SN SciGraph Analytics

Building business tools for the scholarly publishing domain using LOD and the ELK stack

SEMANTiCS
Vienna 2018

Markus Kaindl

Senior Manager Semantic Data
Business Owner SN SciGraph

SPRINGER NATURE

Agenda (25 Min Talk + 5 Min Q&A)

Background

- SN SciGraph: Linked Open Data Motivation
- AAA: Architecture, App stack, Achievements

ELK for Business

- Internal Business Drivers
- Challenges with Triple Stores
- Analytics Dashboard Examples

Current Work

- Known issues are addressed
- SN Insights & ELK supports QA

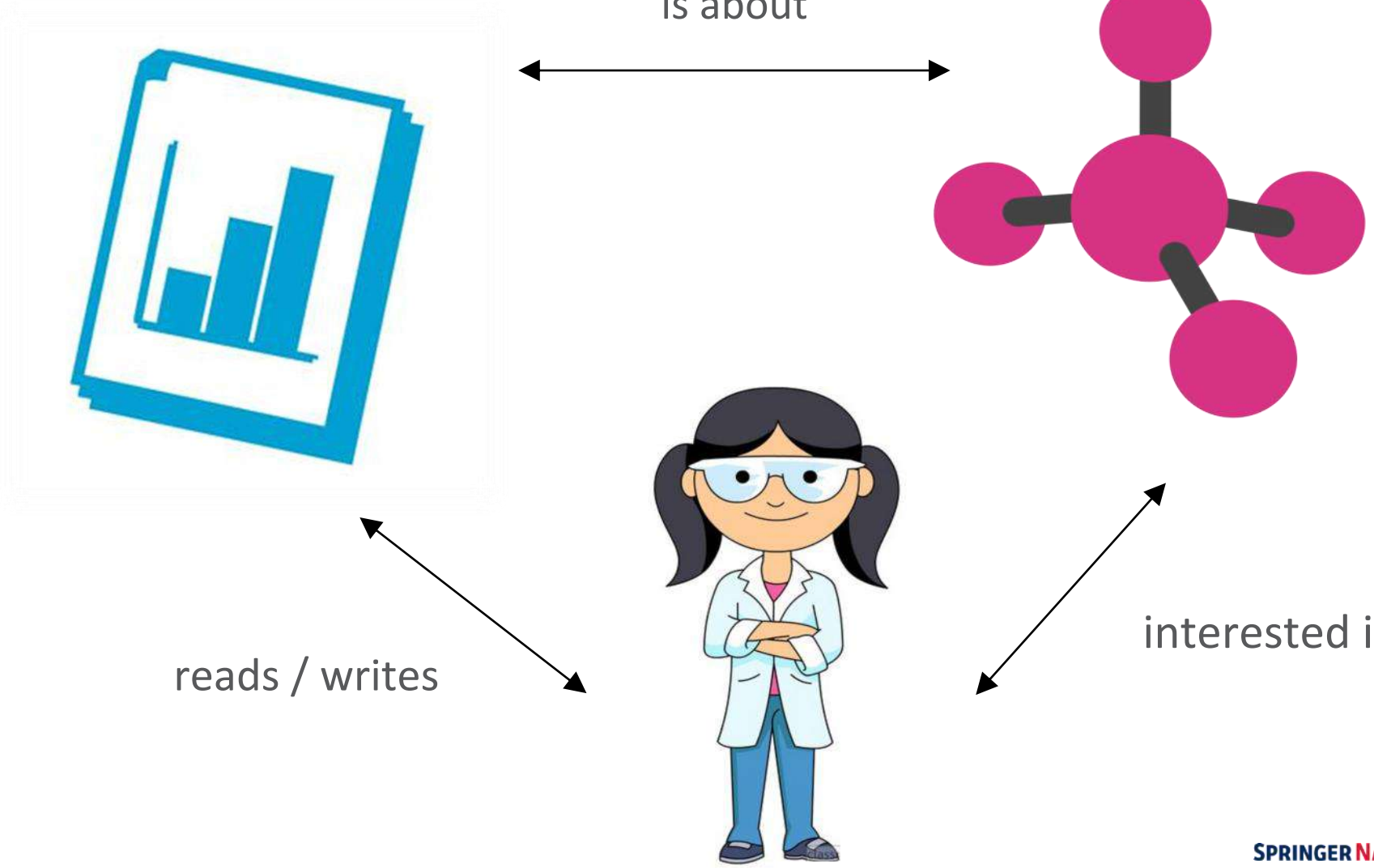
Background

SN SciGraph: Linked Open Data Motivation

SPRINGER NATURE

Formed in **May 2015** through the **merger** of Nature Publishing Group, Palgrave Macmillan, Macmillan Education and Springer Science+Business Media

Three Areas Of Knowledge We Focus On



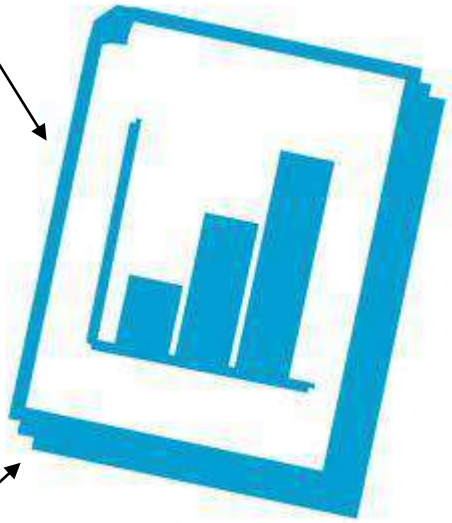
Produces

Has topic

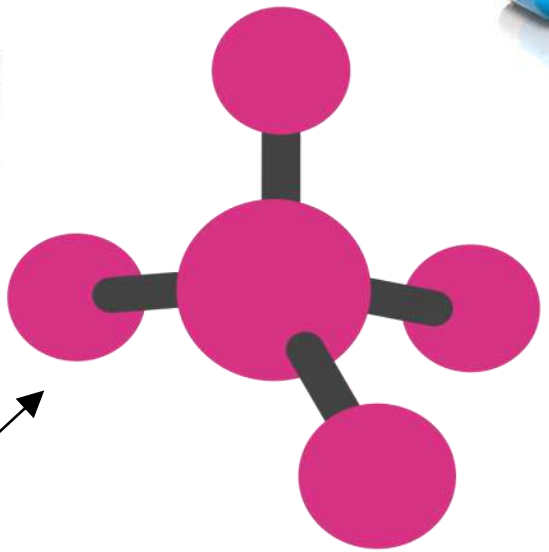
figshare

In proceedings

Describes research data set



Attends



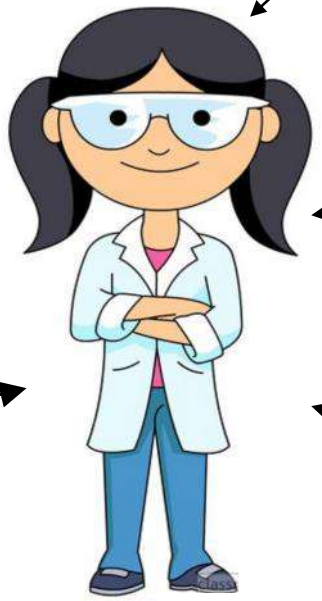
Studies

Produces

Contains

flickr

Reads / Writes



Lead researcher in

Cites

Funds

Works for

Located at

SPRINGER NATURE

Vision: From Content to Data

- We create the largest state-of-the-art linked open data aggregation platform for the scholarly domain from a publisher.
- In doing so, we increase content discoverability and provide data tools and services for researchers, authors, editors, librarians, data scientists, funders, conference organizers, and many others by adding value across all content types.

We publish content

We manage knowledge

SPRINGER NATURE

Stanford Libraries awarded $4 Million grant to implement linked data metadata environment

Stanford, CA—A proposal to dramatically shift how libraries create metadata and greatly improve how users discover library holdings has been accepted and awarded to Stanford Libraries by The Andrew W. Mellon Foundation. In partnership with the libraries of Cornell, Harvard and the University of Iowa, Stanford will lead the effort to integrate library data into the Web, in a semantic way, so it can be discovered intelligently in Web searches as well as in a library's catalogue.

This is a **major development** in the area of LOD and libraries. Up until now a vast number of libraries have adopted Linked Open Data but mainly as part of some kind of data publishing effort – in other words, although their catalogue (or part of) was transformed and made available as LOD, internally they would still rely on traditional data standards like MARC (Machine-Readable Cataloging).

This is now starting to change. There is a **more shared understanding of the opportunities** opened by using Linked Open Data throughout the information management lifecycle.

In **Europe**, a similar project was recently announced by the National Library of Sweden (KB), which has deployed a new version of the Swedish Union Catalogue, Libris, fully based on Bibframe 2.0 and Linked Open Data.

SN SciGraph: Benefits & Applications

Linked Open Data Publishing

- **Researchers** can analyze / build upon our data
- Contributing to Open Research

Content Discoverability

- SN SciGraph Data Explorer
- APIs for better **end user** applications

Business Intelligence and Analytics

- Dashboards for understanding the research landscape
- **Editors**, **Sales**, **Marketing** etc.

AAA: Architecture, App stack, Achievements

ETL Architecture: main features

Airflow

Tech stack

> Airflow framework (Airbnb)

> Amazon S3 to make backups

> GraphDB triplestore (staging and presentation)

> Elastic search and APIs

Components & Principles

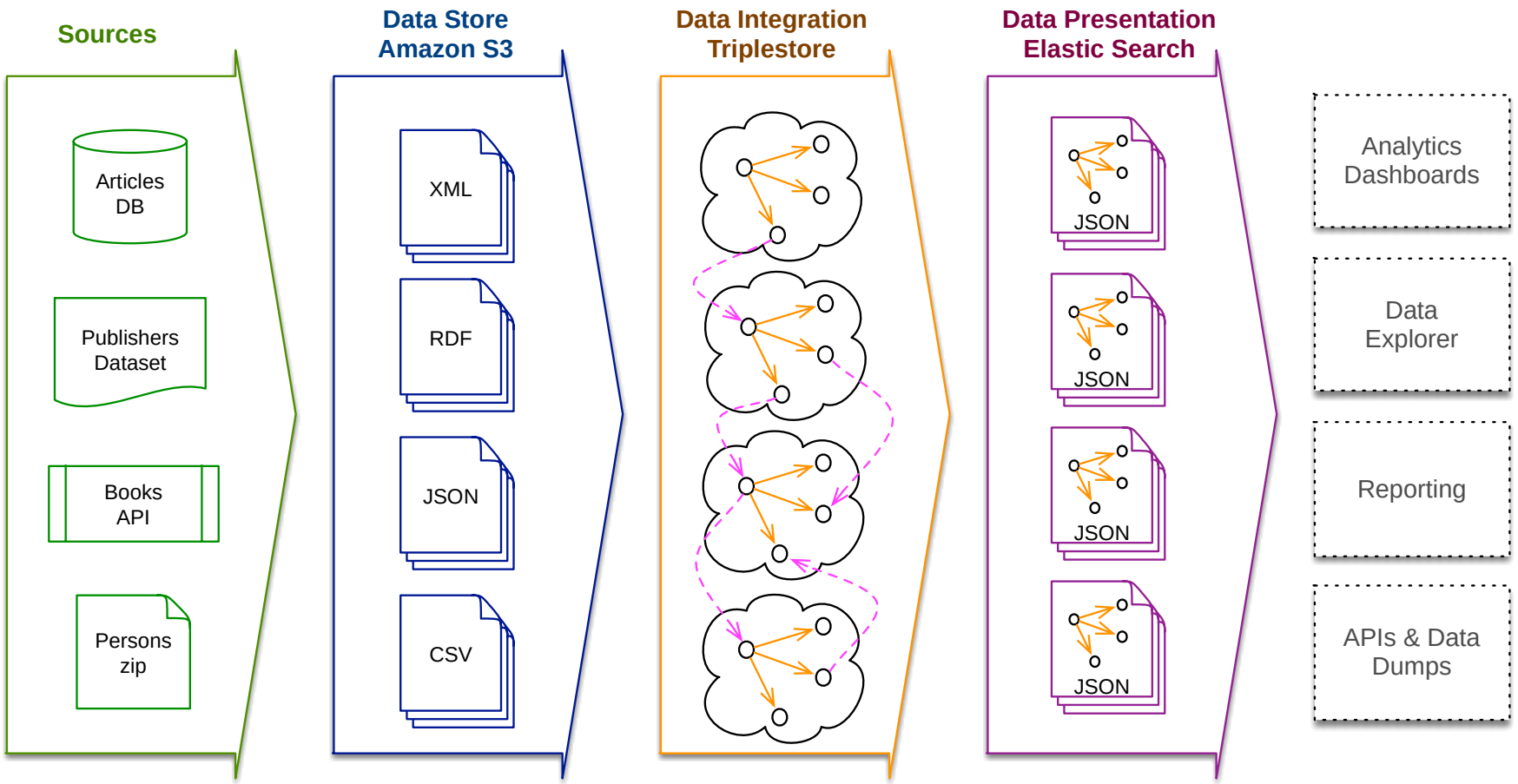
> Graph must be '**ephemeral**'

> Data sources versioning algorithm

> Identity Persistence service

> Validation via SHACL (TopBraid API)

ETL Architecture: data flow



* Versioning service (md5 checksum, timestamps, origin version, etc...)

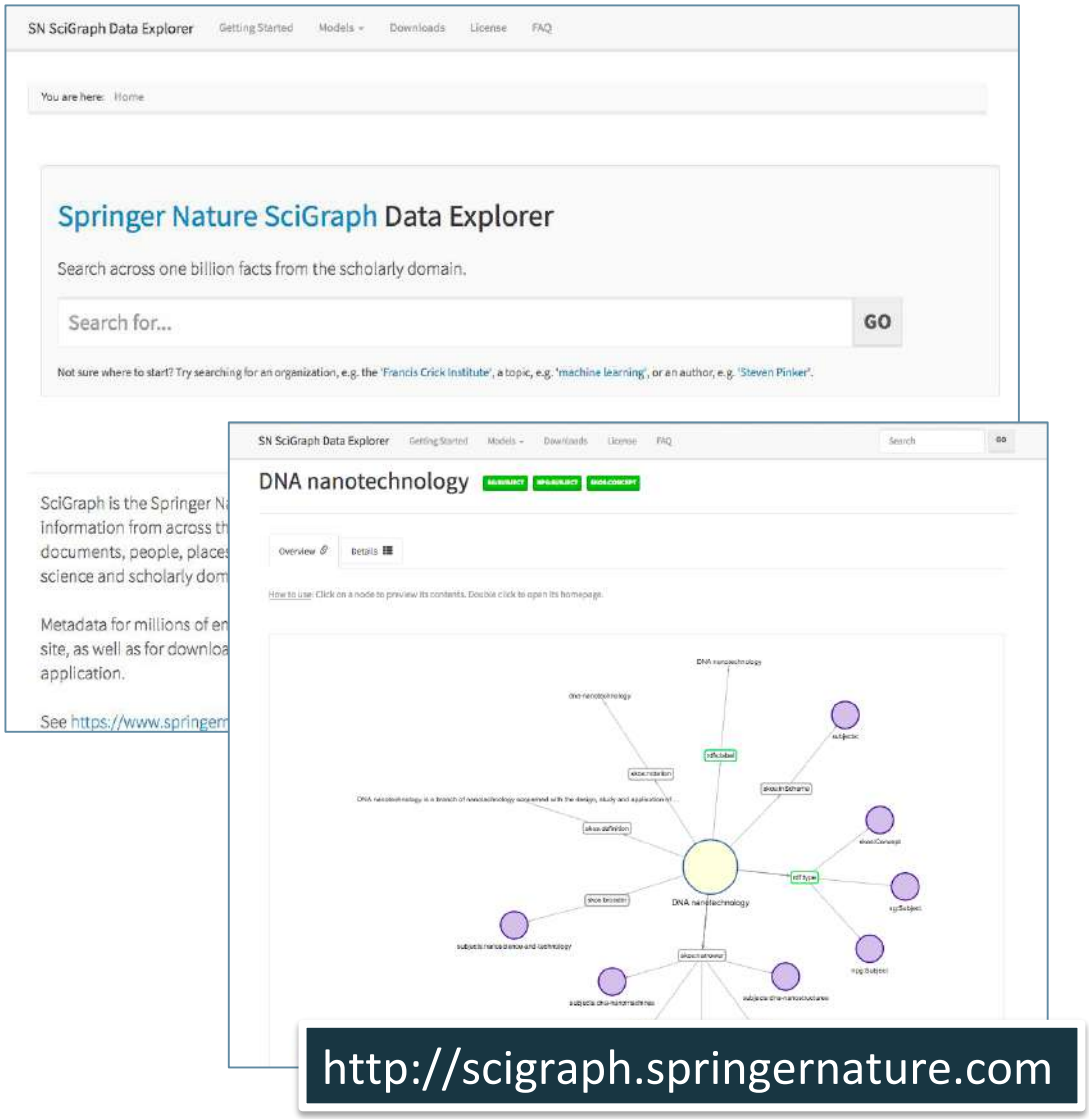
→

* Extraction
* Validation (SHACL)
* Identity Resolution
* Inference (OWL)

→

* Application views
* Search trees (JSON)
* Denormalisation
* Performance

Linked Open Data Publishing So Far



http://scigraph.springernature.com

Nov 2017 Release:

- 1 billion facts / 200 gb download size
- Licensing: CC0, CC-BY-NC & CC-BY

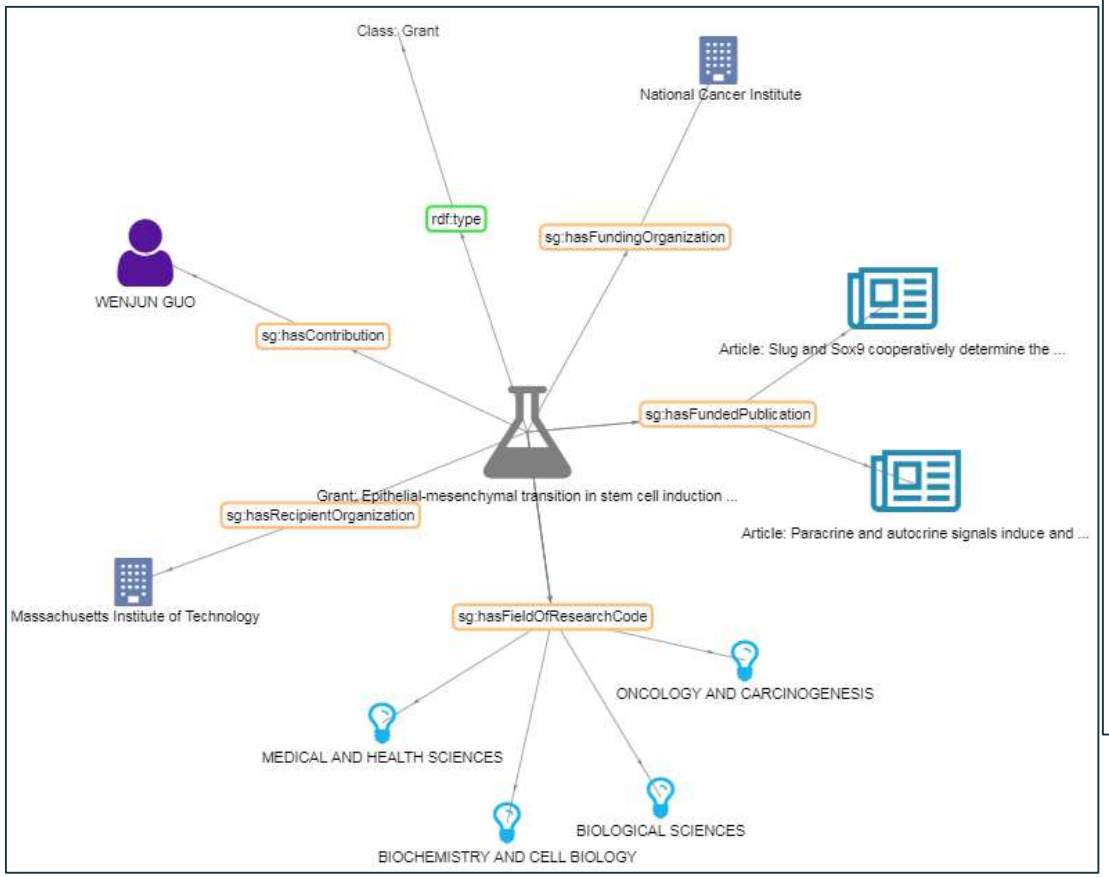
Metadata about:

- Journals & Articles (8M) + Abstracts
- Books & Chapters (4M)
- Grants (200k)
- Subjects, Research Organizations, Conferences, Ontologies (100k)
- Links to to other datasets (20M): Crossref, GRID, MESH, DBpedia, Wikidata

Linked Open Data Explorer

Cell division

Cell division is the process by which a parental cell gives rise to two daughter cells. The process involves both nuclear division and cytokinesis and can either produce two equal cells (symmetric cell division) or two cells with different cellular fates (asymmetric cell division).



Related objects

FUNDER

- National Cancer Institute

RECIPIENT

- Massachusetts Institute of Technology

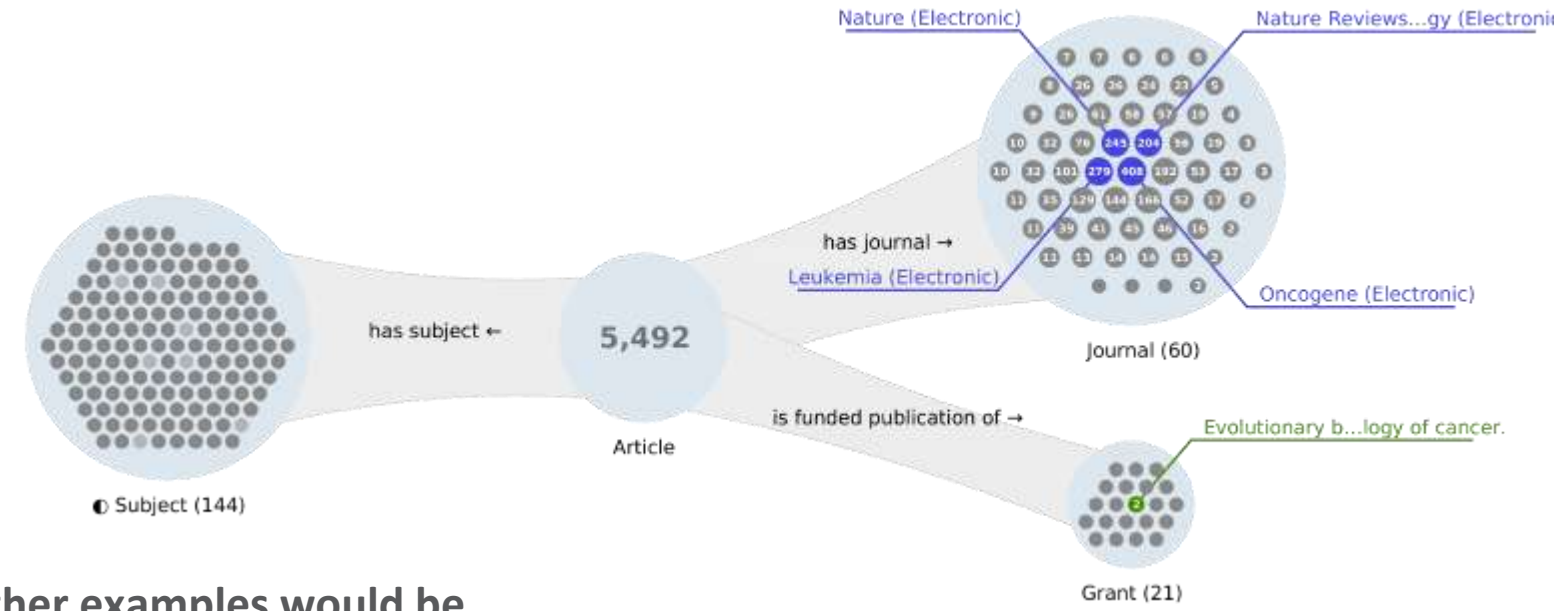
FIELD OF RESEARCH CODES

- Biological Sciences
- Biochemistry And Cell Biology
- Medical And Health Sciences
- Oncology And Carcinogenesis

SciGraph
SPRINGER NATURE

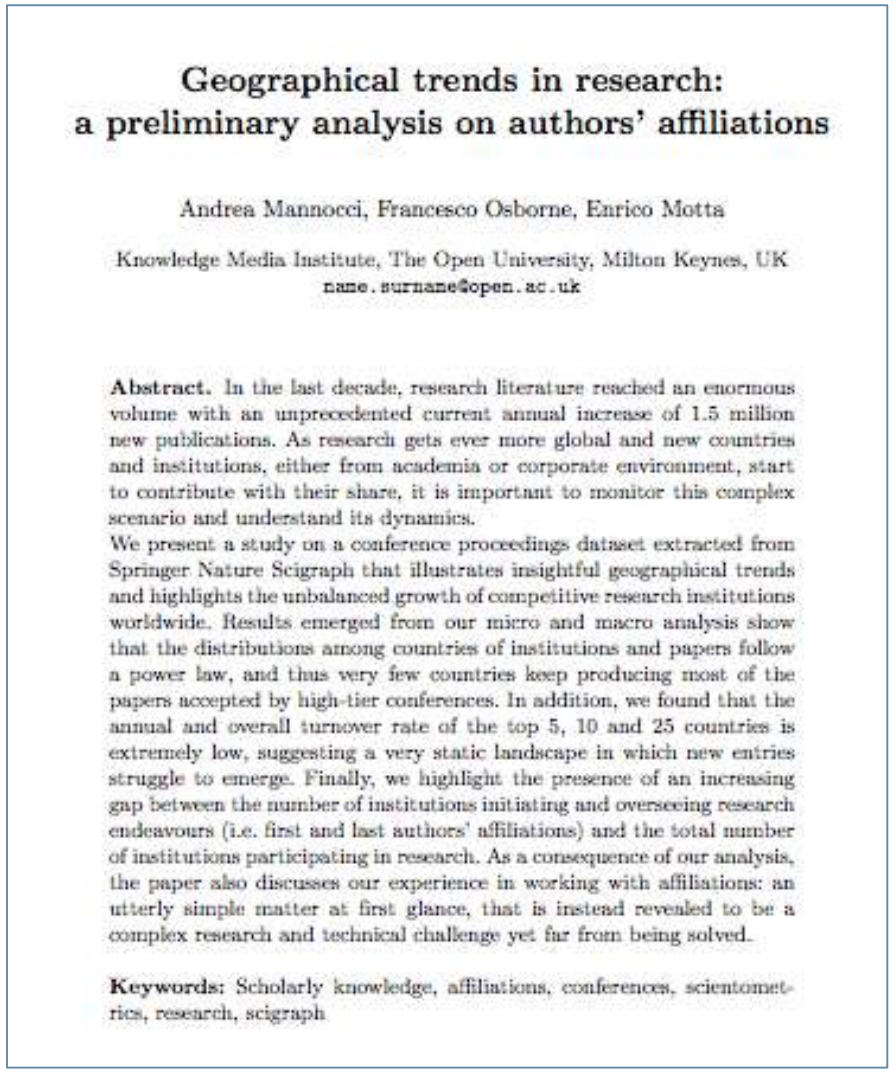
Examples of users already working with SN SciGraph data

- **SemSpect: Uncovering the Hidden in Springer Nature's SciGraph**
  - The essential problem is to get an idea of the queries that deliver real insight.
  - This video shows a sample exploration of SN SciGraph data with SemSpect.



- **Other examples would be**
  - ResearchGraph (Australia)
  - PubChem (USA)

Linked Open Data Put To Use By The Open University, UK



Geographical trends in research: a preliminary analysis on authors' affiliations

Andrea Mannocci, Francesco Osborne, Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, UK
name.surname@open.ac.uk

Abstract. In the last decade, research literature reached an enormous volume with an unprecedented current annual increase of 1.5 million new publications. As research gets ever more global and new countries and institutions, either from academia or corporate environment, start to contribute with their share, it is important to monitor this complex scenario and understand its dynamics.
We present a study on a conference proceedings dataset extracted from Springer Nature Scigraph that illustrates insightful geographical trends and highlights the unbalanced growth of competitive research institutions worldwide. Results emerged from our micro and macro analysis show that the distributions among countries of institutions and papers follow a power law, and thus very few countries keep producing most of the papers accepted by high-tier conferences. In addition, we found that the annual and overall turnover rate of the top 5, 10 and 25 countries is extremely low, suggesting a very static landscape in which new entries struggle to emerge. Finally, we highlight the presence of an increasing gap between the number of institutions initiating and overseeing research endeavours (i.e. first and last authors' affiliations) and the total number of institutions participating in research. As a consequence of our analysis, the paper also discusses our experience in working with affiliations: an utterly simple matter at first glance, that is instead revealed to be a complex research and technical challenge yet far from being solved.

Keywords: Scholarly knowledge, affiliations, conferences, scientometrics, research, scigraph

https://save-sd.github.io/2018/accepted/mannocci/index.pdf

Best paper award at WWW 2018

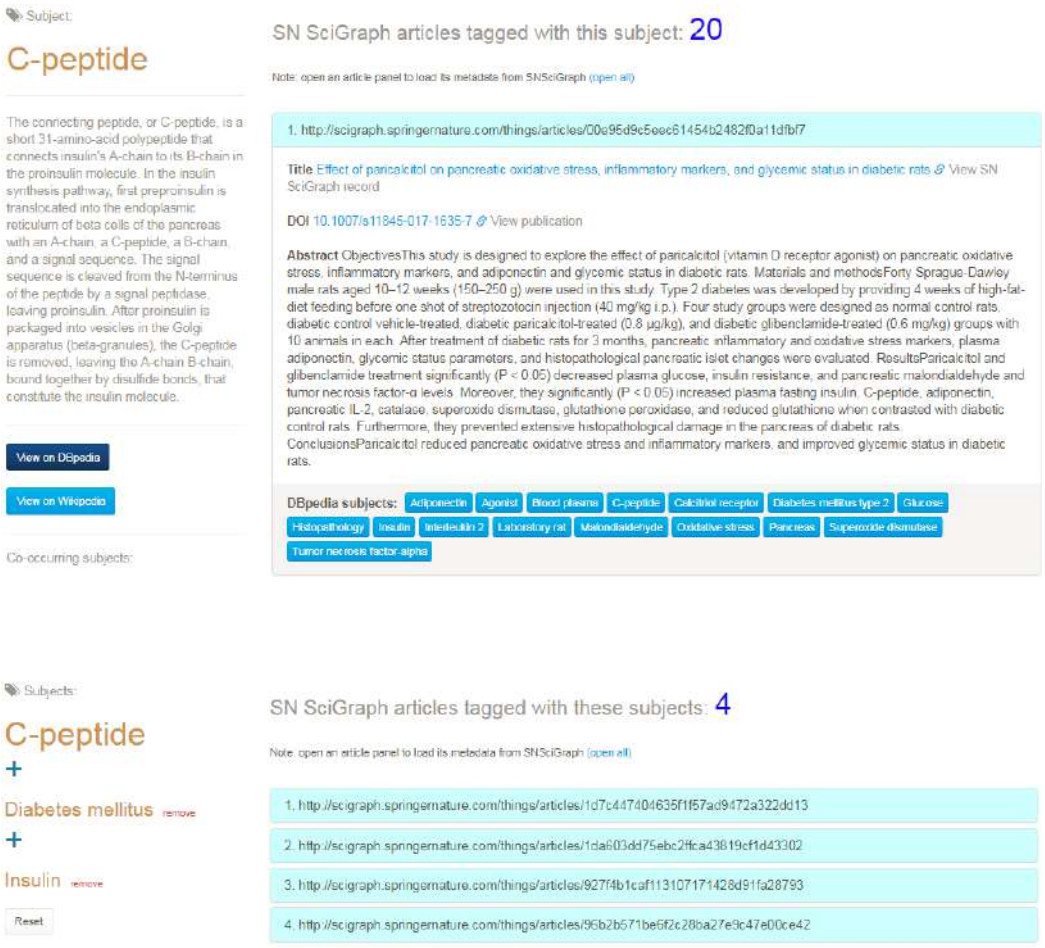
Focus on conference proceedings data in SN SciGraph from 1996 to 2017

Findings :

- "few countries keep producing most of the papers accepted by high-tier conferences.
- "turnover rate [..] is extremely low [..] new entries struggle to emerge.
- "increasing gap between the number of institutions initiating and overseeing research [..] and the total number of institutions participating in research"

SPRINGER NATURE

DBpedia Subjects Browser



What is this?

The prototype allows to search a collection of 2017 SN journal articles which have been semantically enriched using DBpedia subjects.

Purpose and Code:

The purpose of this prototype is to evaluate the quality of the subjects and generate more ideas for future applications.

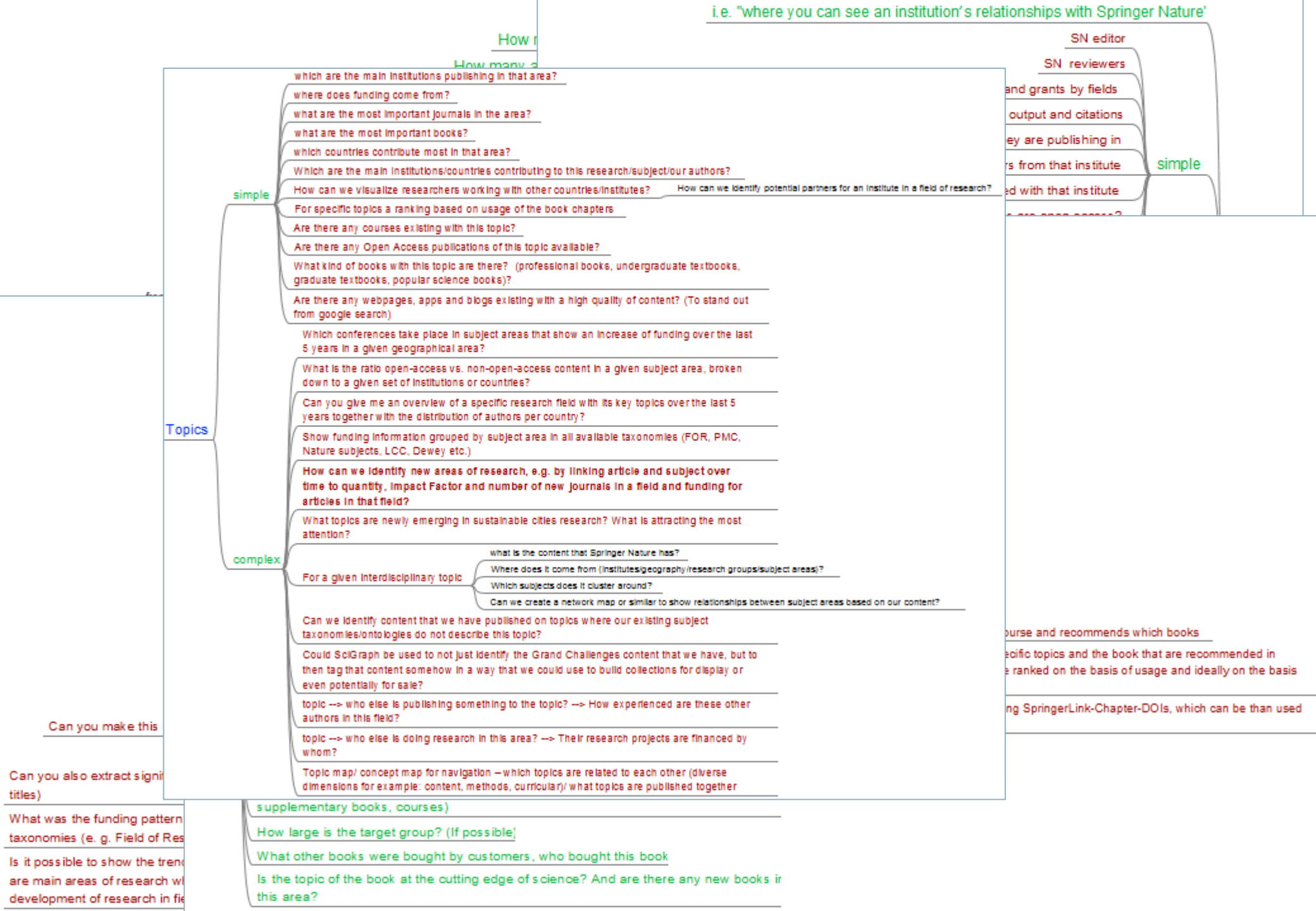
For more info, see the project source code on GitHub

https://github.com/dbpedia/sci-graph-links/

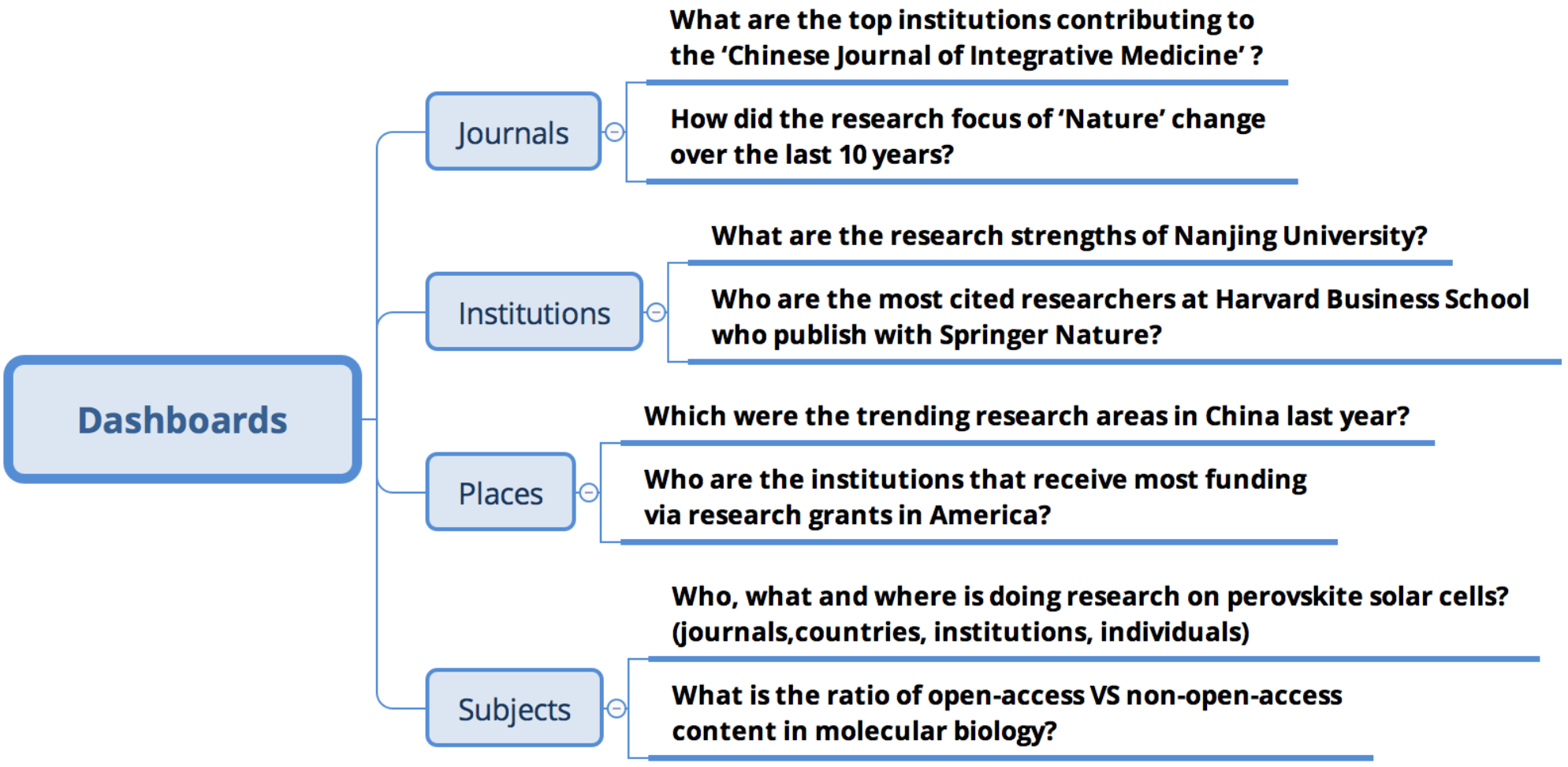
ELK for Business

Internal Business Drivers

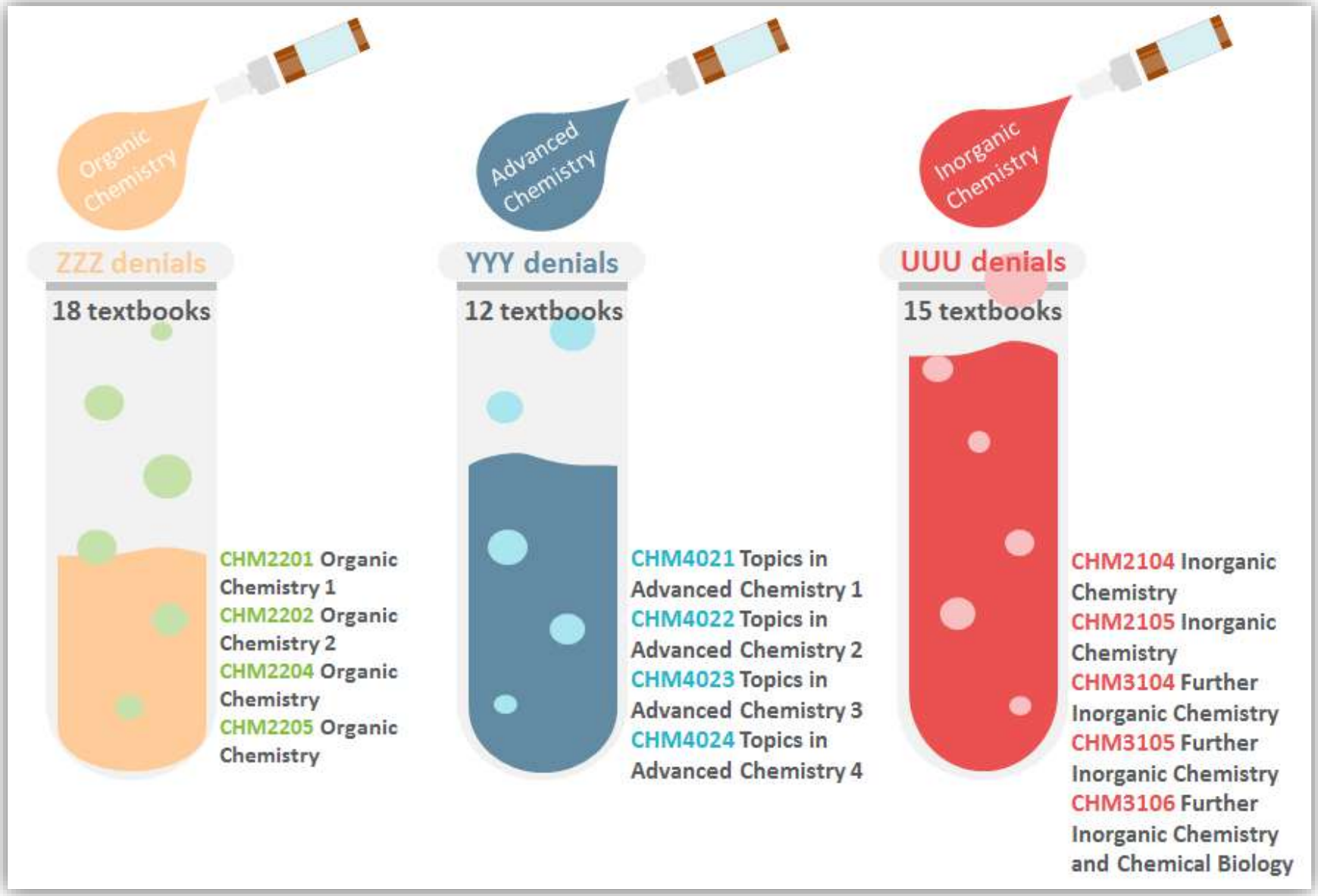
User	Use case examples
Sales	• Understanding how institutes use SN content to **help drive new deals** • Show that SN content leads to commercial applications (patents and grants)
Marketing	• Identifying researchers for better profiling, to support marketing efforts • Understanding how SN content is used at every point of the research process, to **demonstrate the value of being part of the SN world**
Editorial	• Identify potential reviewers for both journals and books • Check the performance of a published title to commission another edition of the same book or commission a different book from the same author • Assessing the **quality of competing journals and book series**, in particular by enabling access to key journal and book metrics
Publishing	• Understand the growth of a given field, or the distribution of sub-fields, to **determine new growth strategies** and portfolio management • A tool to ensure that all publications are properly indexed in 3rd party databases


i.e. "where you can see an institution's relationships with Springer Nature'
SN editor
SN reviewers
and grants by fields
output and citations
ey are publishing in
rs from that institute
ed with that institute
simple
Topics
simple
which are the main institutions publishing in that area?
where does funding come from?
what are the most important journals in the area?
what are the most important books?
which countries contribute most in that area?
Which are the main institutions/countries contributing to this research/subject/our authors?
How can we visualize researchers working with other countries/institutes?
How can we identify potential partners for an institute in a field of research?
For specific topics a ranking based on usage of the book chapters
Are there any courses existing with this topic?
Are there any Open Access publications of this topic available?
What kind of books with this topic are there? (professional books, undergraduate textbooks, graduate textbooks, popular science books)?
Are there any webpages, apps and blogs existing with a high quality of content? (To stand out from google search)
complex
Which conferences take place in subject areas that show an increase of funding over the last 5 years in a given geographical area?
What is the ratio open-access vs. non-open-access content in a given subject area, broken down to a given set of institutions or countries?
Can you give me an overview of a specific research field with its key topics over the last 5 years together with the distribution of authors per country?
Show funding information grouped by subject area in all available taxonomies (FOR, PMC, Nature subjects, LCC, Dewey etc.)
How can we identify new areas of research, e.g. by linking article and subject over time to quantity, Impact Factor and number of new journals in a field and funding for articles in that field?
What topics are newly emerging in sustainable cities research? What is attracting the most attention?
For a given interdisciplinary topic
what is the content that Springer Nature has?
Where does it come from (institutes/geography/research groups/subject areas)?
Which subjects does it cluster around?
Can we create a network map or similar to show relationships between subject areas based on our content?
Can we identify content that we have published on topics where our existing subject taxonomies/ontologies do not describe this topic?
Could SciGraph be used to not just identify the Grand Challenges content that we have, but to then tag that content somehow in a way that we could use to build collections for display or even potentially for sale?
topic --> who else is publishing something to the topic? --> How experienced are these other authors in this field?
topic --> who else is doing research in this area? --> Their research projects are financed by whom?
Topic map/ concept map for navigation – which topics are related to each other (diverse dimensions for example: content, methods, curricular)/ what topics are published together
urse and recommends which books
ecific topics and the book that are recommended in
e ranked on the basis of usage and ideally on the basis
ng SpringerLink-Chapter-DOIs, which can be than used
Can you make this
Can you also extract signi
titles)
What was the funding pattern
taxonomies (e. g. Field of Res
Is it possible to show the tren
are main areas of research wi
development of research in fi
supplementary books, courses)
How large is the target group? (If possible)
What other books were bought by customers, who bought this book
Is the topic of the book at the cutting edge of science? And are there any new books ir
this area?

Dashboards Architecture



Mapping denials to textbook modules



Challenges with Triple Stores

Why the ELK stack

Limitation of triple stores

> Doesn’t support real time analysis

> Too slow with counts & aggregations

> Denormalization leads to a significant increase in data volumes

Pros of ELK stack

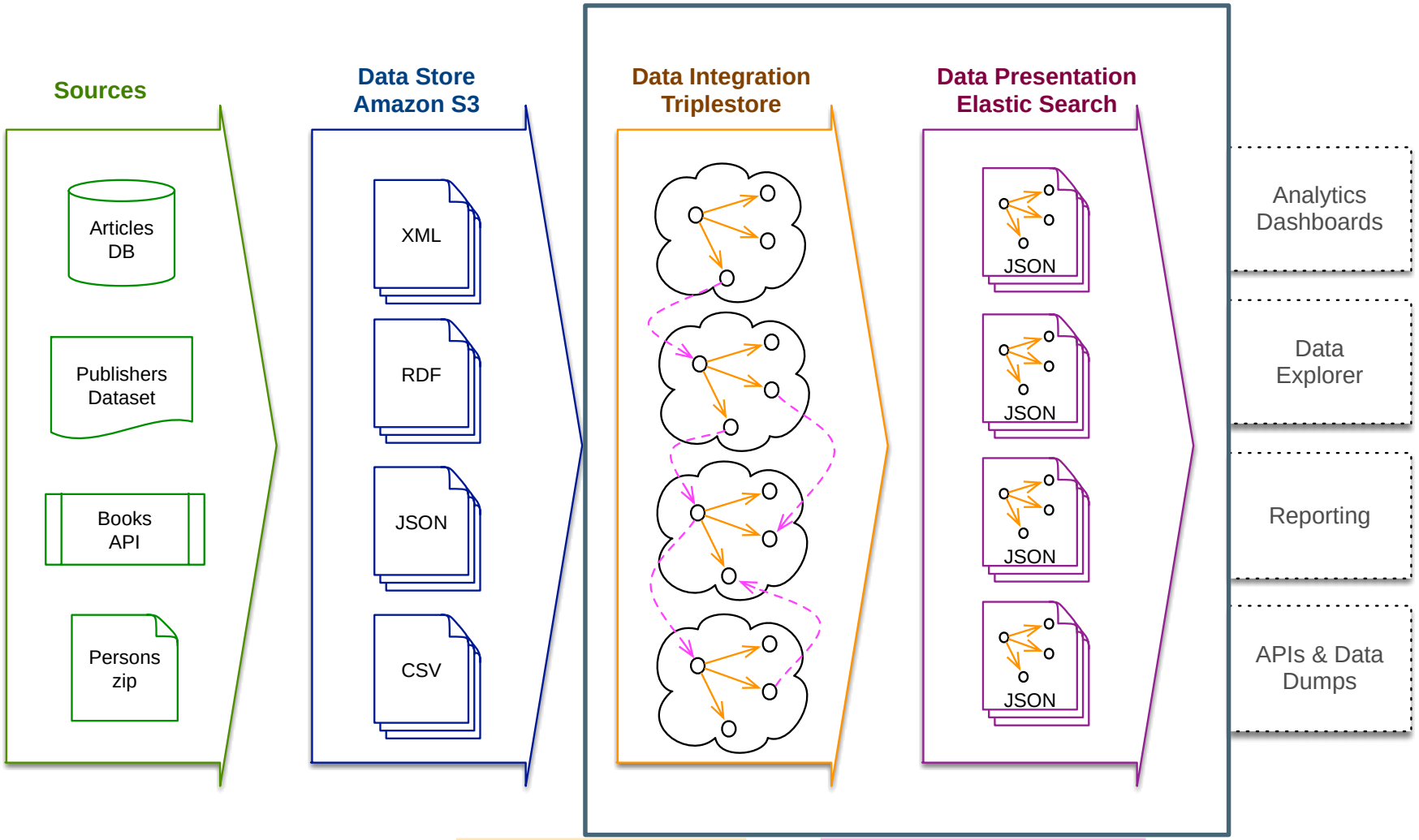
> Very fast (scales linearly)

> GraphDB has built-in connector service

> Kibana offers powerful data visualizations out of the box

> It’s free!

Why the ELK stack [2]



* Versioning service (md5 checksum, timestamps, origin version, etc...)

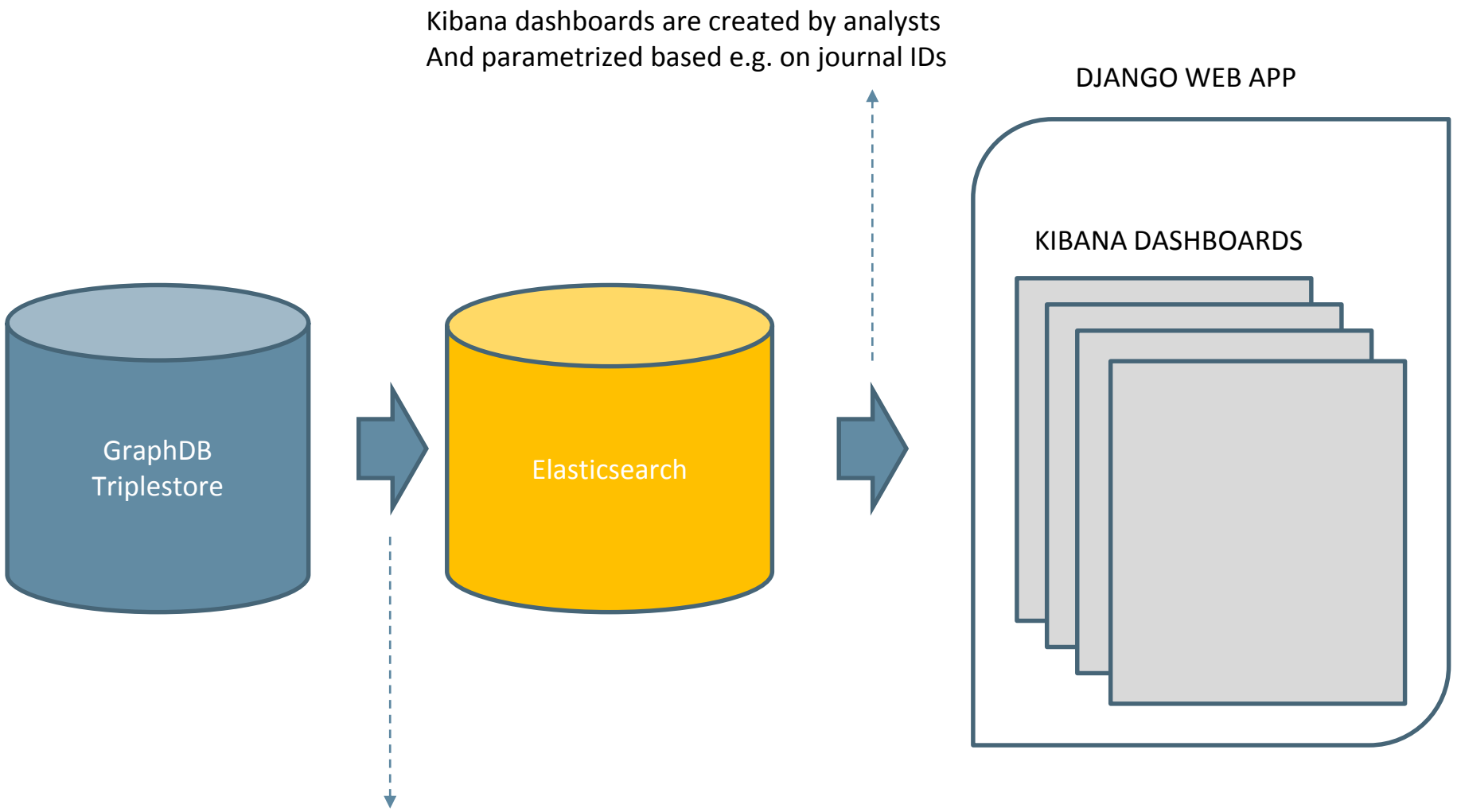
→

* Extraction
* Validation (SHACL)
* Identity Resolution
* Inference (OWL)

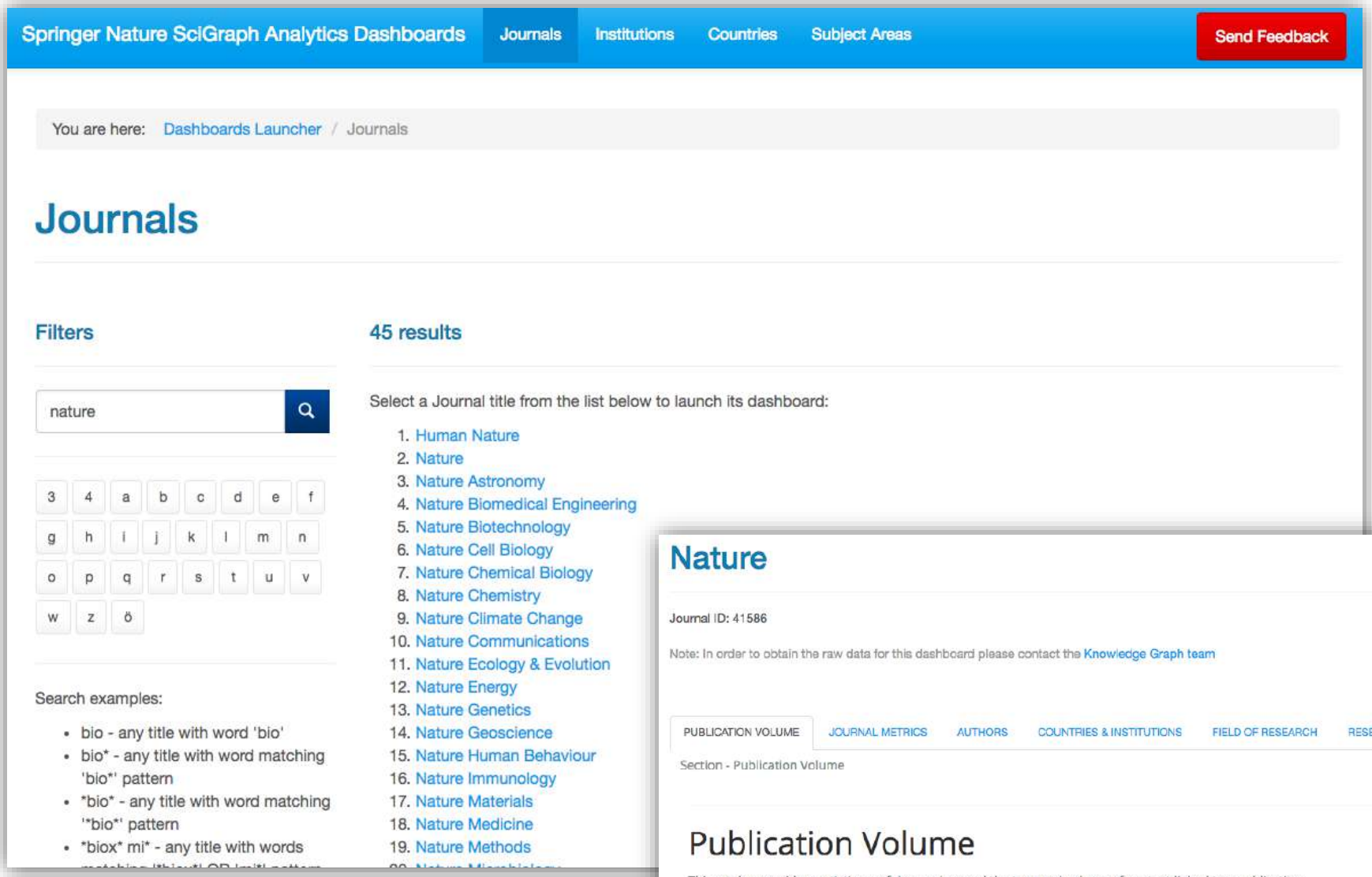
→

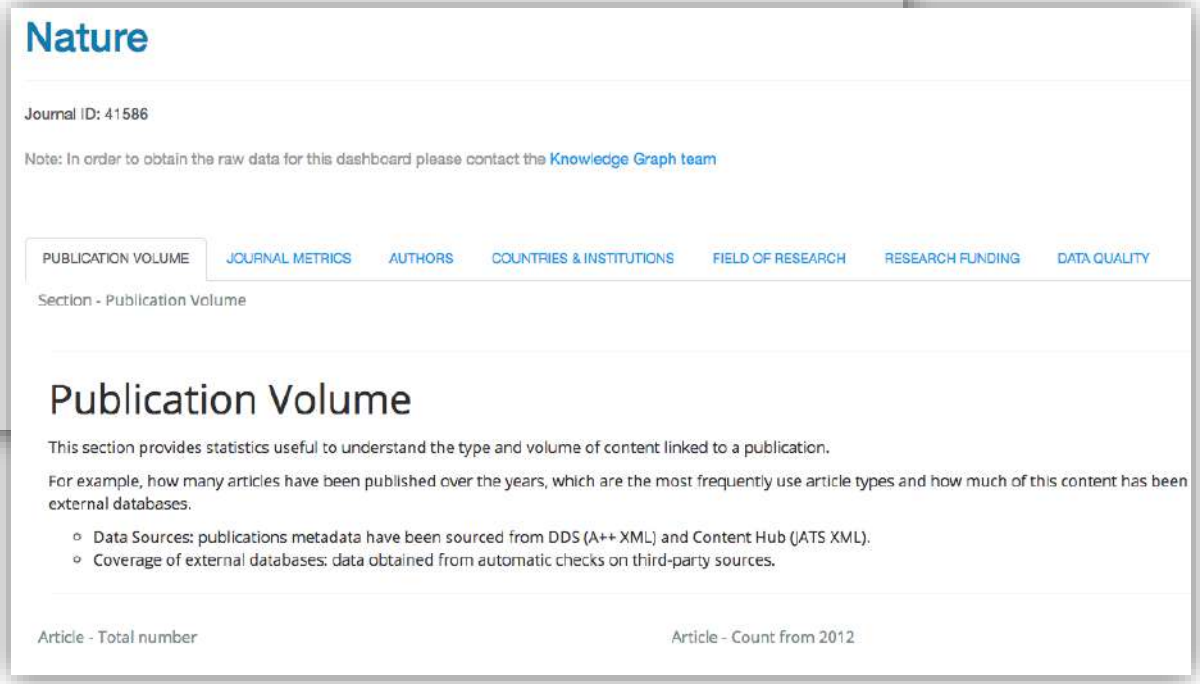
* Application views
* Search trees (JSON)
* Denormalisation
* Performance

System Architecture



Analytics Dashboard Examples


Springer Nature SciGraph Analytics Dashboards
Journals
Institutions
Countries
Subject Areas
Send Feedback
You are here: Dashboards Launcher / Journals
Journals
Filters
nature
3 4 a b c d e f
g h i j k l m n
o p q r s t u v
w z ö
Search examples:
bio - any title with word 'bio'
bio* - any title with word matching 'bio*' pattern
bio - any title with word matching '*bio*' pattern
biox mi* - any title with words
45 results
Select a Journal title from the list below to launch its dashboard:
1. Human Nature
2. Nature
3. Nature Astronomy
4. Nature Biomedical Engineering
5. Nature Biotechnology
6. Nature Cell Biology
7. Nature Chemical Biology
8. Nature Chemistry
9. Nature Climate Change
10. Nature Communications
11. Nature Ecology & Evolution
12. Nature Energy
13. Nature Genetics
14. Nature Geoscience
15. Nature Human Behaviour
16. Nature Immunology
17. Nature Materials
18. Nature Medicine
19. Nature Methods


Nature
Journal ID: 41586
Note: In order to obtain the raw data for this dashboard please contact the Knowledge Graph team
PUBLICATION VOLUME
JOURNAL METRICS
AUTHORS
COUNTRIES & INSTITUTIONS
FIELD OF RESEARCH
RESEARCH FUNDING
DATA QUALITY
Section - Publication Volume
Publication Volume
This section provides statistics useful to understand the type and volume of content linked to a publication.
For example, how many articles have been published over the years, which are the most frequently use article types and how much of this content has been external databases.
Data Sources: publications metadata have been sourced from DDS (A++ XML) and Content Hub (JATS XML).
Coverage of external databases: data obtained from automatic checks on third-party sources.
Article - Total number
Article - Count from 2012

Publication Volume

Argumentation

Journal ID: 10503

Note: In order to obtain the raw data for this dashboard please contact the Knowledge Graph team

PUBLICATION VOLUME

Section - Publication Vol

Publicati

This section provides s

For example, how man
external databases.

- Data Sources: pu
- Coverage of exte

Article - Total number

Article - Total number

1,145

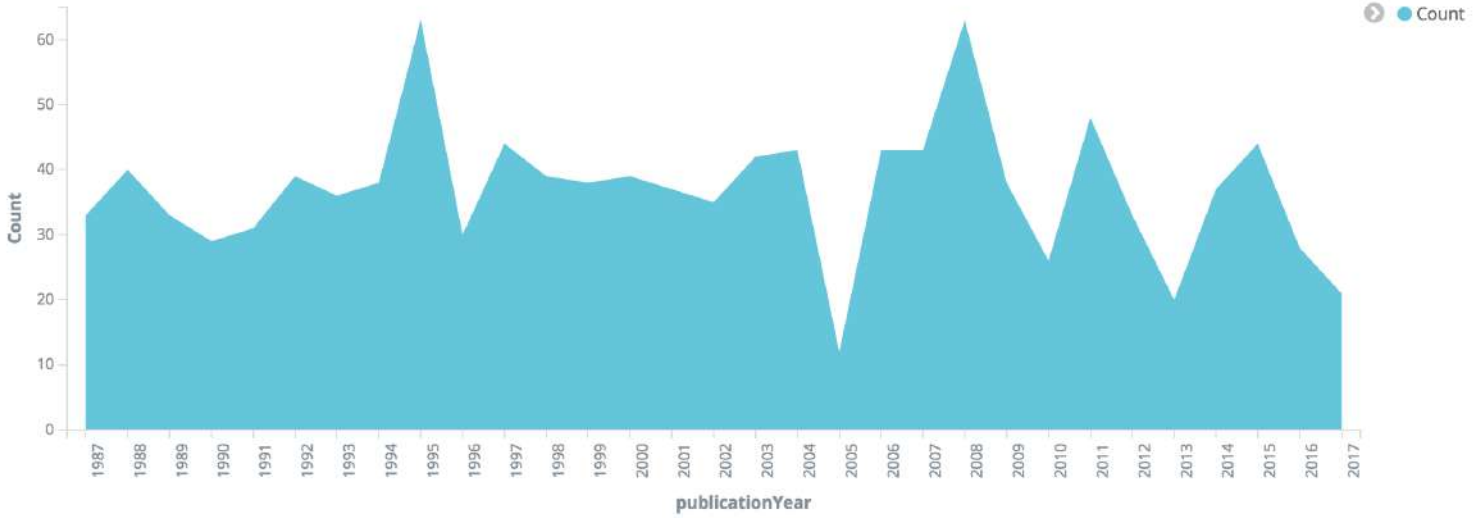
Articles in Total

Article - Count from 2012

150

Articles Published Last 5 Years

Article - Count by publication year



Indexing Status

Journal Metrics

Analytical and Bioanalytical Chemistry

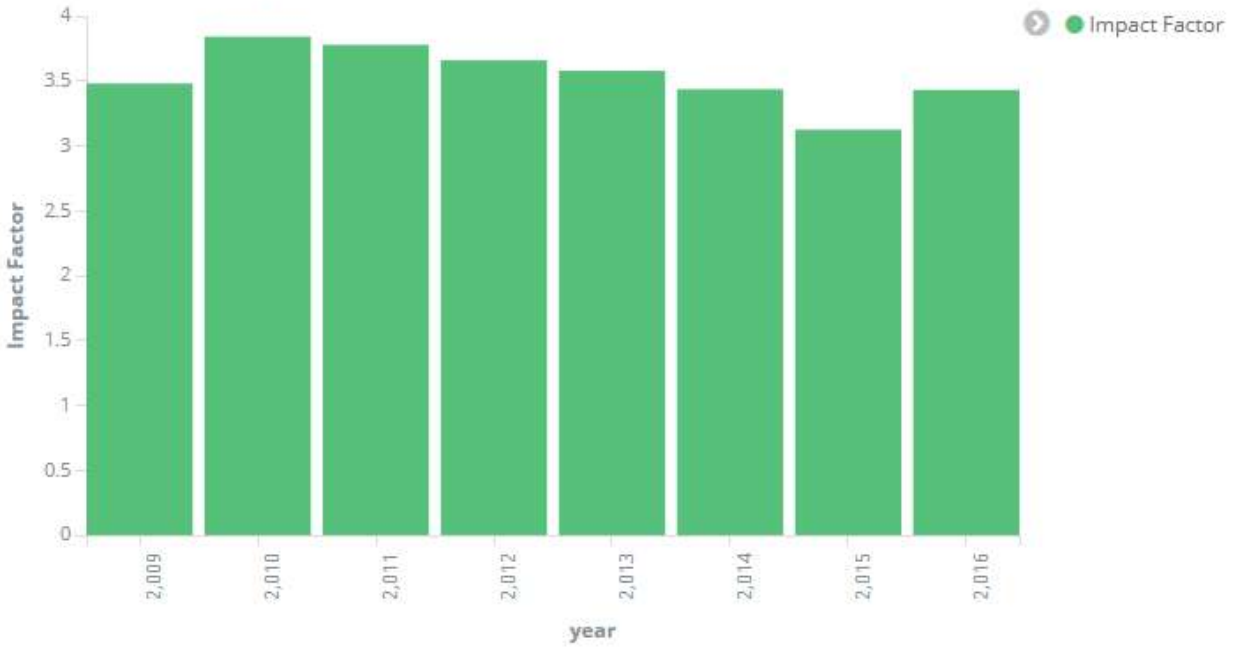
Journal ID: 216

Note: In order to obtain the raw data for this dashboard please contact the Knowledge Graph team

Metric - IF per year

journalTitle	year	impactFactor
Analytical and Bioanalytical Chemistry	2,016	3.431
Analytical and Bioanalytical Chemistry	2,015	3.125
Analytical and Bioanalytical Chemistry	2,014	3.436
Analytical and Bioanalytical Chemistry	2,013	3.578
Analytical and Bioanalytical Chemistry	2,012	3.659
Analytical and Bioanalytical		

Metric - Impact factor by year



Geographical Patterns

BMC Women's Health

Journal

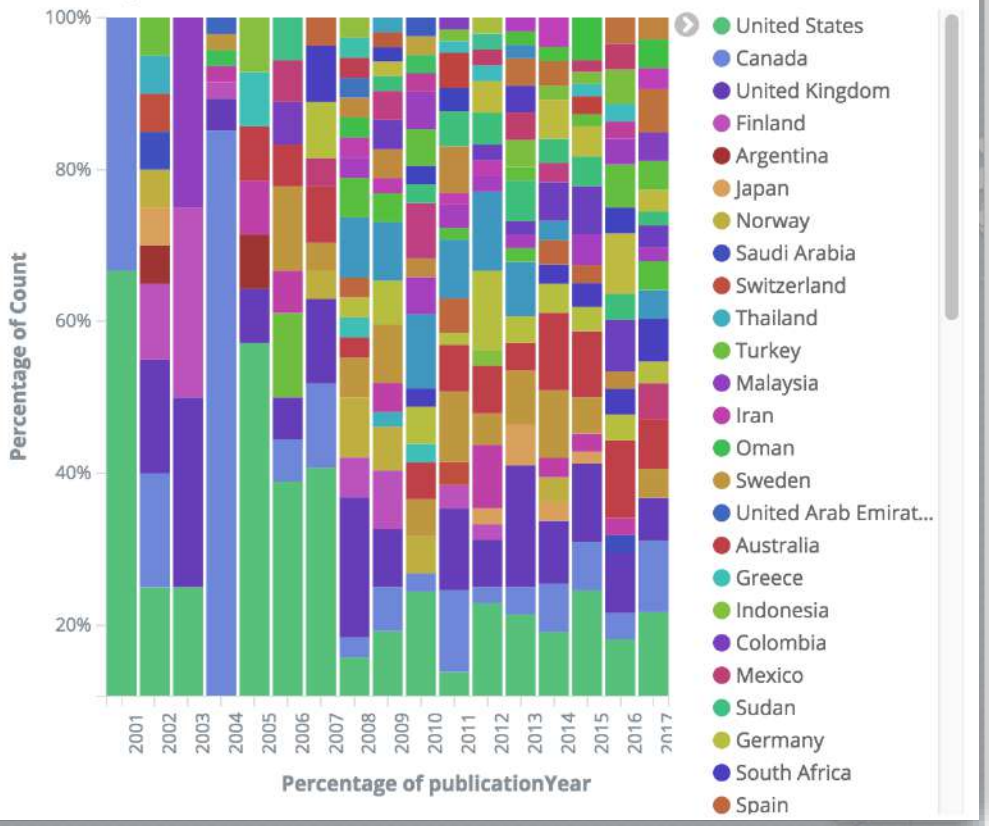
Article - by country

contribution.affiliation.countryName: Descending	Count
United States	197
Canada	89
United Kingdom	81
Australia	59
Sweden	50
Netherlands	35
Germany	32
China	29
Ethiopia	26
Brazil	25

Export: Raw Formatted

1 2 3 4 5 ...9 »

Article - top ten countries distribution as a timeline



Leaflet | © OpenStreetMap contributors | Elastic Maps Service

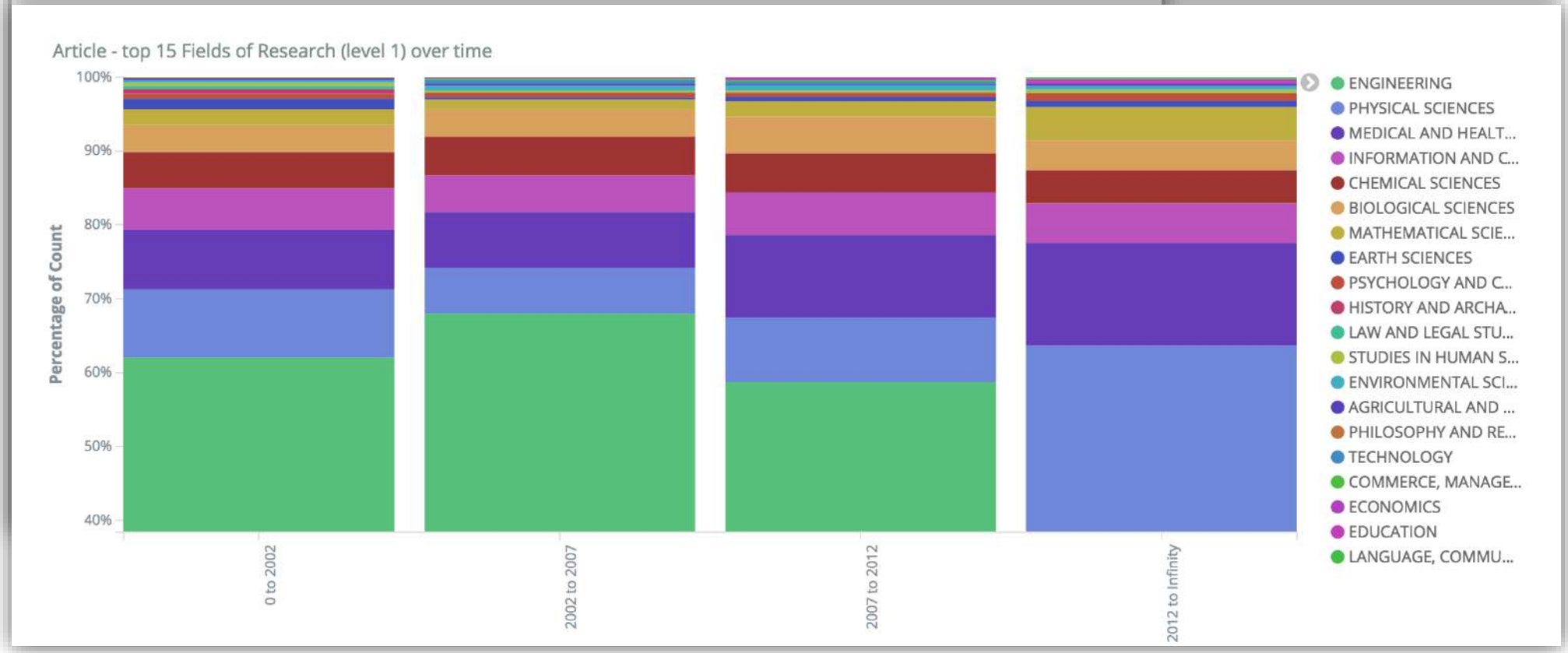
Fields of Research over time

You are here: Dashboards Launcher / Journals / Experiments in Fluids

Experiments in Fluids

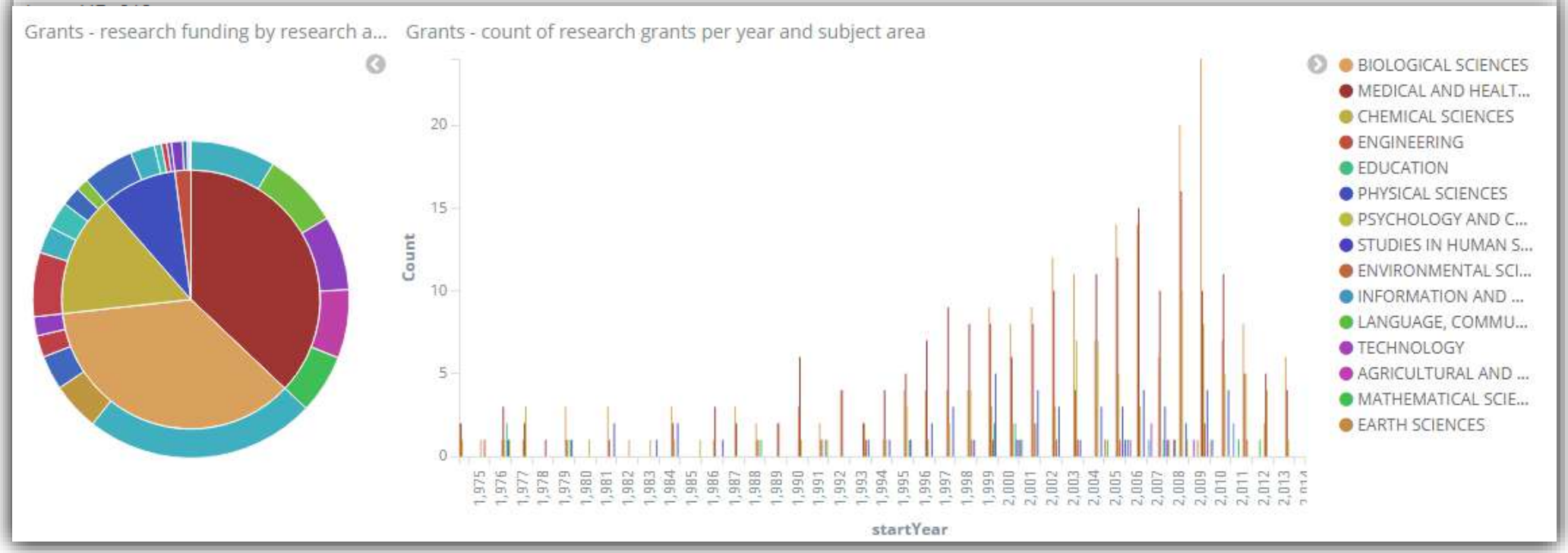
Journal ID: 348

Note: In order to obtain the raw data for this dashboard please contact the Knowledge Graph team



Research Funding

Analytical and Bioanalytical Chemistry



Current Work

Known issues are addressed

SN SciGraph: upcoming data release to increase usage

Going native with schema.org and JSONLD

- Integration with de-facto linked data standard on the web
- Quality metadata straight from internal data delivery system
- Millions of links to relevant scholarly objects
  - Grants related to a publication
  - Clinical trials related to a publication
  - Patents related to a publication
  - Policy documents related to a publication
  - Persons (ORCID and Dimensions IDs)
  - Disambiguated Organizations (GRID)

Google

springer book "teaching mindfulness"

All Shopping Videos Images News More Settings Tools

About 8,890 results (0.63 seconds)

Teaching Mindfulness - A Practical Guide for Clinicians and ... - Springer
www.springer.com/gp/book/9780387094830
The first academic text on **teaching mindfulness** across a broad range of professional clinical settings; Written ... About this **book**; About the authors; Reviews.

Teaching Mindfulness: A Practical Guide for Clinicians and Educators ...
https://www.amazon.com/Teaching-Mindfulness-Practical-Clinicians.../1461402409
Teaching Mindfulness: A Practical Guide for Clinicians and Educators: 8601405091178: Medicine & Health Science **Books** @ Amazon.com. ... Paperback: 250 pages; Publisher: **Springer**; 2010 edition (June 23, 2011); Language: English ...

Fundamentals of Complementary and Alternative Medicine - E-Book
https://books.google.com/books?isbn=032329894X
Marc S. Micozzi - 2014 - Medical
McCown DM, Reibel D, Micozzi MS: **Teaching mindfulness**, New York, 2009, **Springer**. McMahan DL: Repackaging Zen for the West. In Prebish CS, Baumann M ...

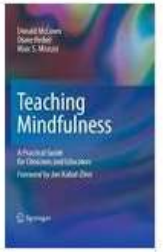
About | Mindfulness and More
www.mindfulnessandmore.com/about/
She is coauthor of the **book Teaching Mindfulness**: A Practical Guide for Clinicians and Educators, a **Springer** publication (2010). Diane is a long-time student of ...

Mindfulness and Buddhist-Derived Approaches in Mental Health and ...
https://books.google.com/books?isbn=3319222554
Edo Shonin, William Van Gordon, Mark Griffiths - 2015 - Psychology
Competence in **teaching mindfulness**-based courses: Concepts, Development and assessment. Mindfulness ... New York: Bantam **Books**. ... New York: **Springer**.

Training mindfulness teachers › Research Explorer
research.bangor.ac.uk/portal/en/publications/training-mindfulness.../export.html
Nov 2, 2016 - Resources for **Teaching Mindfulness**: A cross-cultural and ... **Springer**, 2016. ... UR - http://www.**springer**.com/gb/**book**/9783319300986.

Resources for Teaching Mindfulness: An International Handbook

Teaching Mindfulness: A Practical Guide for Clinicians and ...
Book by Diane Reibel and Marc Micozzi



Preview book

The applications and use of mindfulness-based interventions in medicine, mental health care, and education have been expanding as rapidly as the empirical evidence base that is validating and recommending them. ... Google Books

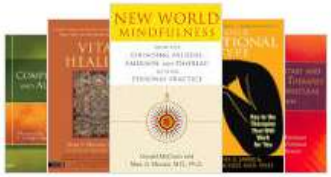
Originally published: 2010

Authors: Marc Micozzi, Diane Reibel

Buy ebook

Google Play Books	$18.25
Kobo	$53.99

People also search for



Marc Micozzi books

Medicine books

Feedback

Beyond the ELK stack: Challenges

Current setup does not scale well (for 1000s of users) esp. when attempting to make it externally available

- UI can be a bit fiddly to use for non tech-savvy colleagues (e.g. filtering)
- Password protection or hiding sensitive data/visualizations can be difficult
- Saving / versioning the visualizations in Kibana requires ad hoc work
  (we were using GitHub to save state, but tricky to work with)

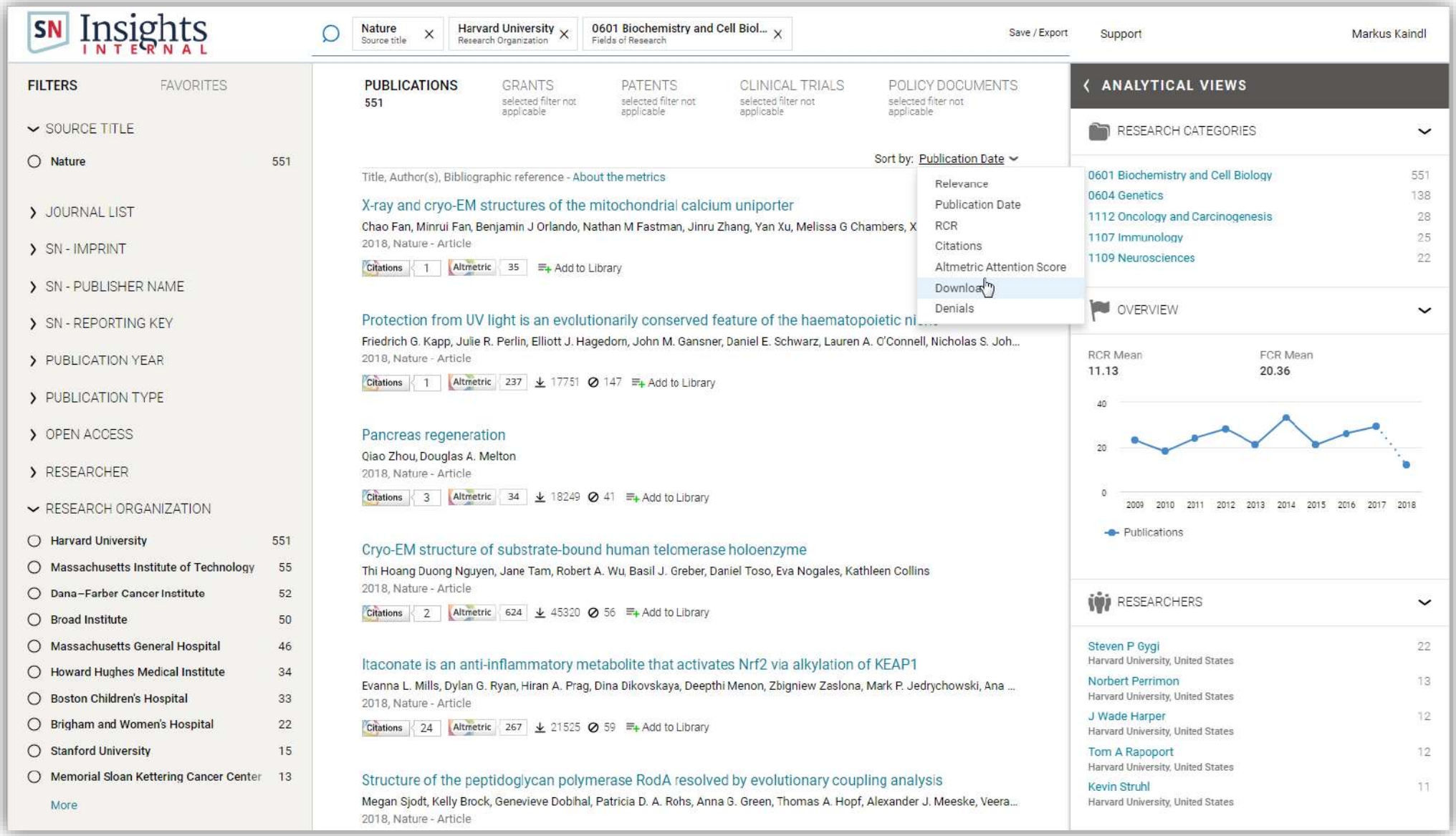
+ Need to enable analytics beyond SN on pan-publisher level

+ Add a lot more dynamic data (e.g. citations, usage statistics)

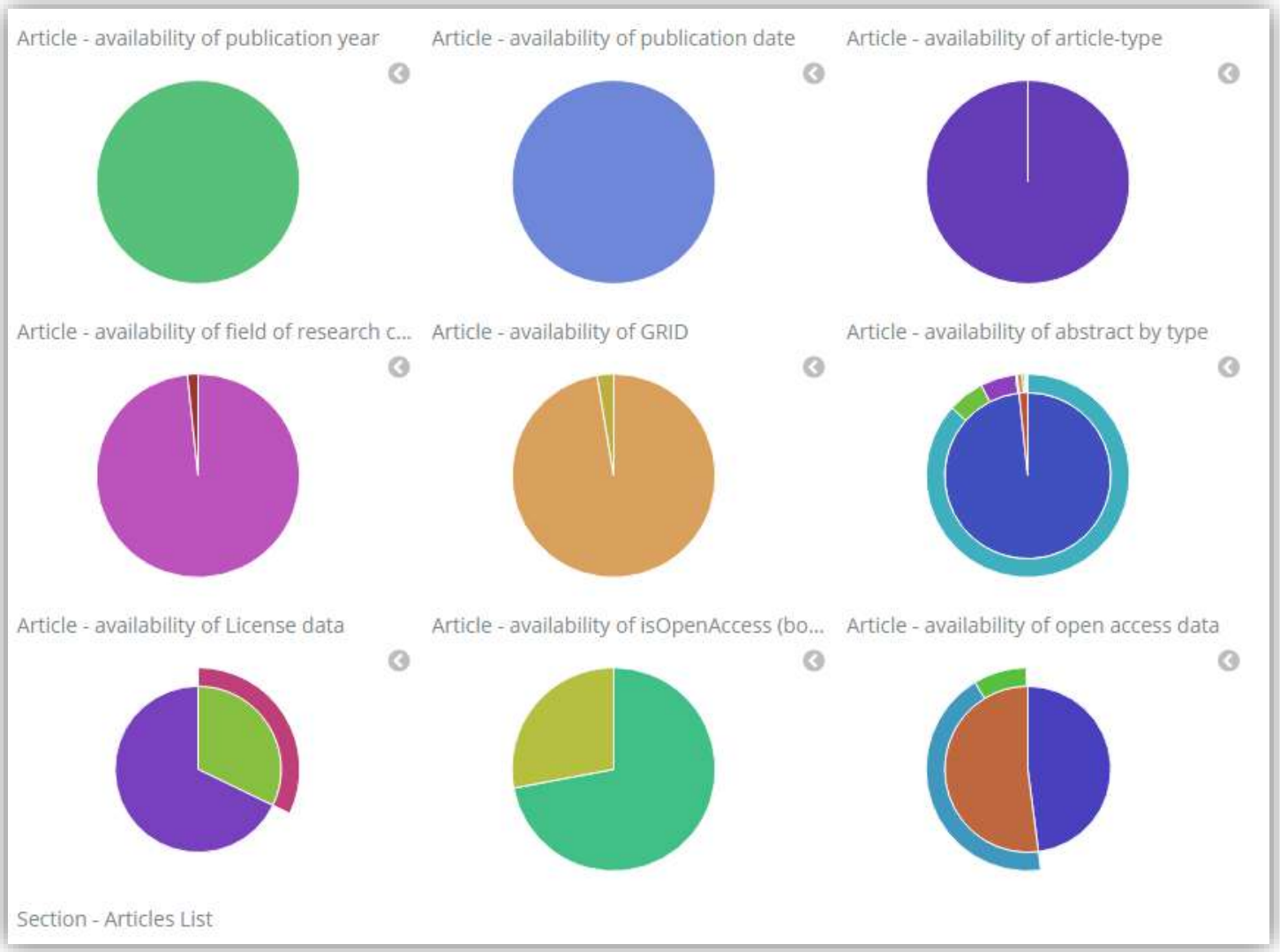
SN Insights & ELK supports QA

SN Insights is powered by Dimensions

SN Insights: Sneak Peek



Using ELK for internal QA going forward



Thank you!

How to keep in touch:

- **SN SciGraph Product Manager**
  markus.kaindl@springernature.com

- **SN SciGraph Portal**
  https://www.springernature.com/scigraph