.org/10.1002/env.3170050203

[RB21] Rykova V., Busygina T.: Bibliometric analysis of a research field "paleopedology". *Arabian Journal of Geosciences 14*, 18 (2021), 1939.

[RMB19] Rosenthal P., Müller N. H., Bolte F.: Visual analytics of bibliographical data for strategic decision support of university leaders: A design study. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2019, Volume 3: IVAPP* (2019), Kerren A., Hurter C., Braz J., (Eds.), SciTePress, pp. 297–305. https://doi.org/10.5220/0007396302970305

[RV24] Rafieian B., Vázquez P.: Evaluating the suitability of long document embeddings for classification tasks: A comparative analysis. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K* (2024), F. Coenen, A. Fred, J. Bernardino (Eds.), SCITEPRESS, pp. 320–327. https://doi.org/10.5220/0012950400003838

[RWVW23] Raval S., Wang C., Viégas F., Wattenberg M.: Explain-and-test: An interactive machine learning framework for exploring text embeddings. In *2023 IEEE Visualization and Visual Analytics (VIS)* (2023), pp. 216–220. https://doi.org/10.1109/VIS54172.2023.00052

[SDMT16] Stahnke J., Dörk M., Müller B., Thom A.: Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 629–638. https://doi.org/10.1109/TVCG.2015.2467717

[SOR*09] Strobelt H., Oelke D., Rohrdantz C., Stoffel A., Keim D. A., Deussen O.: Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1145–1152. https://doi.org/10.1109/TVCG.2009.139

[SRA22] Soufan A., Ruthven I., Azzopardi L.: Searching the literature: An analysis of an exploratory search task. In *CHIIR: ACM SIGIR Conference on Human Information Interaction and Retrieval* (2022), D. Elsweiler (Ed.), ACM, pp. 146–157. https://doi.org/10.1145/3498366.3505818

[VdMH08] Van der Maaten L., Hinton G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 11 (2008), 2579–2605.

[Won18] Wong D.: VOSviewer. *Technical Services Quarterly 35*, 2 (2018), 219–220.

[WQQ*24] Wang Y., Qian Y., Qi X., Cao N., Wang D.: Innovationinsights: A visual analytics approach for understanding the dual frontiers of science and technology. *IEEE Transactions on Visualization and Computer Graphics 30*, 1 (2024), 518–528. https://doi.org/10.1109/TVCG.2023.3327387

[WRLW21] Wen Q.-J., Ren Z.-J., Lu H., Wu J.-F.: The progress and trend of bim research: A bibliometrics-based visualization analysis. *Automation in Construction 124* (2021), 103558. https://doi.org/10.1016/j.autcon.2021.103558

[WTP*95] Wise J., Thomas J., Pennock K., Lantrip D., Pottier M., Schur A., Crow V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Visualization 1995 Conference* (1995), pp. 51–58. https://doi.org/10.1109/INFVIS.1995.528686

[WW16] Wang Q., Waltman L.: Large-scale analysis of the accuracy of the journal classification systems of web of science and scopus. *Journal of Informetrics 10*, 2 (2016), 347–364.

[XZS*21] Xia J., Zhang Y., Song J., Chen Y., Wang Y., Liu S.: Revisiting dimensionality reduction techniques for visual cluster analysis: an empirical study. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2021), 529–539.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Video S1