

SN SciGraph

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

Los Angeles
2018

Michele Pasin

Lead Data Architect
Technical Manager SN SciGraph

SPRINGER NATURE

Agenda

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

SPRINGER NATURE

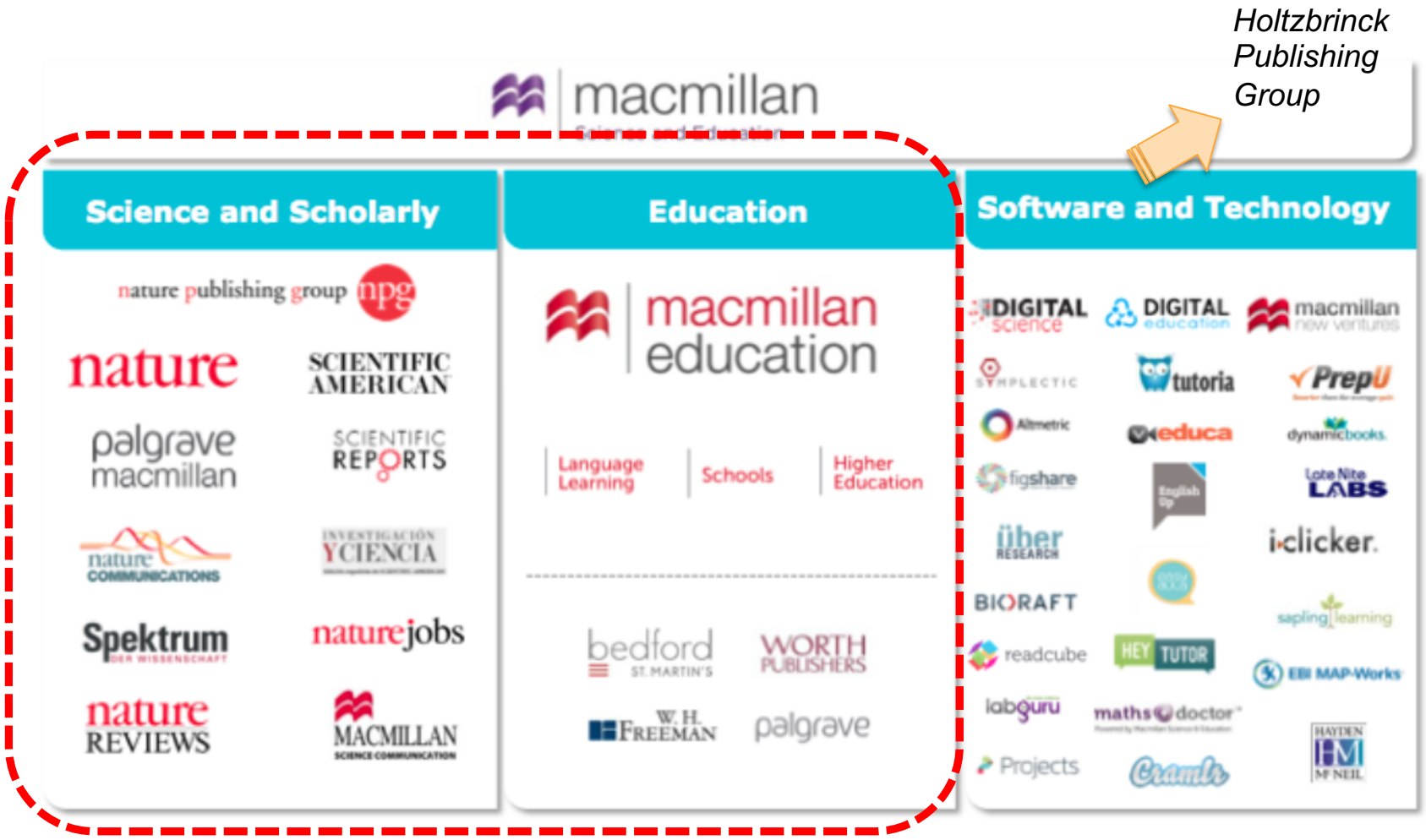


Formed in **May 2015** through the **merger** of Nature Publishing Group, Palgrave Macmillan, Macmillan Education and Springer Science+Business Media

[Pre-Merger] Springer Science + Business Media brands



[Pre-Merger] Macmillan Science & Education brands

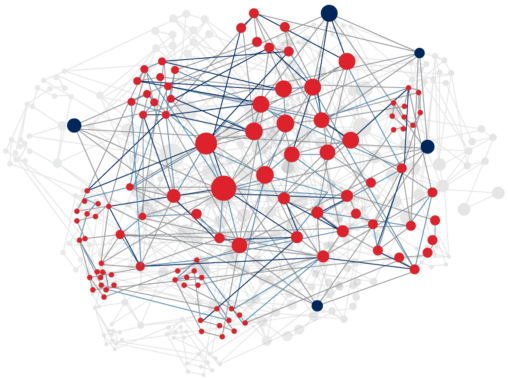


SPRINGER NATURE

Springer Nature SciGraph

A Linked Open Data platform for the scholarly domain

SN SciGraph



> Collaborative effort between Springer Nature and Digital Science (mid 2016)

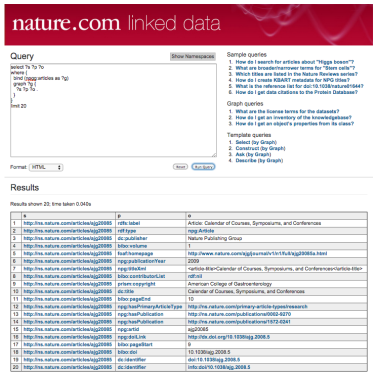
> Increasing discoverability of content by using linked data and semantic technologies

> Supporting internal use cases,but also contributing to an emerging web of **linked science data**

www.springernature.com/scigraph

A Rich History..

NPG Linked Data Platform



CURI Semantic Annotation Project

Subject Pages

Scigraph prototype

2012

2013

2014

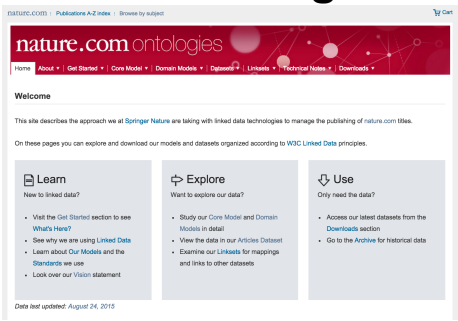
2015

2016

Linnaeus Project



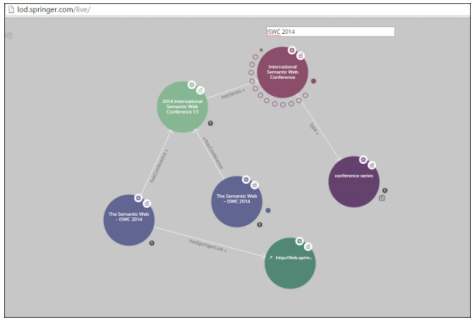
Nature Ontologies Portal





Nero Project

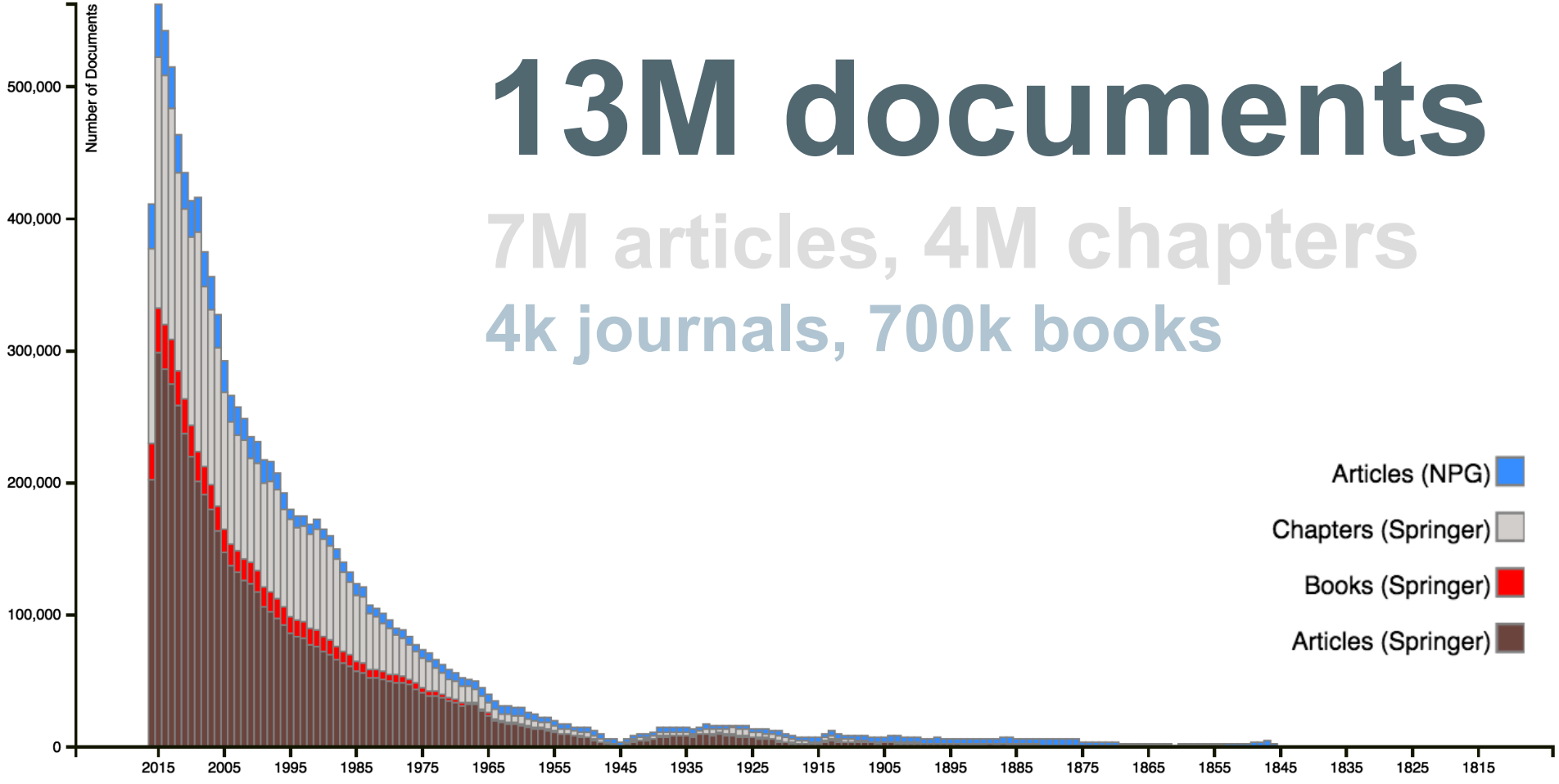
Springer Conferences



Springer Protocols



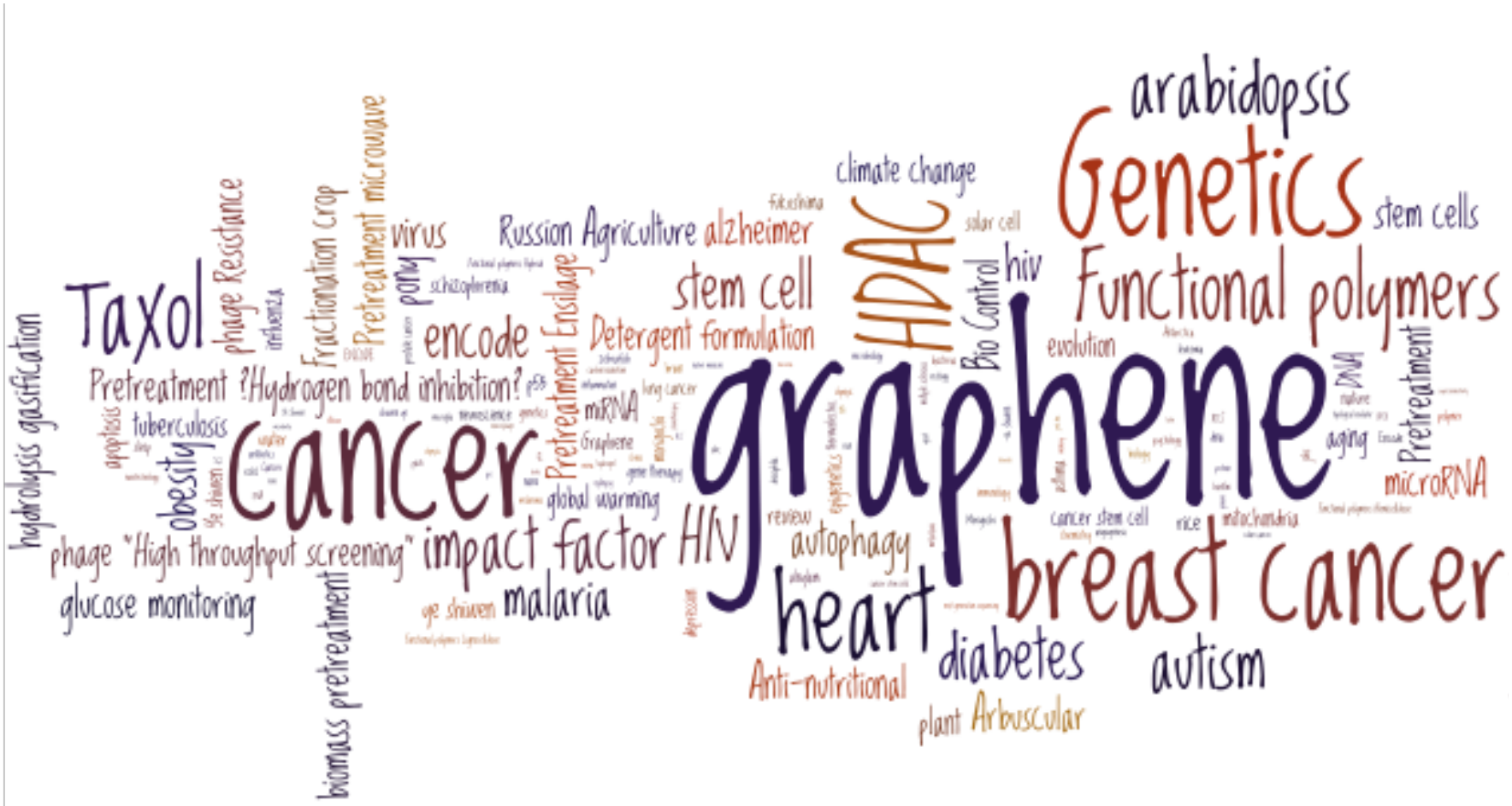
We publish a lot of science! (since 1815)



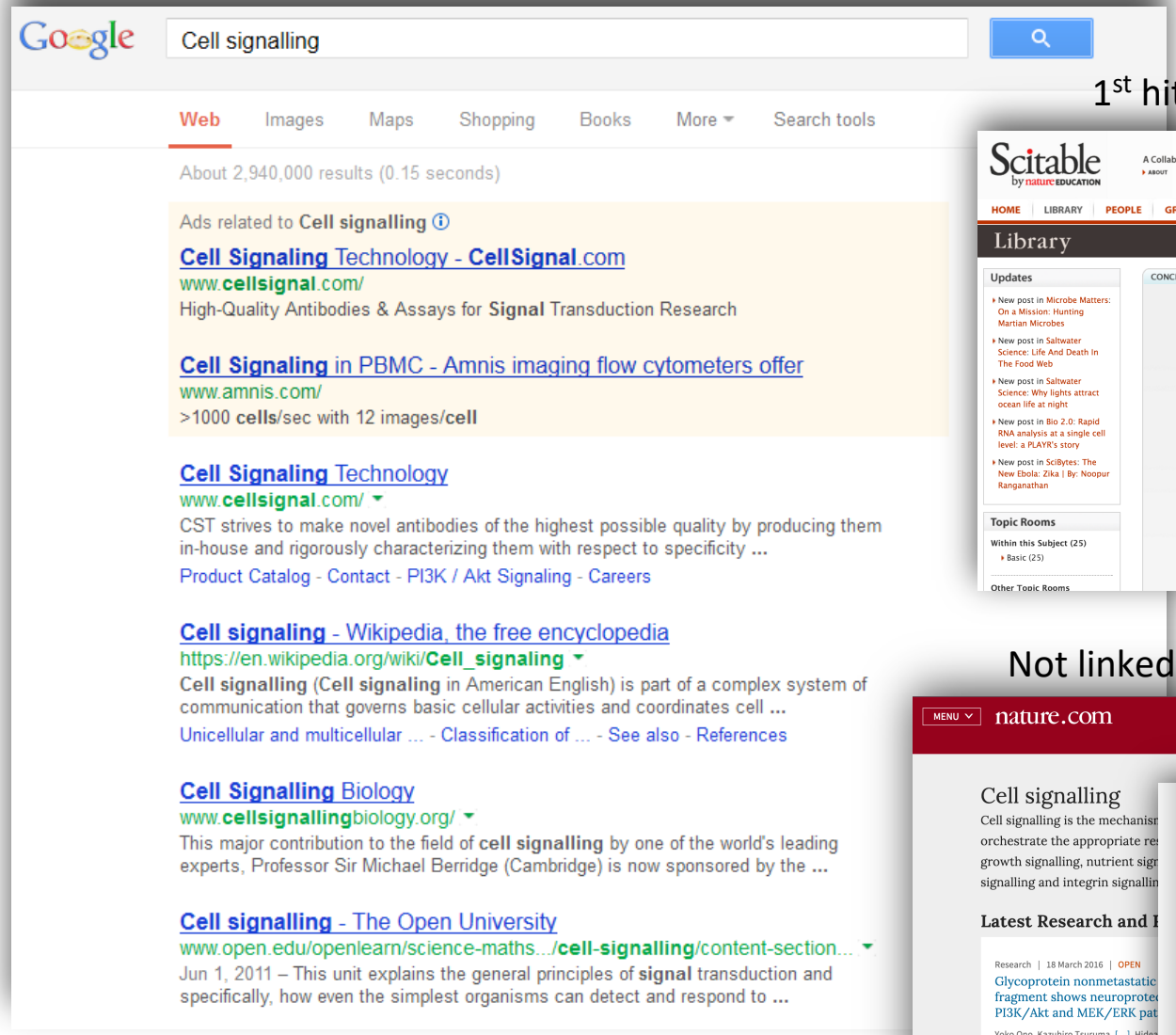
For example: our sites are currently organised around articles, journals and issues…



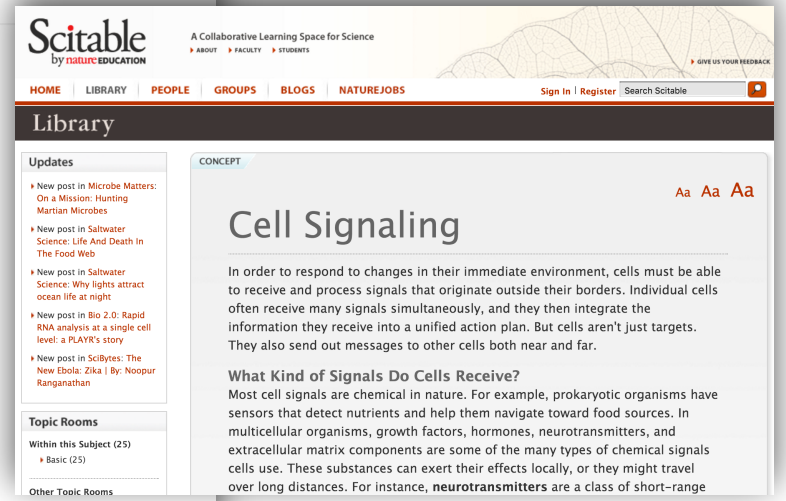
However, scientists are interested in answering questions about real world things…



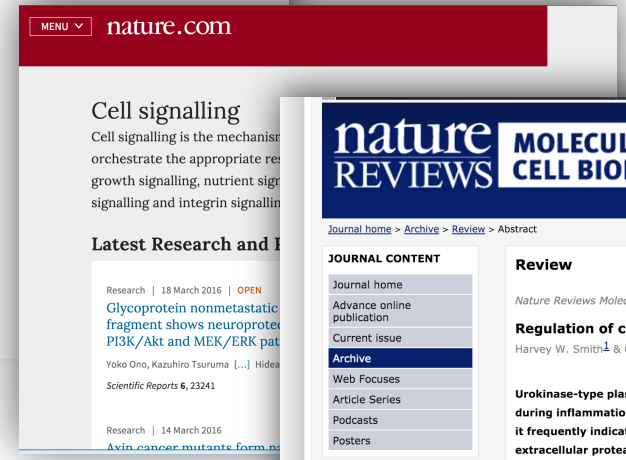
Search engines do not know we have content about these things…

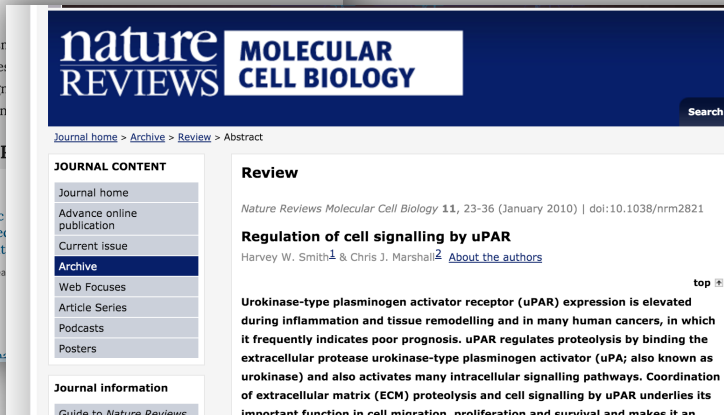


1st hit from nature.com…



Not linked to/from..

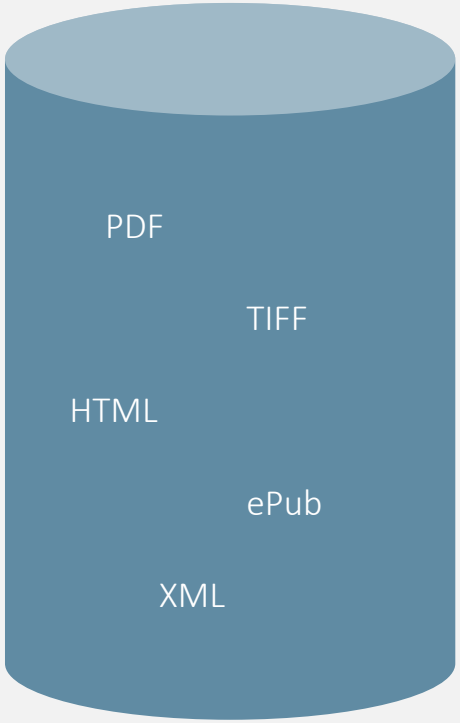




..and neither do we!

Vision

Today: Content base



We publish science

Tomorrow: Knowledge Graph

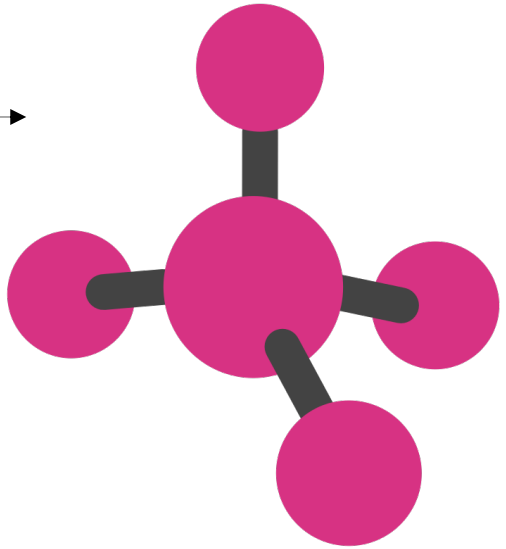


We manage knowledge

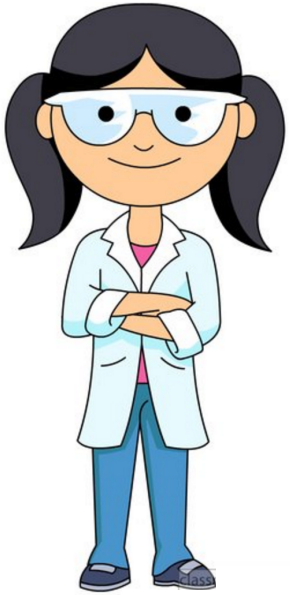
Three areas of knowledge we care about



is about



reads / writes



interested in



Produces

Has topic

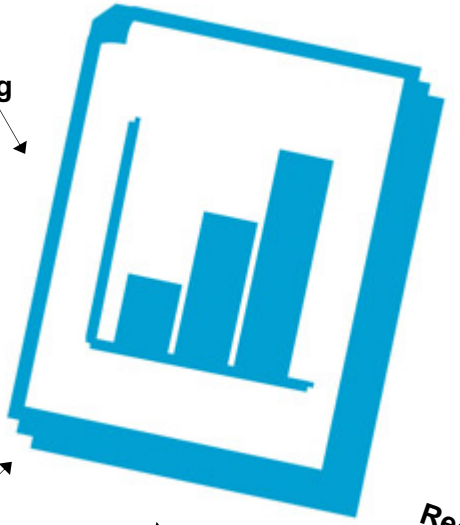


YouTube

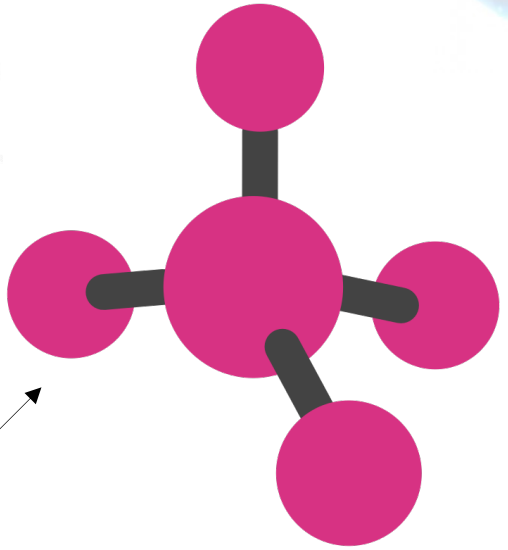
In proceedings



Has learning resource



Attends



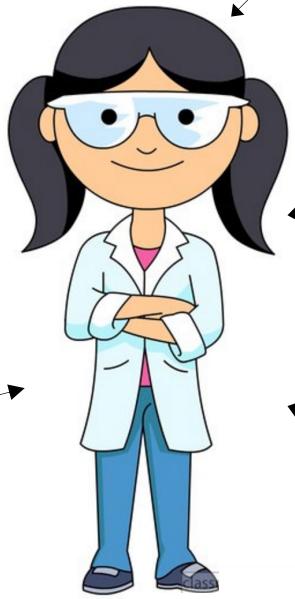
Studies

Contains

Produces

flickr

Reads / Writes



Lead researcher in



Cites



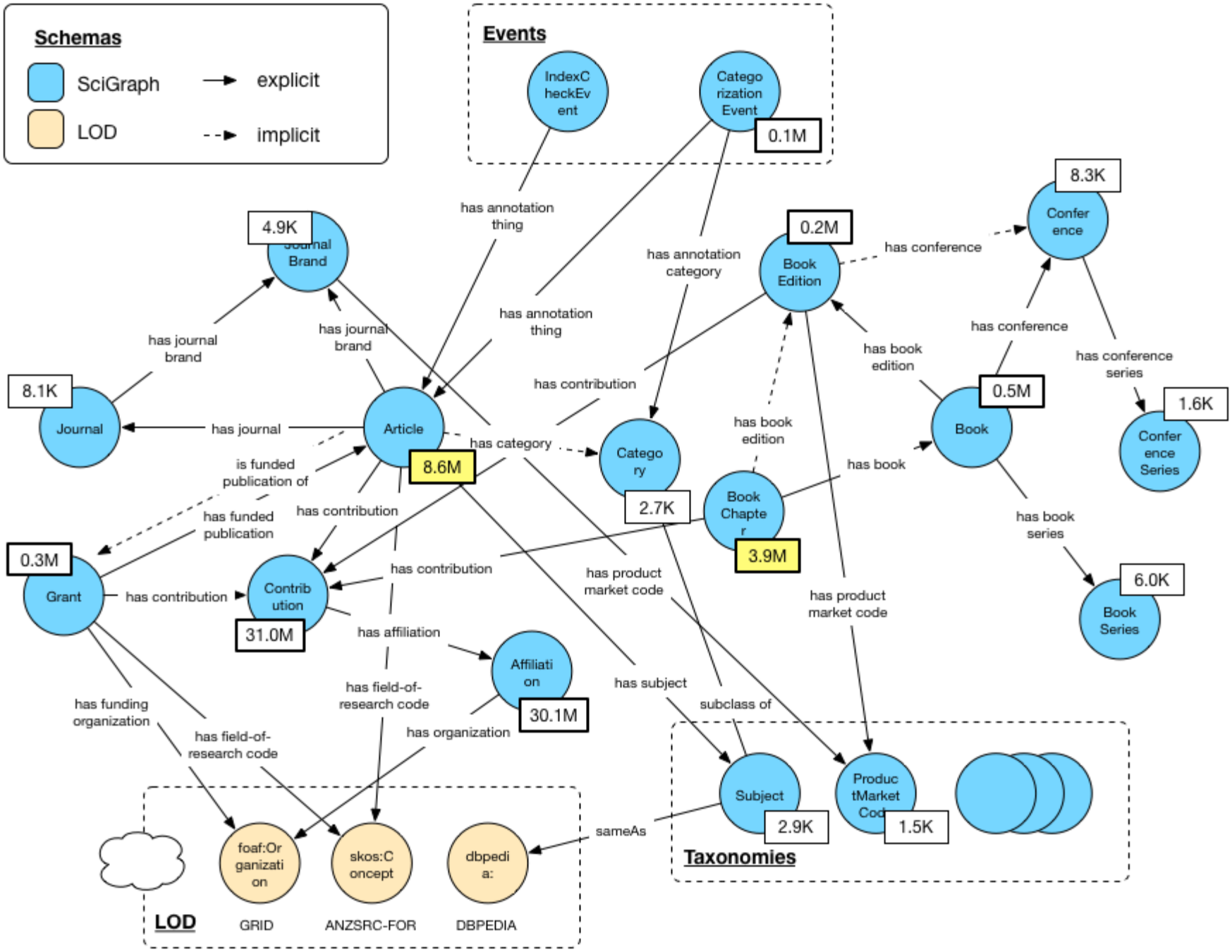
Funds

Works for

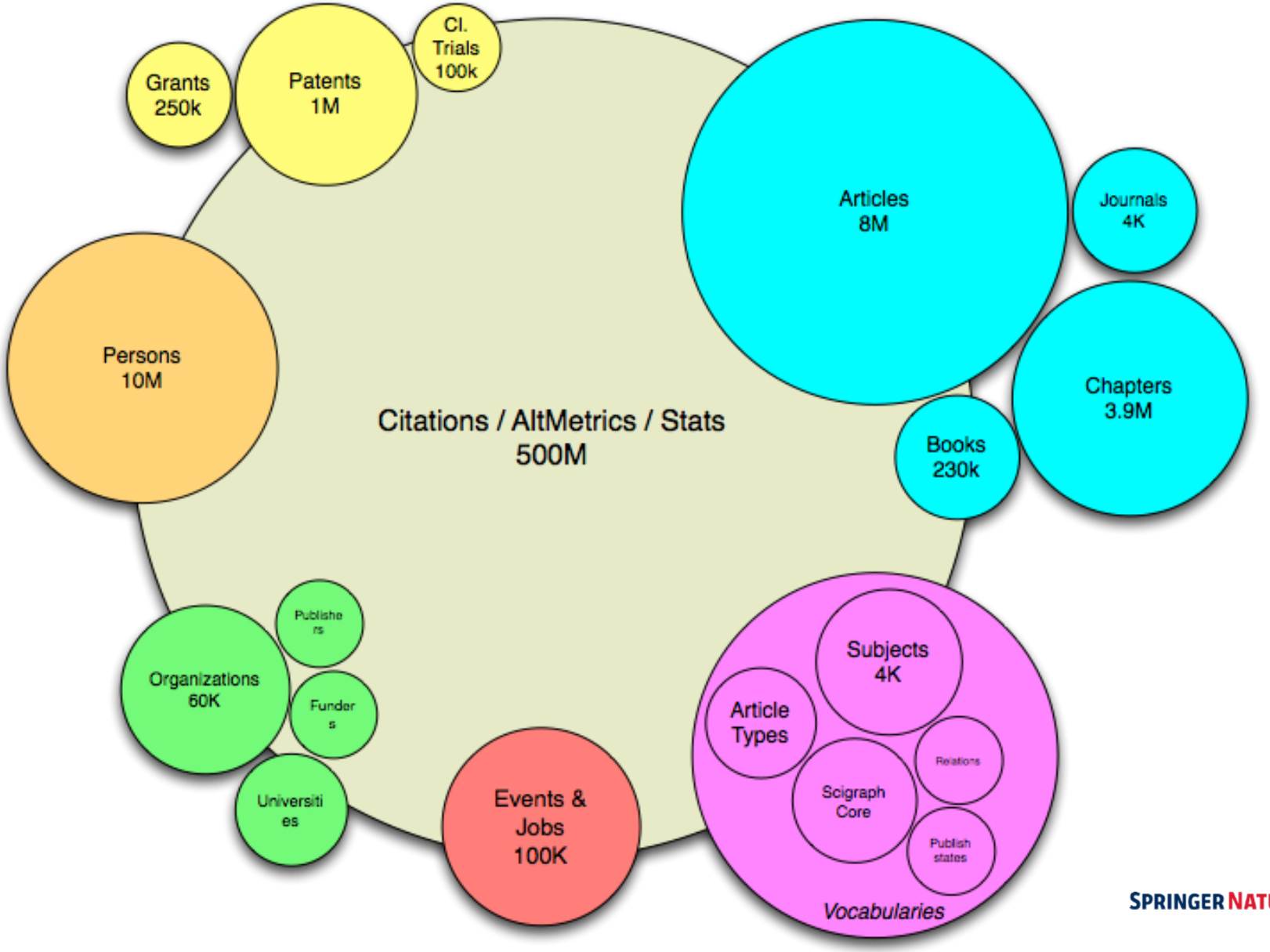


Located at

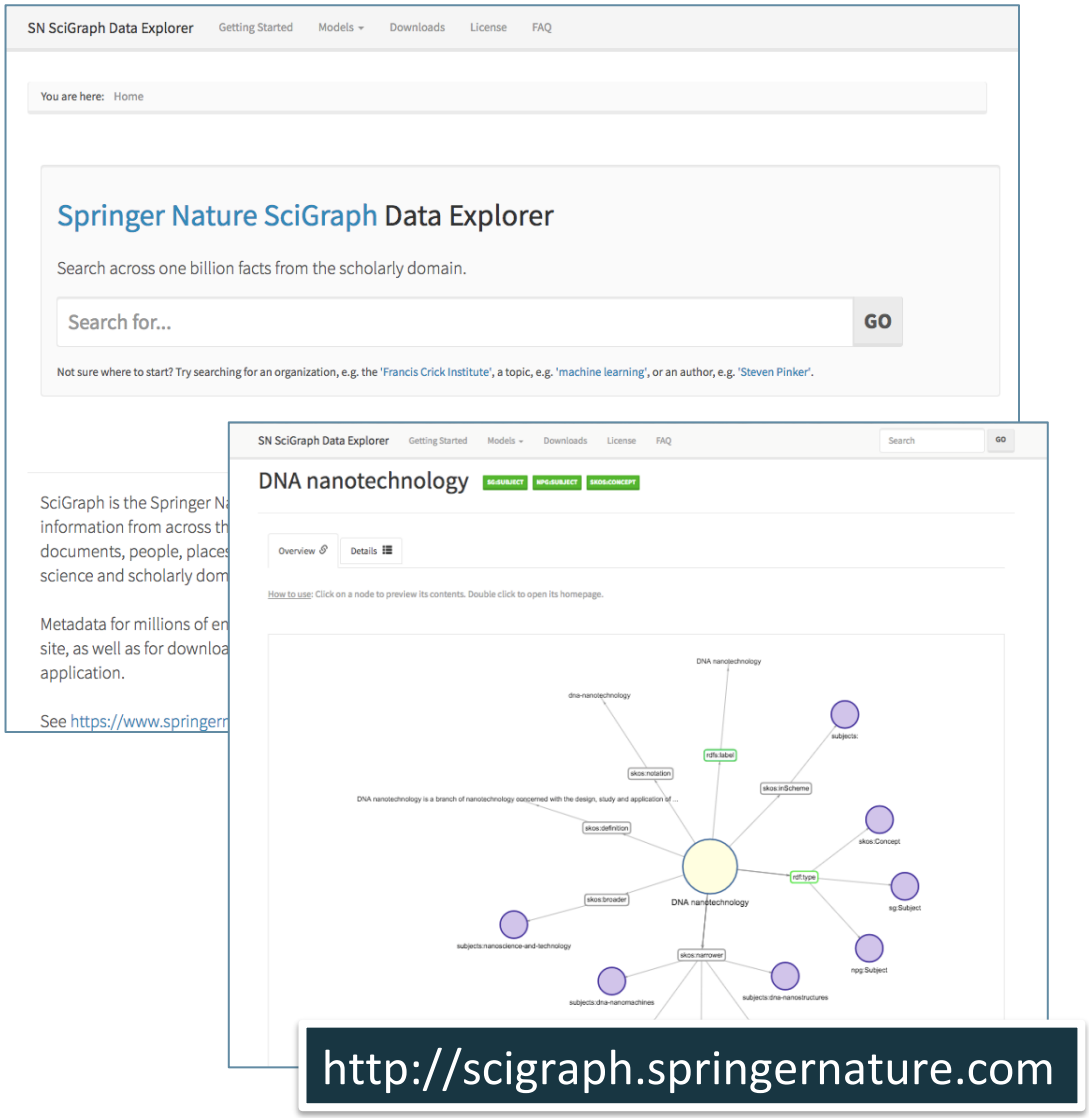
NATURE


Schemas
SciGraph
LOD
explicit
implicit
Events
IndexCheckEvent
Categorization Event
0.1M
has annotation thing
has annotation category
4.9K
Journal Brand
0.2M
Book Edition
8.3K
Conference
has conference
has journal brand
has book edition
has conference
has conference series
8.1K
Journal
has journal
Article
8.6M
has contribution
has category
Category
2.7K
Book Chapter
3.9M
0.5M
Book
1.6K
Conference Series
is funded publication of
has funded publication
has book
has book series
0.3M
Grant
has contribution
Contribution
31.0M
has product market code
6.0K
Book Series
has affiliation
Affiliation
30.1M
has subject
subclass of
has funding organization
has field-of-research code
has organization
Subject
2.9K
ProductMarketCod
1.5K
Taxonomies
sameAs
foaf:Organization
skos:Concept
dbpedia:
LOD
GRID
ANZSRC-FOR
DBPEDIA

SN SciGraph Data Landscape



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:

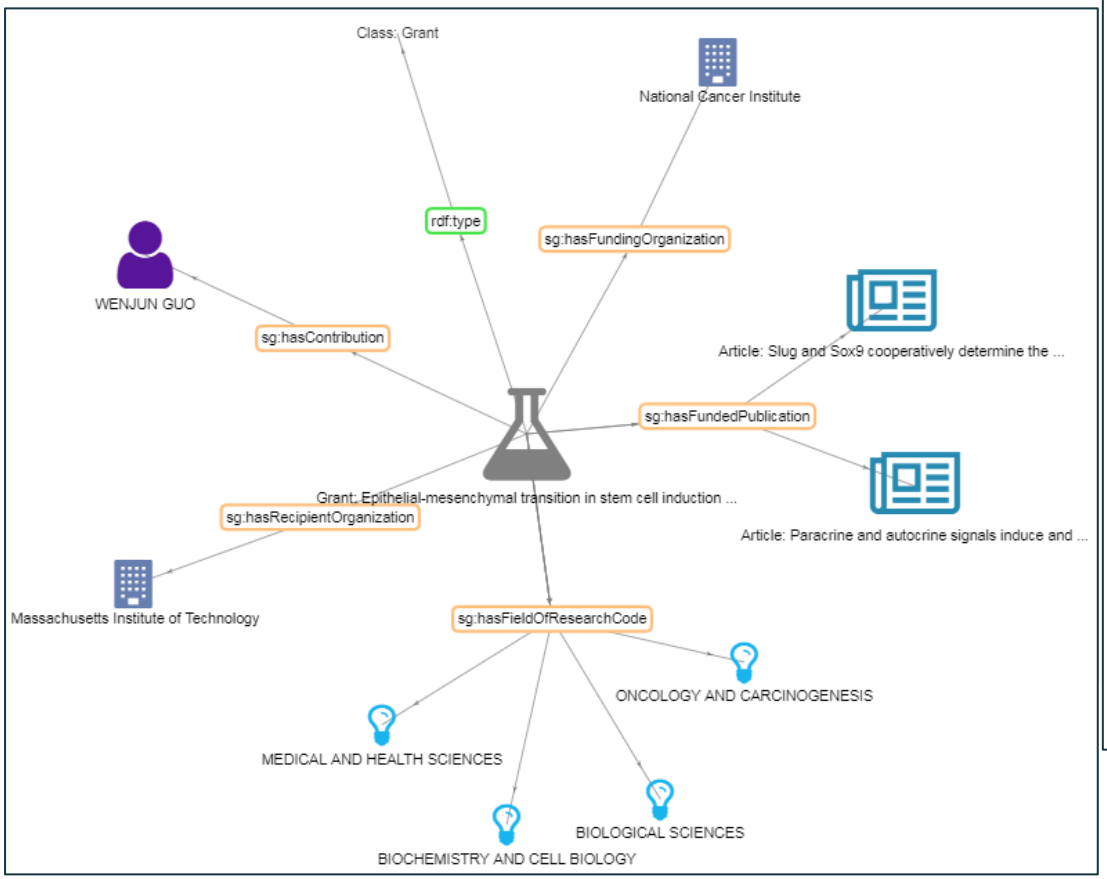
- Journals & Articles (8M) + Abstracts
- Books & Chapters (4M)
- Grants (200k)
- Subjects, Research Organizations, Conferences, Ontologies (100k)
- Links to to other datasets (20M): Crossref, GRID, MESH, DBpedia, Wikidata

SPRINGER NATURE

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

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**'

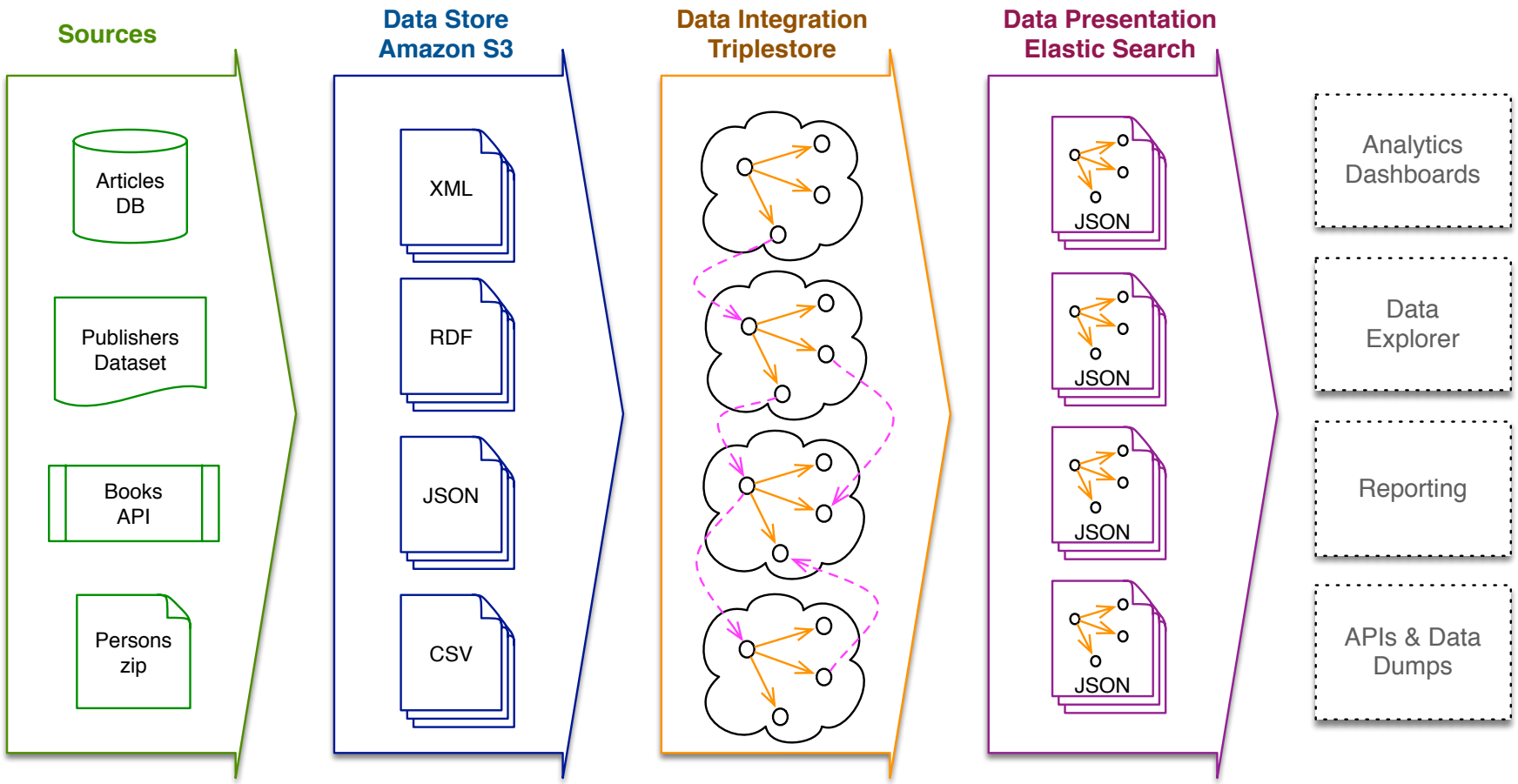
> Data sources versioning algorithm

> Identity Persistence service

> Validation via SHACL (TopBraid API)



ETL Architecture: data flow



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

→

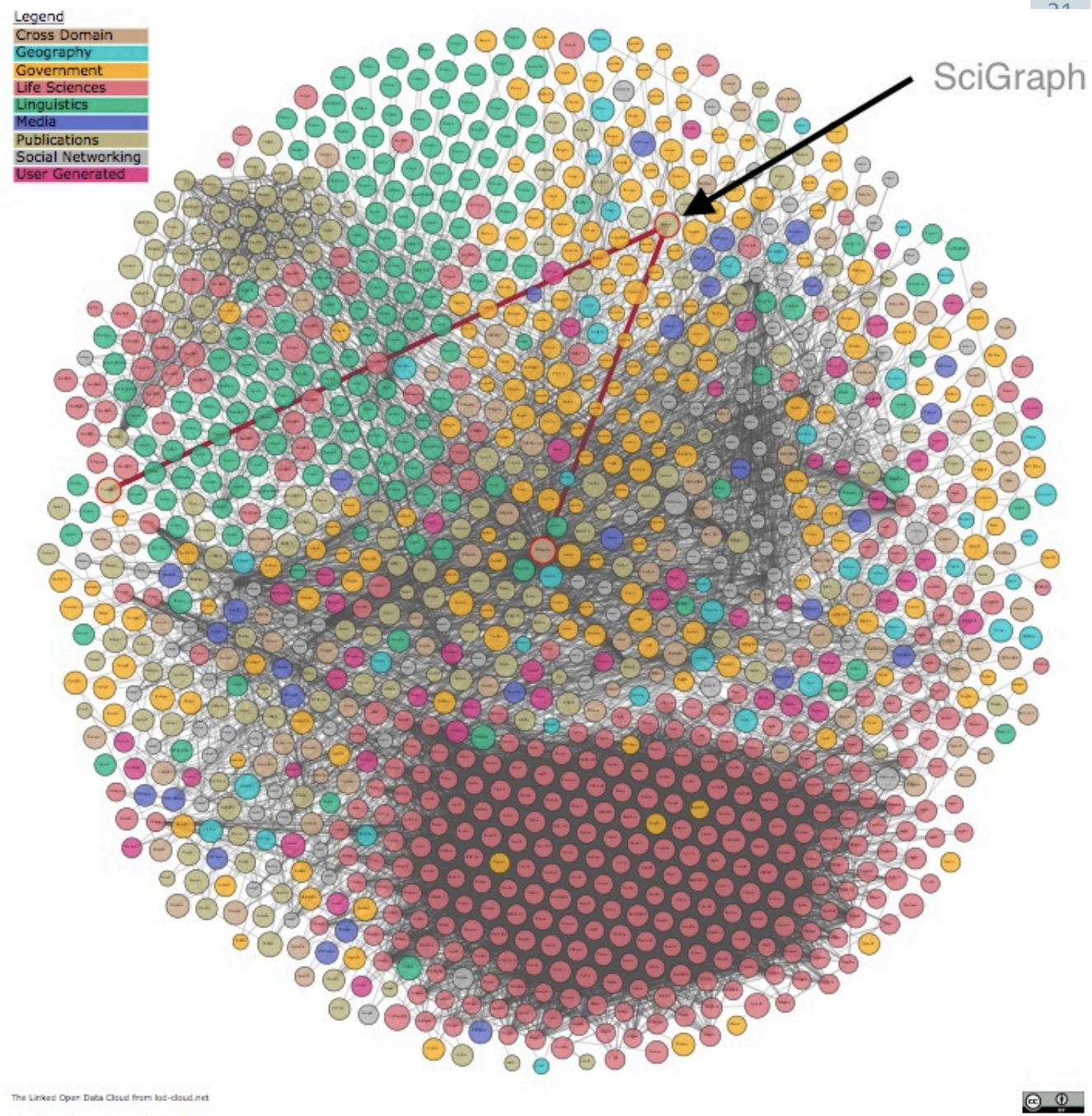
* Extraction
* Validation (SHACL)
* Identity Resolution
* Inference (OWL)

→

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

Web of Data

- Be part of the LOD cloud!
- The future library and metadata catalog is the Internet!



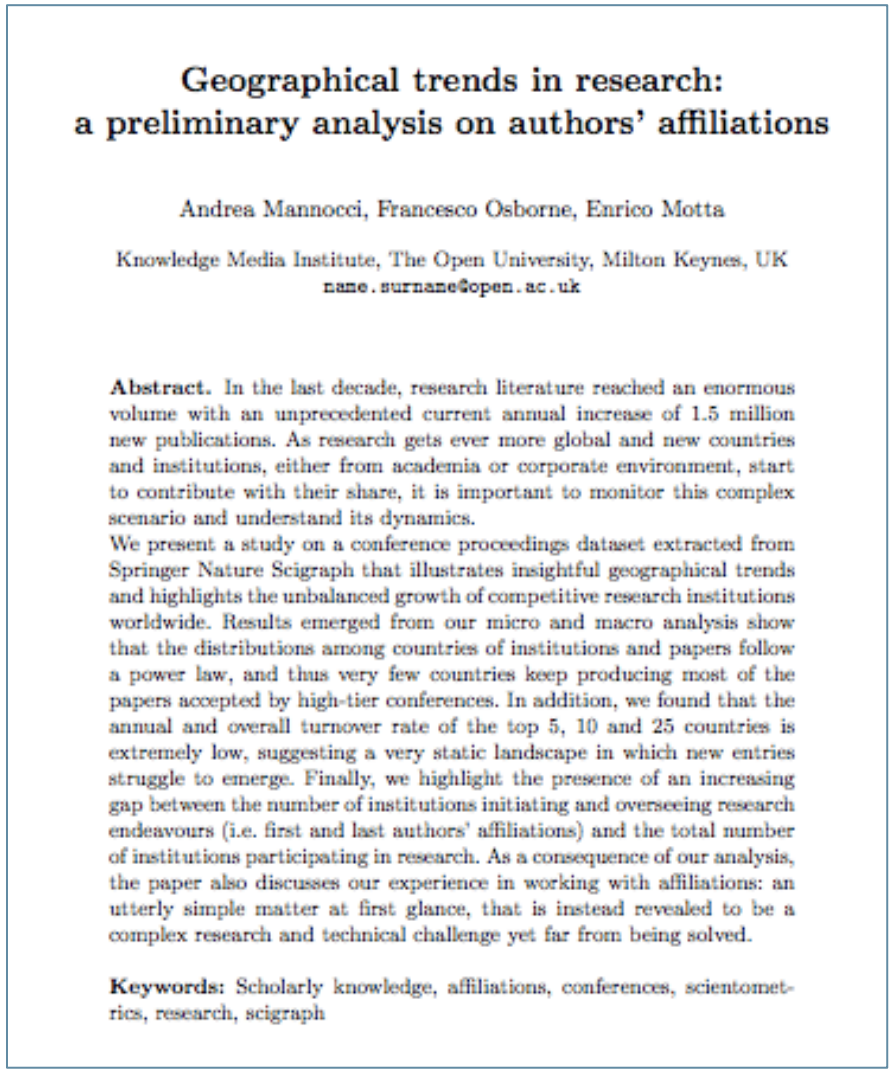
Springer Nature Hack Day Series Started 2017 In London

- **Aims and Scope**
  - Position ourselves as Open Data research publisher
  - Gather first-hand feedback from potential users of our data
  - Allow trial & error with swift prototypes and proofs of concept
  - Serve as innovation hub for colleagues and industry partners
  - Build strong relationships with top research institutes



Hopper Hosts Hackathon with Elasticsearch, Igor Motov. https://www.elastic.co

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

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 oth catalogue (or part of) was transformed and made a uld still rely on traditional data standards like M

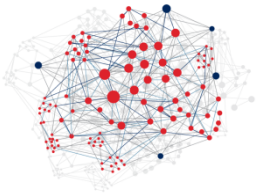
This is now starting to **standing of the opportunities** opened t the information management lifecycle.

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

Can LOD help sharing data with Libraries?

ELK for Business

SN SciGraph



Recap: SN SciGraph 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.

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 book or journal • 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

Challenges with triplestores / 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

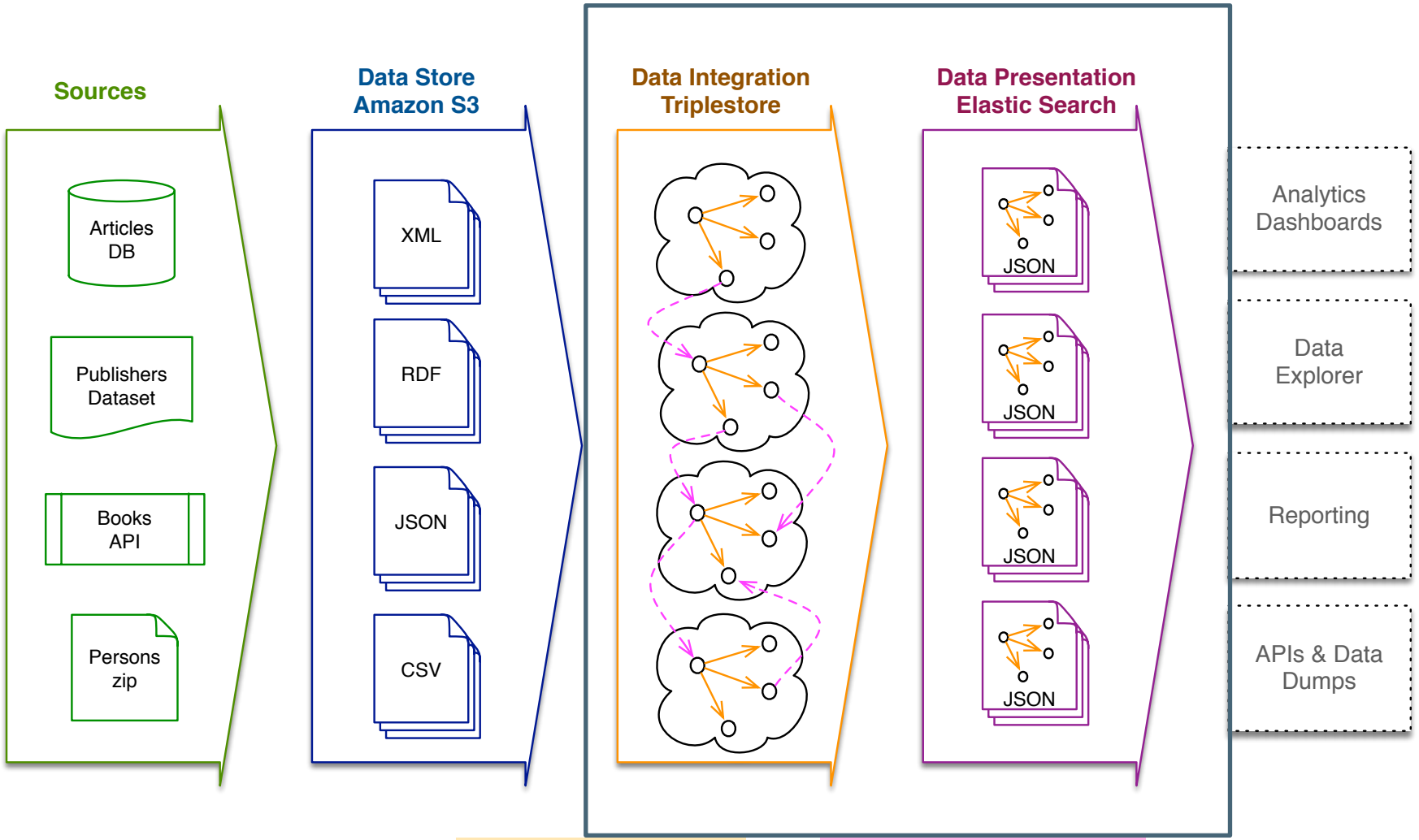
> Very fast (scales linearly)

> GraphDB has built-in connector service

> Kibana offers powerful data visualizations out of the box

> It’s free!

Why the ELK stack [2]



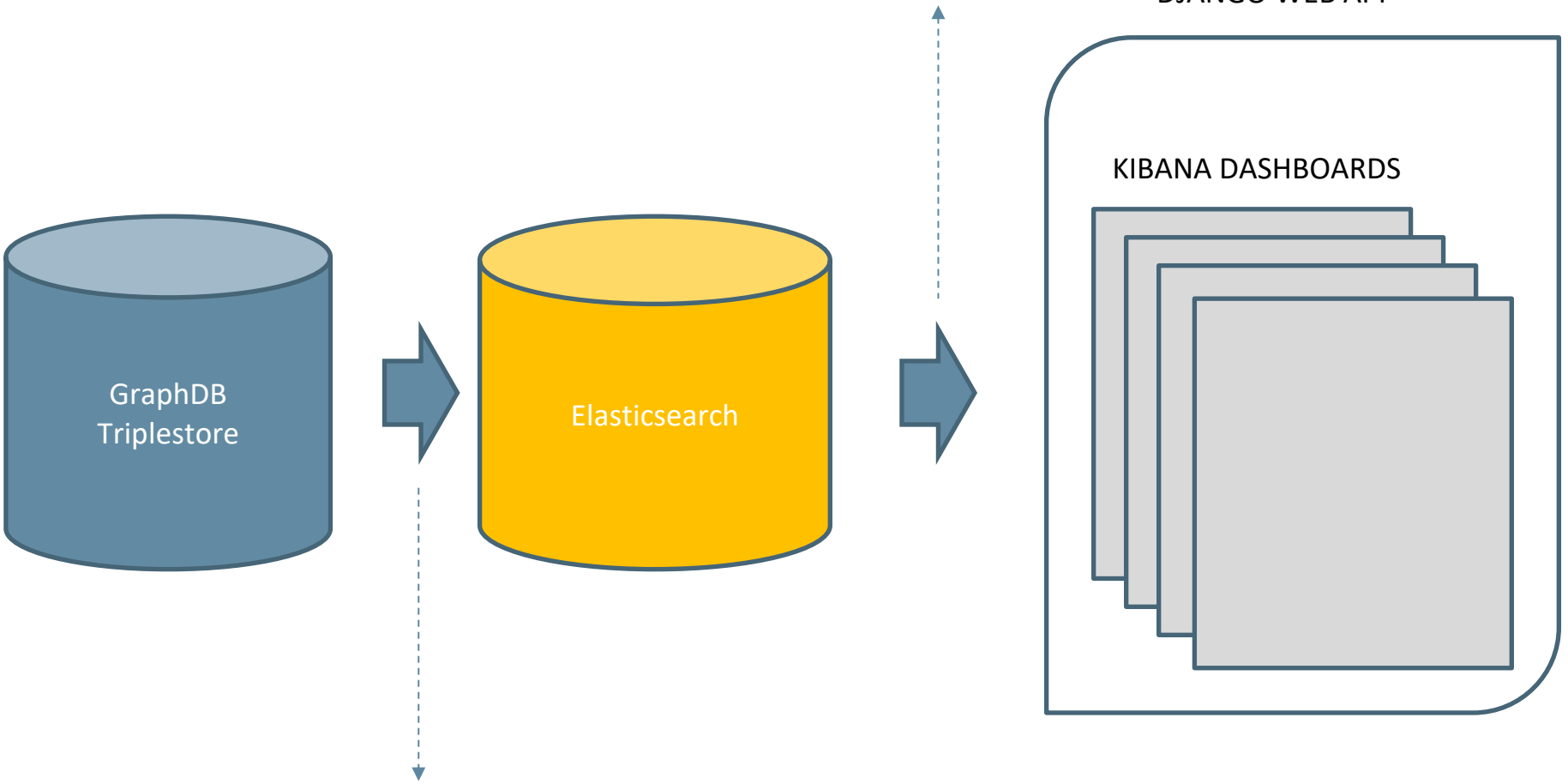
* Versioning service (md5 checksum, timestamps, origin version, etc...)

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

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

Solution: embedded & dynamic Kibana dashboards

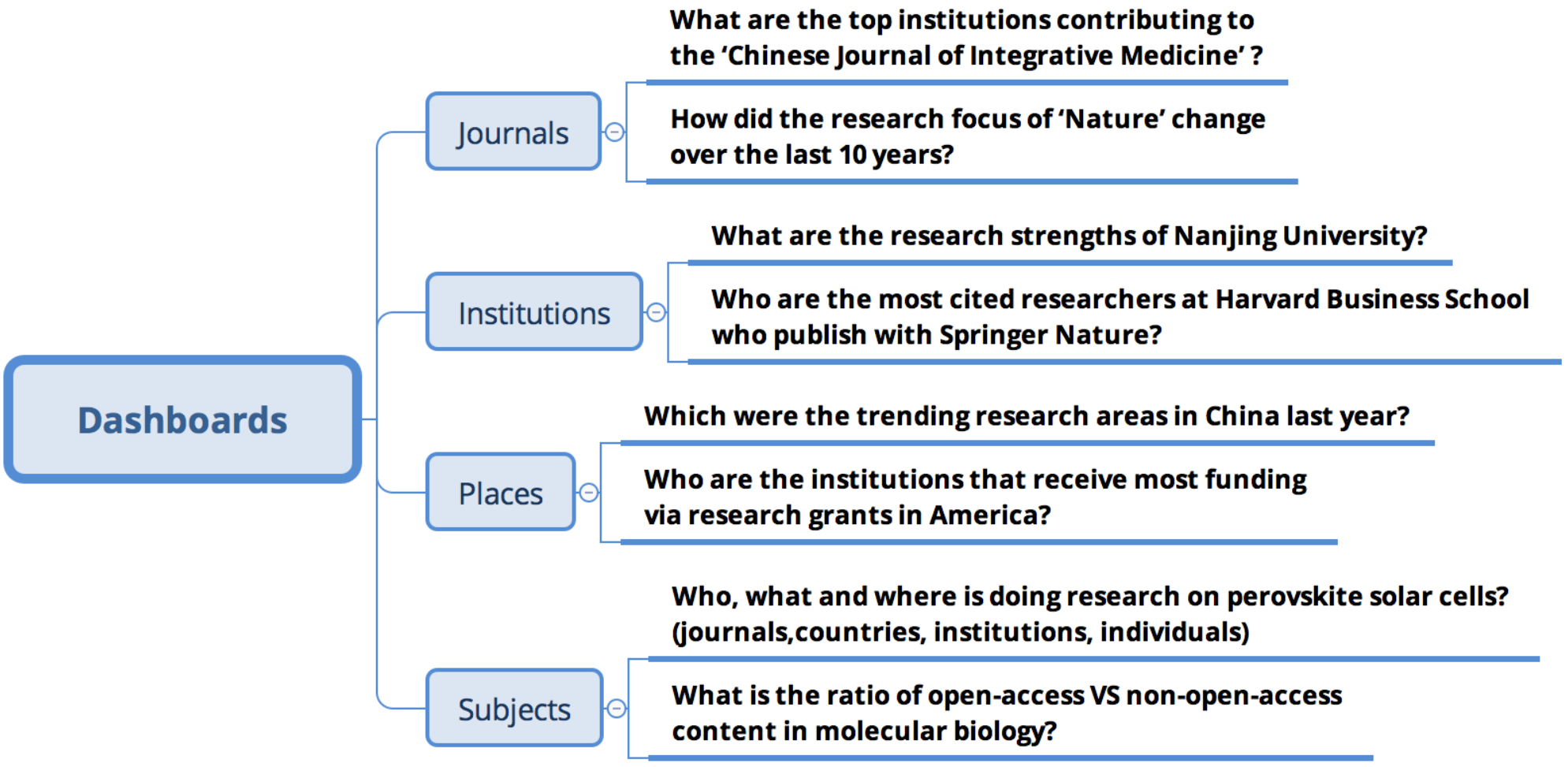
Kibana dashboards are created by analysts
And parametrized based e.g. on journal IDs

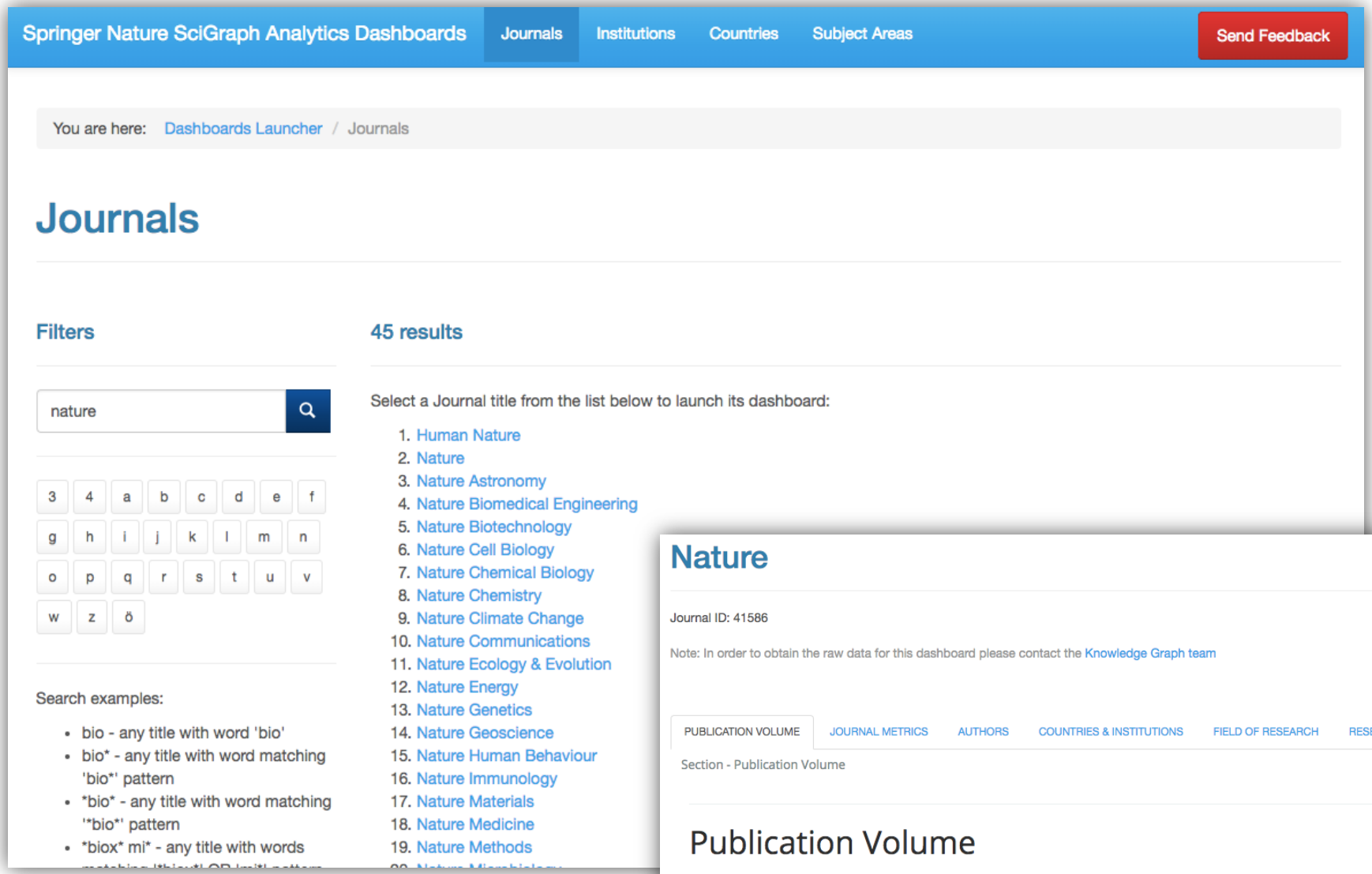


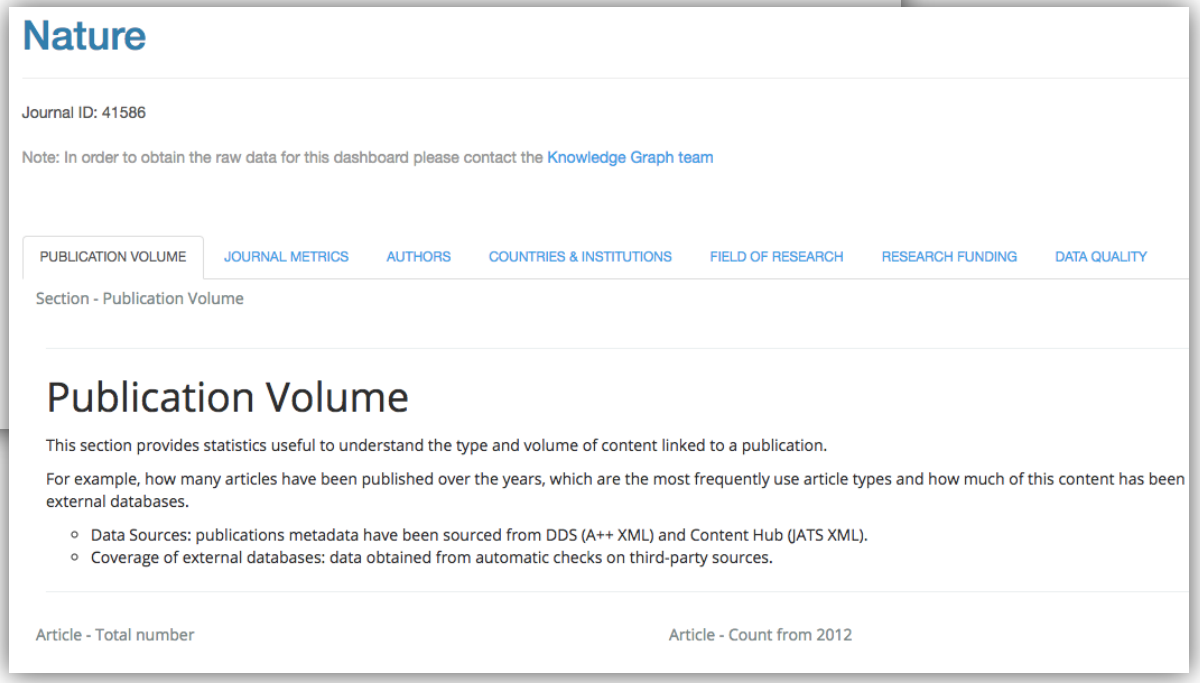
Elasticsearch connectors
(automatic updates / deletions after each ingest)

Analytic Dashboards Examples

High level architecture




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

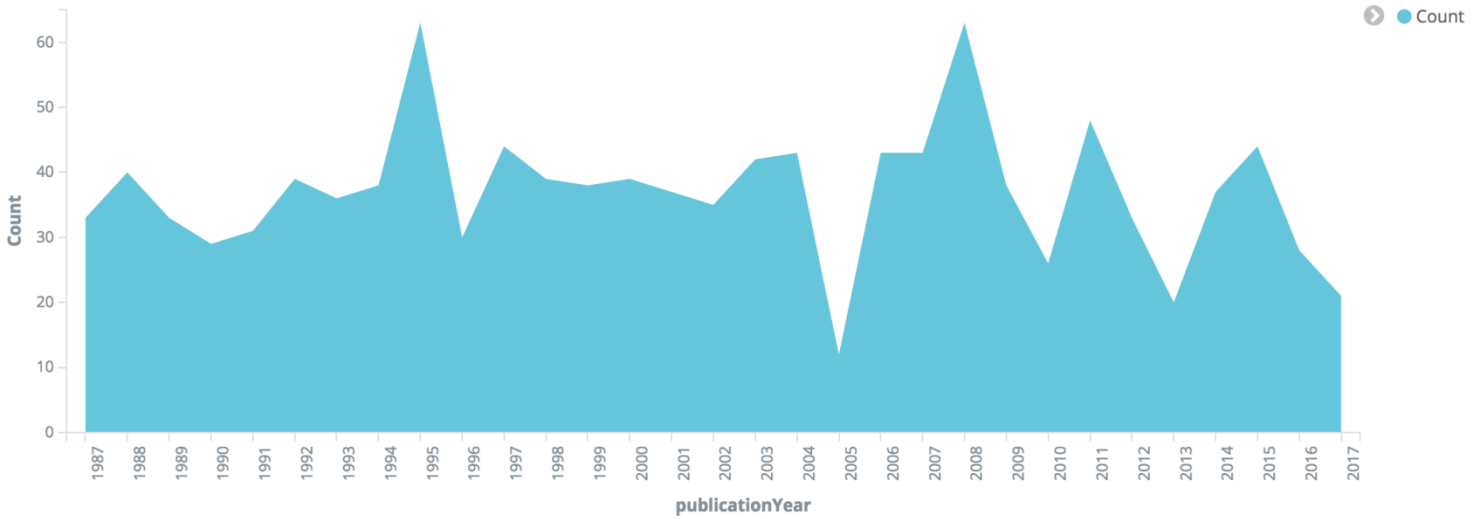
Articles in Total

Article - Count from 2012

150

Articles Published Last 5 Years

Article - Count by publication year



Geographical Patterns

BMC Women's Health

Journal ID: 12905

Note: In order to

PUBLICATION VO

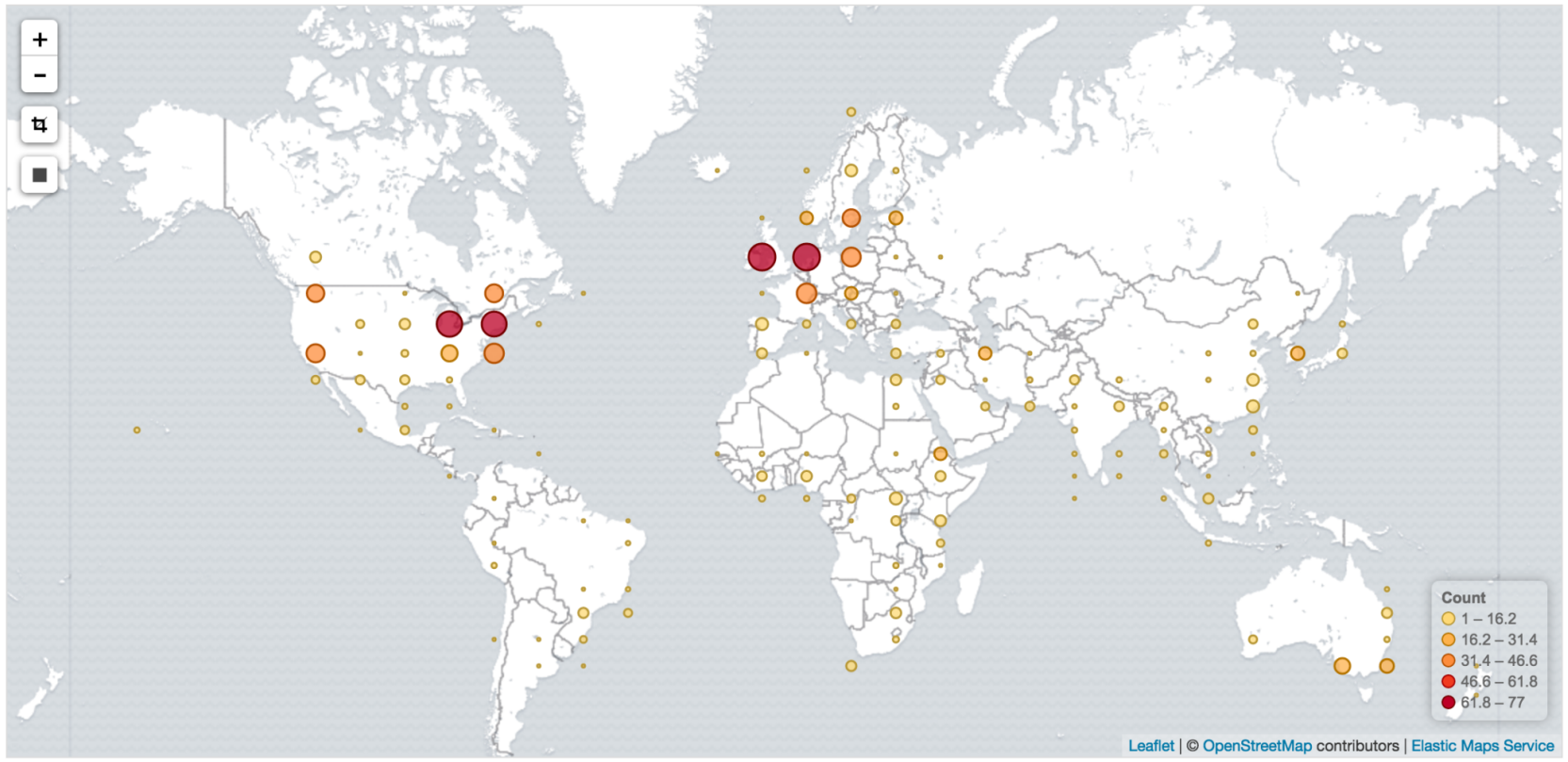
Section - Count

Coun

Use this secti

- Data Sc

Article - map view



Geographical Patterns over time

BMC Women's Health

Journal ID: 12905

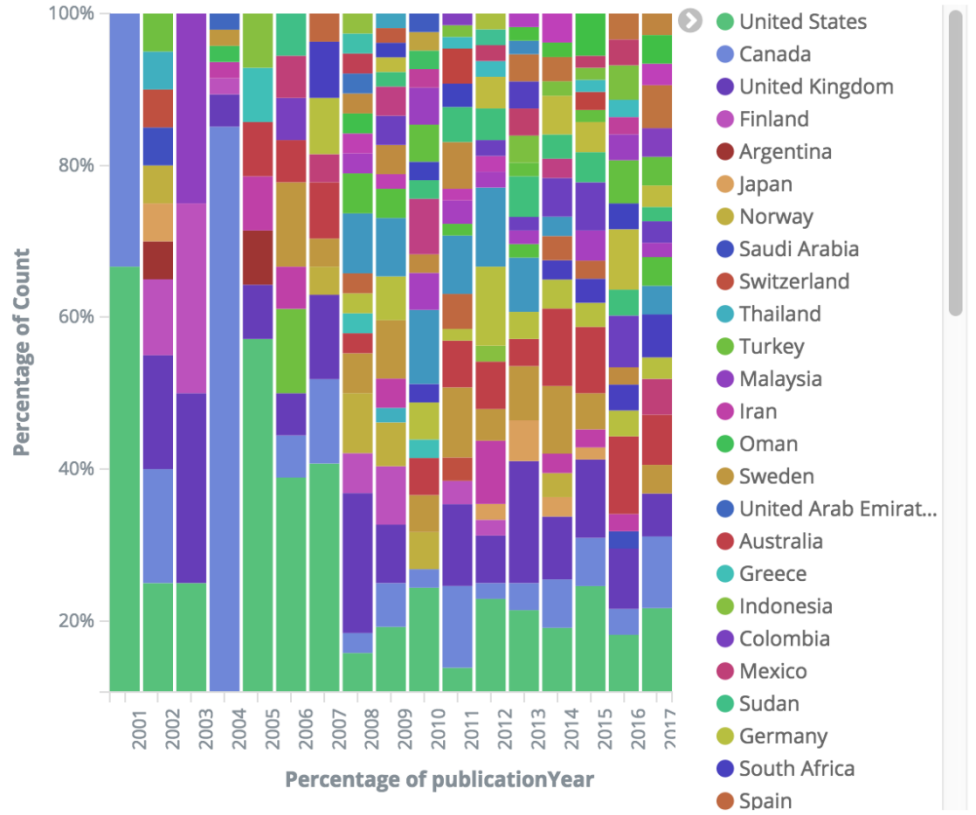
Article - by country

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 »

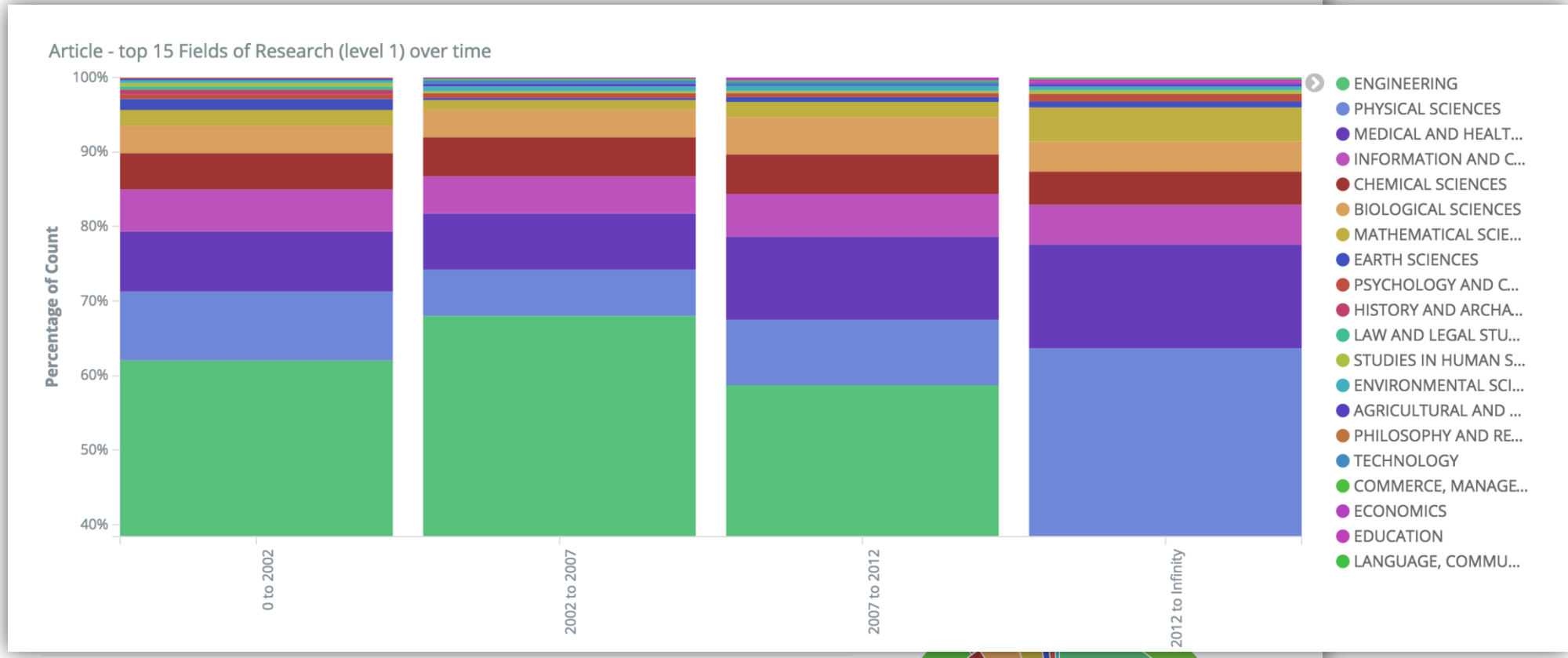
Article - top ten countries distribution as a timeline



Fields of Research over time

Experiments in Fluids

Journal ID: 348



Research Funding

Analytical and Bioanalytical Chemistry

Journal ID: 216

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

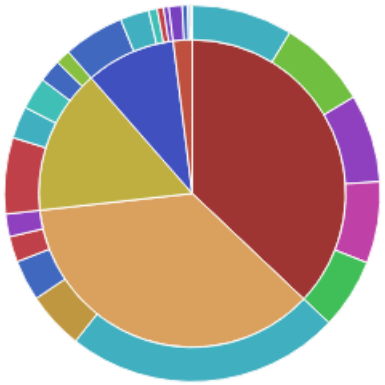
Section - Rese

Rese

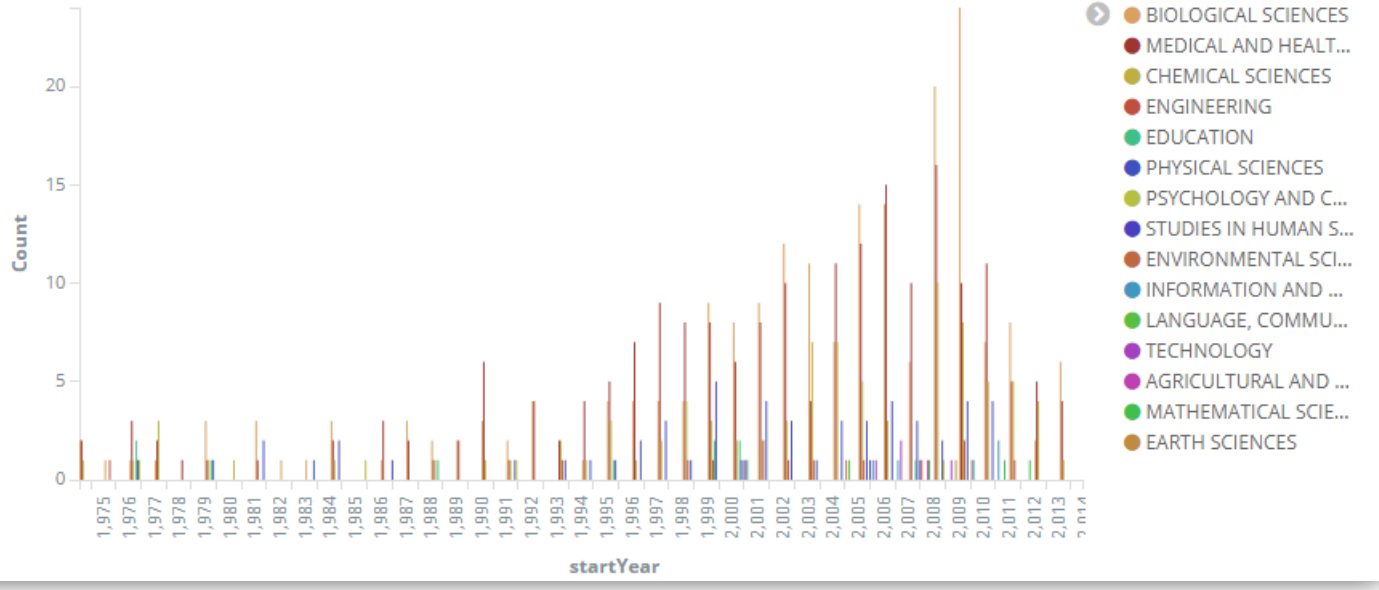
This section

- Data S linked

Grants - research funding by research a...



Grants - count of research grants per year and subject area



Current Work

1.
Schema.org and JSONLD

2.
Scaling up analytics using Dimensions

SN SciGraph: increasing reusability

Going native with schema.org

- Integration with de-facto linked data standard on the web
- Simpler model for documents (vs FRBR)
- Simpler URIs (e.g. use DOIs where possible)

Going native with JSONLD

- Easier to work with and consume
- Fully RDF-compatible, but can hide the complexities of RDF for those who don’t need it
- Better support for lists, intuitive way to embed related objects

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 ...

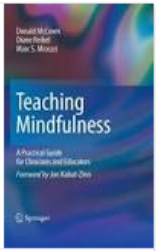
Mindfulness and Buddhist-Derived Approaches in Mental Health and ...
https://books.google.com/books?isbn=3319222554
Edo Shonin, William Van Gordon, Mark Griffiths - 2015 - Psychology
Competence in **teaching mindfulness**-based courses: Concepts, Development and assessment. Mindfulness ... New York: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

Analytics dashboards: scaling up

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)

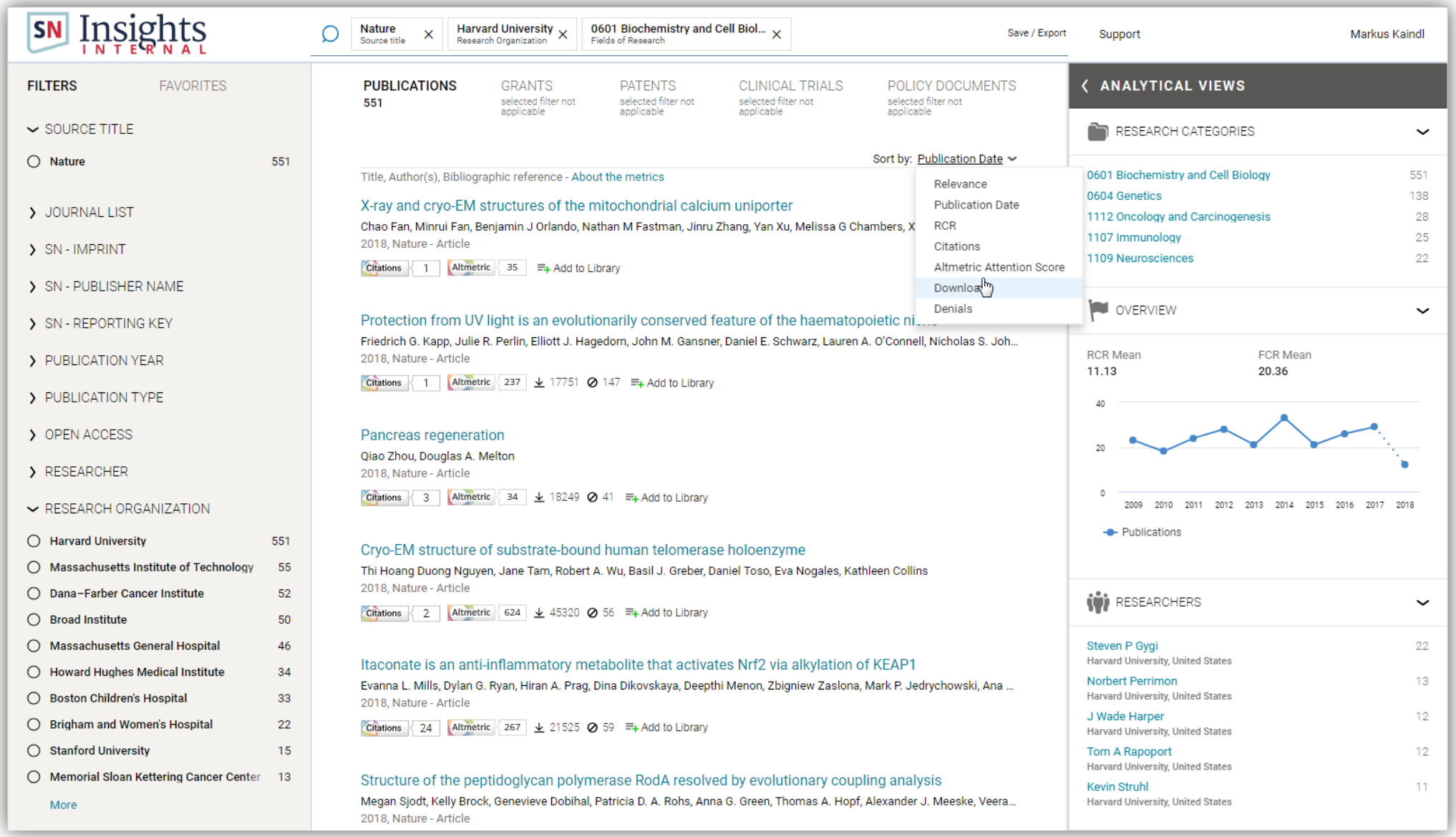
+ Need to enable analytics beyond SN on pan-publisher level

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

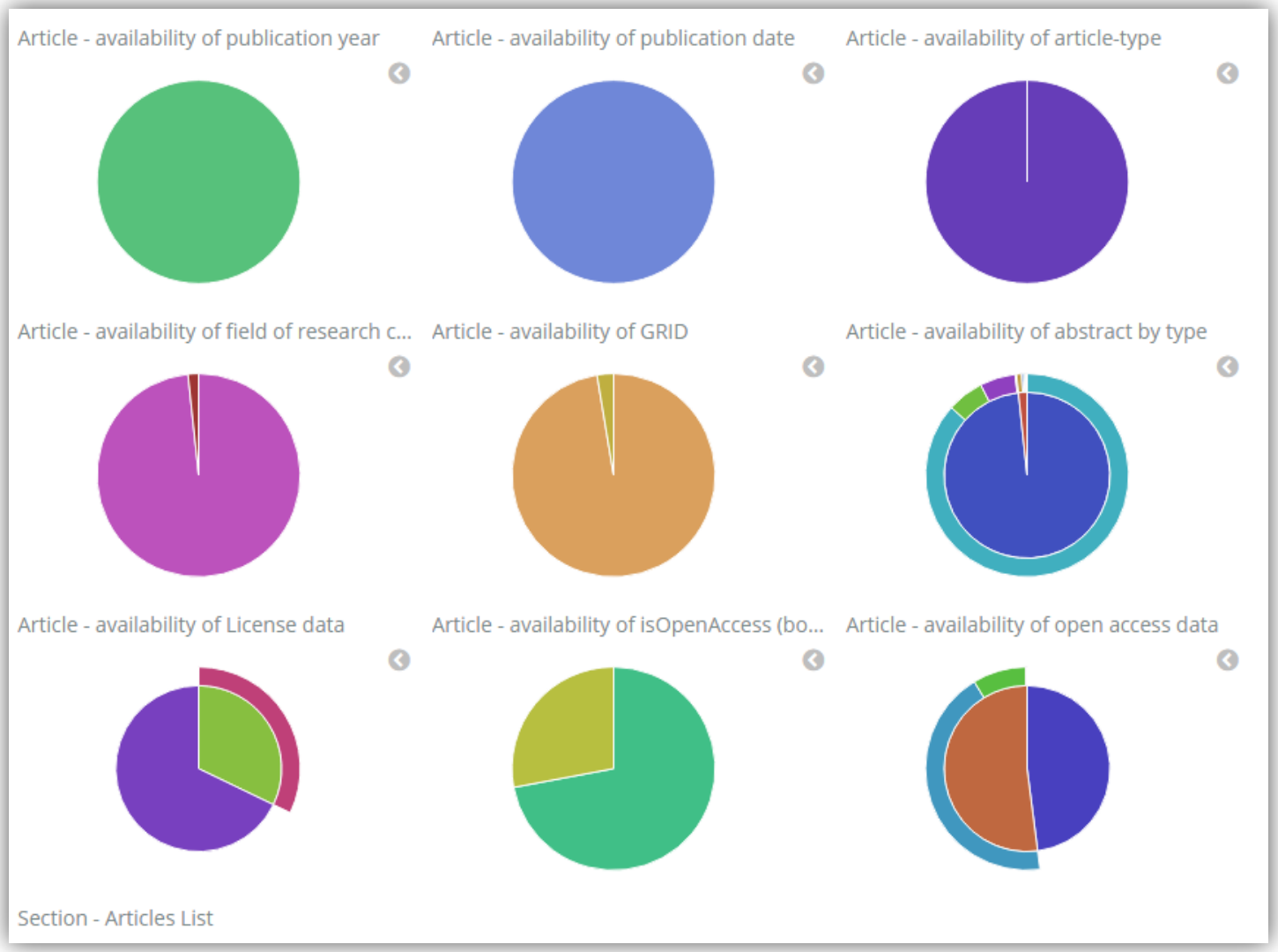
Solution: SN Insights powered by Dimensions



SN Insights: Sneak Peek



Using ELK for internal QA going forward



Thank you!



How to keep in touch:

- **SN SciGraph Technical Manager**
  michele.pasin@springernature.com

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