Springer Nature SciGraph

Building a high quality semantic graph for linked science

Sept 2017

Michele Pasin

Lead Data Architect / Tech Product Owner

Knowledge Graph Team

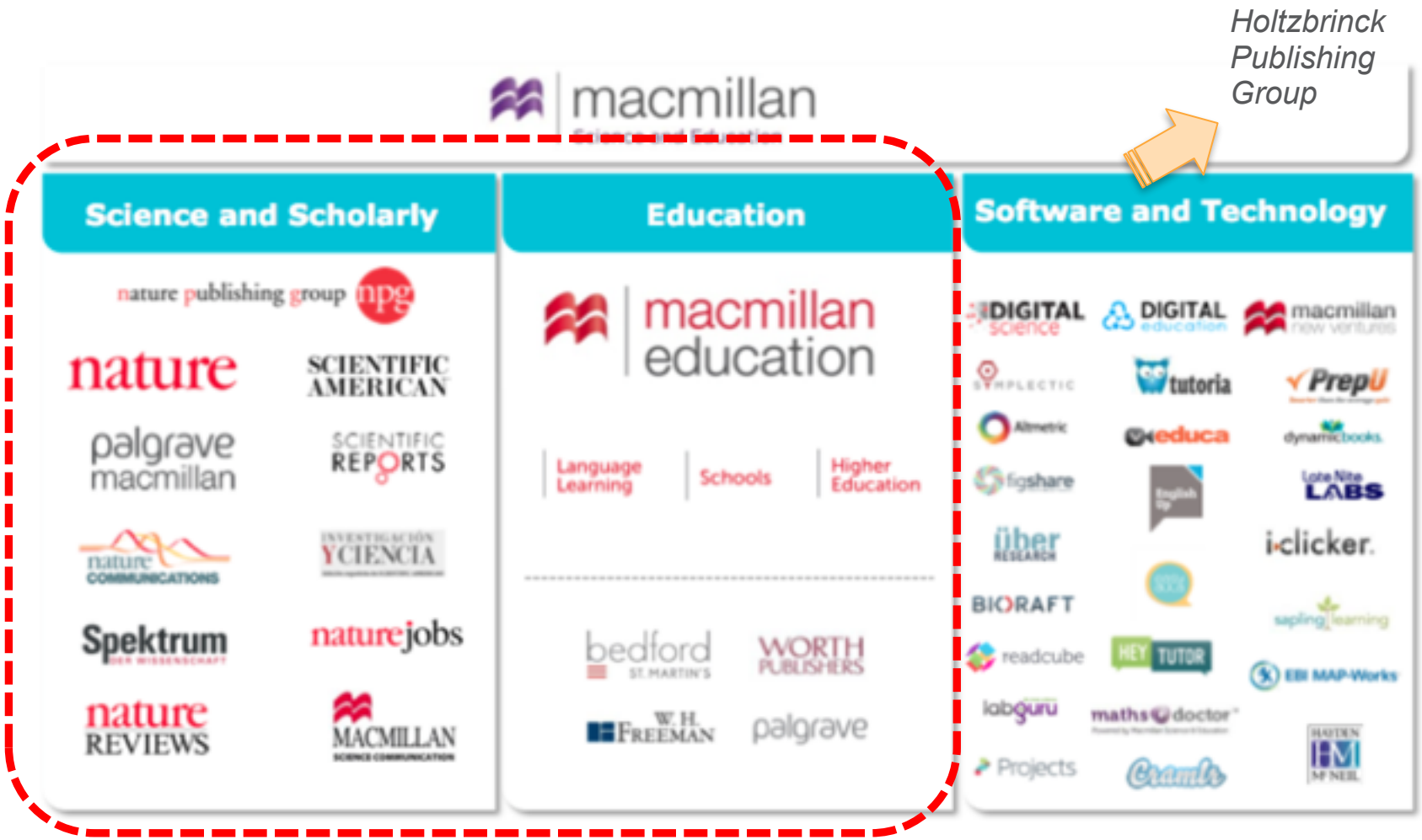
SPRINGER NATURE

SPRINGER NATURE

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



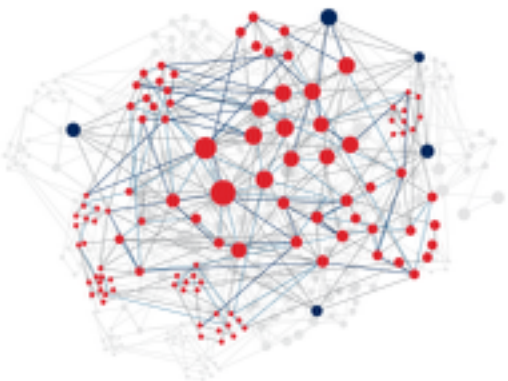
SPRINGER NATURE

just released

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

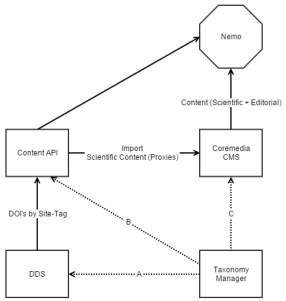
2013

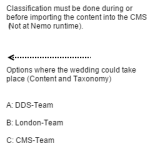
2014

2015

2016

Linnaeus Project





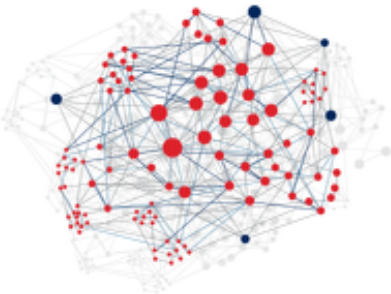
Nature Ontologies Portal

Nero Project

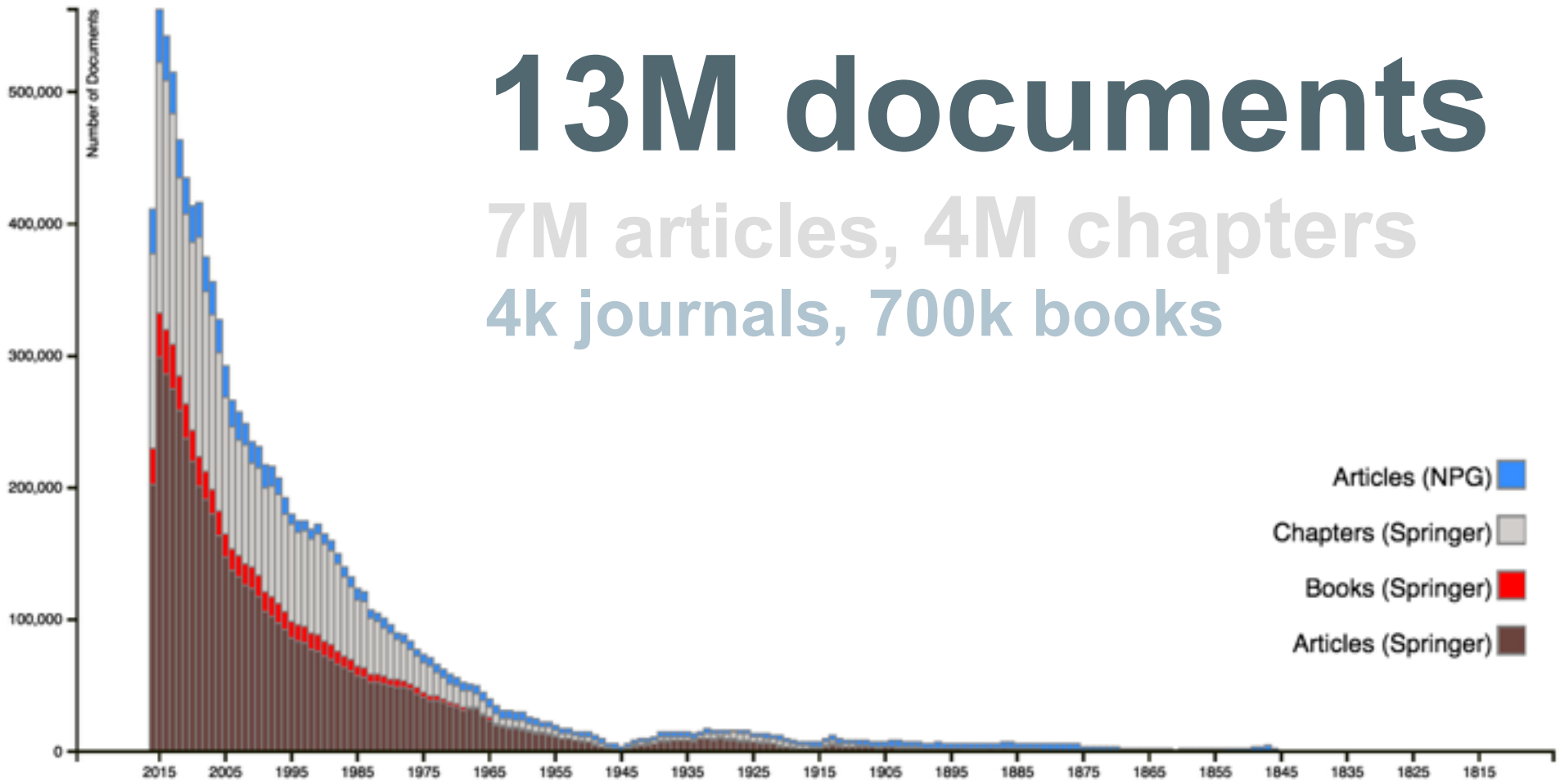
Springer Conferences

Springer Protocols

SN SciGraph



We publish a lot of science! (since 1815)



13M documents

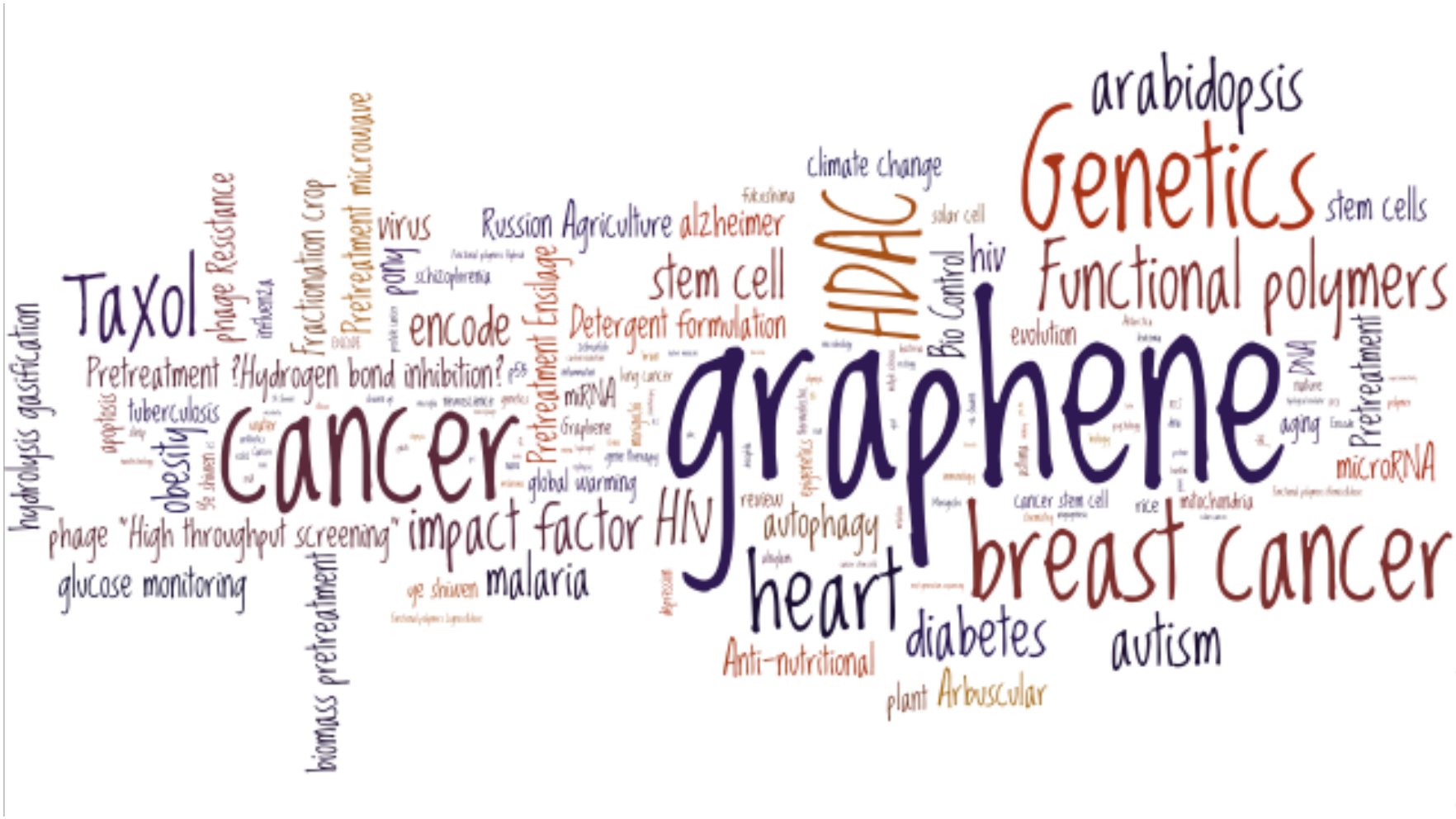
7M articles, 4M chapters

4k journals, 700k books

SPRINGER NATURE

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…

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.**cellsignalling**biology.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

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

Latest Research and

nature REVIEWS MOLECULAR CELL BIOLOGY

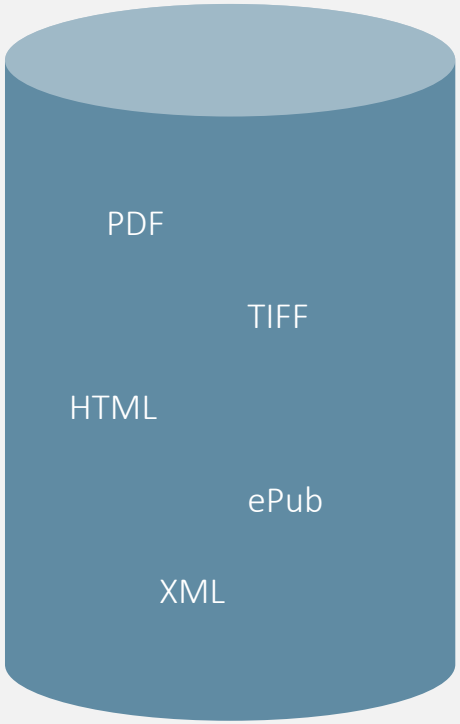
Review

Regulation of cell signalling by uPAR

..and neither do we!

Vision

Today: Content base



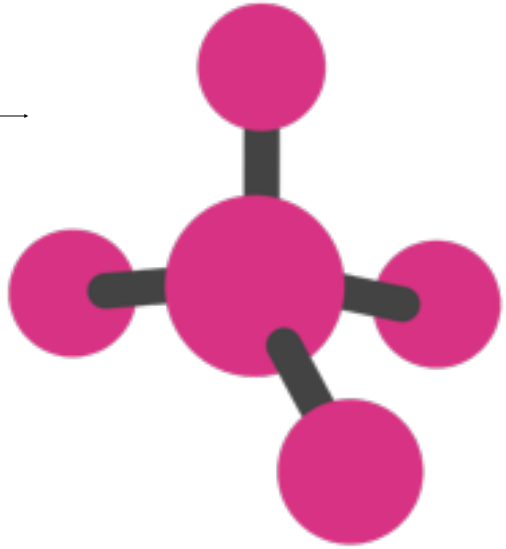
We publish science

Tomorrow: Knowledge Graph

We manage knowledge

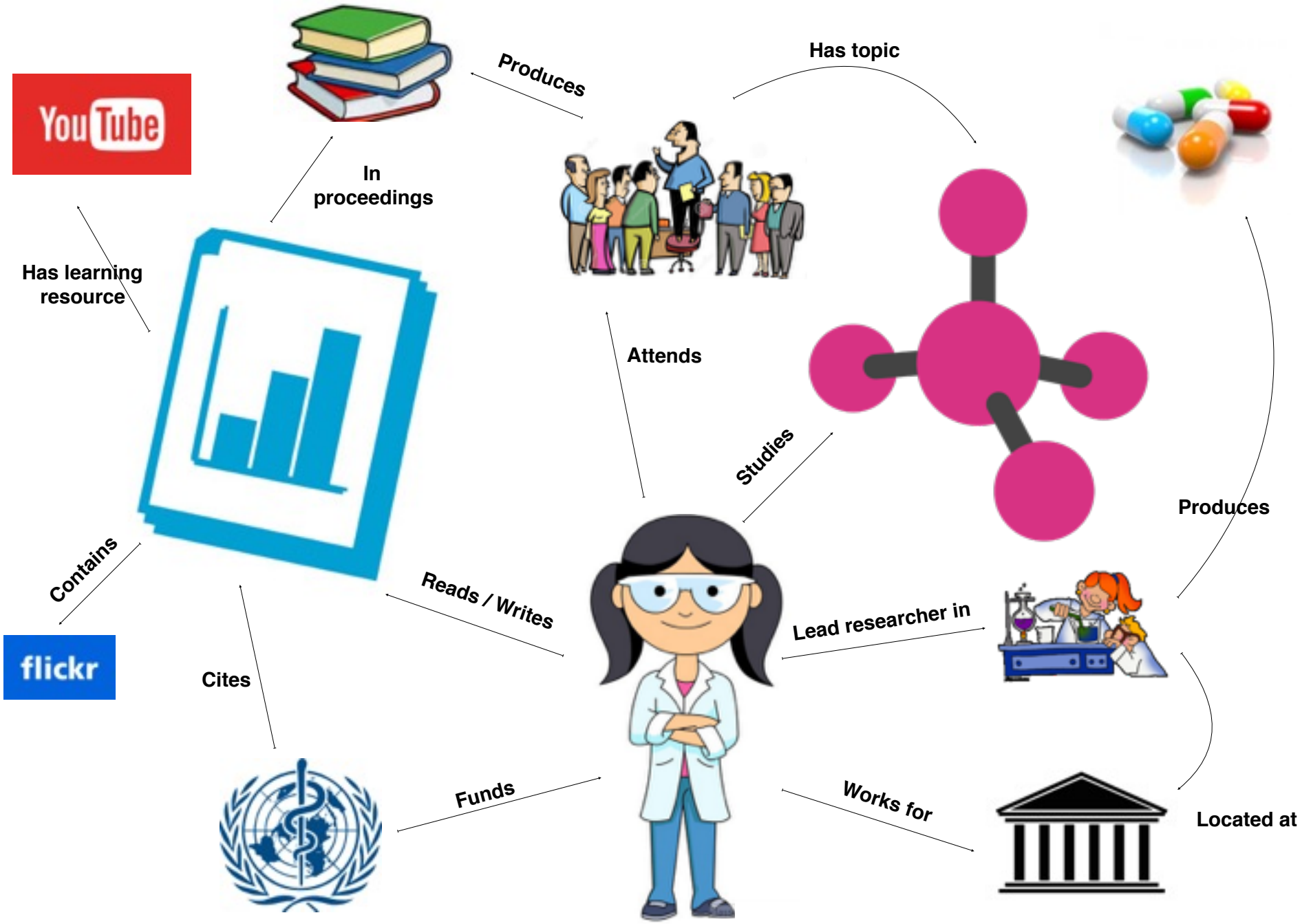
Three areas of knowledge we care about

is about

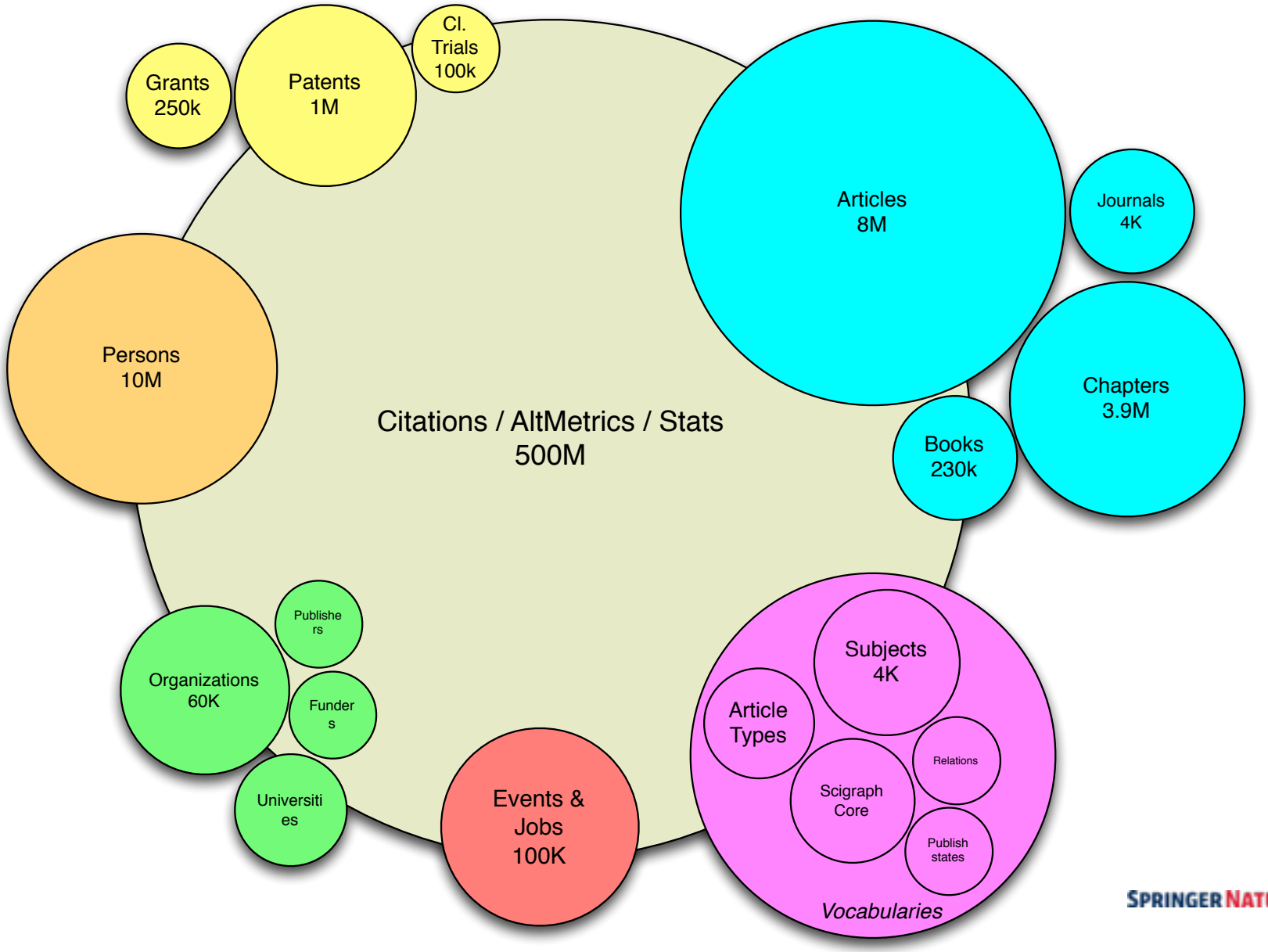


reads / writes

interested in


Produces
Has topic
In proceedings
Has learning resource
Attends
Studies
Contains
Reads / Writes
Lead researcher in
Produces
Cites
Funds
Works for
Located at

SciGraph Data Landscape



SPRINGER NATURE

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

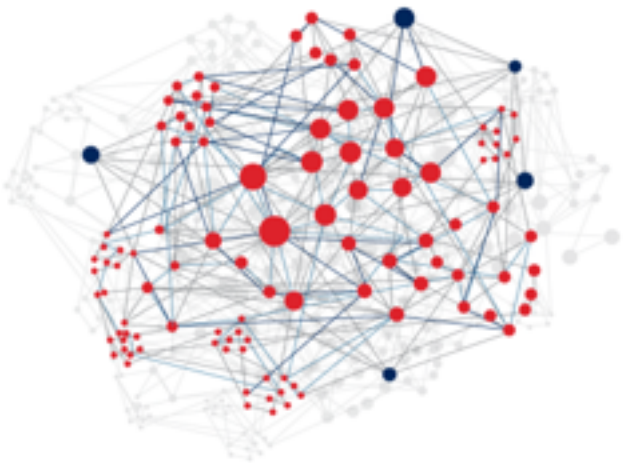
Applications: Discoverability & Analytics

> Better end user applications

> Business analytics dashboards

> Open Linked Data publishing

SN SciGraph



ETL Architecture: main features [in evolution]

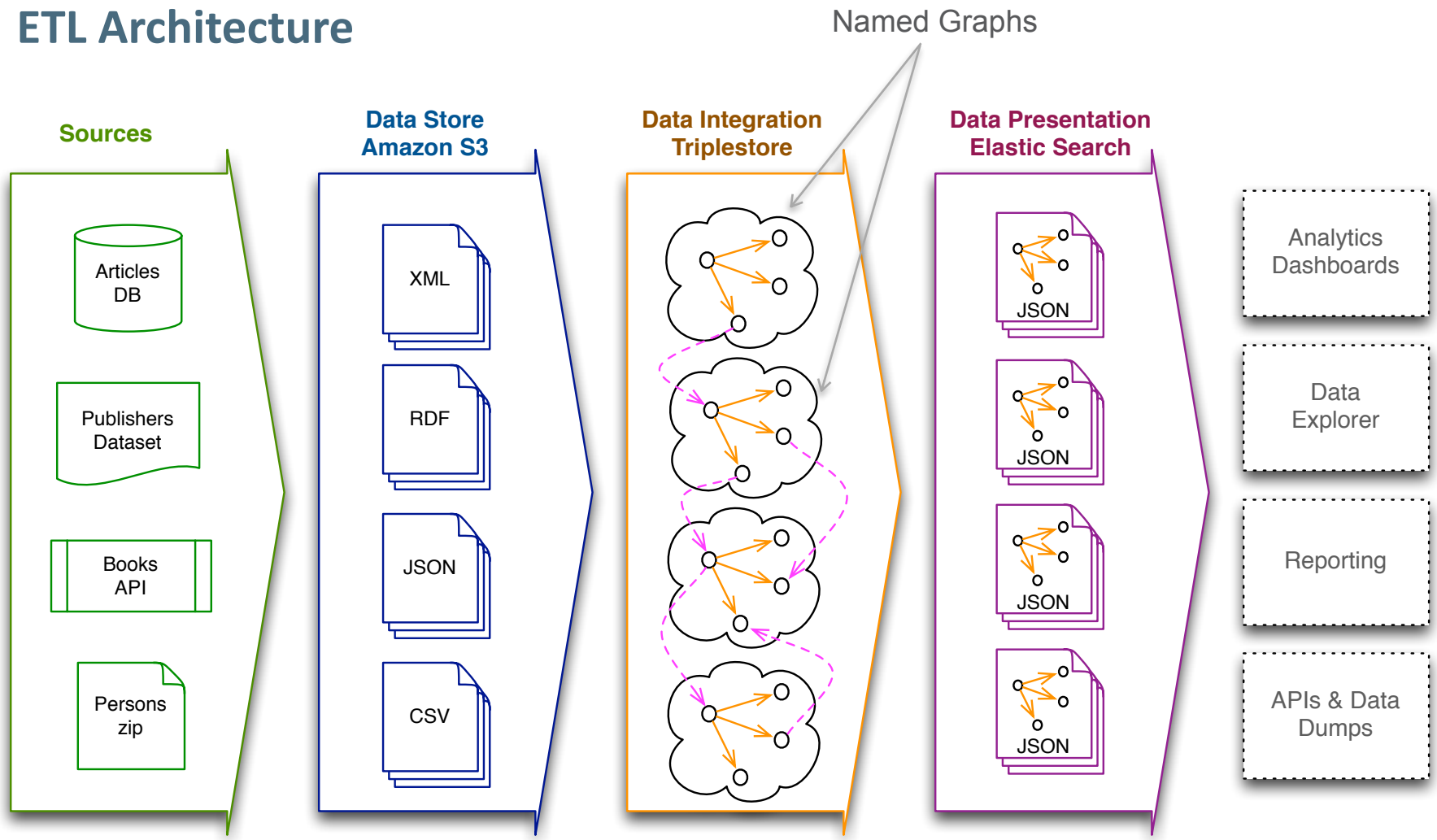
Tech stack

> 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

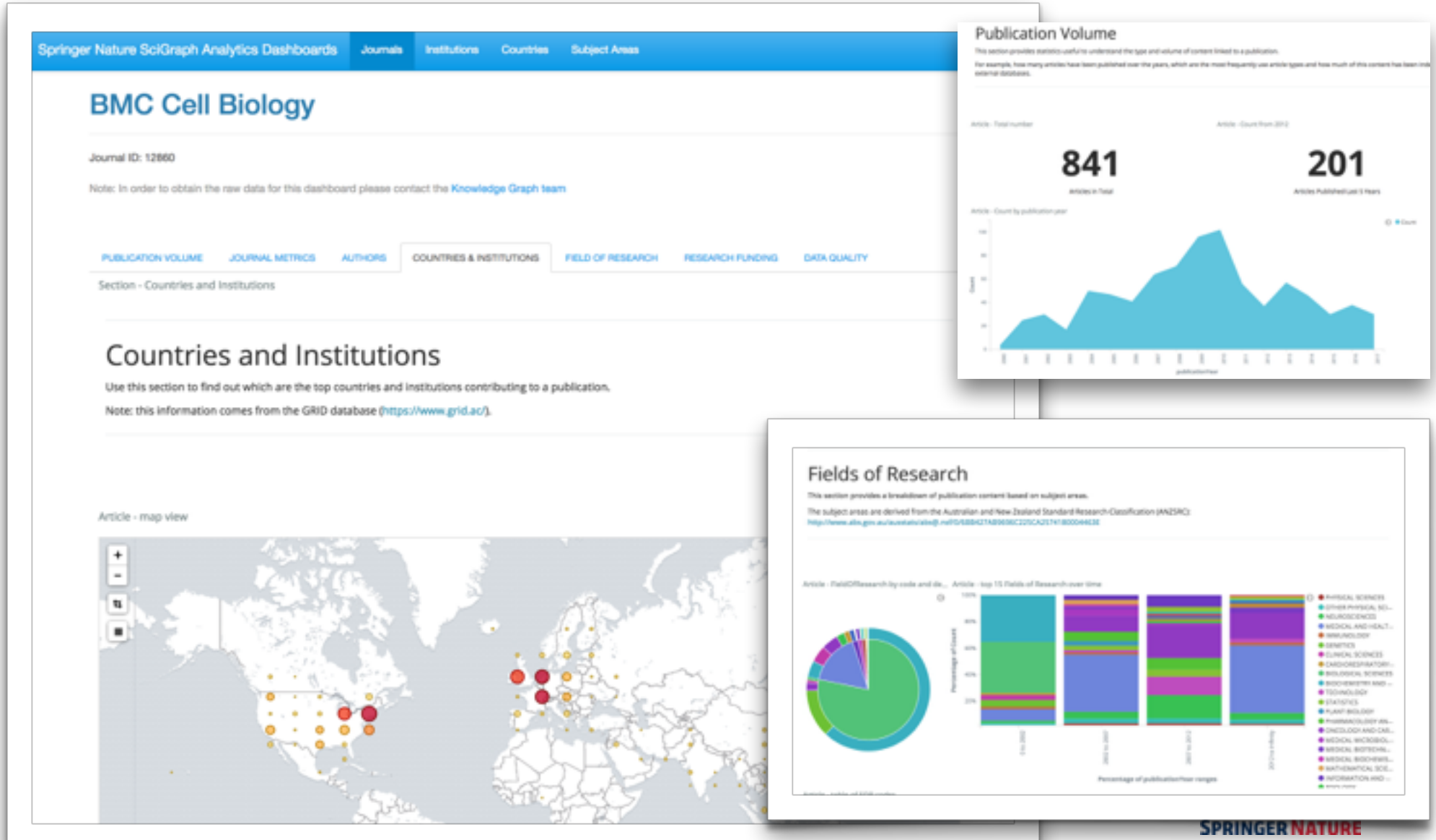
Related Subjects

Cell division
CRISPR-Cas systems
DNA recombination
Non-coding RNAs
Protein folding
Riboswitches
RNAi

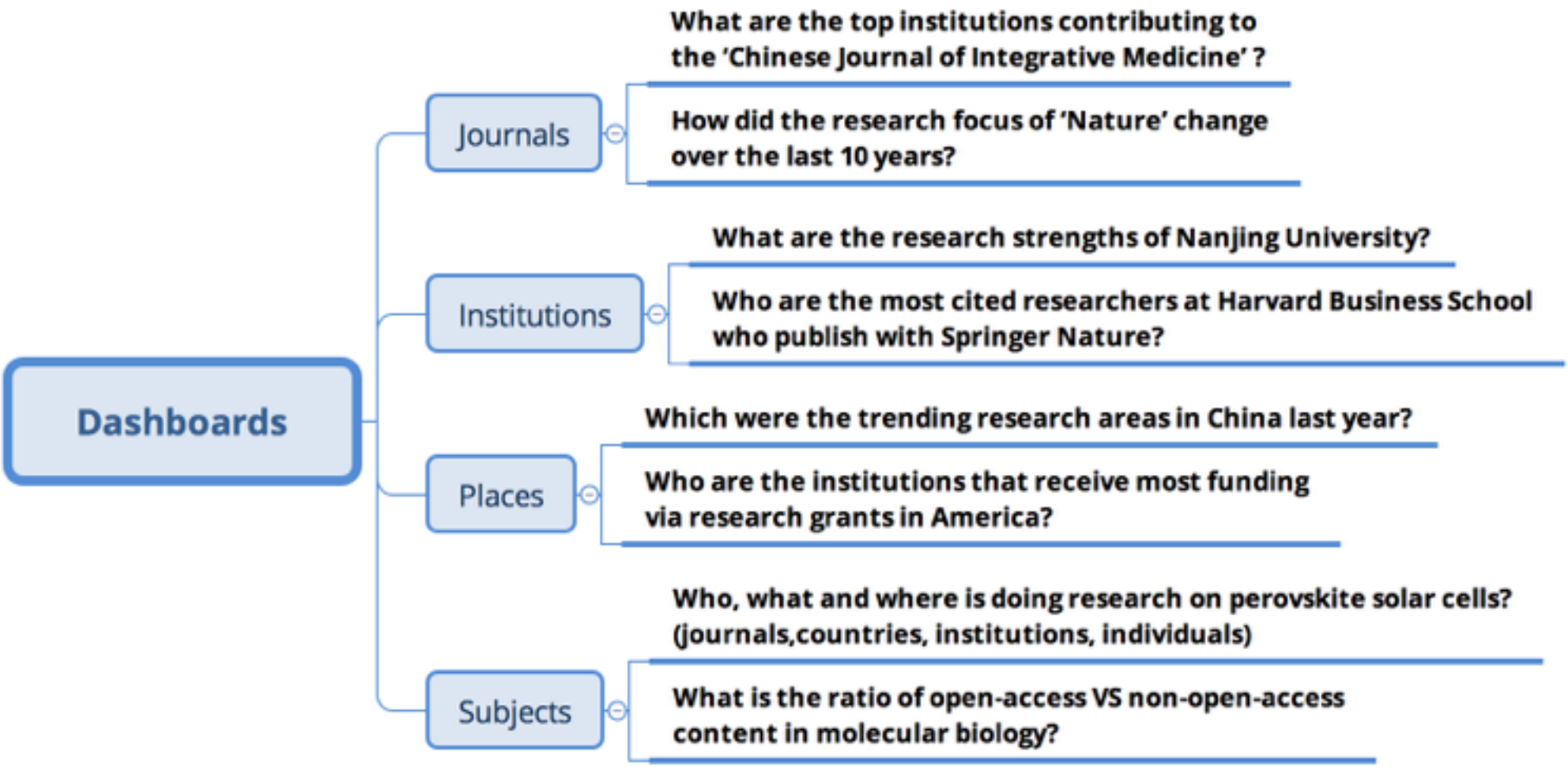
Chromatin
DNA damage and repair
DNA replication
Nuclear organization
Proteolysis
Ribozymes
Single-molecule biophysics

Chromosomes
DNA metabolism
Epigenetics
Post-translational modifications
Proteomics
RNA metabolism
Transcription

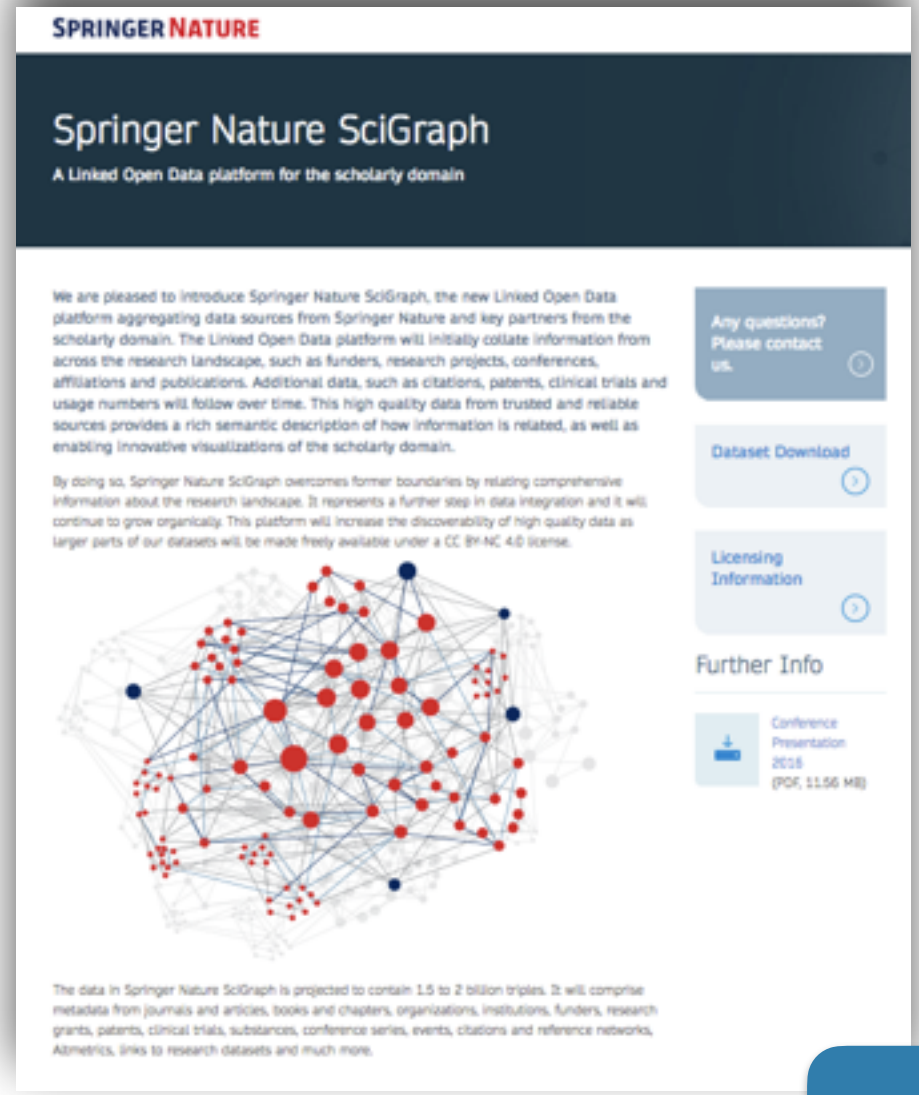
SciGraph Analytics: Dashboards



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

www.springernature.com/scigraph

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