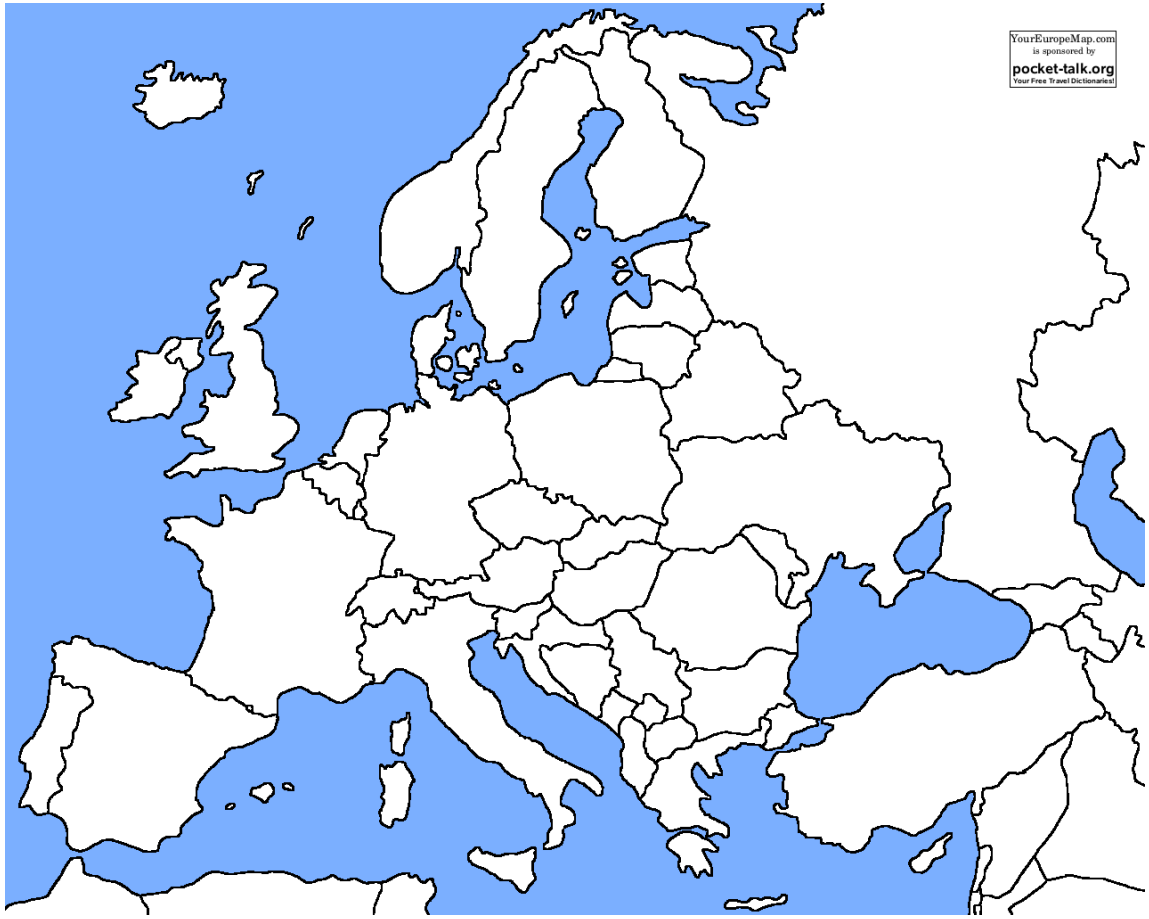
Biblium: A Python-Based Framework for Advanced Bibliometric Analysis

Authors:
Lan Umek, Dejan Ravšelj
Presenter:
Lan Umek
lan.umek@fu.uni-lj.si

BRIC 2025 Conference
Montreal, Canada – June 5th 2025

Where is Slovenia?



Some data about Slovenia

- > 2.1 M inhabitants
- Largest city: Ljubljana (282,994 inhabitants)
- Area: 20,273 km^2
- Slovenian language (slovenščina)
- Dual form (between singular and plural)

Landscape of Slovenia

University of Ljubljana

- established in 1919
- 40,000 students
- 300 study programs
- 23 faculties, 3 arts academies
- 6,000 higher education teachers, researchers, assistants and administrative staff
- listed amongst the top 500 universities in many lists

Faculty of Public Administraion

- one of the youngest members of the University of Ljubljana as a fully faculty
- established in 1956 as School of Public Administration
- since 2003 the Faculty of (Public) Administration
- 3 undergraduate and 2 master study programs, 1 PhD programe
- app. 1,000 students
- 70 higher education teachers, researchers, assistants and administrative staff
- EAPAA accreditation - European Association for Public Administration Accreditation

Motivation

- build a comprehensive Python library for bibliometric analysis that reflects and extends the capabilities of R library bibliometrix
- develop and implement methods for (sub)group analysis in bibliometric research
- link prediction models (classification, regression) to bibliometric analysis

About biblum

- Available at (from June 1st)

https://github.com/lan-umek/biblium

- Main classes:
  - BiblioStats, BiblioAnalysis
  - BiblioGroup, BiblioGroupAnalysis
  - BiblioGroupClassifier
- Special classes:
  - ScienceGroupAnalysis, AreaGroupAnalysis, FiledGroupAnalysis
  - SDGGroupAnalysisGoals, SDGGroupAnalysisPerspective, SDGGroupAnalysisDimension

Python libraries used

- Standard Library
  - os
  - re
  - datetime
  - math
  - collections
  - itertools
  - textwrap

Python libraries used

- Third-Party

- pandas
- numpy
- openpyxl
- scipy
- sklearn
- nltk
- sentence_transformers
- networkx
- matplotlib
- seaborn
- wordcloud
- squarify
- venn
- upsetplot
- holoviews
- pptx
- docx

Bibliometric Analysis

- Initialization parameters

```
In [1]: import biblium as bb
   ...: ba = bb.BiblioAnalysis(f_name="dataset.csv", db="scopus", res_folder="results",
   ...:                        preprocess_level=0, exclude_list_kw=None,
   ...:                        synonyms_kw=None, lemmatize_kw=False, default_keywords="author",
   ...:                        asjc_map_df=None, lang_of_docs="en", fancy_output=False, label_docs=True,
   ...:                        dpi=600, cmap="viridis", default_color="lightblue")
```

- Simpler version

```
In [1]: import biblium as bb
   ...: ba = bb.BiblioAnalysis(f_name="dataset.csv", db="scopus")
```

Dataset for demonstration

- Top 200 documents (in terms of global citations) from Scopus related to the query:

TITLE-ABS-KEY(bibliometric*)

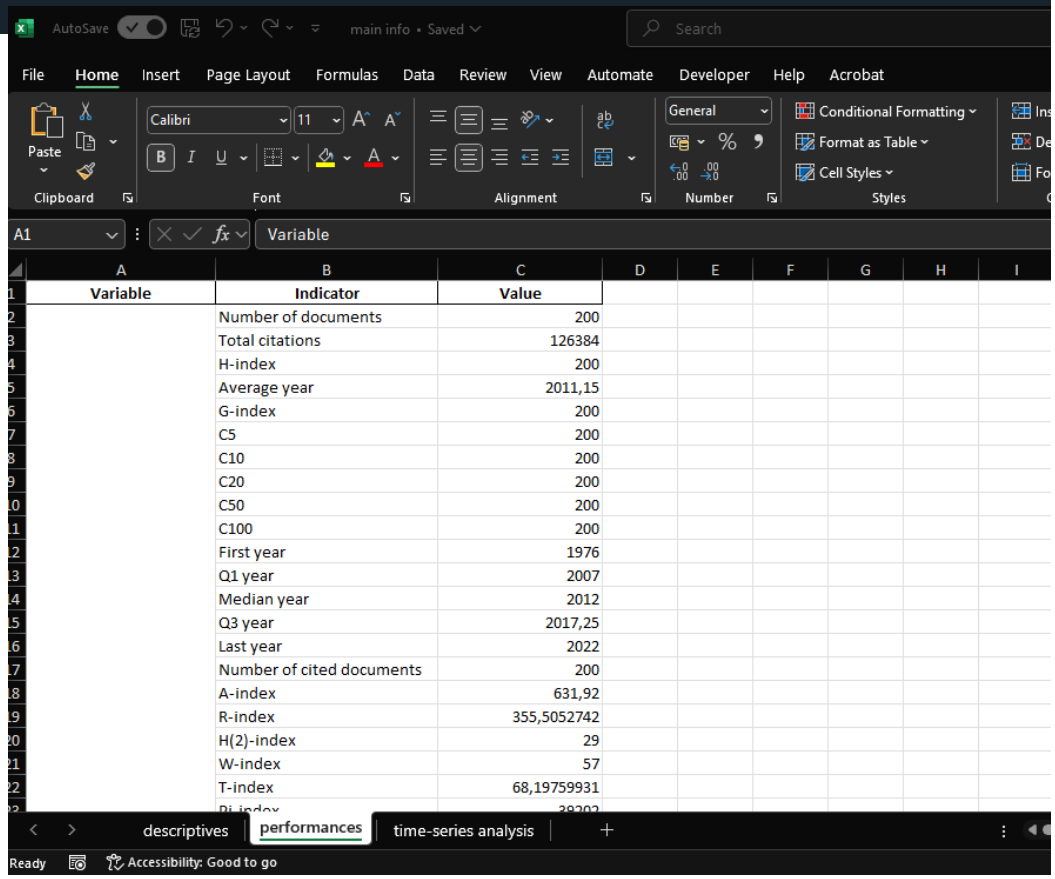
Main info

- Calling

```
In [2]: ba.get_main_info(include=["descriptives", "performance", "time series"], performance_mode="full", stopwords=None,
excluded_sources_references=None)
```

- Result

in xlsx format



	A	B	C
1	Variable	Indicator	Value
2		Number of documents	200
3		Total citations	126384
4		H-index	200
5		Average year	2011,15
6		G-index	200
7		C5	200
8		C10	200
9		C20	200
10		C50	200
11		C100	200
12		First year	1976
13		Q1 year	2007
14		Median year	2012
15		Q3 year	2017,25
16		Last year	2022
17		Number of cited documents	200
18		A-index	631,92
19		R-index	355,5052742
20		H(2)-index	29
21		W-index	57
22		T-index	68,19759931

Preprocessing of variables

- Processing of keywords

```
In [3]: ba.process_keywords(exclude_list=["research", "impact"], synonyms={"bibliometric analysis": "bibliometrics"},
lemmatize=False)
```

- Processing of title and abstract

```
In [4]: ba.process_text_vars(stopwords_file=None, lang="en", remove_numbers=True, remove_two_letter_words=True)
```

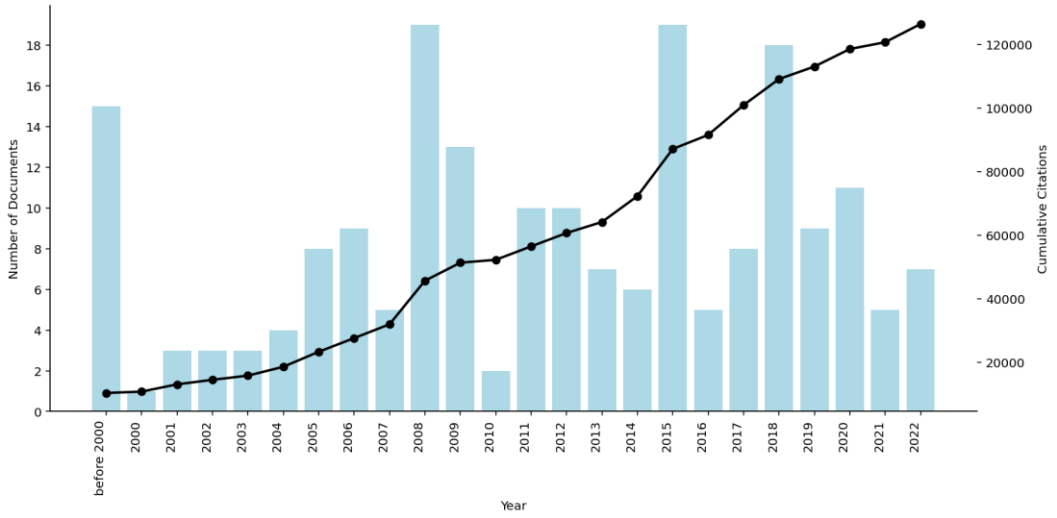
Scientific production

- Computation of production (number of documents, total number of citations)

```
In [2]: ba.get_production(relative_counts=True, cumulative=True, predict_last_year=False)
```

- Plotting

```
In [4]: ba.plot_scientific_production(cut_year=2000)
```



Counting occurrences

- Initialization parameters

```
In [5]: ba.count_sources() # single type
   ...: ba.count_author_keywords() # list type
   ...: ba.count_ngrams_abstract(ngram_range=(1, 2)) # text type
```

- Result

	Word - Phrase	...	Number of occurrences
15408	research	...	352
805	analysis	...	218
17650	study	...	266
15729	result	...	114
9750	journal	...	286
...	...	...	...
12	1960s term	...	1
11	1960s approximately	...	1
9	1940s schlesinger	...	1
8	1940s	...	1
7	1930s evolved	...	1

[20309 rows x 5 columns]

Performance indicators

- Initialization parameters

```
In [6]: ba.get_sources_stats(items_of_interest=None, exclude_items=None, top_n=20,
   ...: counts_df=None, count_method=None,
   ...: regex_include=None, regex_exclude=None,
   ...: value_type="string", indicators=False,
   ...: missing_as_zero=False, mode="full")
```

- Result (automatically saved to xlsx format)

```
In [7]: ba.sources_stats_df.columns
Out[7]:
Index(['Source', 'Number of documents', 'Total citations', 'H-index',
       'Average year', 'G-index', 'C5', 'C10', 'C20', 'C50', 'C100',
       'First year', 'Q1 year', 'Median year', 'Q3 year', 'Last year',
       'Number of cited documents', 'A-index', 'R-index', 'H(2)-index',
       'W-index', 'T-index', 'Pi-index', 'Gini index', 'HG-index', 'Chi-index',
       'Tapered H-index'],
      dtype='object')
```

LLM description

- Initialization parameters

```
In [7]: ba.llm_describe_df(ba.sources_counts_df.head(5), provider="openai", model="gpt-3.5-turbo", prompt="Describe the dataframe
in one paragraph (with numbers and top sources).")
```

- Result

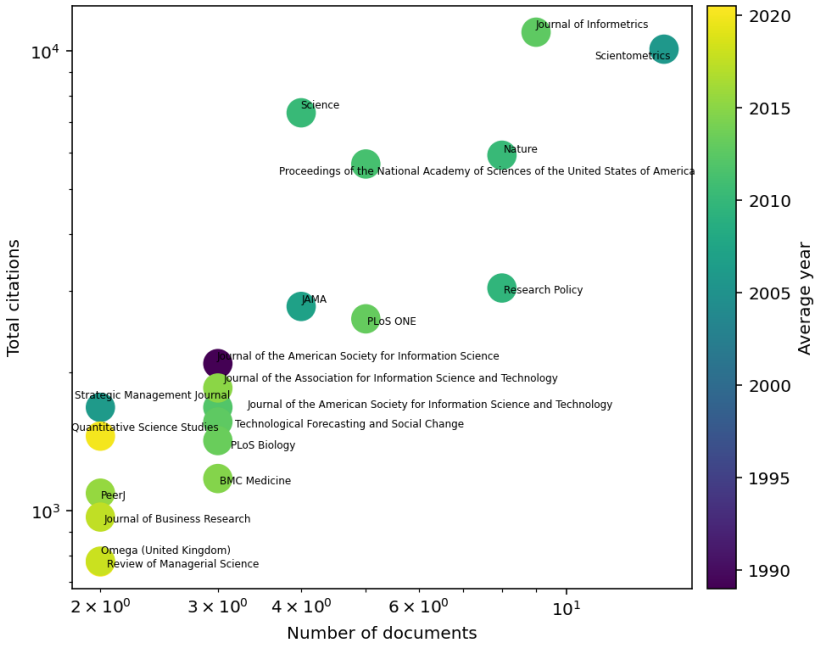
The collection is dominated by a handful of key outlets: Scientometrics contributes the most documents with 14 (7 %), followed by the Journal of Informetrics with 9 (4.5 %). Both Nature and Research Policy each account for 8 documents (4 % apiece), while the Proceedings of the National Academy of Sciences rounds out the top five with 5 documents (2.5 %). Together, these five sources represent nearly a quarter of the total document set.

Plotting top items

- Example of scatter plot

```
In [8]: ba.plot_top_items_multi(["sources"], kind="scatter",
   ...: y="Total citations", color_col="Average year",
   ...: label_col="Source") # many **kwargs possible
   ...: # different possibilites for different kinds
   ...: # (scatter, barh, lollipop)
```

- Result



Co-occurrences

- Construction of network

```
In [12]: ba.get_author_keyword_cooccurrence(items_of_interest=None, top_n=20, normalization=False, network=True,
    ...: top_items_df=None, top_items_col=None, count_func=None, count_attr=None,
    ...: output_attr_prefix=None, value_type="list", separator="; ")
```

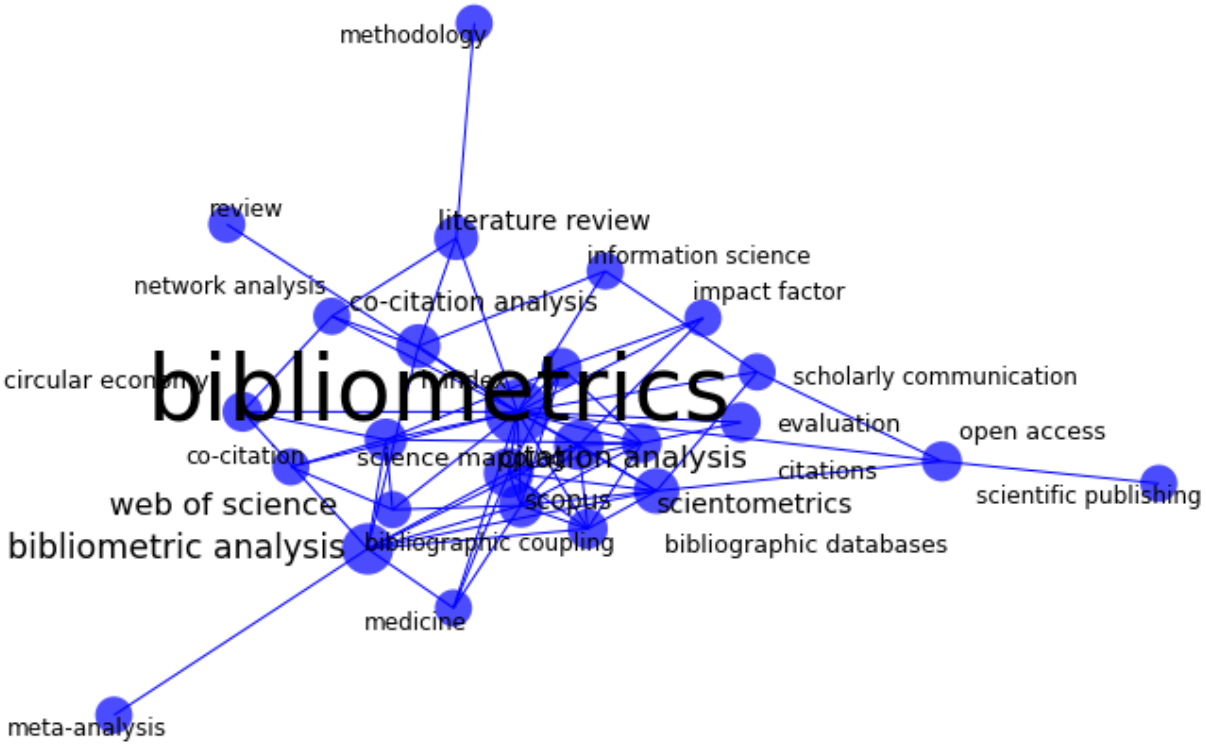
- Setting partitions

```
In [5]: ba.add_partitions_to_network(louvain_kwargs=None, greedy_kwargs=None, label_kwargs=None,
   ...: girvan_kwargs=None, k_clique_kwargs=None, kernighan_kwargs=None, walktrap_kwargs=None,
   ...: edge_betweenness_kwargs=None, infomap_kwargs=None, leading_kwargs=None, leiden_kwargs=None,
   ...: spinglass_kwargs=None)
```

- Setting vectors

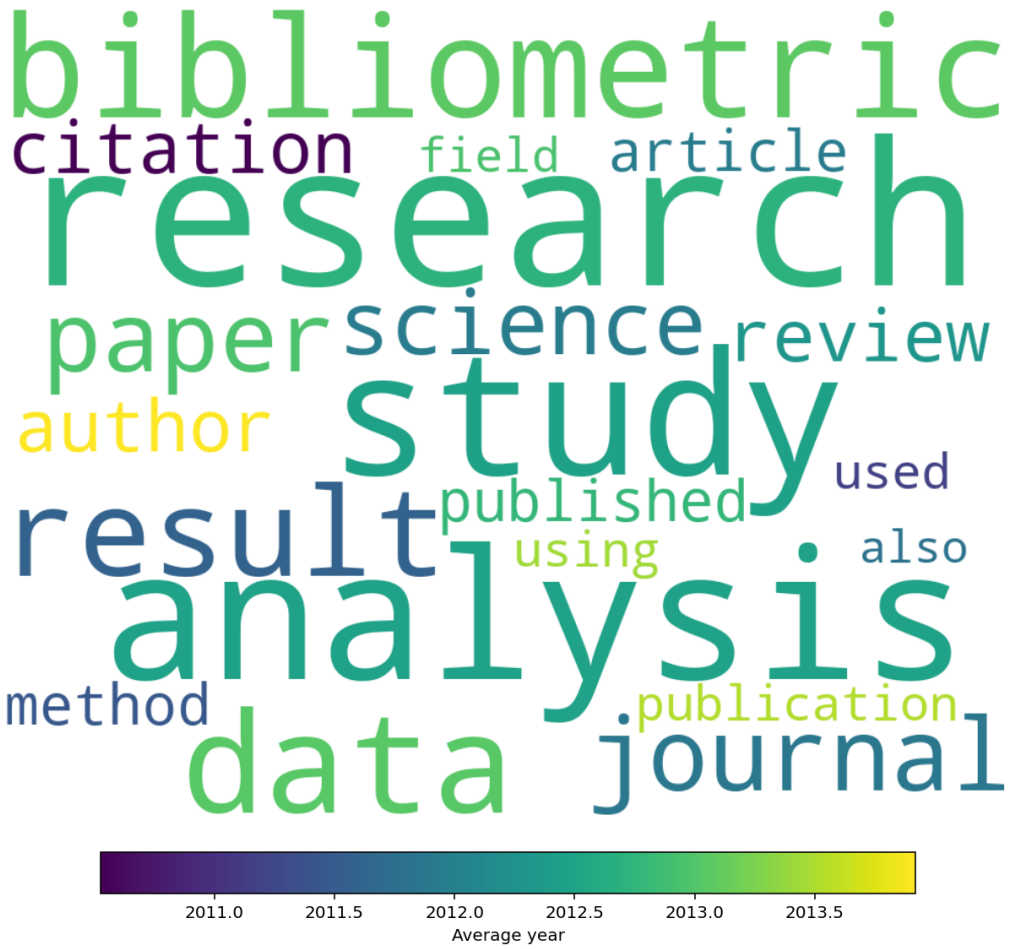
```
In [6]: ba.add_vectors_from_dataframe(["Average year", "H-index"])
```

Keyword co-occurrence network



Wordcloud

```
In [47]: ba.visualize_text("words from abstract",color_by="Average year")
```



Advanced features

- Topic modelling
- Factorial approach
- Sentiment analysis
- Bibliometric laws: Lotka, Bradford, Zipff
- K-field plot (extension of Sakey diagram 3-field plot)
- Citation network of documents with main path analysis

Bibliometric group analysis

- Initialization parameters

```
In [7]: bg.BiblioGroupAnalysis(f_name=None, db="", df=None, group_desc=None,
   ...: res_folder="results-groups", preprocess_level=0, exclude_list_kw=None,
   ...: synonyms_kw=None, lemmatize_kw=False, default_keywords="author",
   ...: lang_of_docs="en", fancy_output=False, label_docs=True,
   ...: group_colors=None, **kwargs)
```

- Main difference: group_desc parameter
  - a categorical variable
  - „list-like“ variable
  - binary dataframe
  - Year (with n_perdiods or cut_points)

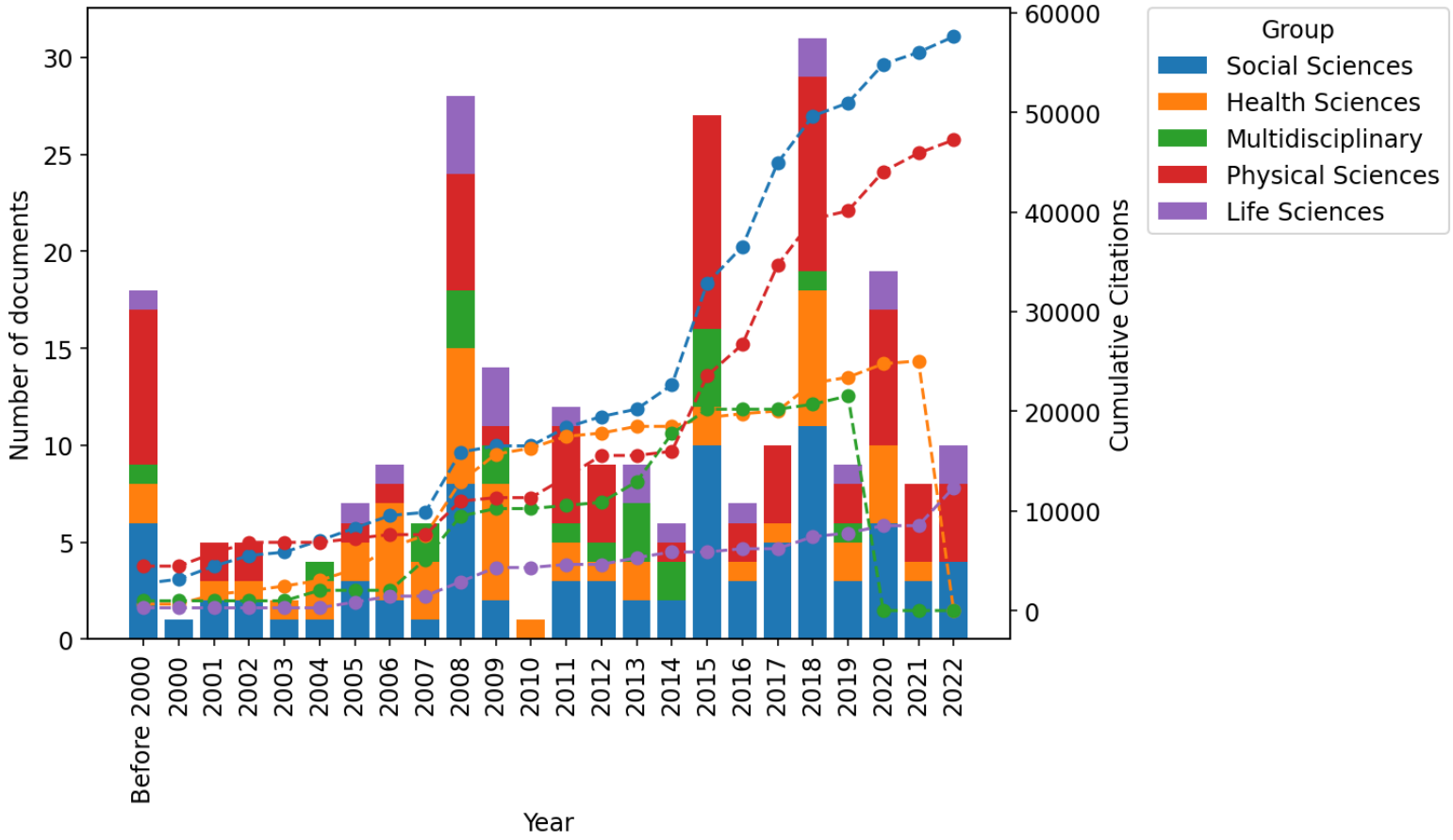
Bibliometric group analysis

- Example of initialization

```
In [16]: ba.add_sciences_scopus()
    ...: bg = bb.BiblioGroupAnalysis(db="scopus", df=ba.df, group_desc="Science")
```

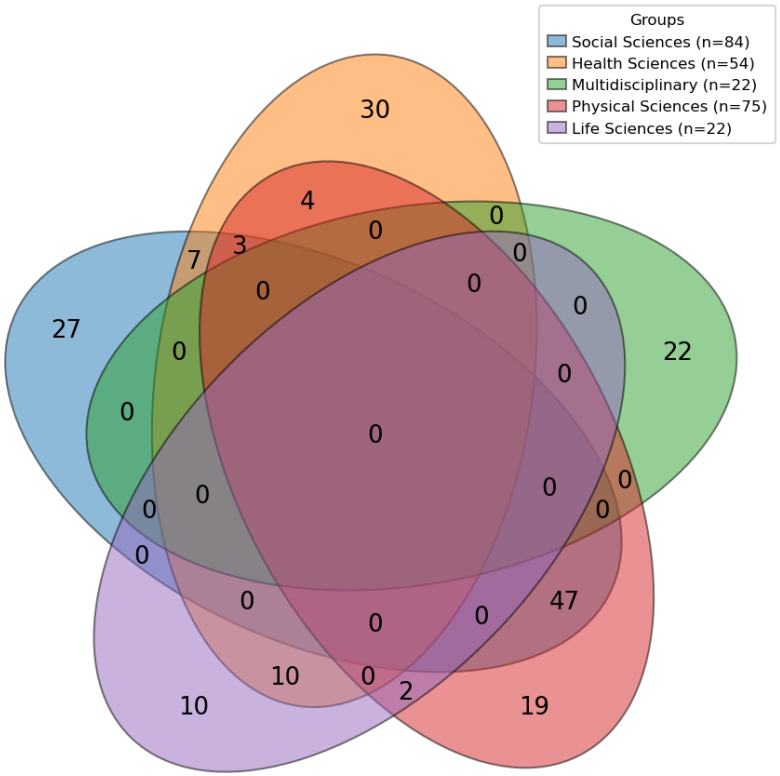
Scientific production

```
In [53]: bg.plot_stacked_production_by_group(cut_year=2000)
```



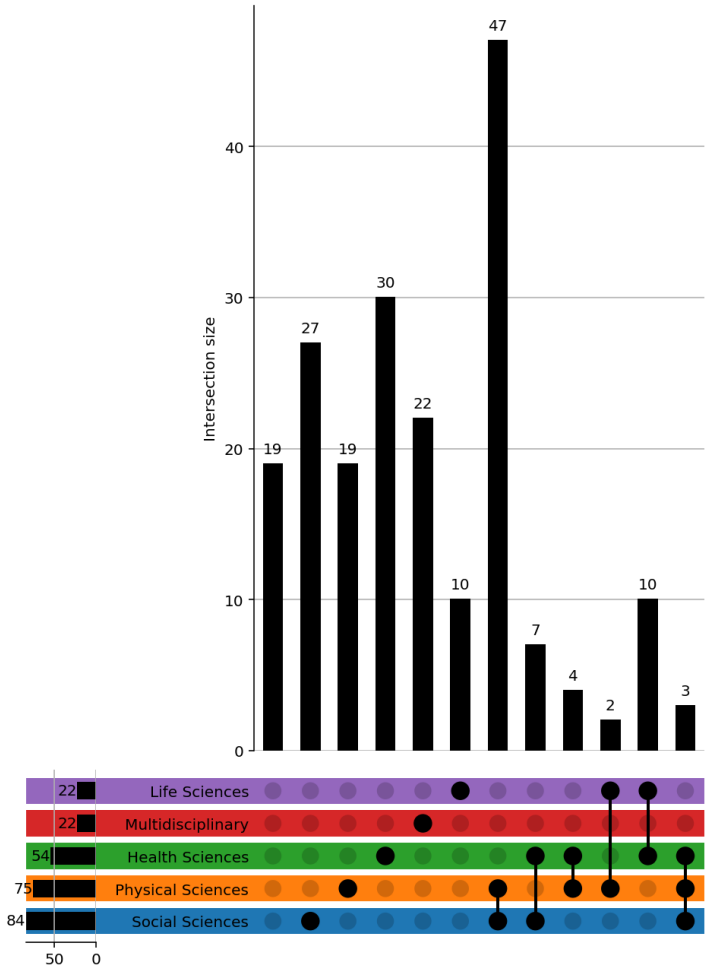
Overlap visualization - Venn

```
In [20]: bg.plot_group_venn()
```



Overlap visualization – upset plot

```
In [23]: bg.plot_group_upset()
```

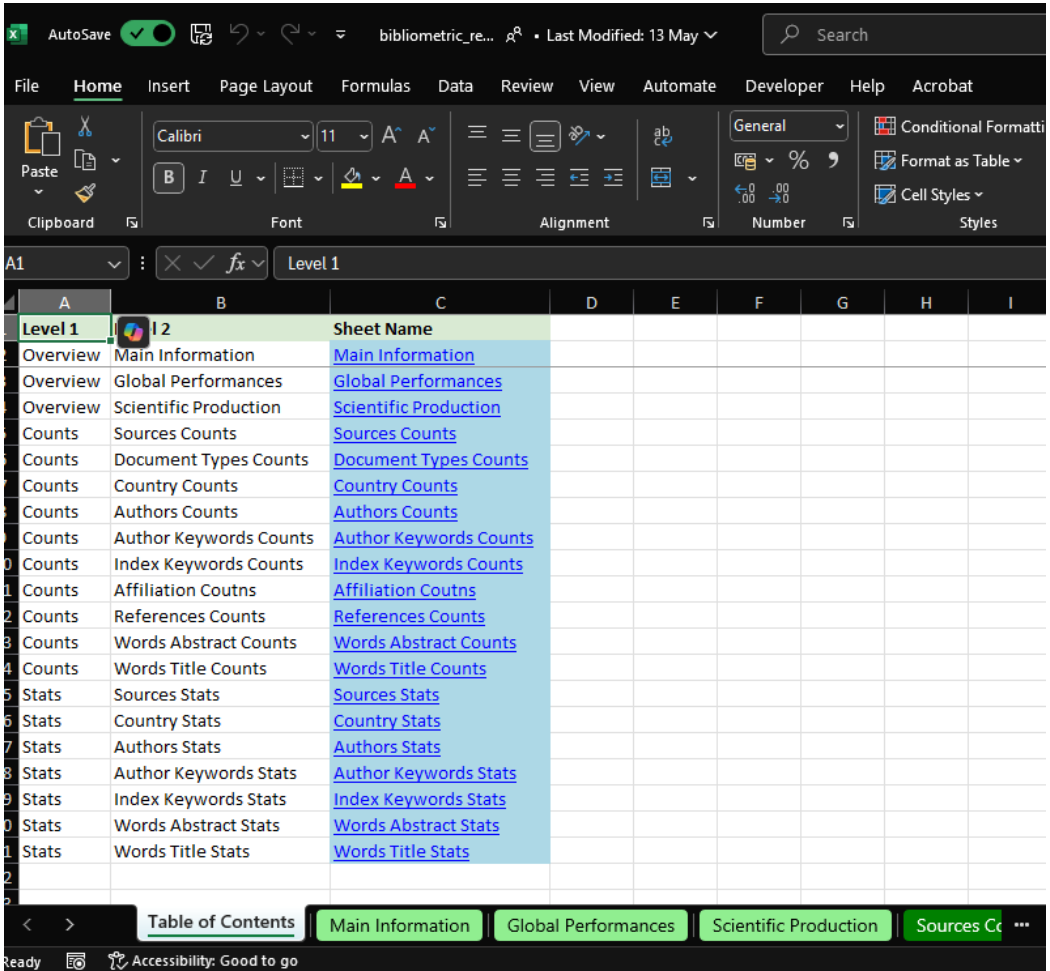


Counting and performance analysis (in development)

- same idea as for the global analysis
- user can select from long and wide output

Saving the results

```
In [1]: ba.save_reports(formats=["docx", "xlsx", "pptx", "tex"], f_name="bibliometric report")
```



AutoSave bibliometric_re... • Last Modified: 13 May Search

File Home Insert Page Layout Formulas Data Review View Automate Developer Help Acrobat

Paste Clipboard Calibri 11 Font Alignment General Number Conditional Formatti Format as Table Cell Styles Styles

A1 Level 1

	A	B	C	D	E	F	G	H	I
1	Level 1	l 2	Sheet Name						
2	Overview	Main Information	Main Information						
3	Overview	Global Performances	Global Performances						
4	Overview	Scientific Production	Scientific Production						
5	Counts	Sources Counts	Sources Counts						
6	Counts	Document Types Counts	Document Types Counts						
7	Counts	Country Counts	Country Counts						
8	Counts	Authors Counts	Authors Counts						
9	Counts	Author Keywords Counts	Author Keywords Counts						
10	Counts	Index Keywords Counts	Index Keywords Counts						
11	Counts	Affiliation Coutns	Affiliation Coutns						
12	Counts	References Counts	References Counts						
13	Counts	Words Abstract Counts	Words Abstract Counts						
14	Counts	Words Title Counts	Words Title Counts						
15	Stats	Sources Stats	Sources Stats						
16	Stats	Country Stats	Country Stats						
17	Stats	Authors Stats	Authors Stats						
18	Stats	Author Keywords Stats	Author Keywords Stats						
19	Stats	Index Keywords Stats	Index Keywords Stats						
20	Stats	Words Abstract Stats	Words Abstract Stats						
21	Stats	Words Title Stats	Words Title Stats						
22									

Table of Contents | Main Information | Global Performances | Scientific Production | Sources Cc ...

Ready Accessibility: Good to go

Association analysis (in development)

- analysis of two binary dataframes
  - group_matrix: description of groups
  - df_items: binary dataframe of items of interest
- matrix multiplication provides 2x2 tables
  - compute associations from these tables
    - Jaccard, Yule‘s Q, Sokal-Michener
  - statistical tests for association
    - Fisher‘s exact test, chi-squared test

Future work

- develop biblium as pip package
- Tkinter app
- Integration to open source data minig software Orange
- https://orangedatamining.com/

Future work

- presentation at ISSI conference (June 2025 in Yerevan, Armenia)
- publication in Scientometric journal
- implementation of subgroup discovery algorithms to biblium