

Springer Nature SciGraph ¹

Building a high quality semantic graph
for linked science


Sept 2017

Michele Pasin

Lead Data Architect / Tech Product Owner
Knowledge Graph Team

SPRINGER NATURE

SPRINGER NATURE



A world-leading
research, educational
and professional
publisher

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

macmillan
Science and Education

Science and Scholarly

- nature publishing group npg
- nature
- palgrave macmillan
- nature COMMUNICATIONS
- Spektrum DER WISSENSCHAFT
- nature REVIEWS
- SCIENTIFIC AMERICAN
- SCIENTIFIC REPORTS
- INVESTIGACIÓN Y CIENCIA
- naturejobs
- MACMILLAN SCIENCE COMMUNICATION

Education

- macmillan education
- Language Learning | Schools | Higher Education
- bedford ST. MARTIN'S
- WORTH PUBLISHERS
- W. H. FREEMAN
- palgrave

Software and Technology

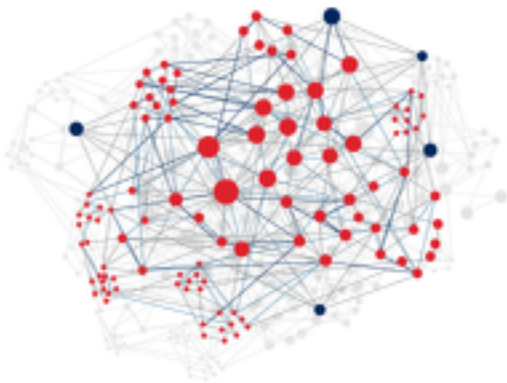
- DIGITAL science | DIGITAL education | macmillan new ventures
- SYMPLECTIC | tutoria | Prepu
- Altmetric | educa | dynamicbooks
- figshare | English On | Late Nite LABS
- über RESEARCH | i-clicker.
- BIORAFIT | Key TUTOR | sapling learning
- readcube | HEY TUTOR | EBI MAP-Works
- labguru | maths doctor | HADEN MFNEIL
- Projects | Cramble

Holtzbrinck Publishing Group

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**

A Rich History..

NPG Linked Data Platform



CURI Semantic Annotation Project

Subject Pages

Scigraph prototype

2012

2013

2014

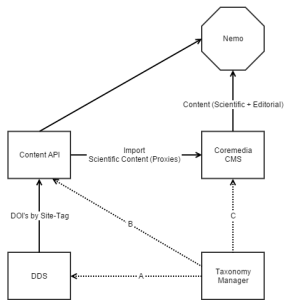
2015

Nature Ontologies Portal



2016

Linnaeus Project



Classification must be done during or before importing the content into the CMS (not at Nemo runtime).

Options where the wedding could take place (Content and Taxonomy)

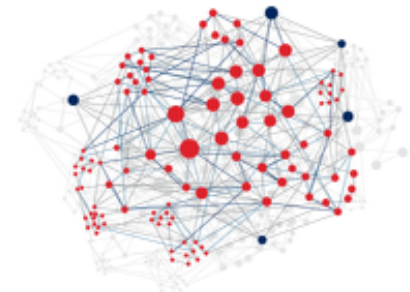
- A: DDS-Team
- B: London-Team
- C: CMS-Team

Springer Conferences



Springer Protocols

SN SciGraph

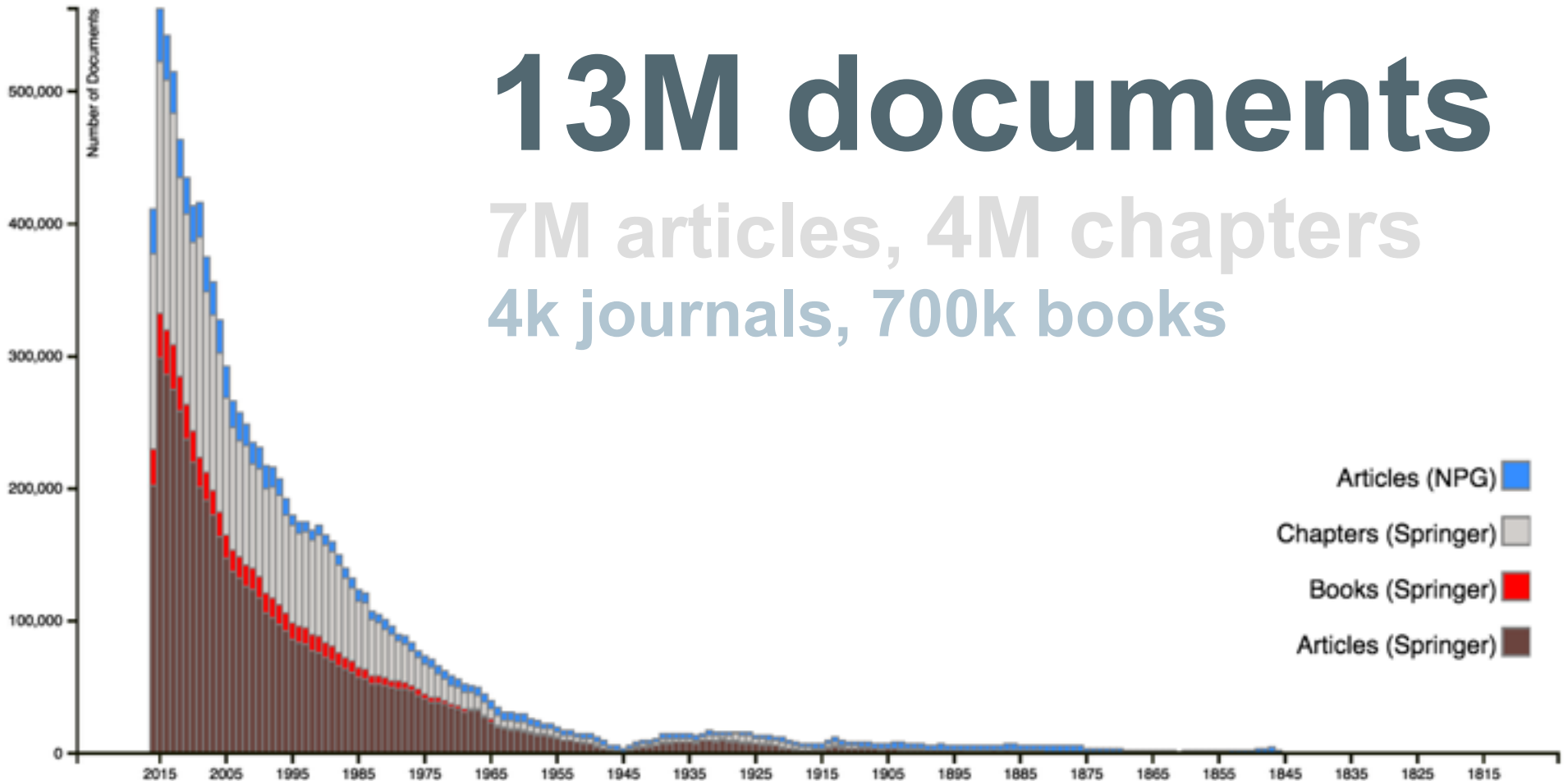


We publish a lot of science! (since 1815)

13M documents

7M articles, 4M chapters

4k journals, 700k books



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



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

Google Cell signalling

Web Images Maps Shopping Books More Search tools

About 2,940,000 results (0.15 seconds)

Ads related to Cell signalling

- Cell Signaling Technology - CellSignal.com**
www.cellsignal.com/
High-Quality Antibodies & Assays for Signal Transduction Research
- Cell Signaling in PBMC - Amnis imaging flow cytometers offer**
www.amnis.com/
>1000 cells/sec with 12 images/cell

Cell Signaling Technology
www.cellsignal.com/

CST strives to make novel antibodies of the highest possible quality by producing them in-house and rigorously characterizing them with respect to specificity ...
Product Catalog - Contact - PI3K / Akt Signaling - Careers

Cell signaling - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Cell_signaling

Cell signalling (Cell signaling in American English) is part of a complex system of communication that governs basic cellular activities and coordinates cell ...
Unicellular and multicellular ... - Classification of ... - See also - References

Cell Signalling Biology
www.cellsignallingbiology.org/

This major contribution to the field of cell signalling by one of the world's leading experts, Professor Sir Michael Berridge (Cambridge) is now sponsored by the ...

Cell signalling - The Open University
www.open.edu/openlearn/science-maths.../cell-signalling/content-section...

Jun 1, 2011 – This unit explains the general principles of signal transduction and specifically, how even the simplest organisms can detect and respond to ...

1st hit from nature.com...

Scitable by nature education

A Collaborative Learning Space for Science

HOME LIBRARY PEOPLE GROUPS BLOGS NATURE JOBS Sign In Register Search Scitable

Library

Cell Signaling

In order to respond to changes in their immediate environment, cells must be able to receive and process signals that originate outside their borders. Individual cells often receive many signals simultaneously, and they then integrate the information they receive into a unified action plan. But cells aren't just targets. They also send out messages to other cells both near and far.

What Kind of Signals Do Cells Receive?

Most cell signals are chemical in nature. For example, prokaryotic organisms have sensors that detect nutrients and help them navigate toward food sources. In multicellular organisms, growth factors, hormones, neurotransmitters, and extracellular matrix components are some of the many types of chemical signals cells use. These substances can exert their effects locally, or they might travel over long distances. For instance, neurotransmitters are a class of short-range

Not linked to/from..

nature.com

Cell signalling

Cell signalling is the mechanism that orchestrates the appropriate response to growth signalling, nutrient signalling and integrin signalling

Latest Research and Reviews

- Research | 28 March 2012 | OPEN
Glycoprotein nonmetastatic fragment shows neuroprotection: PDGF/Akt and MEK/ERK pathways
Yoko Shim, Kazuhiko Tsushima & ...
Scientific Reports 4, 2206
- Research | 24 March 2012
Acyl-lysine acetyltransferase forms a ...

nature REVIEWS MOLECULAR CELL BIOLOGY

Journal home | Review | Article | Review

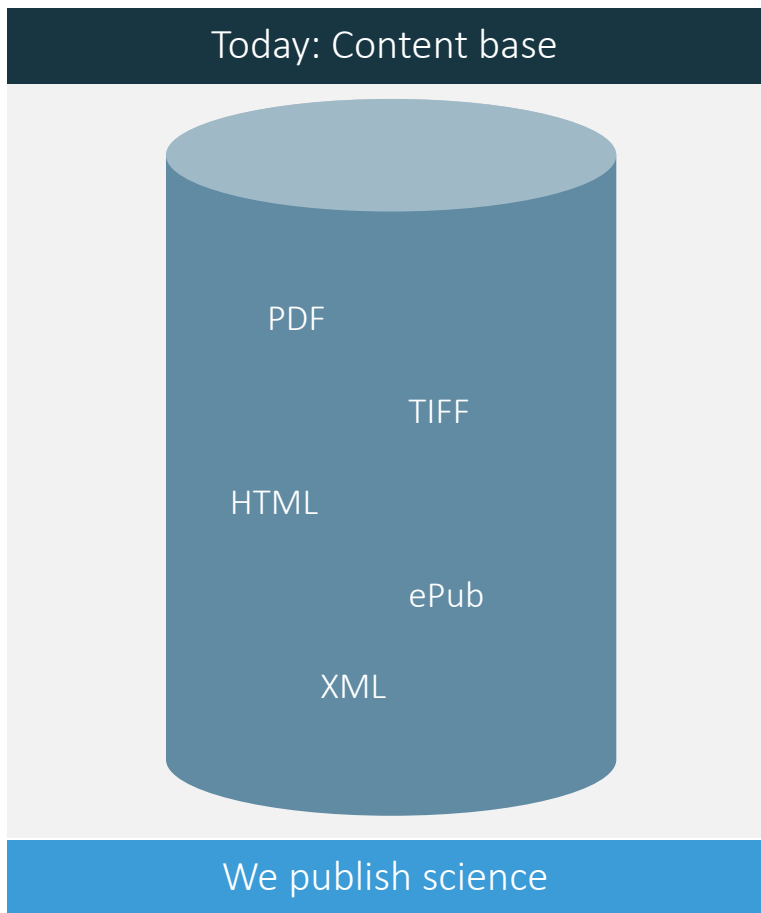
Regulation of cell signalling by uPAR

Harvey W. Smith & Chris J. Marshall | About the authors

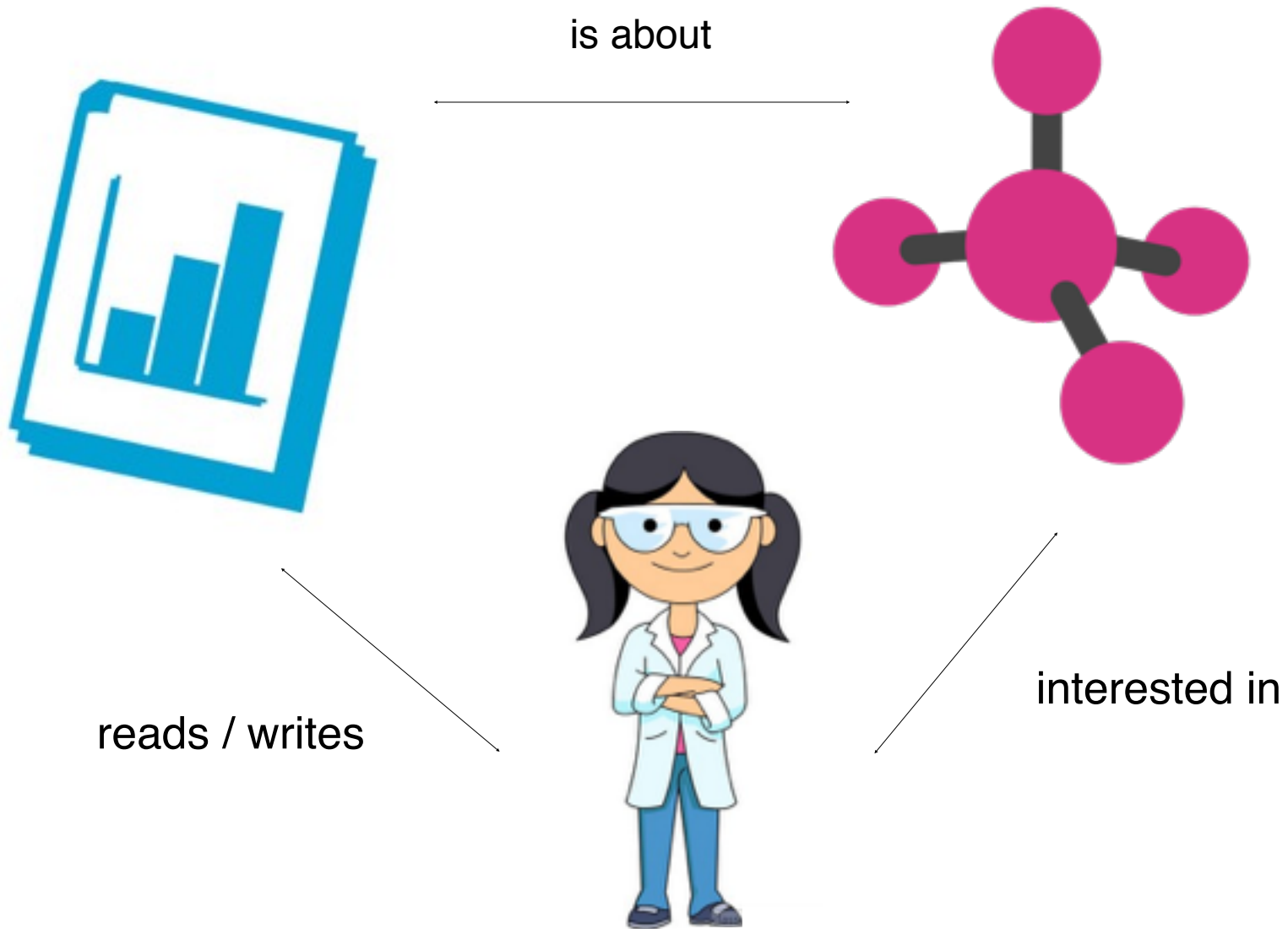
Uniknase-type plasminogen activator receptor (uPAR) expression is elevated during inflammation and tissue remodelling and in many human cancers, in which it frequently indicates poor prognosis. uPAR regulates proteolysis by binding the extracellular protease urokinase-type plasminogen activator (uPA; also known as urokinase) and also activates many intracellular signalling pathways. Coordination of extracellular matrix (ECM) proteolysis and cell signalling by uPAR underlies its important function in cell migration, proliferation and survival and makes it an

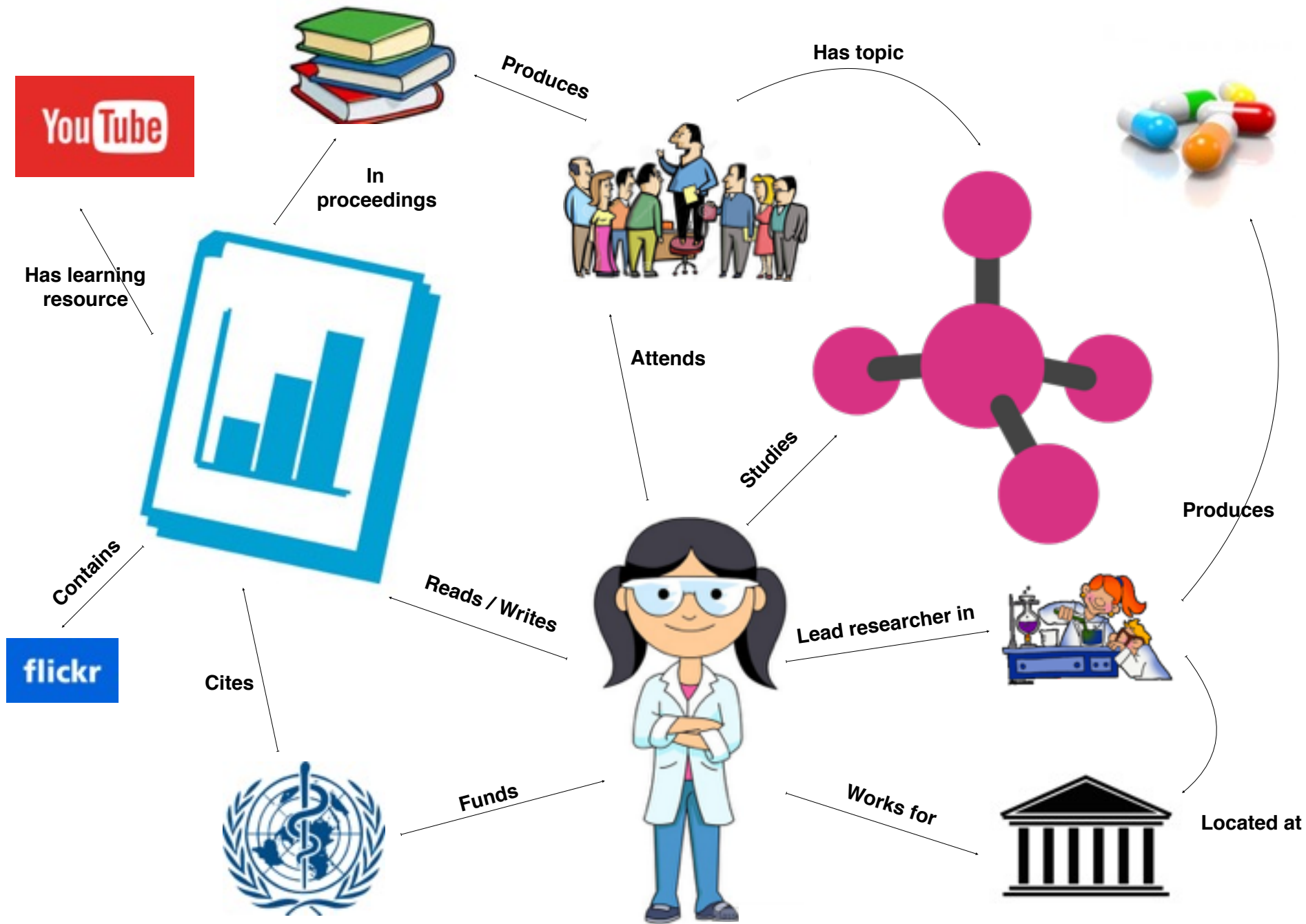
..and neither do we!

Vision

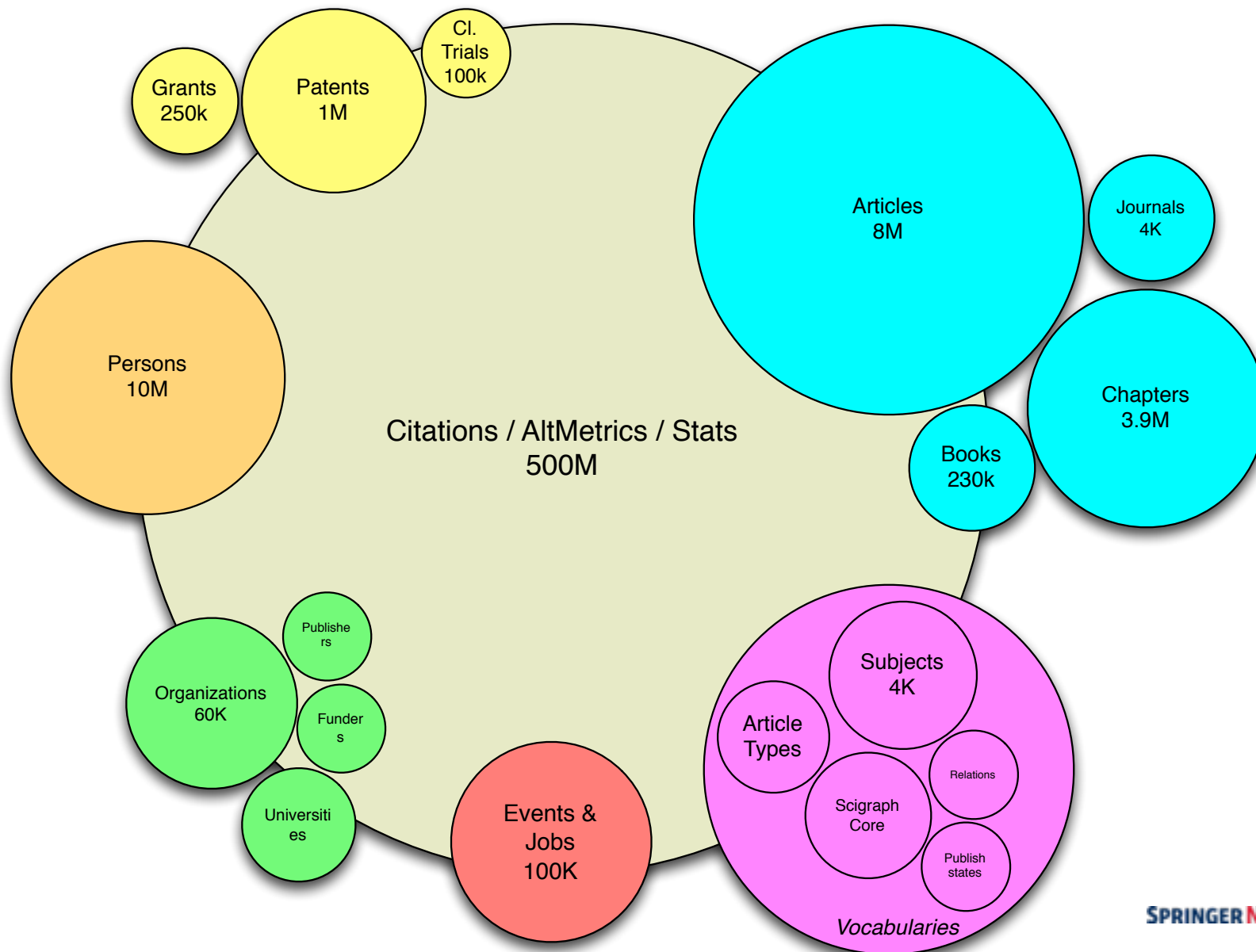


Three areas of knowledge we care about





SciGraph Data Landscape



A Closer Look At SciGraph

Semantic integration

Content Enrichment

SciGraph Project: Capabilities and Applications

Capabilities: Data Integration & Semantic Enrichment

- > Consolidation of existing LD efforts via a single domain mode
- > Ingestion and normalisation of third party datasets
- > Data mining and entity extraction

Applications: Discoverability & Analytics

- > Better end user applications
- > Business analytics dashboards
- > Open Linked Data publishing



ETL Architecture: main features [in evolution]

Tech stack

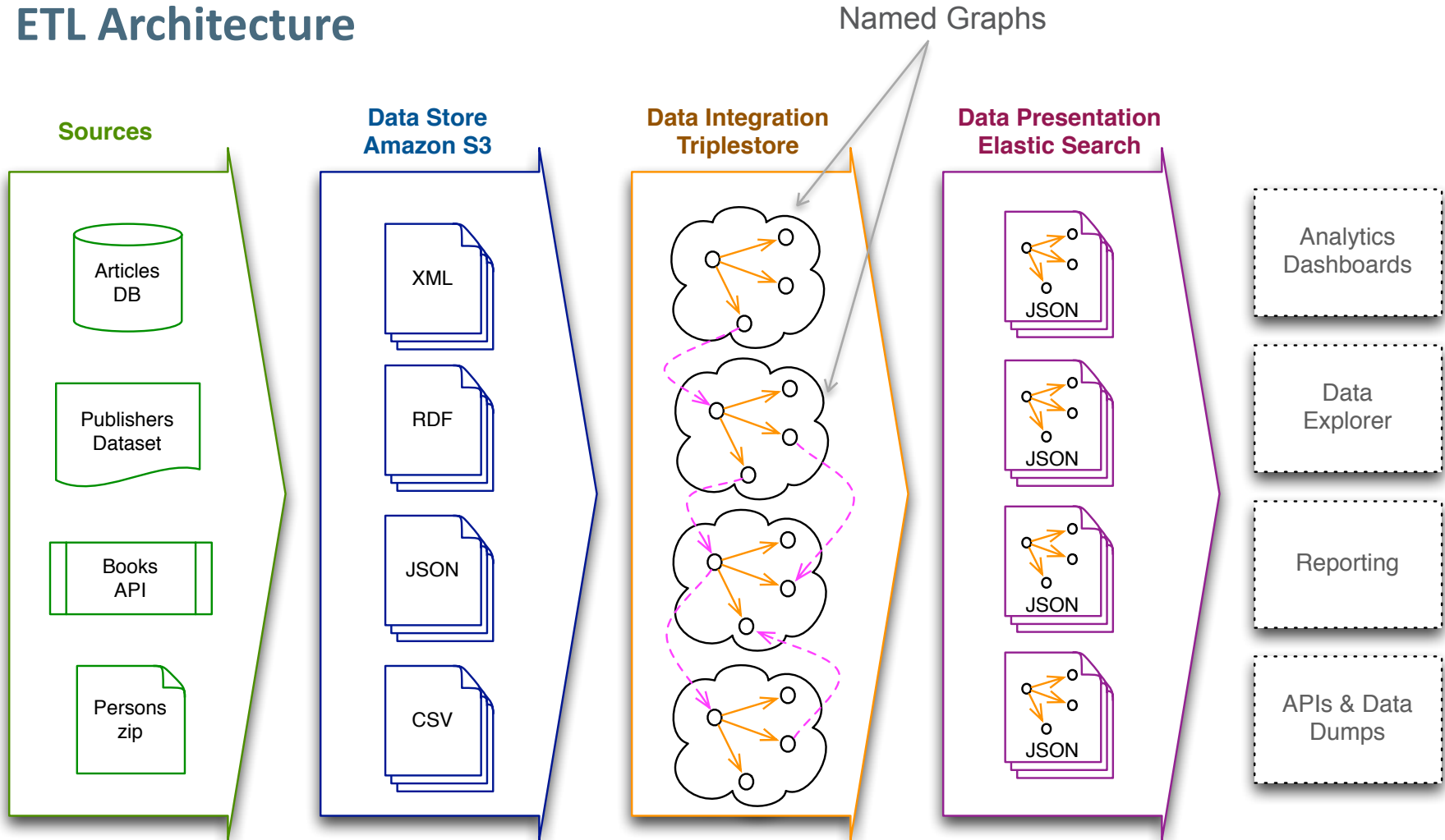
- > Airflow framework (Airbnb)
- > Amazon S3 to make backups
- > GraphDB triplestore (staging and presentation)
- > Elastic search and APIs

Components & Principles

- > Graph must be **'ephemeral'**
- > Data sources versioning algorithm
- > Identity Persistence service
- > Validation via SHACL (TopBraid API)



ETL Architecture



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

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

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


Applications



Discovery Tools

Analytics Tools



Linked Data Publishing

Discovering Content: Subject Pages

MENU  nature.com

 Search  Login

Molecular biology

 Atom  RSS Feed

Molecular Biology is the field of biology that studies the composition, structure and interactions of cellular molecules such as nucleic acids and proteins that carry out the biological processes essential for the cells functions and maintenance.


Featured

News and Views | 08 March 2017

Molecular biology: A hidden competitive advantage of disorder

P. Andrew Chong & Julie D. Forman-Kay

Nature




News and Views | 08 March 2017

Non-coding RNA: More uses for genomic junk

Karen Adelman & Emily Egan

Nature **543**, 183–185




News and Views | 01 March 2017

Cancer epigenetics: Reading the future of leukaemia

Alex W. Wilkinson & Or Gozani

Nature **543**, 186–188



Related Subjects

- Cell division
- CRISPR-Cas systems
- DNA recombination
- Non-coding RNAs
- Protein folding
- Riboswitches
- DNA
- Chromatin
- DNA damage and repair
- DNA replication
- Nuclear organization
- Proteolysis
- Ribozymes
- Single-molecule biophysics
- Chromosomes
- DNA metabolism
- Epigenetics
- Post-translational modifications
- Proteomics
- RNA metabolism
- Transcription

SciGraph Analytics: Dashboards

BMC Cell Biology

Journal ID: 12860

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

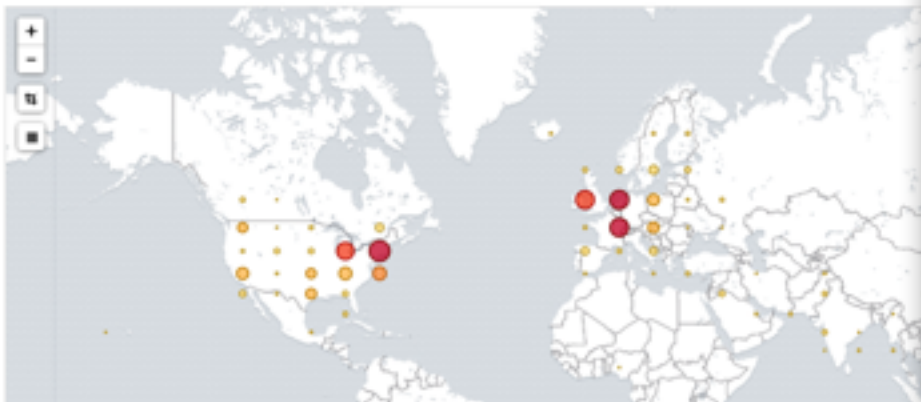
Section - Countries and Institutions

Countries and Institutions

Use this section to find out which are the top countries and institutions contributing to a publication.

Note: this information comes from the GRID database (<https://www.grid.ac/>).

Article - map view



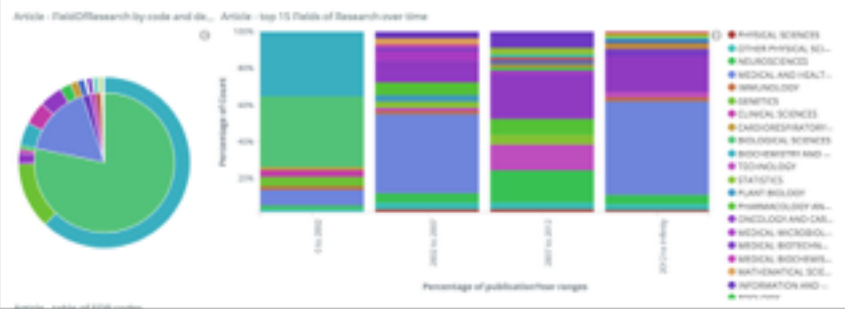
Publication Volume

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 used article types and how much of this content has been cited in external databases.

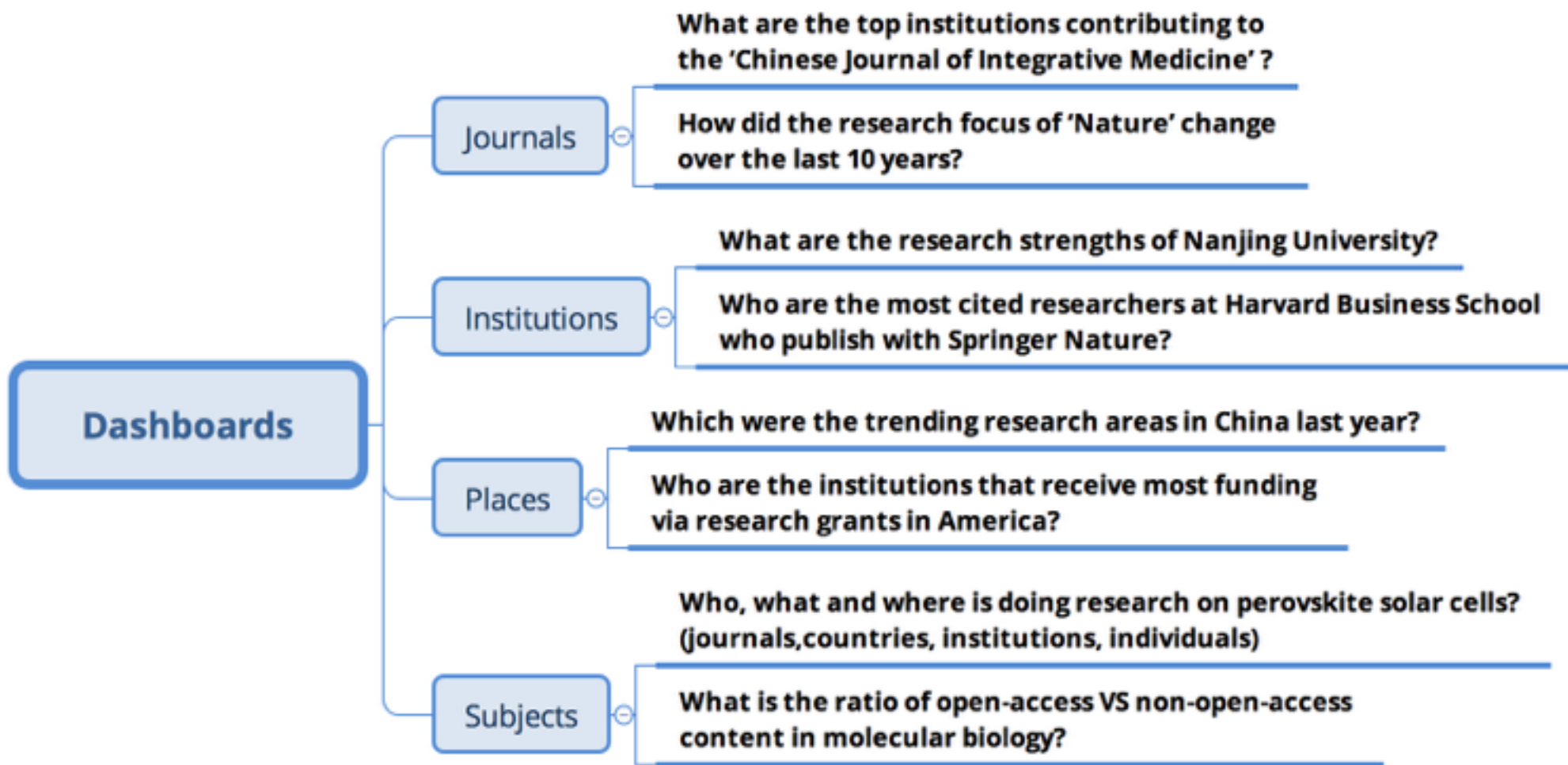


Fields of Research

This section provides a breakdown of publication content based on subject areas. The subject areas are derived from the Australian and New Zealand Standard Research-Classification (ANZSRC) <http://www.anzsrc.gov.au/australian-research-classification/>



SciGraph Analytics: Supporting Data Driven Decisions



Open Linked Data publishing (Feb 2017)

SPRINGER NATURE

Springer Nature SciGraph

A Linked Open Data platform for the scholarly domain

We are pleased to introduce Springer Nature SciGraph, the new Linked Open Data platform aggregating data sources from Springer Nature and key partners from the scholarly domain. The Linked Open Data platform will initially collate information from across the research landscape, such as funders, research projects, conferences, affiliations and publications. Additional data, such as citations, patents, clinical trials and usage numbers will follow over time. This high quality data from trusted and reliable sources provides a rich semantic description of how information is related, as well as enabling innovative visualizations of the scholarly domain.

By doing so, Springer Nature SciGraph overcomes former boundaries by relating comprehensive information about the research landscape. It represents a further step in data integration and it will continue to grow organically. This platform will increase the discoverability of high quality data as larger parts of our datasets will be made freely available under a CC BY-NC 4.0 license.



The data in Springer Nature SciGraph is projected to contain 1.5 to 2 billion triples. It will comprise metadata from journals and articles, books and chapters, organizations, institutions, funders, research grants, patents, clinical trials, substances, conference series, events, citations and reference networks, Altmetrics, links to research datasets and much more.

Any questions?
Please contact us.

Dataset Download

Licensing Information

Further Info

Conference Presentation 2016 (PDF, 11.56 MB)

At a glance:

- 300 M triples / 32G downloads
- CC-BY-NC License

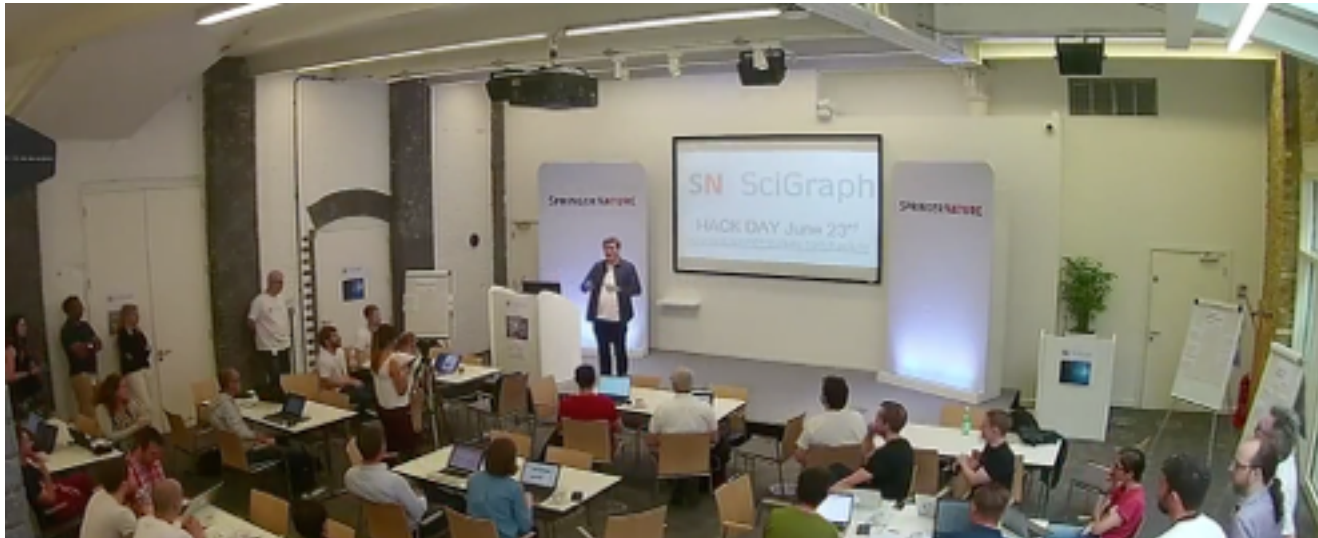
Metadata about:

- Articles 2012-2016 (5M)
- Grants (200k)
- Journals (3k)
- Subjects (3k)
- Core Ontology

Open Data Events: Hack Day June 23rd in London

Aims and Scope

- > Engagement with Linked Data Researcher Community
- > Encourage developers to build cool tools with our data
- > Position ourselves as Open Data research publisher
- > Gather first-hand feedback from potential users of our data



Summary

What's next

Looking Ahead

Summary

- Scigraph is our latest LOD platform: focus on data integration and enrichment
- Collaboration between SN and Digital Science (other partners too)
- Internal use cases: discoverability, analytics dashboards
- Data publishing: ~300M triples released in February, supporting Open Science

Next Steps

- Data publishing: new release towards complete archive, hybrid license model
- Tools for analytics, reporting, visualisation, interactive exploration of the graph
- Entities extraction: scientific entities, places, people, events etc..
- Collaboration with DBpedia: funding internship in London/Leipzig this autumn

Thanks

Email:

michele.pasin@springernature.com

Project Homepage:

<http://www.springernature.com/scigraph>