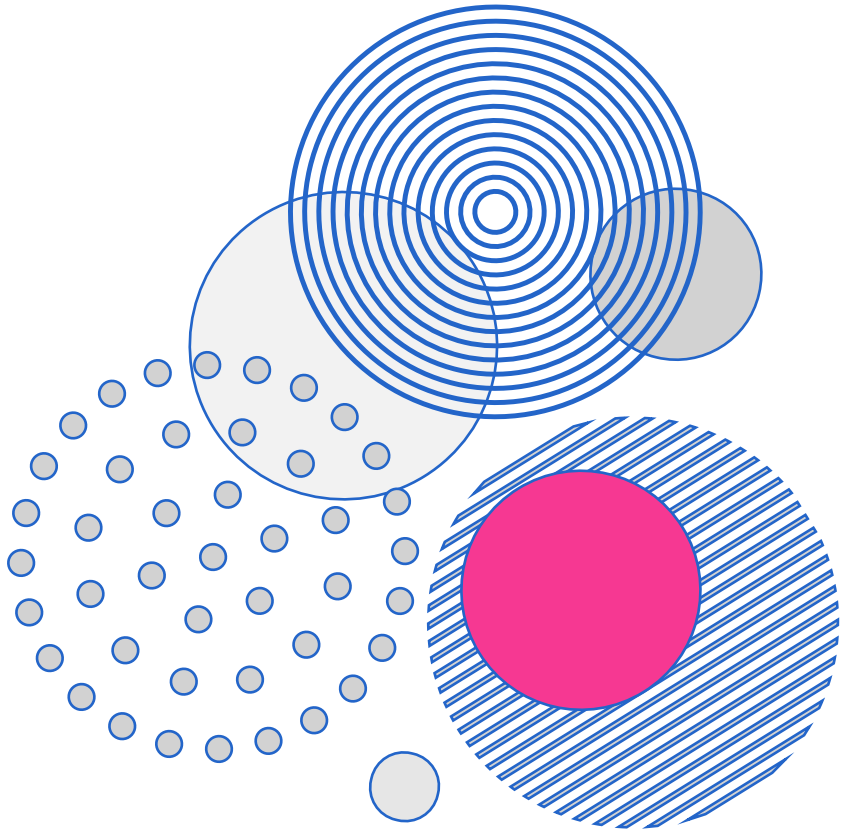
Final Portfolio for CodeCademy Career Path
Data Scientist: Analytics Specialist

Analysis of Scientific Article Metadata

Delphine Vincent

Table of Content

Table of content

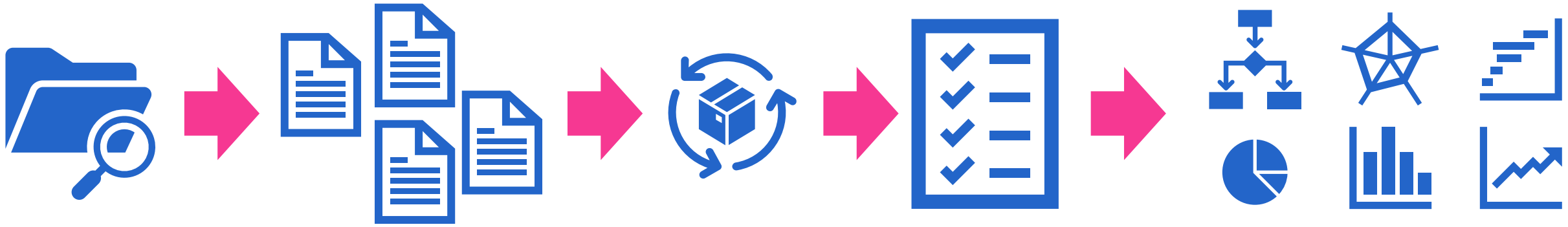


Introduction

Introduction

Why is metadata analysis important?

- Metadata is defined as the information that describes and explains data in a detailed and unambiguous way.
- It provides context with details to understand the relevance of a dataset and how to use it.
- It has various applications in different research areas.
- In this project, I chose a dataset detailing metadata from scientific publications in various science fields (natural sciences, engineering and technologies, medical and health sciences, agricultural sciences, social and behavioral sciences, humanities, arts, etc…).
- I did not produce the dataset but retrieved it from AI Engineer Salman alsheikh via Kaggle (source on see dataset slide).
- Analysing this data will help me understand publication trends and topics of interest.



Dataset

(retrieved from Kaggle)

Dataset

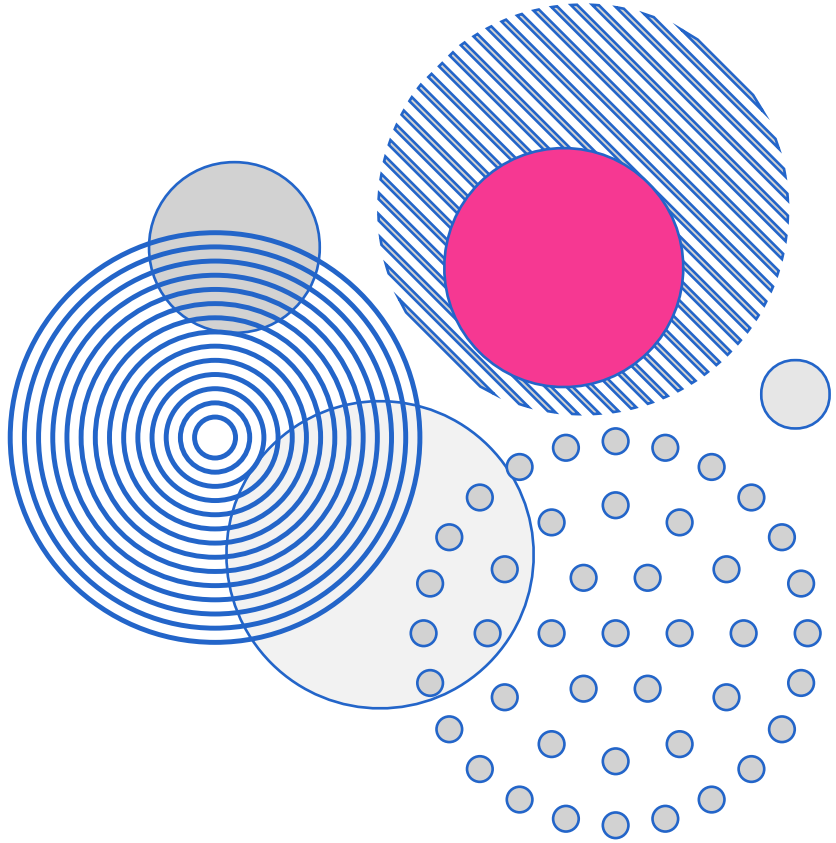
Source: https://www.kaggle.com/datasets/s4lman/scientific-articles-metadata-dataset

Description: This dataset contains detailed metadata on scientific research articles, lending itself to bibliometric analysis, and tracking research trends via Exploratory Data Analysis (EDA).

Raw Dataset dimensions: 120,000 rows x 16 columns

Dataset variables:

- DOI: Digital Object Identifier, the unique ID of a publication
- title: Title of the article
- author: Name of first author
- issued: Date of publication
- abstract: Abstract of the article
- publisher: Publisher of the scientific journal
- container-title: Name of the scientific journal
- volume: Volume number of the scientific journal
- issue: Issue number of the scientific journal
- URL: URL of DOI
- score: ?
- references-count: Number of references used in the article
- language: Language used in the article (2 letters abbreviations)
- subject: Topic of the article
- type: Type of article
- indexed: Date of article indexing

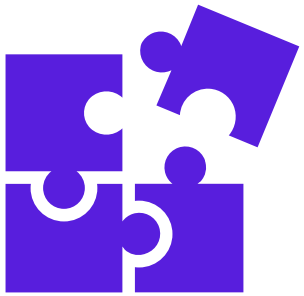


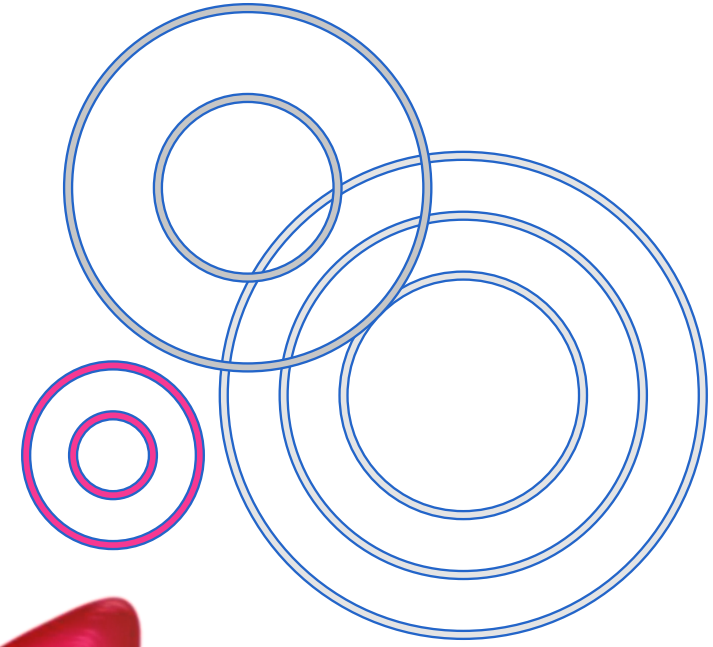
Questions

Questions

What would be interesting to explore?

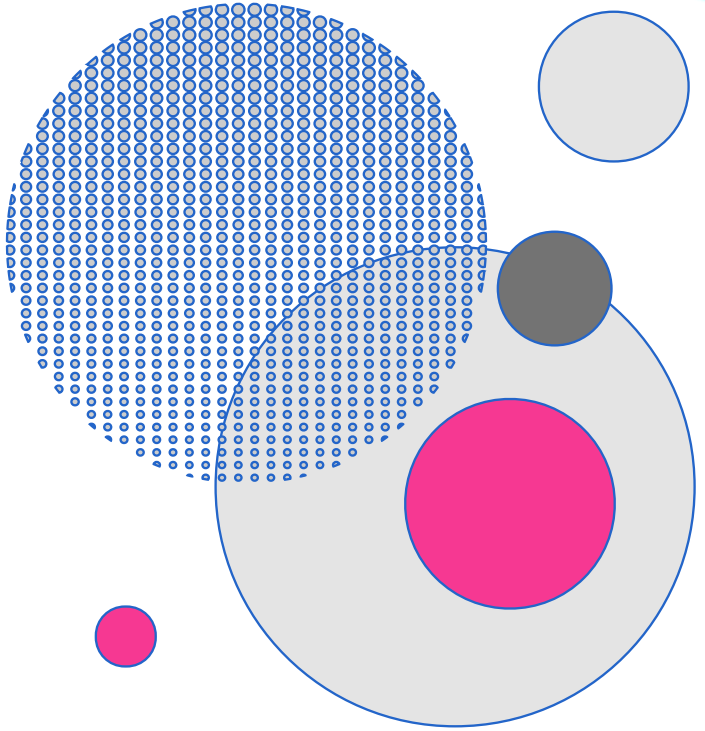
- Publication rate over time
- Successful publishers and favored journals
- Prolific authors
- Number of references and length of title over time
- Preferred publication types
- Languages used
- Most popular subjects





Data Handling

(using Python3 and Excel)



Data Handling

Data assessment and subsetting (KEEP / REMOVE):

- DOI: looks fine, keep as it is. KEEP
- title: few missing values, capitals, special characters, non sensical words, incomplete title, date instead of title etc... KEEP
- author: many missing values, inconsistent format (separators, capitalisation, first name initialed, etc...). KEEP
- issued: the dates formats are varied, needs to be uniformized, extract Years, Months. Convert 'None' to missing values. KEEP
- abstract: mostly missing values, don't use. REMOVE
- publisher: looks fine, keep as it is. KEEP
- container-title: few missing values, otherwise it looks fine, keep as it is, rename header as 'journal'. KEEP
- volume: missing values, inconsistent format (numbers, letters, dates), unnecessary field. REMOVE
- issue: missing values, inconsistent format (numbers, letters, dates), maybe I don't need to use them. REMOVE
- URL: redundant information with DOI, can be removed. REMOVE
- score: always 0, can be removed. REMOVE
- references-count: looks fine (integers), keep as it is, only numerical variable. KEEP
- language: missing values but otherwise it looks fine, keep as it is. KEEP
- subject: missing values but otherwise it looks fine, keep as it is. KEEP
- type: looks fine, keep as it is. KEEP
- indexed: looks fine, keep as it is (dates from 2022-2023). Not sure how relevant it is. REMOVE

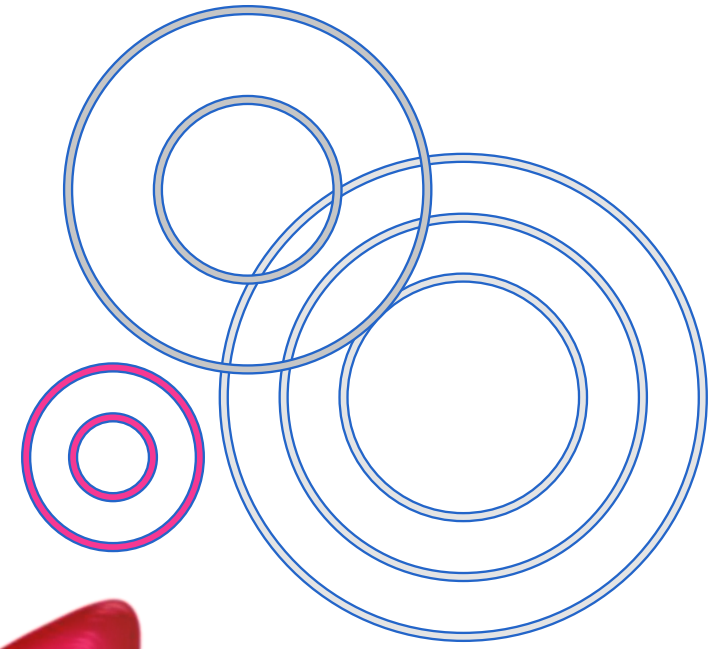
All the variables are categorical, except for 'references-count' which is numerical (integer).

Data Handling

Data wrangling and cleaning:

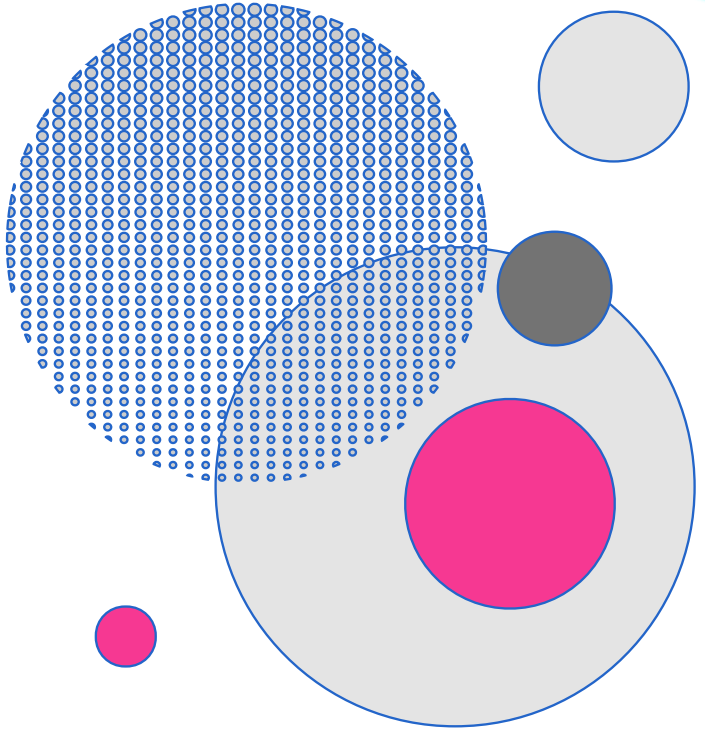
- author: the last names were extracted using str.extract() into new column ‘author_lastname’. Must keep ‘author’ as well.
- issued: the dates were coerced into datetime formats using pd.datetime() into new column ‘publication_date’; the year was extracted using str.extract() into new column ‘publication_year’. The ‘issued’ column was removed.
- container-title: renamed header as 'journal’
- title: it required a multistep approach:
    - remove rows with no authors as they do not correspond to real scientific articles; this eliminates title like dates and numbers as well as some of the empty titles.
    - copy the titles into a new column ‘better_title’
    - using str.lower() convert ‘better_title’ to lower case
    - using str.l replace('[^a-zA-Z0-9\s]', '') remove special characters
    - using str.len() compute title length into new column ‘title_length’
    - using fillna('Unknown', inplace=True), replace missing values with unknown

Clean Dataset dimensions: 103,282 rows x 13 columns

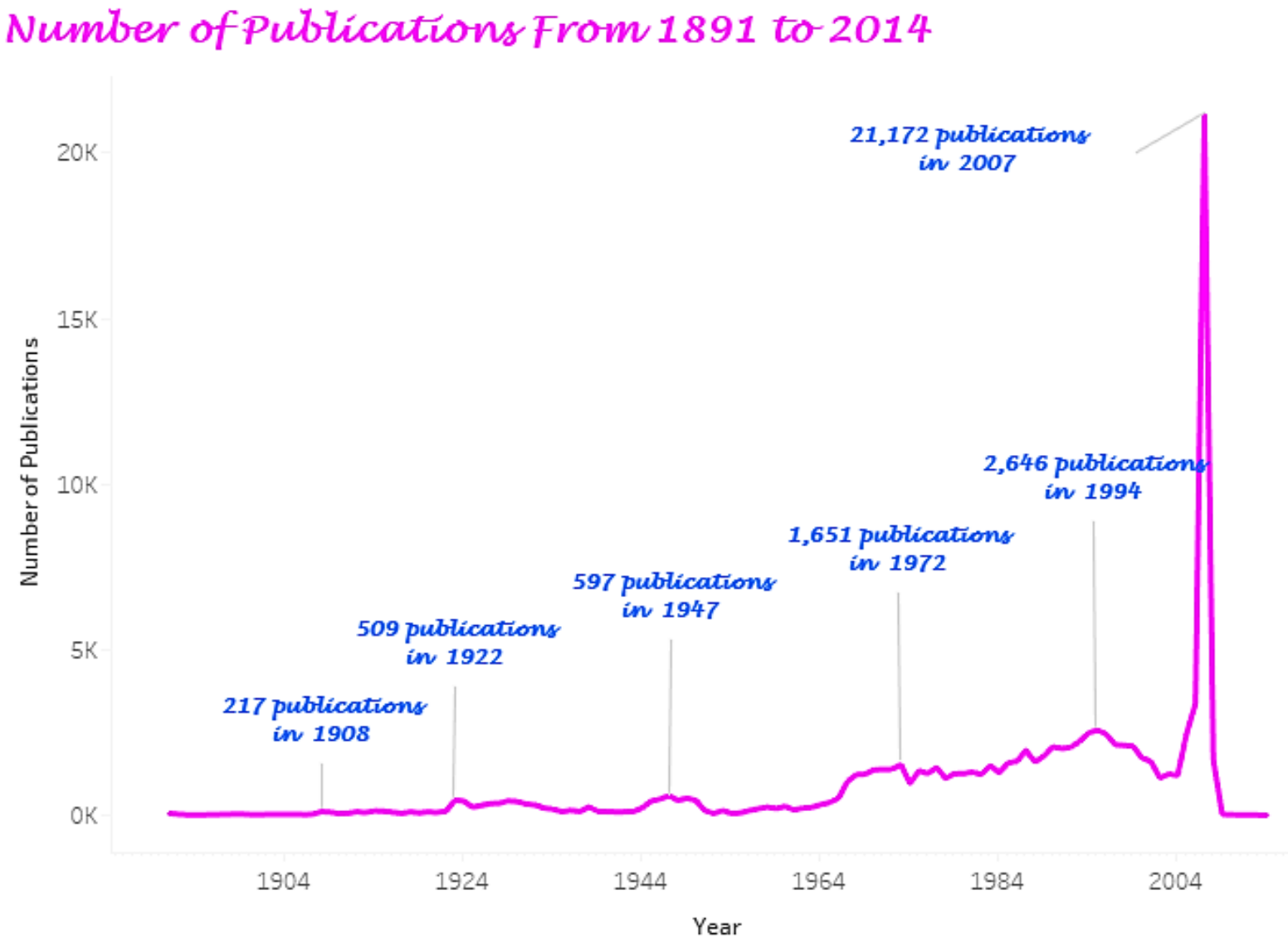


Exploratory Data Analysis (EDA)

(using Python3 and Tableau)



Publication trend



How do publication numbers track over time?

This dataset ranges from 1891 to 2014 with some gaps.

The yearly rate of publication is fairly constant from 1891 to 1960, after which it gradually increases to reach 2,500 publications per year in the mid-90's.

There is a huge spike in 2007 featuring 21,153 publications. Is it a dataset bias or did something happen to explain such spike?

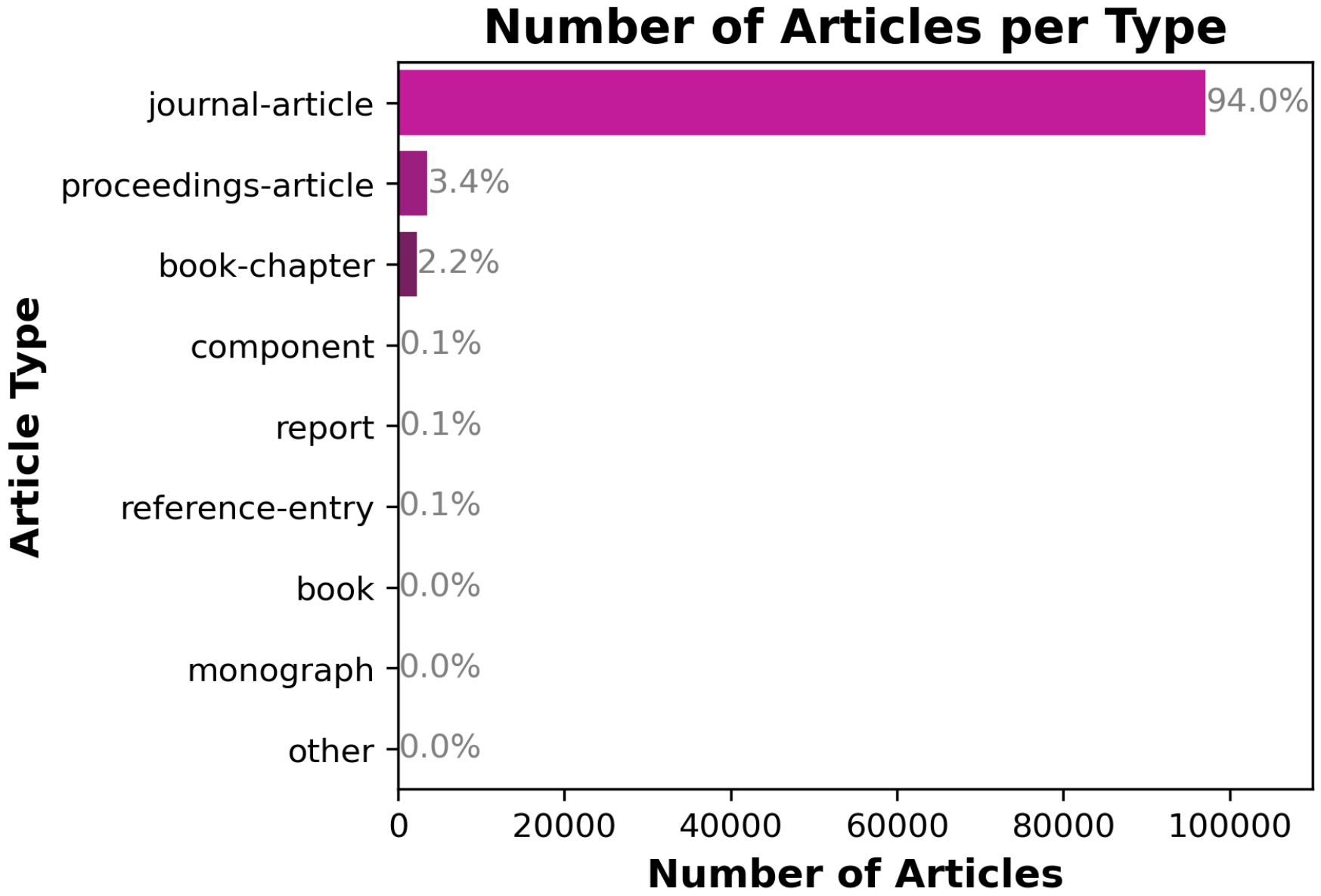
There is a gap between 2010 and 2014.

Years 2009, 2010, and 2014 list only 22, 12, and 1 publications, respectively. Perhaps, those years are outliers and can be ignored.

Article types

Which article types are favored?

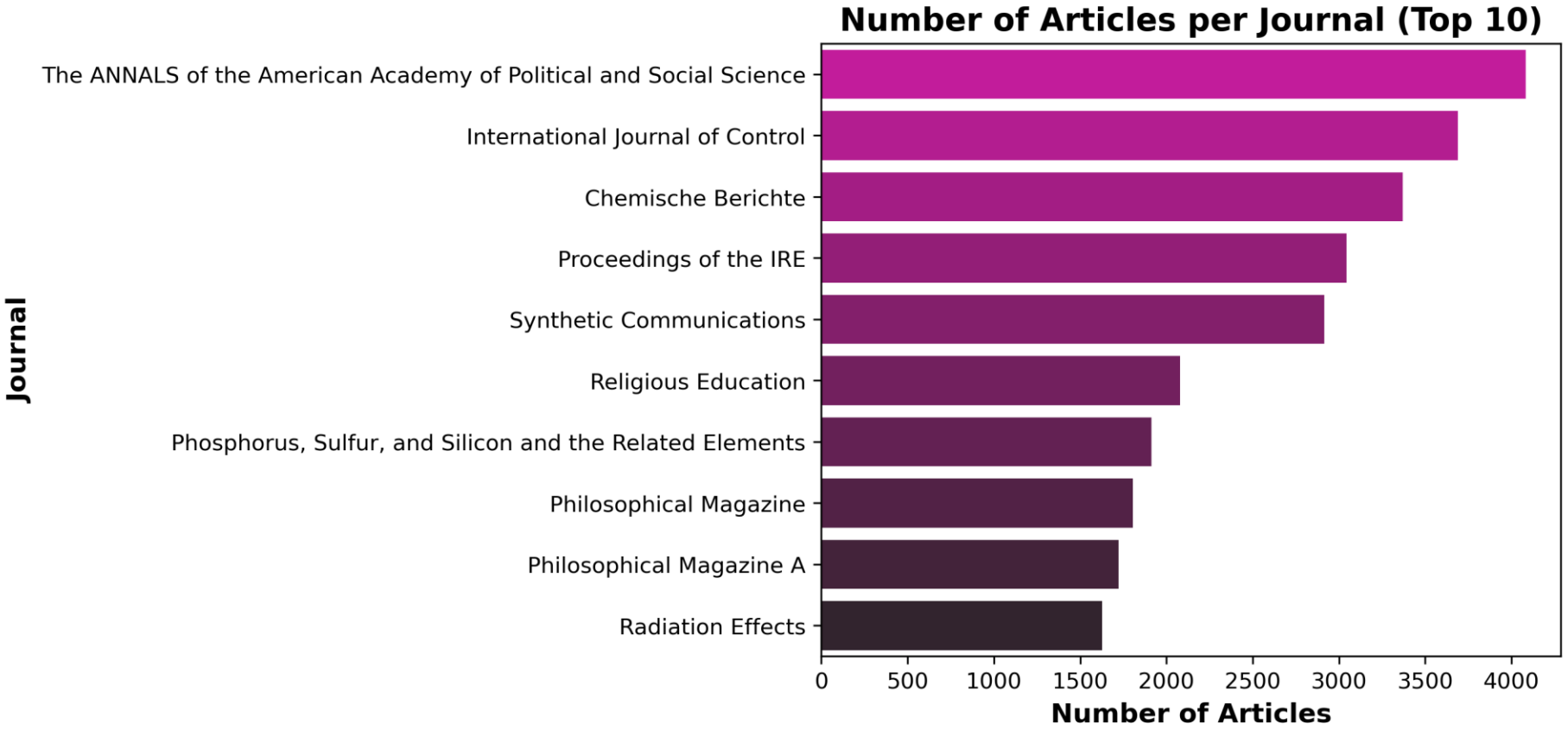
Among the 9 types of article, the most favored by far are journal article (94%,) followed by proceeding articles (3%), and book chapters (2%).



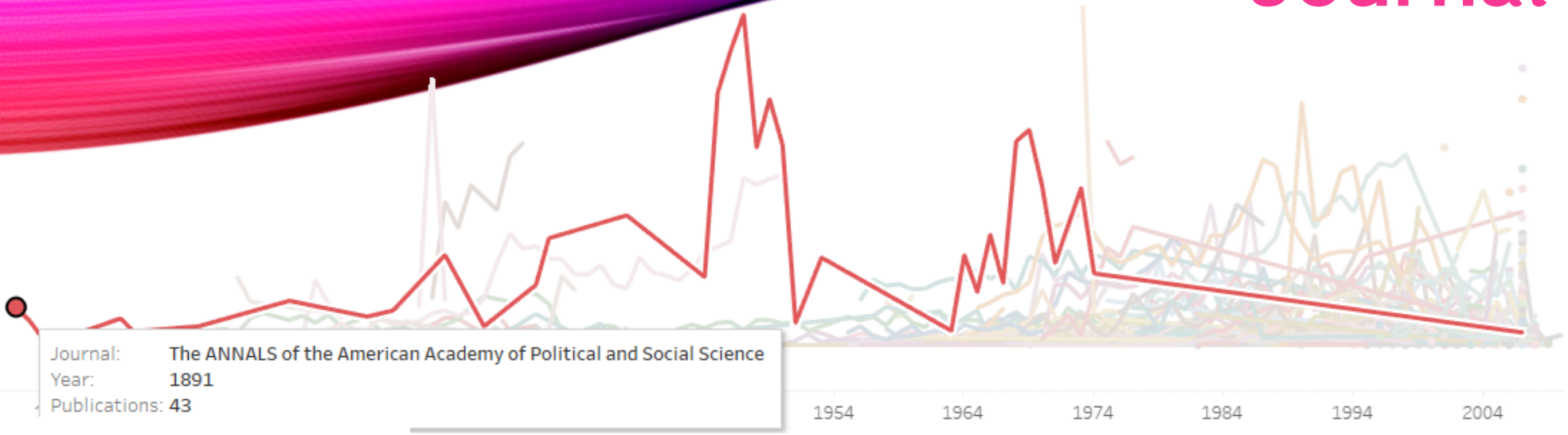
Journals

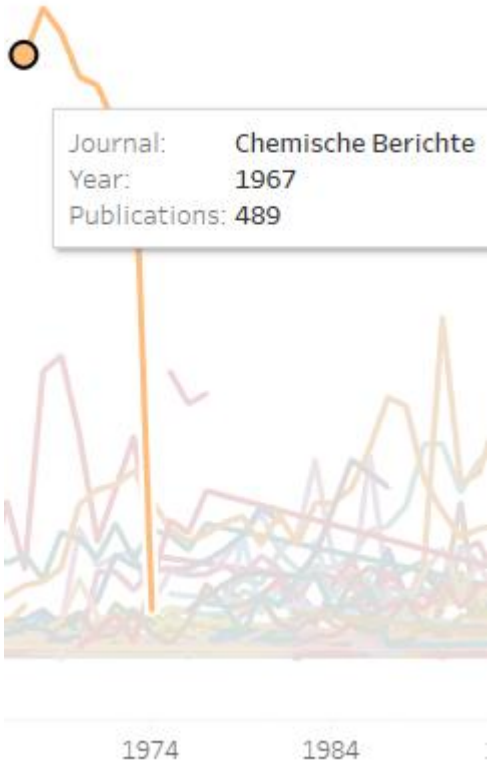
In which journals are most articles published?

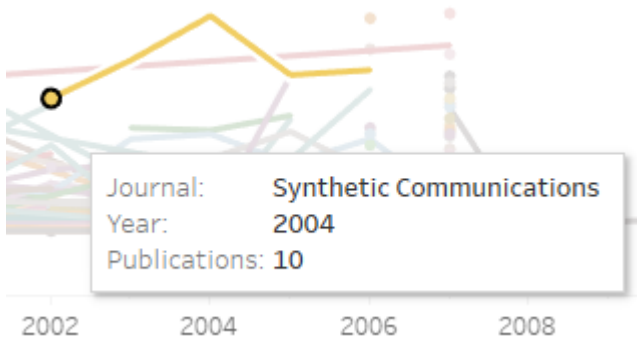
Among the 3,717 journals, the best publishing ones are The ANNALS of the American Academy of Political and Social Science (4082 publications), International Journal of Control (3689 publications), Chemische Berichte (3370 publications), and Proceedings of the IRE (3043 publications).
The top journal deals with sociology, social and political sciences.



Journal trends







How do journals evolve over time?

Some journals span the whole timeline (e.g. “The ANNALS of the American Academy of Political and Social Science”).

Others have appeared later and are durable (e.g. “Philosophical Magazine”).

Some are short-lived (e.g. “Chische Berichte” 1967-1974, “National Municipal Review” 1923-1930).

The journal “Synthetic Communications” emerged in 2004.

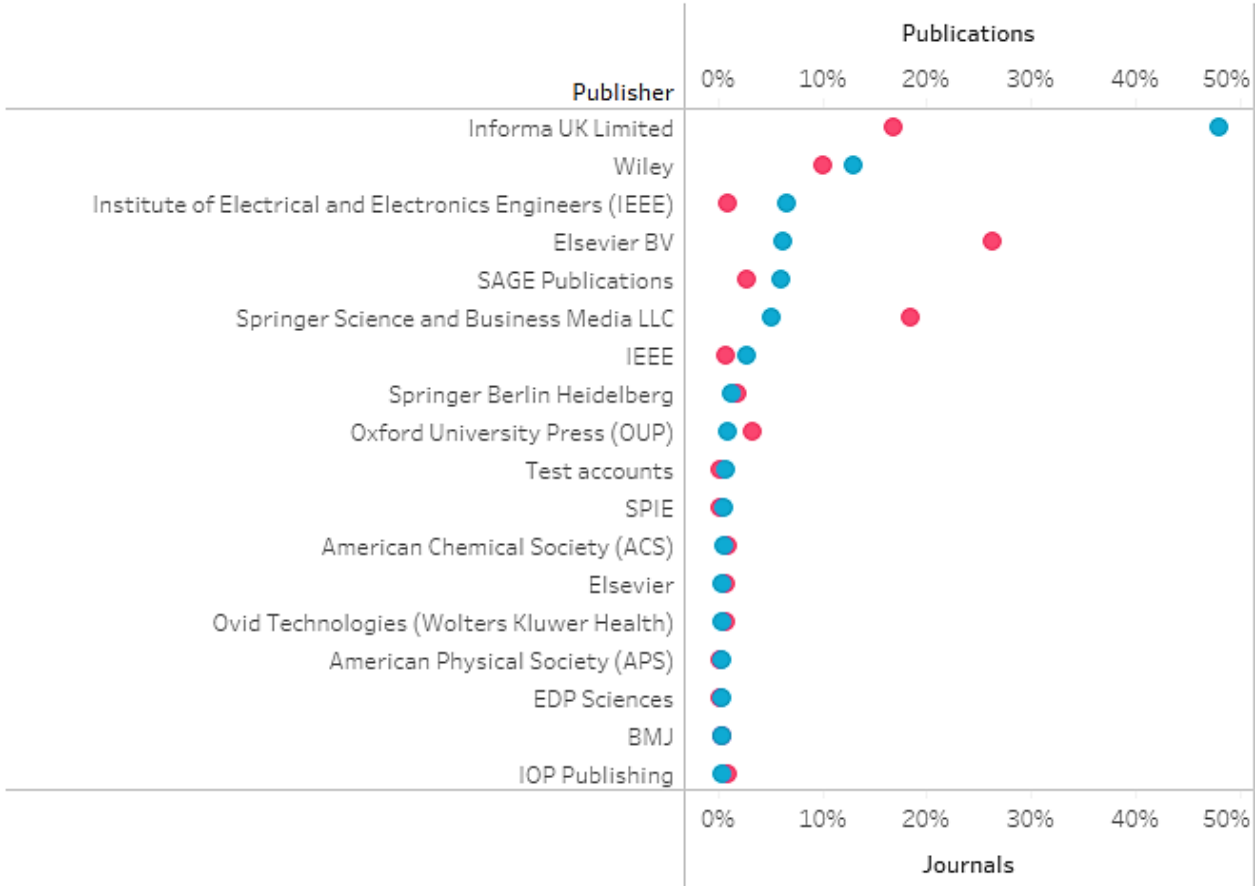
Publishers & Journals

Is there a relationship between publisher and journals??

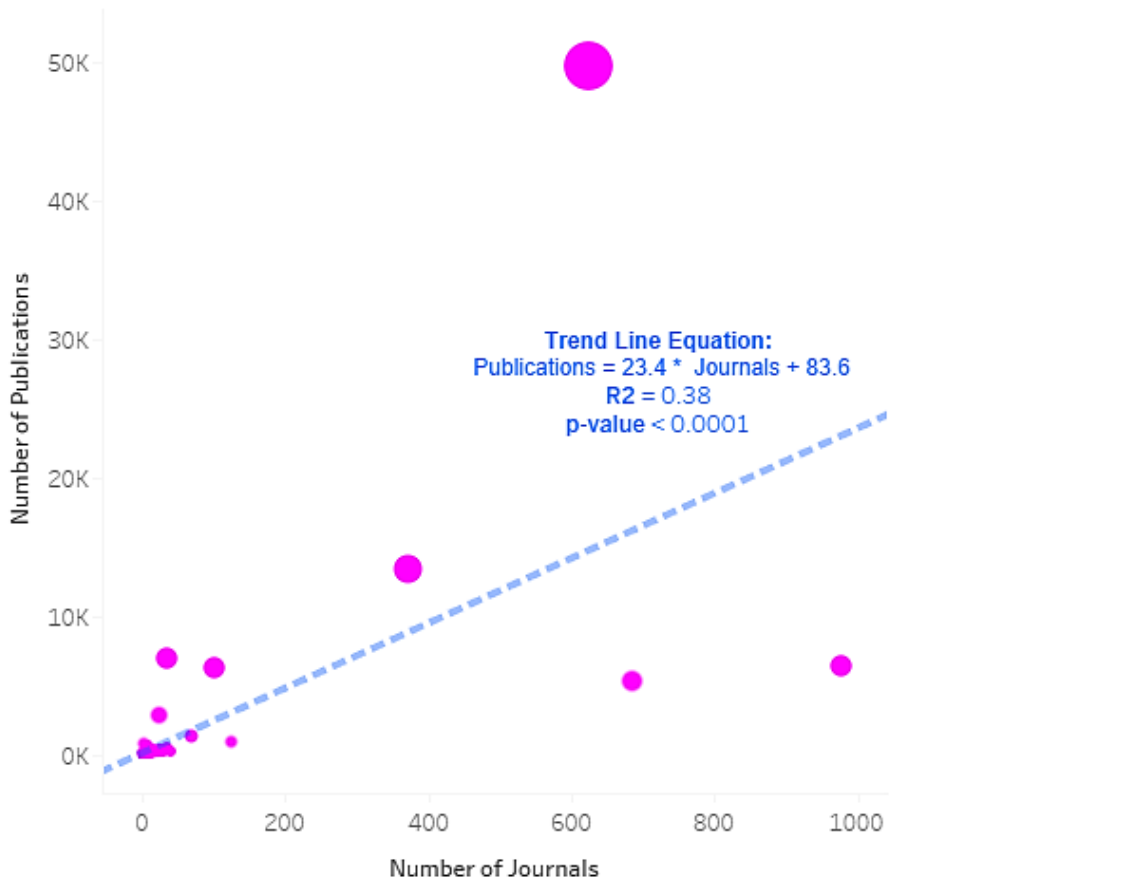
Among the 194 publishers, the most successful are Informa UK Limited (49628 publications, 48%) and Wiley (13324 publications, 13%). Publishers that host the largest number of journals are Elsevier BV (977 journals), Springer Science and Business Media LLC (685 journals), Informa UK Limited (623 journals), and Wiley (372 journals).

There is a positive relationship (R^2=0.4) between number of publications and number of journals per publisher.

Number of Journals and Publications per Publisher



Number of Publications and Journals per Publisher

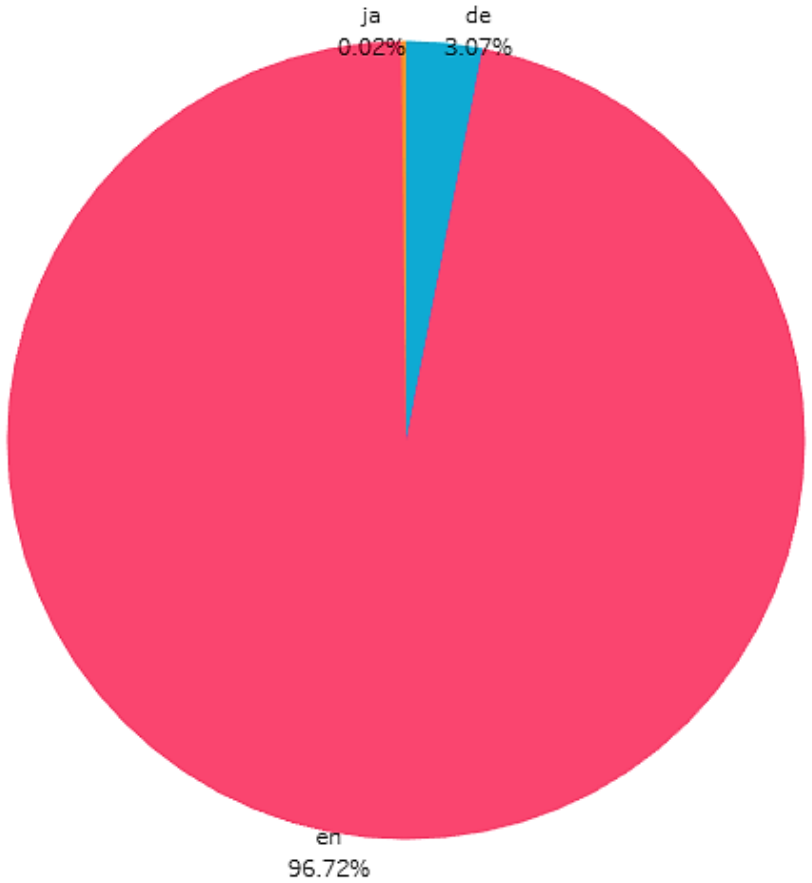


Language

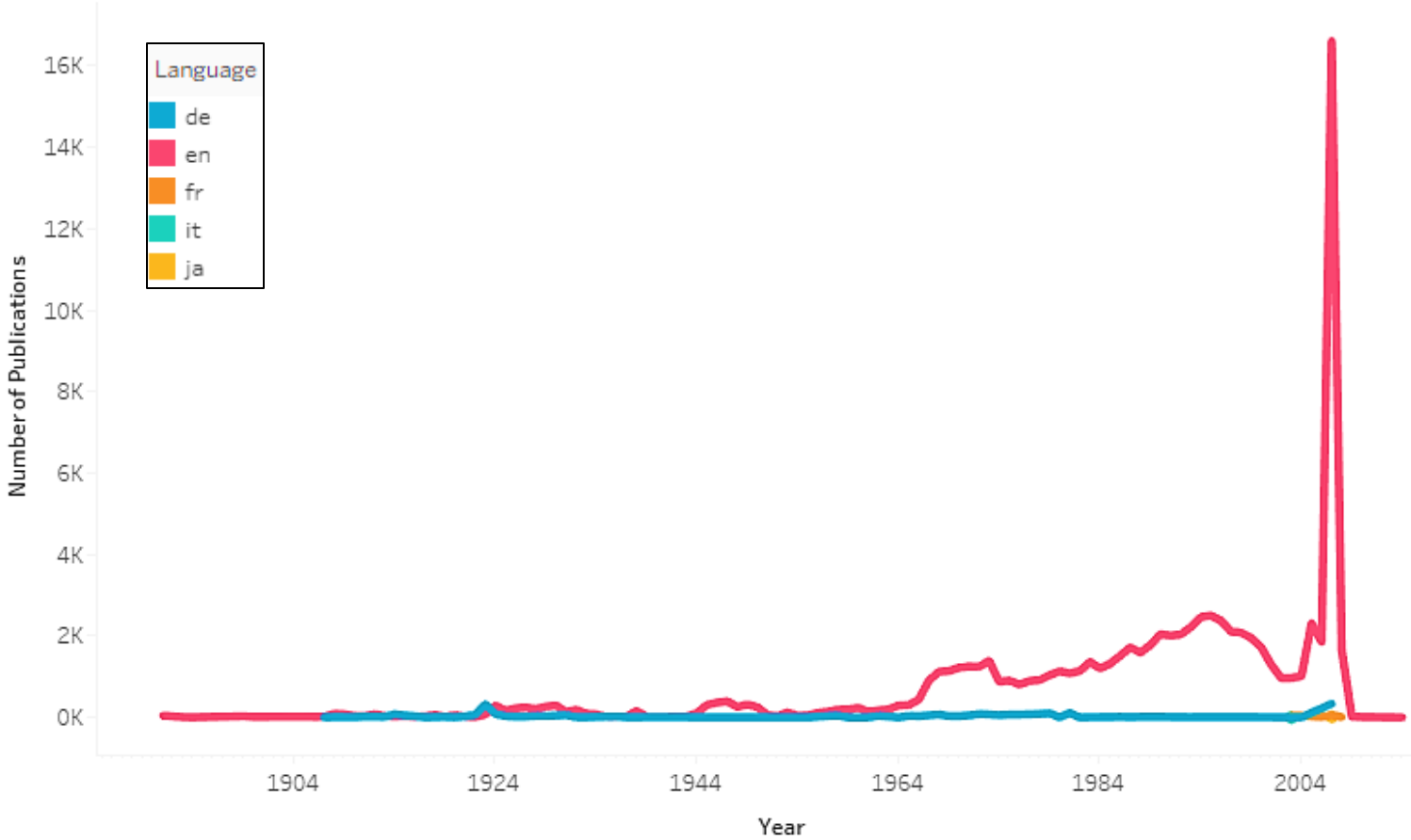
Which languages are used?

A total of 5 languages are listed: en, de, fr, it, ja.English is overwhelmingly (97%) the language used for publication purpose, followed by Deutch (3%). French is only used from 2003-2008, Italian in 2003 and Japanese in 2007.

Percentage of Publications per Language



Publications per Language from 1891 to 2014

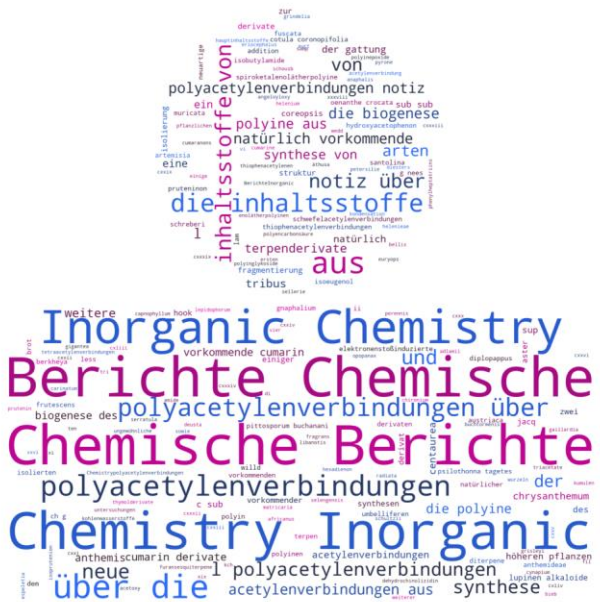


Prolific Authors

Which authors are the most prolific?

Among the 81,743 first authors, the most prolific ones are Ferdinand Bohlmann a chemist interested in inorganic chemistry with 124 publications, followed by D.A. Michalopoulos interested in Computer Science with 68 publications, John Bauer a municipal public servant with 60 publications, and D.J. Evans a mathematician with 54 publications.

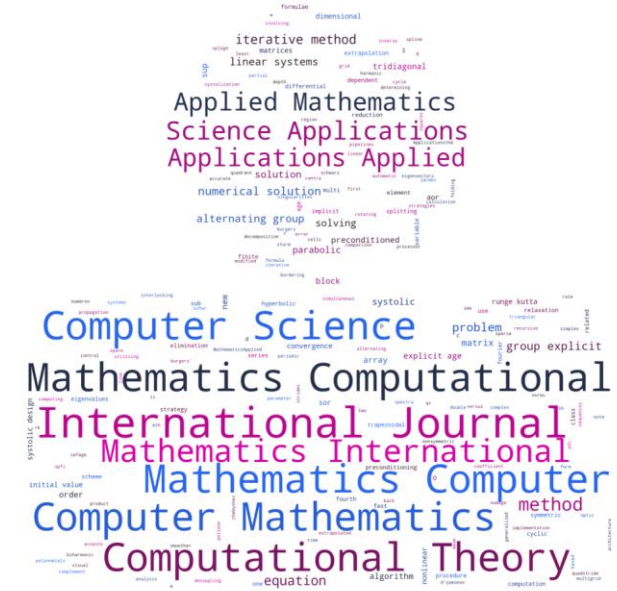
1. Ferdinand Bohlmann



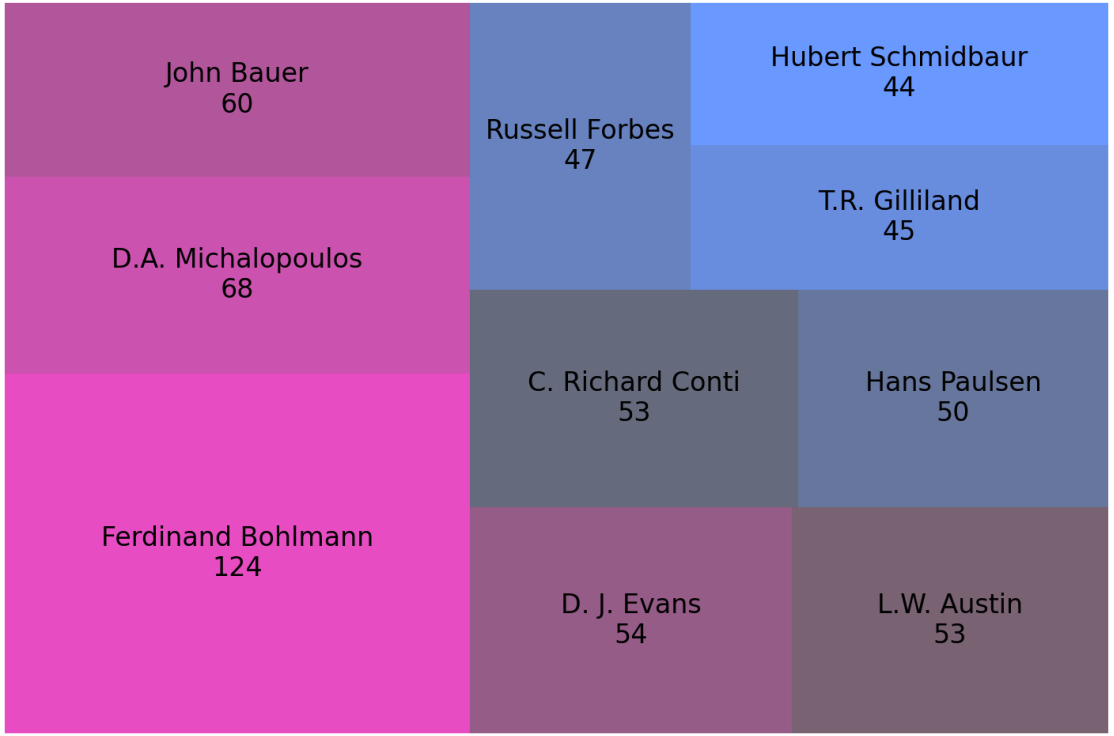
2. D.A. Michalopoulos

3. John Bauer

4. D.J. Evans



Number of Articles per Author (top 10)

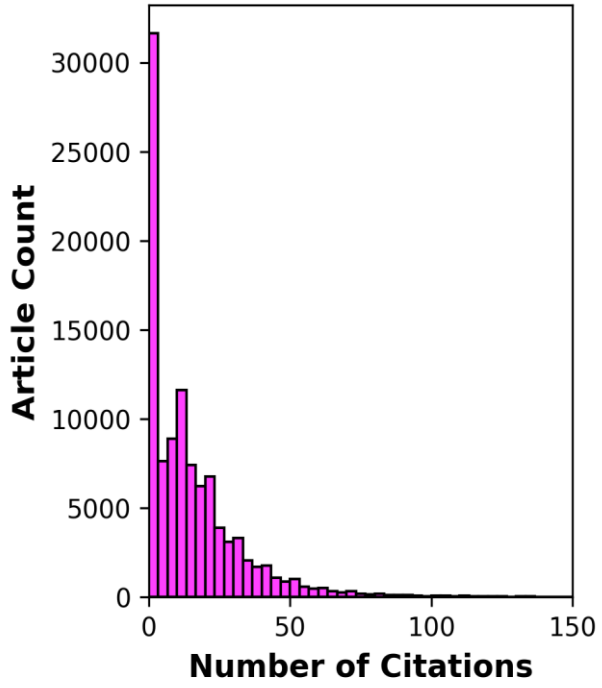


Citations

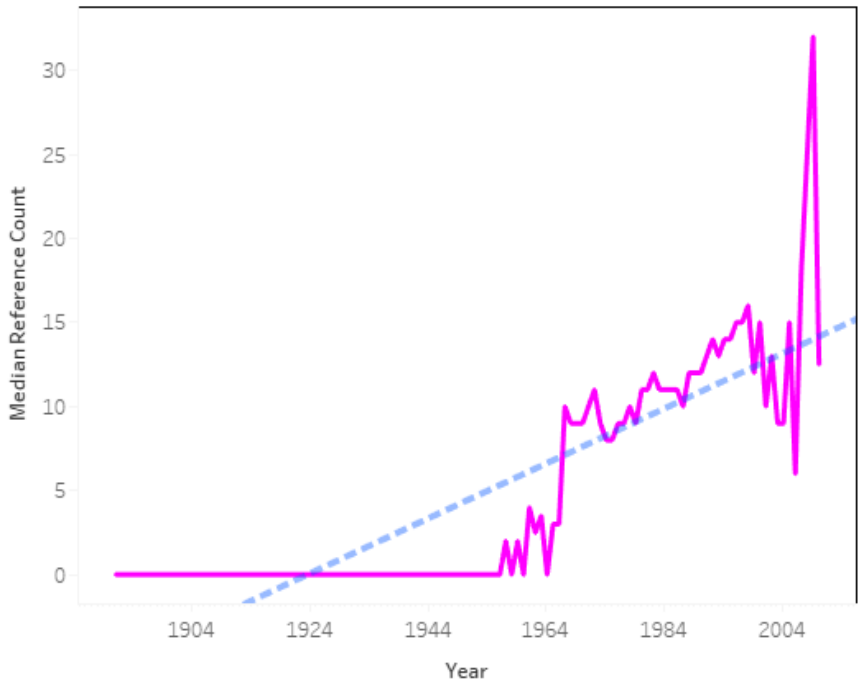
How do citation numbers track over time?

The distribution is highly right-skewed; 75% articles have < 22 references, with a few outliers featuring > 200 citations. The average number of references per article was almost non-existent from 1891-1955, bar 2 spikes in 1913-1916 (33 citations in 1914) and 1934 (5 citations). It steadily increased from 1956 to 2010 with a peak at 35 in 2009. Reference entries exhibits the widest range of citation numbers. Book chapters typically list more references than other publications.

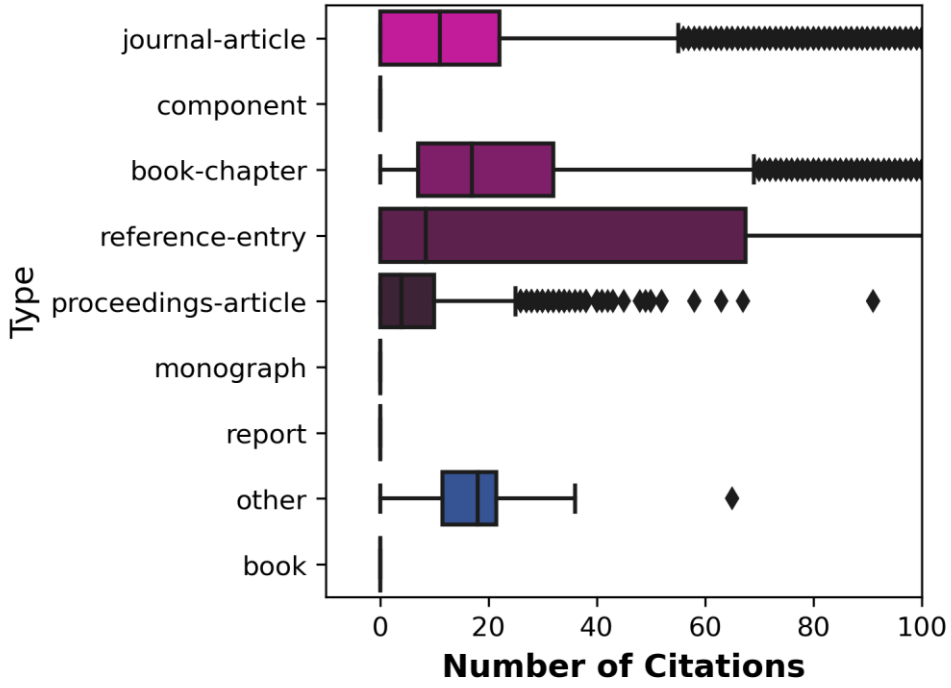
Distribution of citation numbers



Median Citation Number from 1891 to 2014



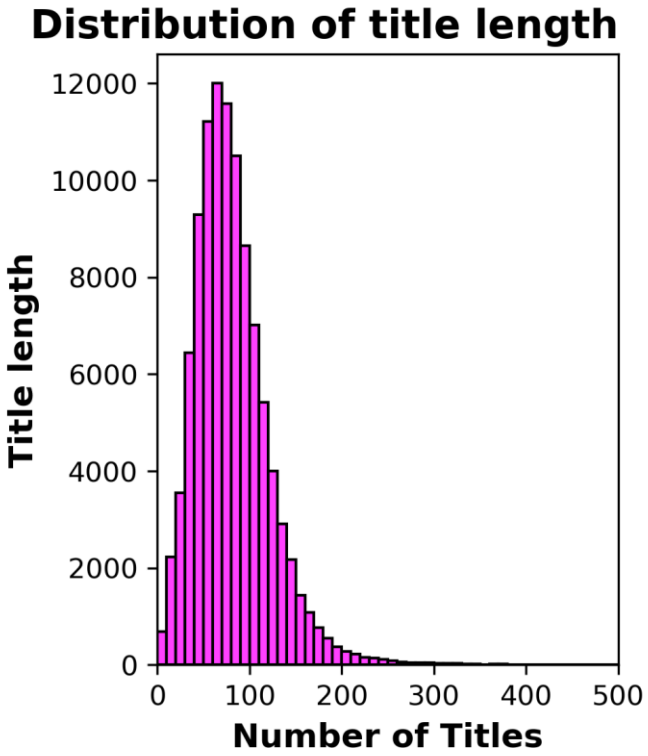
Distribution of citation numbers per type

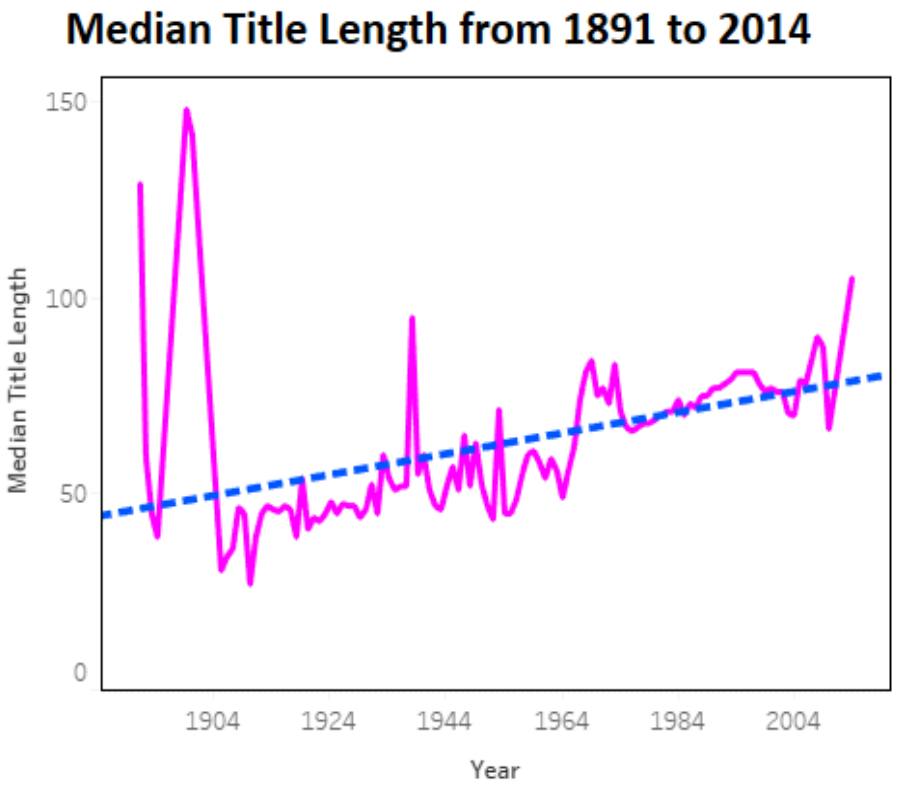


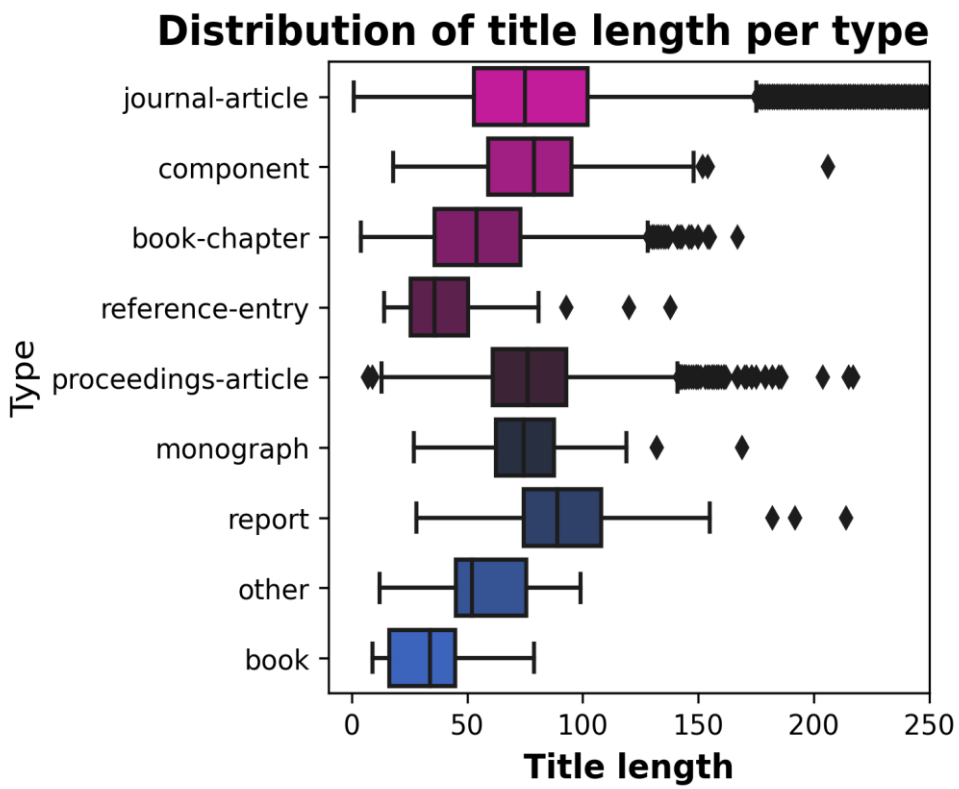
Titles

How do title lengths track over time?

The distribution is right-skewed; 75% articles have titles with < 101 characters, with a few outliers featuring > 400 characters. The average length of title is 80. While fluctuating, title lengths steadily increase from the beginning of the 20th century to 2014. Journal article entries exhibits the widest range of title length. Conversely reference entries and book feature the shortest titles.



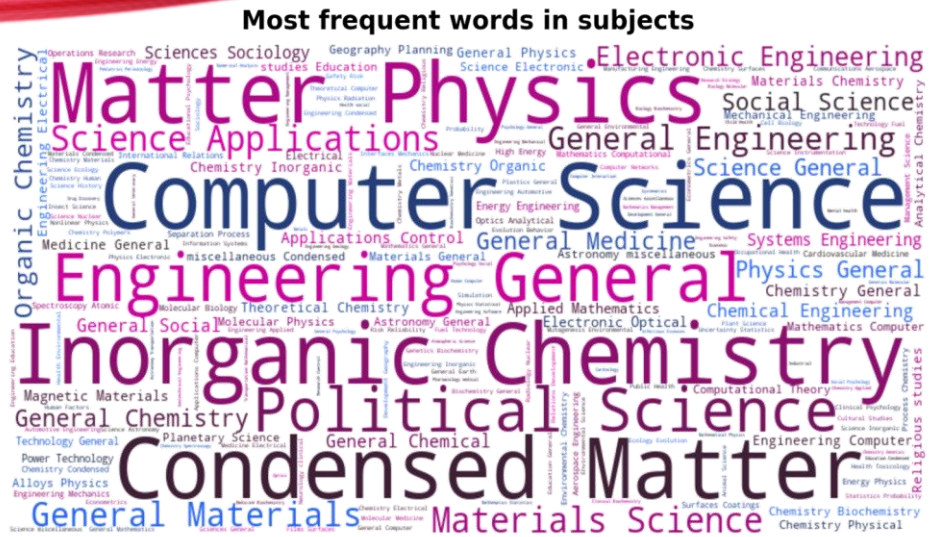




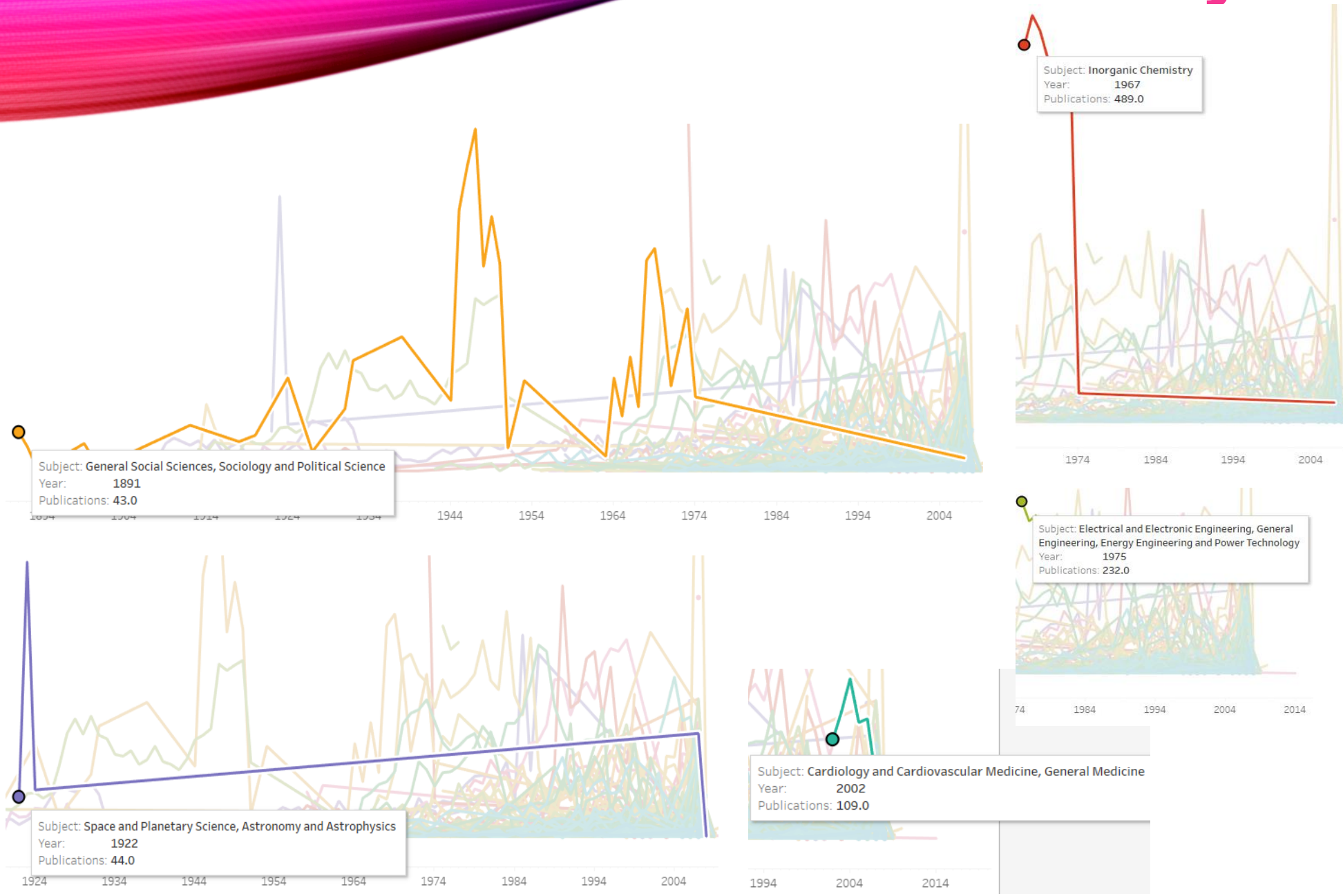
Subjects

What are popular subjects?

Among the 1265 subjects, the most popular ones deal with social sciences, computer science, electrical and general engineering as well as chemistry. Most common subject words reflect this as well as matter physics.



Subject trend



How do subjects evolve over time?

Some subjects span the whole timeline (e.g. “General Social Sciences, Sociology and Political Science”).

Others have appeared later and are durable (e.g. “Space and Planetary Science, Astronomy and Astrophysics” from 1922, “Inorganic Chemistry” from 1967).

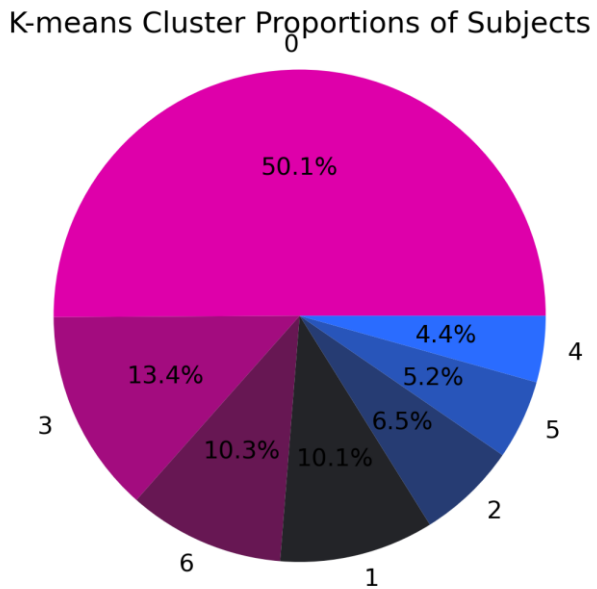
Some are short-lived (e.g. “Electrical and Electronic Engineering, General Engineering, Energy Engineering and Power Technology” 1975-1977).

The subject “Cardiology and Cardiovascular Medicine, General Medicine” emerged in 2002.

Categorising Subjects

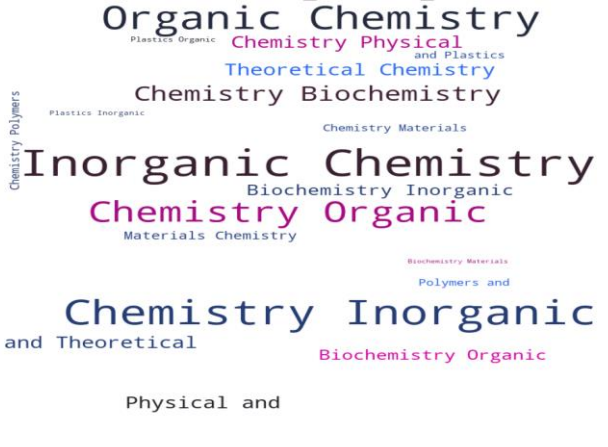
Can we group the 1265 subjects into broader categories?

Using text preprocessing, vectorization and k-means clustering (TfidfVectorizer() Normalizer() Kmeans() from sklearn), subjects were grouped into 7 clusters of varying specificity levels (e.g. cluster 0 = multidisciplinary, cluster 1 = highly specific to chemistry, cluster 2 = specific to computer science).

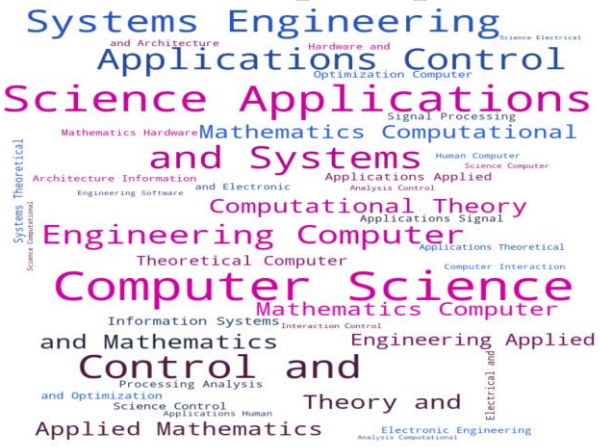


Wordcloud subject_cluster_kmeans 0

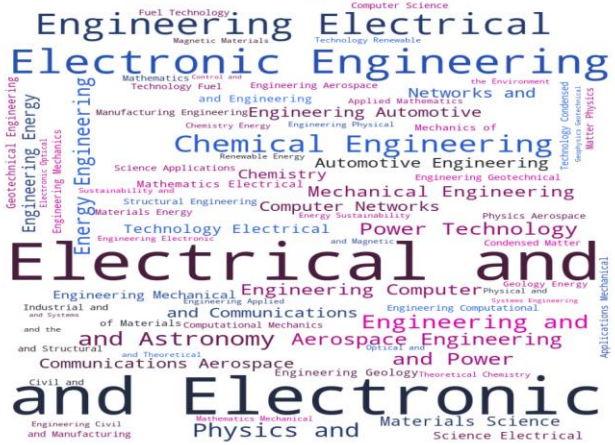
Wordcloud subject_cluster_kmeans 1



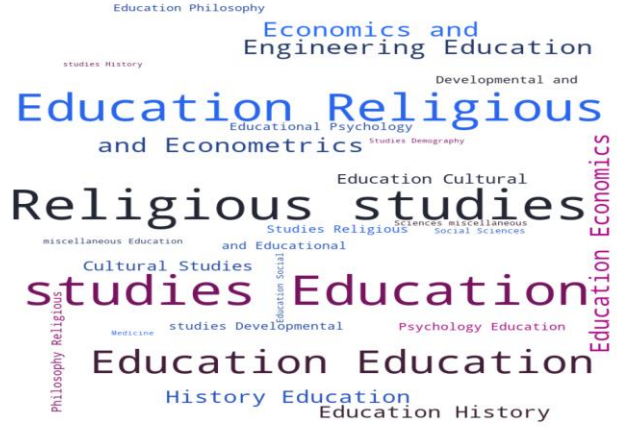
Wordcloud subject_cluster_kmeans 2



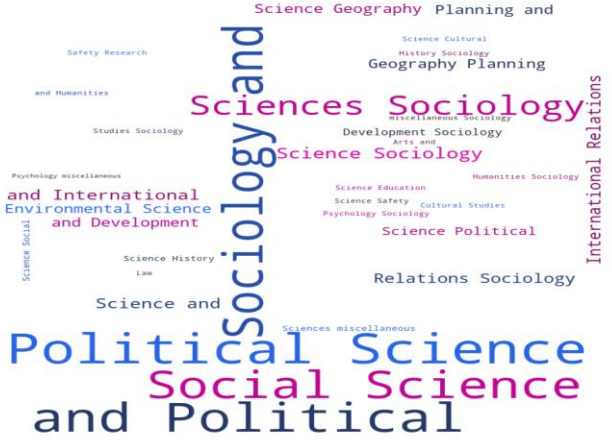
Wordcloud subject_cluster_kmeans 3



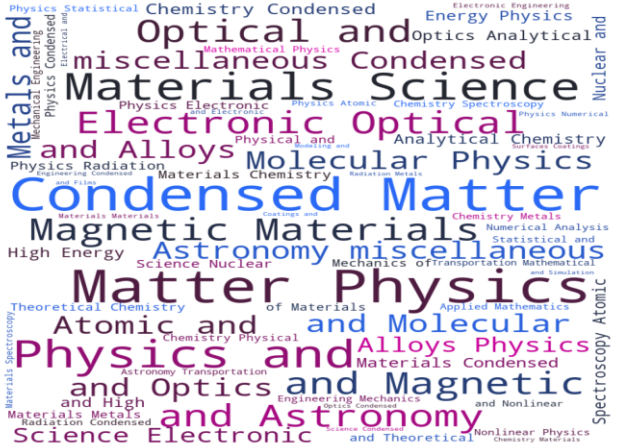
Wordcloud subject_cluster_kmeans 4



Wordcloud subject_cluster_kmeans 5



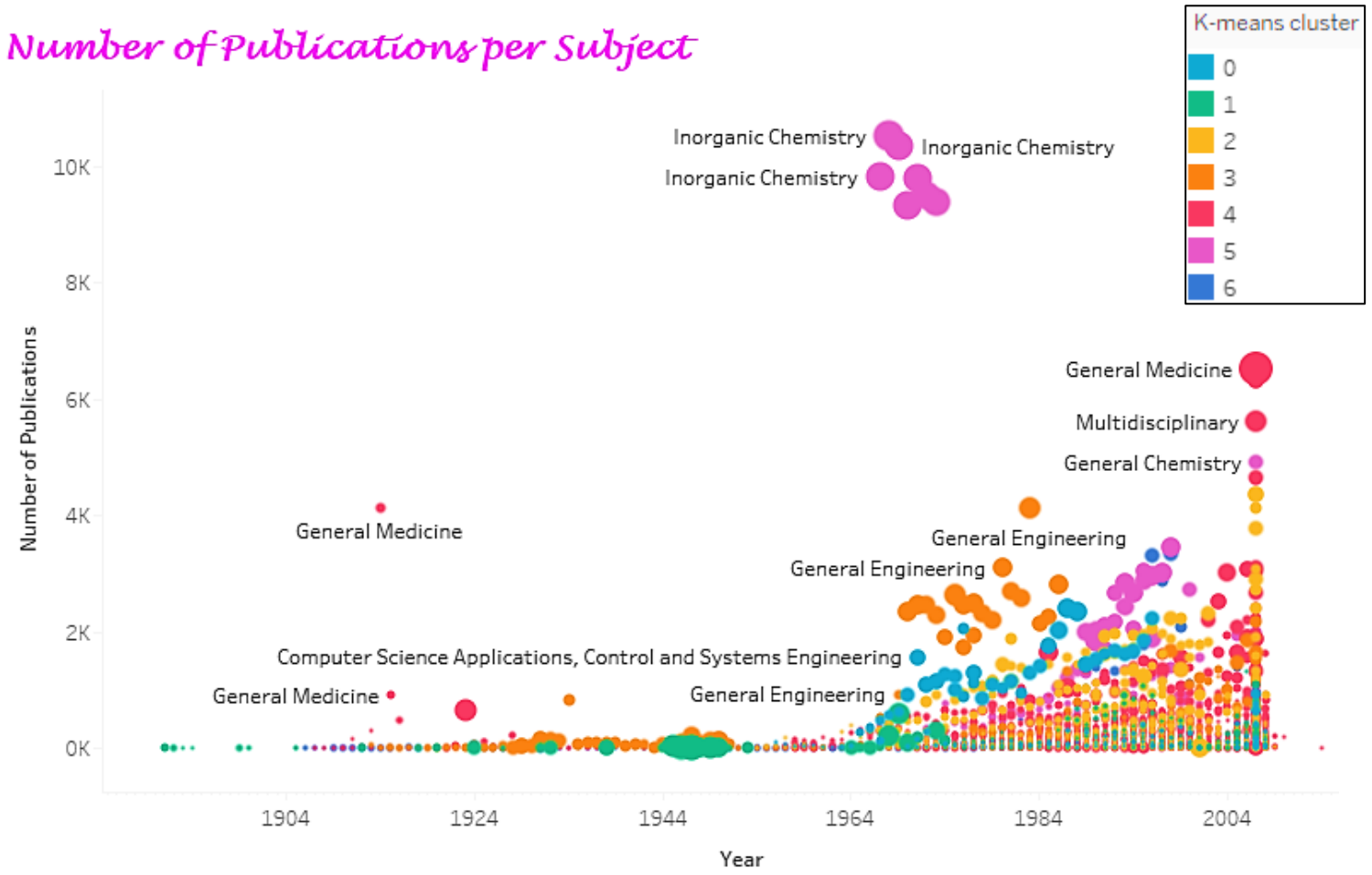
Wordcloud subject_cluster_kmeans 6

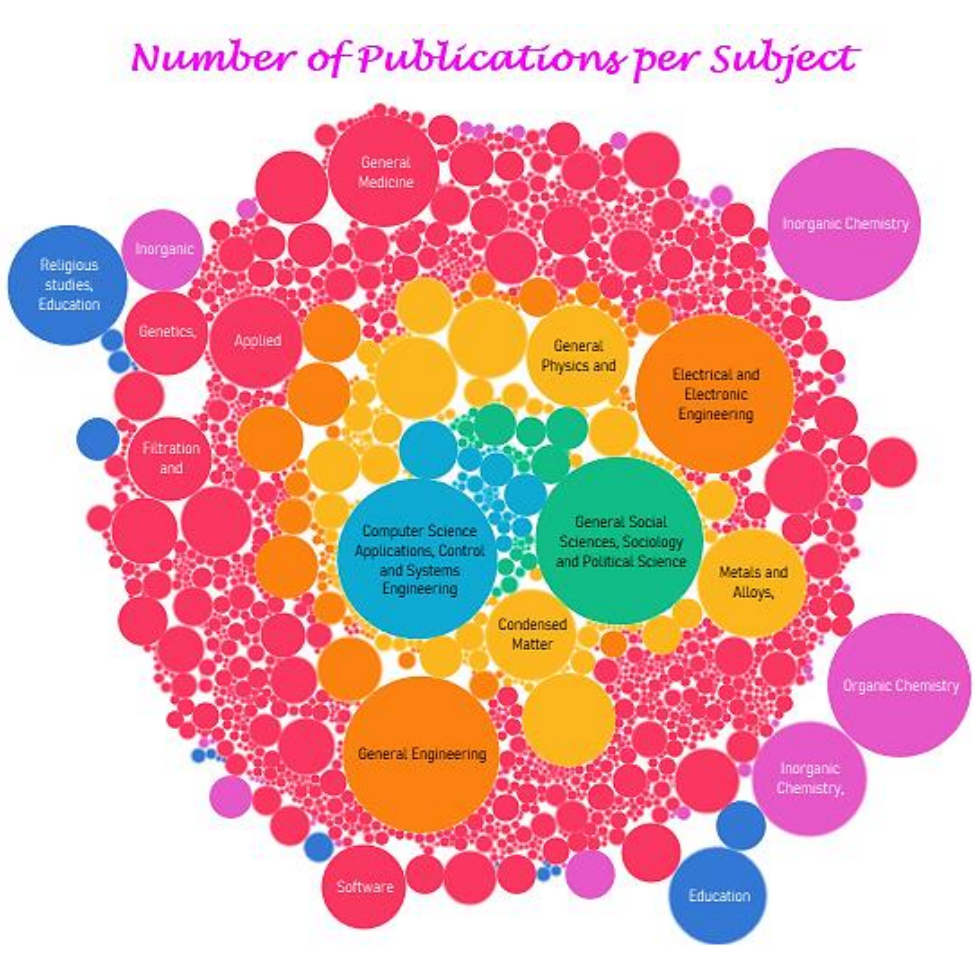


Categorising Subjects

What are some of k-means cluster trends?

Having less categories (7 instead of 1265) helps visualize this complex data. A timeline shows that chemistry (K-means cluster 5) became hugely popular in the 70's but were less fashionable later on. Computer science (K-means cluster 0) surged in the 60's, with increased interest until the late 90's, coincidently with general engineering (K-means cluster 3).

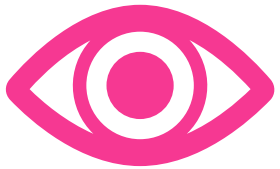


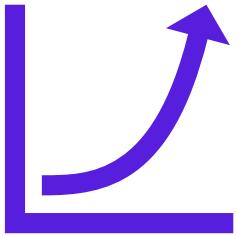


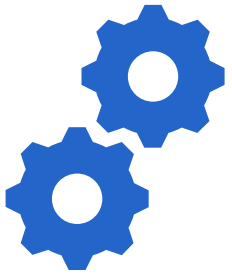
Conclusions

Conclusions

What did we learn?



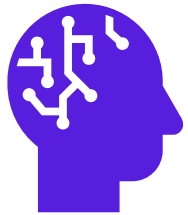




- Yearly rate of publication gradually increases from the 60's onward to reach 2,500 publications per year in the mid-90's
- 2007 featured many more publications than other years. Is it a dataset bias or did something happen to explain such spike?
- Most publications are journal article (94%,) followed by proceeding articles (3%), and book chapters (2%)
- The popularity and longevity of journals vary immensely
- Most successful publishers host a larger number of journals
- We covered examples of most prolific authors interested in different academic fields
- Publications are overwhelmingly written in English
- The number of citations as well as the title length are both on the rise
- The most popular subjects deal with social sciences, computer science, electrical and general engineering as well as chemistry
- Some subjects are long-lasting while others are short-lived or emerging
- As no broad subject categories were available, we made our own using text-mining and clustering which helped further analyse trends

Future Directions

Future Directions



What could we do next?

- Create more sophisticated visualization (e.g. Sankey or chord diagrams) to better explore relationship across variables
- Refine the text-mining models using other LNP methods
- Fetch more metadata:
  - co-authors' names to establish scientist networks
  - UNESCO academics field of knowledge to properly categorise the data
  - More numerical variables (how many times the publication was cited, submission/acceptation/publication dates to track how fast reports are processed and published)
  - Keywords for easier text-mining analysis
  - All the abstracts for more in-depth text mining
  - More detailed publication types for articles (reviews, short communication, technical note, etc…)
  - Authors' geographical area (country, region, town) to map where research hotspots are

Links

Links to files

The following files are publicly available:

- Raw dataset: scientific_papers_metadata_20240709.csv (GitHub)
- Python code: Scientific_Articles_Metadata_dlf2024.ipynb (GitHub)
- Report: scientific_papers_metadata_dlf2024.pdf (GitHub)
- Tableau data visualization: scientific_papers_metadata_dlf2024.twb (GitHub and Tableau Public)

Repositories:

- GitHub: https://github.com/dlf2024?tab=repositories
- Tableau Public: https://public.tableau.com/app/profile/delphine.vincent/vizzes