SN SciGraph

Building a Linked Data Knowledge Graph for the Scholarly Publishing Domain

Markus Kaindl

Senior Manager Semantic Data &
SN SciGraph Product Owner

LDBC Technical User Community (TUC) Meeting; Munich, September 1st 2017

SPRINGER NATURE

Agenda

Intro
- Springer Nature SciGraph
- Linked Open Data Publishing

Status
- SN SciGraph Hack Day
- Analytics Dashboards

Data
- Roadmap EOY and beyond

SPRINGER NATURE

Intro:
- Springer Nature SciGraph
- Linked Open Data Publishing

#1

Intro:
- Springer Nature SciGraph
- Linked Open Data Publishing

#1.1

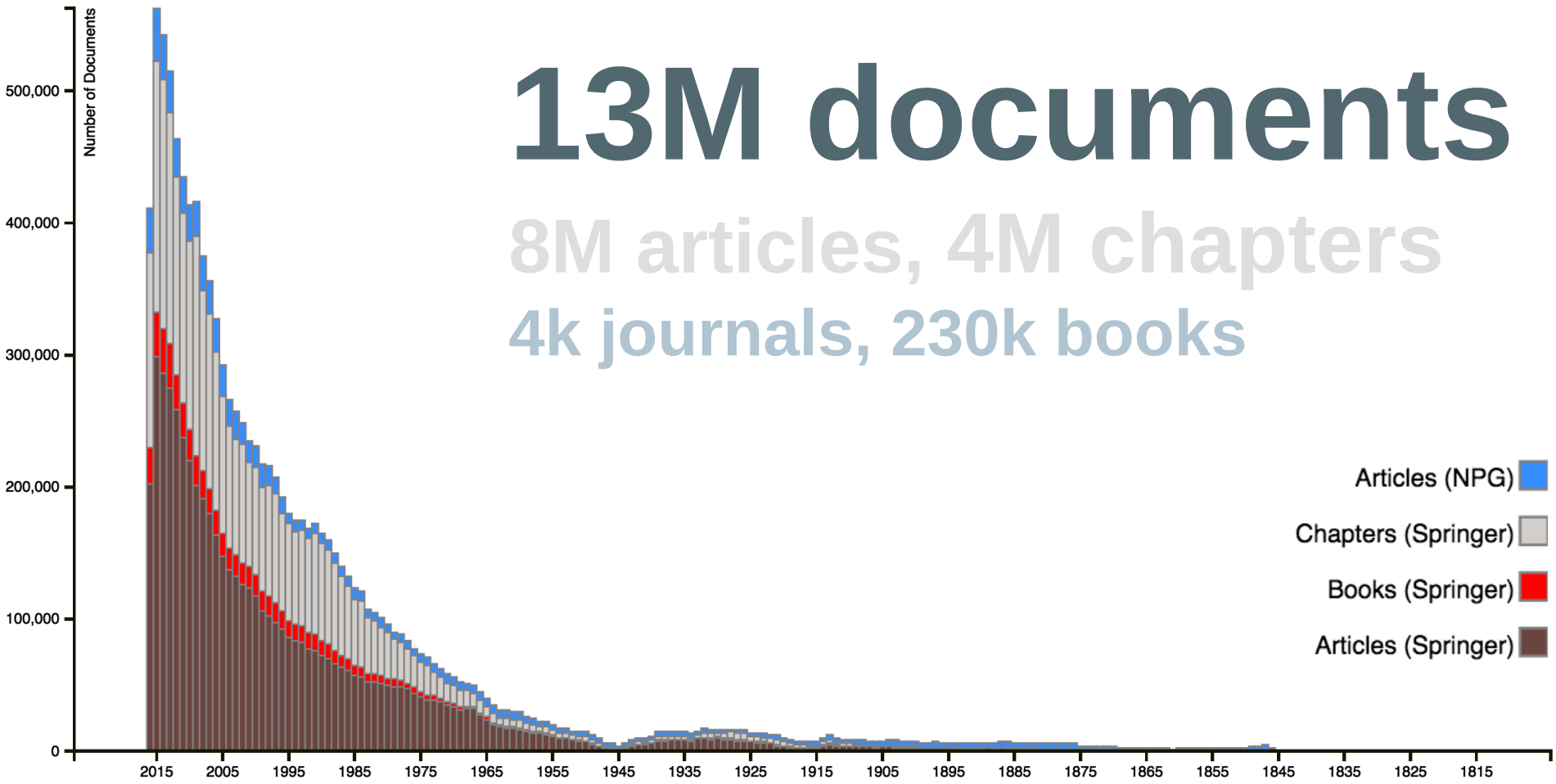
SPRINGER NATURE

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

We publish a lot of science (since 1815)



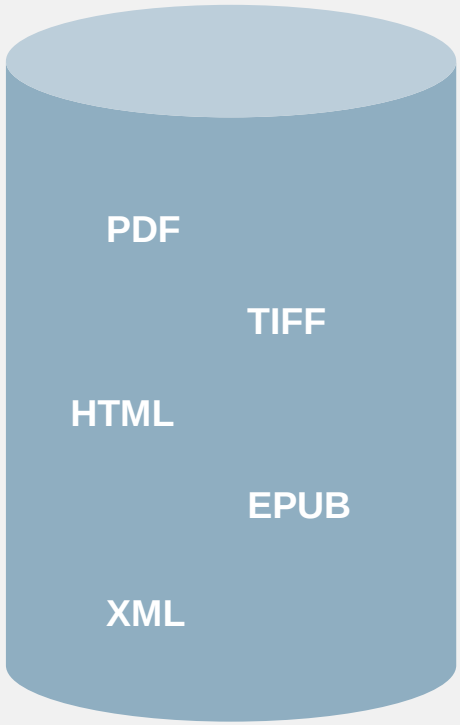
13M documents

8M articles, 4M chapters

4k journals, 230k books

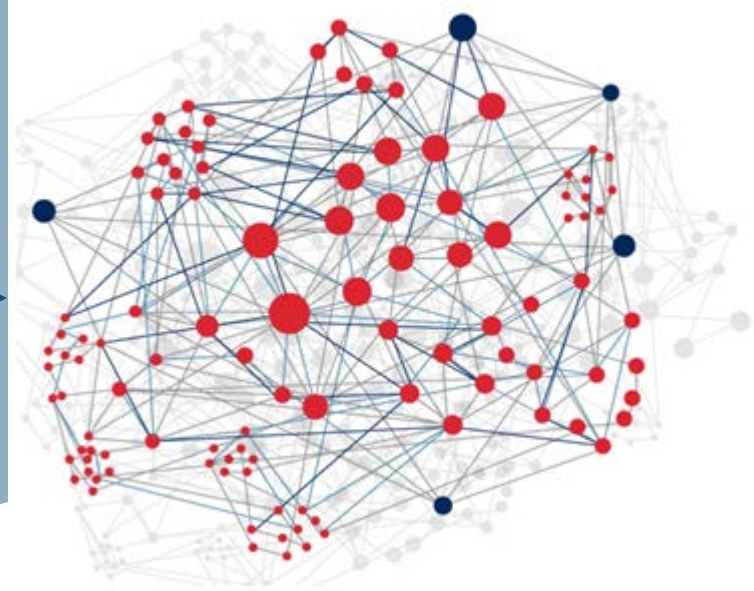
From Content to Data

Content base



We publish content

SN SciGraph



We manage knowledge

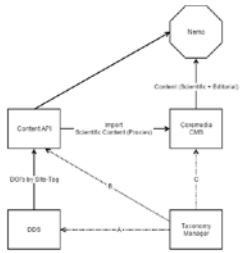
SPRINGER NATURE

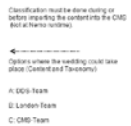
Integration of Various Semantic Data Initiatives

Our Work So Far

2012

Linnaeus Project





2013

Nero Project

2014

NPG Linked Data Platform

CURI Semantic Annotation Project

Scigraph prototype

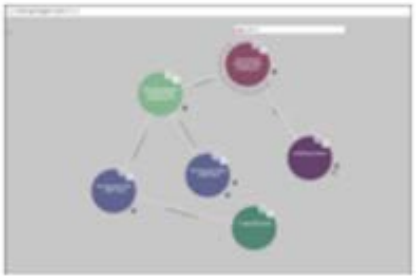
Subject Pages

2015

Nature Ontologies Portal

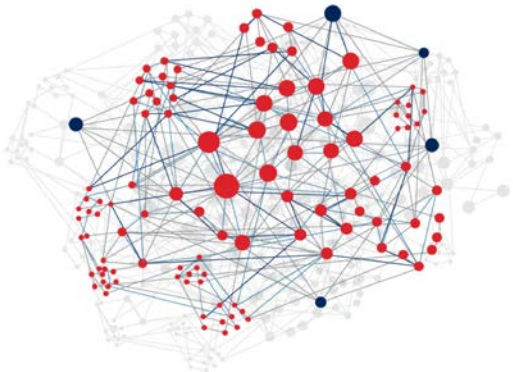
2016

Springer Conferences



Springer Protocols

SN SciGraph



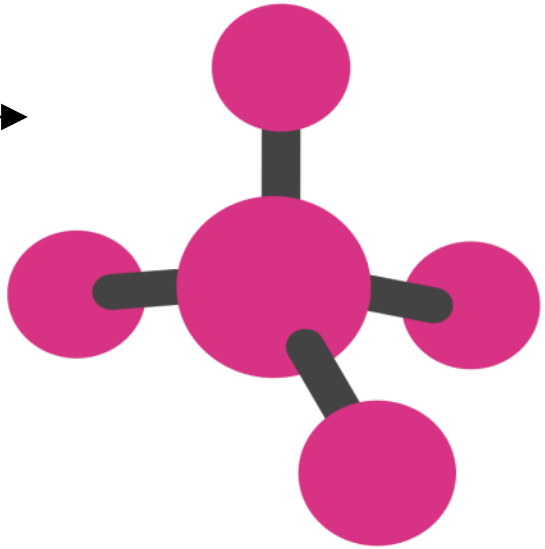
Product Vision

We create the largest state-of-the-art linked open data aggregation platform in the scholarly domain. 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.

SN SciGraph

Three areas of knowledge we care about

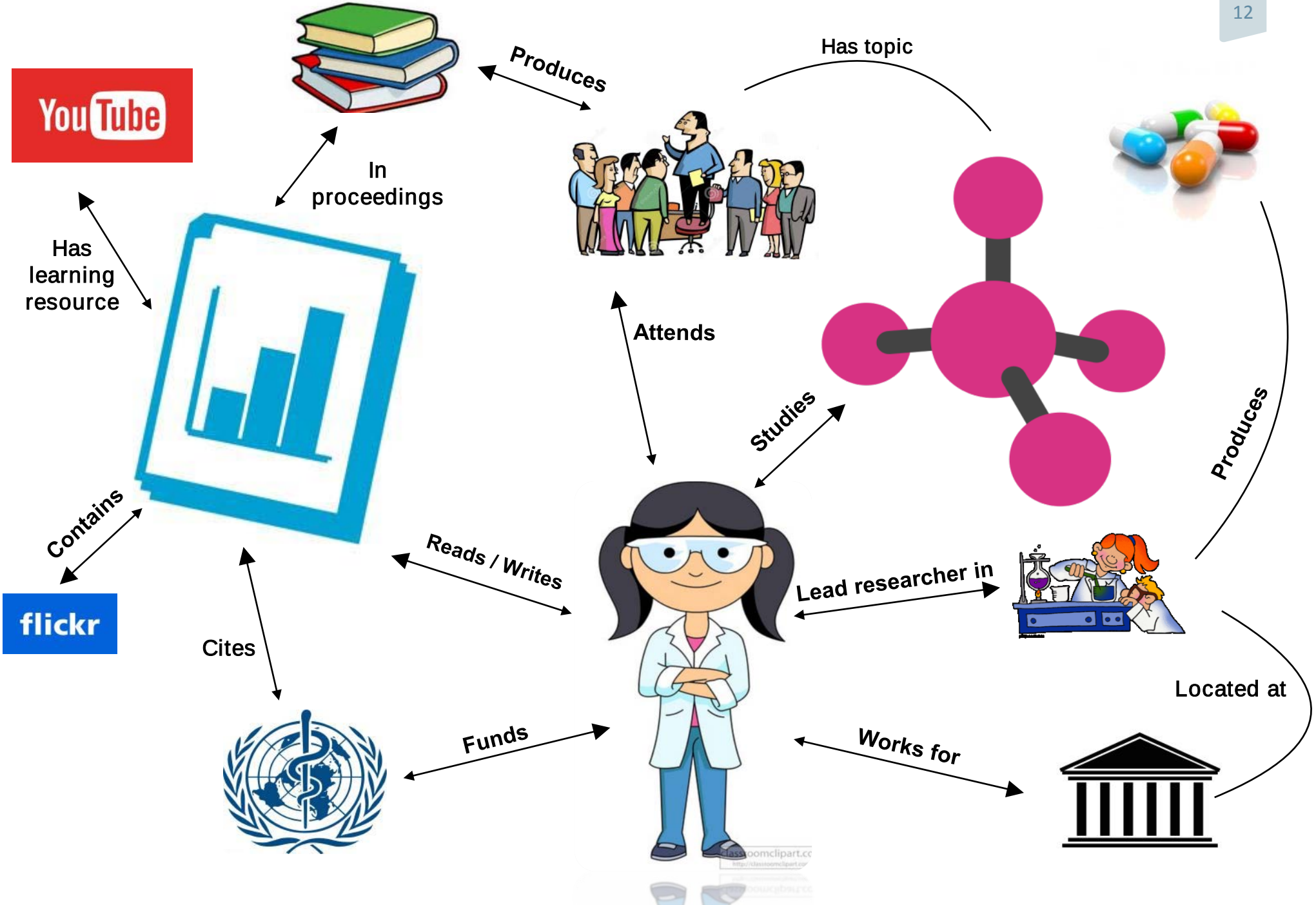
is about



reads / writes

interested in

SPRINGER NATURE


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

ETL Architecture: main features [in evolution]

Tech stack

> Airflow framework (Airbnb)*

> Amazon S3 to make backups

> GraphDB triplestore (staging and presentation)

> Elasticsearch and Kibana*

Components & Principles

> Graph must be '**ephemeral**'

> Data sources versioning algorithm

> Identity Persistence service

> Validation via SHACL (TopBraid API)*

Airflow

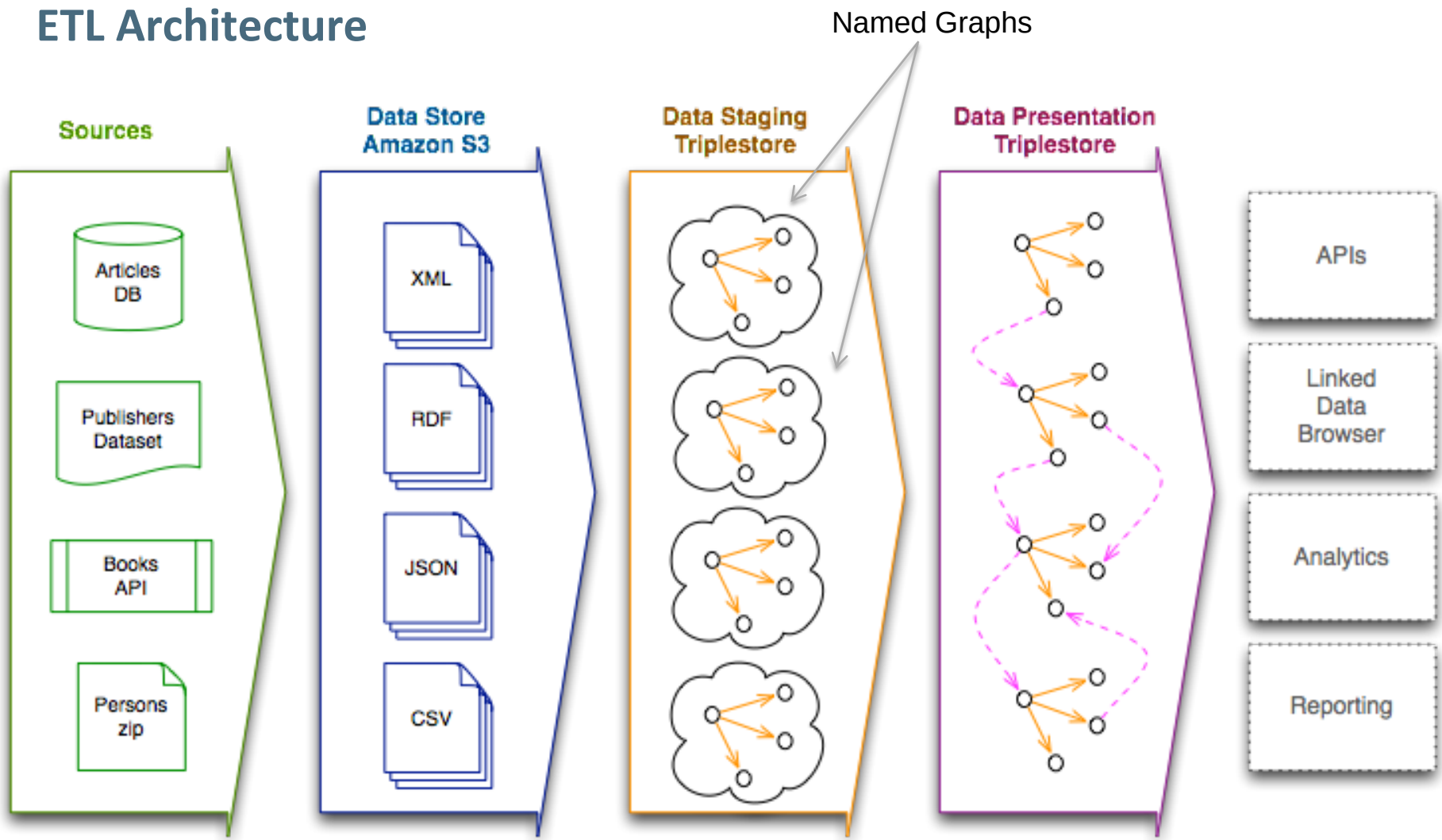
amazon web services S3

ontotext

elastic

*Open Source

ETL Architecture



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

→

- Extraction
- Validation
- Identity Persistence
- Updating / Replacing named graphs

→

- Integration (union graph)
- Inference

SPRINGER NATURE

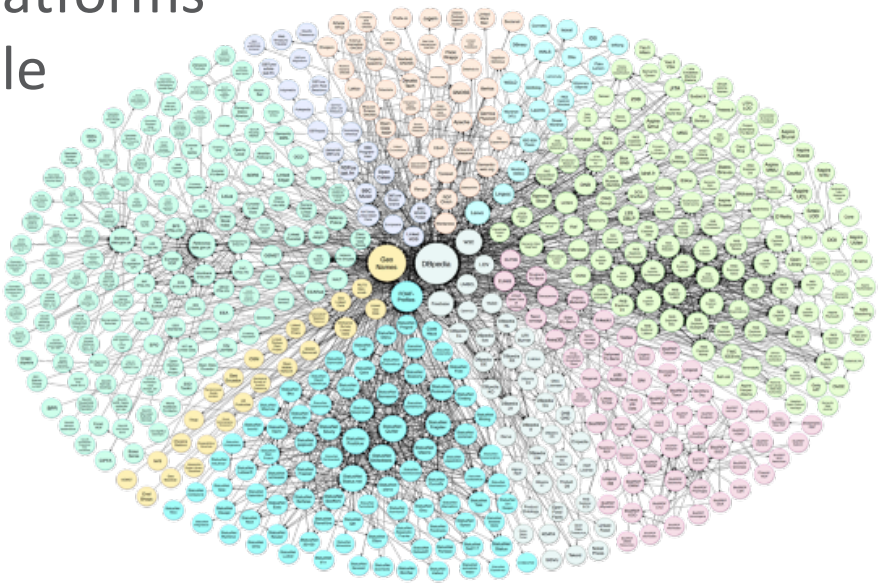
Intro:
- Springer Nature SciGraph
- Linked Open Data Publishing

#1.2

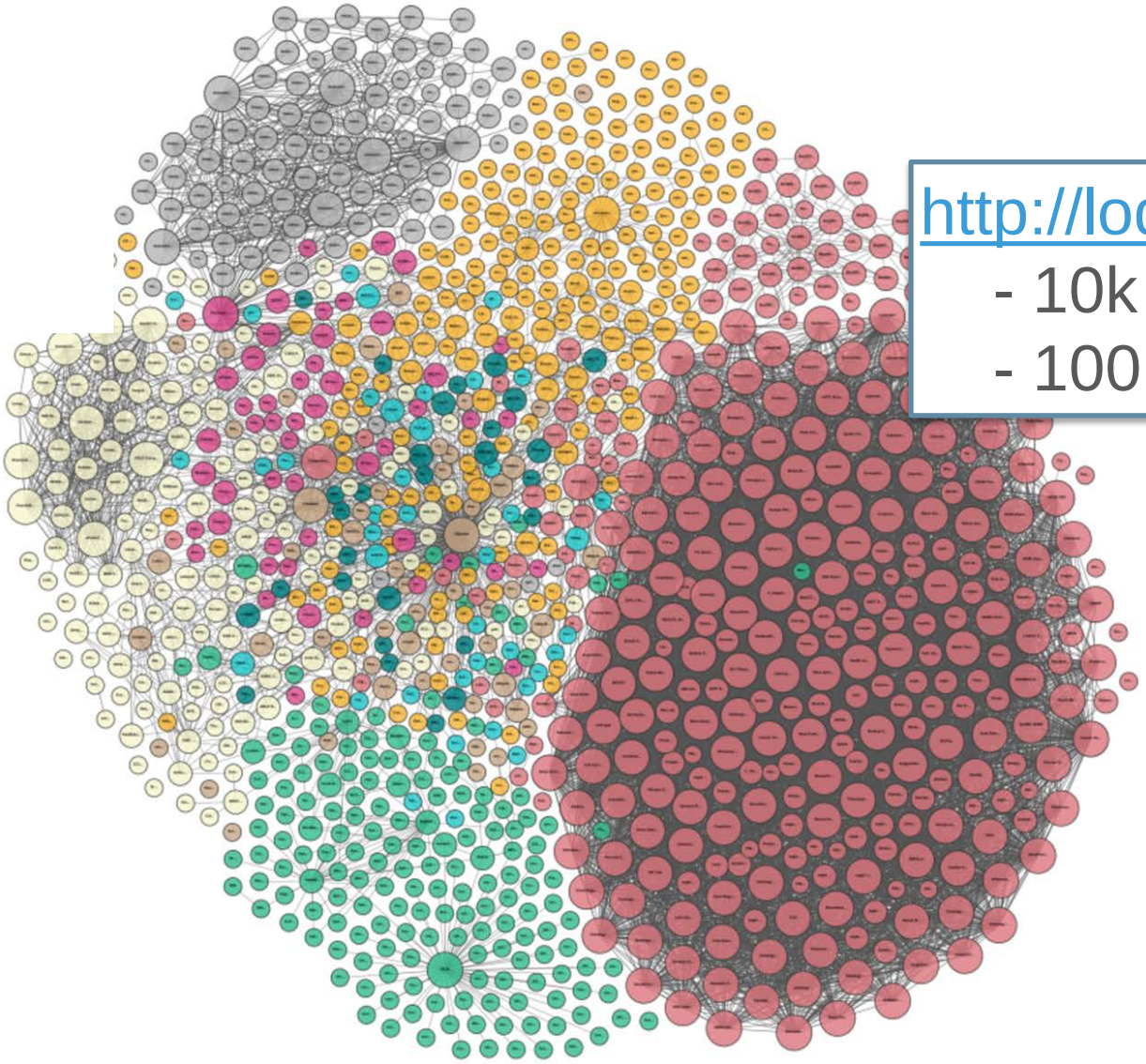
The Web of Data: Benefits of Linked Data at a glance

Why publish Linked Open Data?

- Increase discoverability and ranking through linking
- “The future library catalogue is the Internet!”
- Drive traffic and usage to our platforms
- Make our data machine-readable
- Be part of the LOD cloud



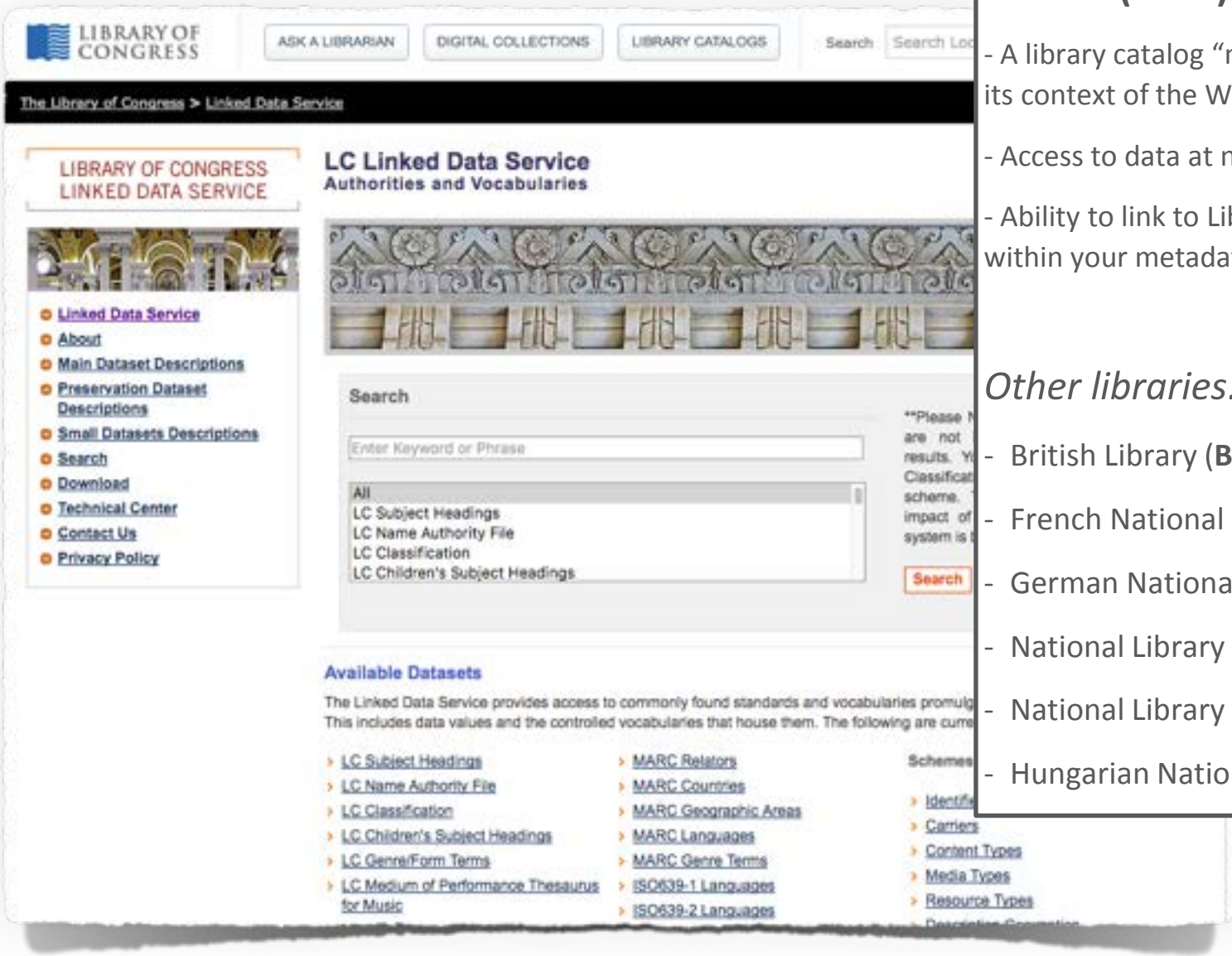
The Global Linked Data Cloud (2017)



http://lod-cloud.net/
- 10k datasets
- 100 BL triples

Legend
- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated
- Incoming Links
- Outgoing Links

Libraries using Linked Data



Library of Congress Linked Data Service (2009)

- A library catalog "must be designed by considering its context of the Web"

- Access to data at no cost.

- Ability to link to Library of Congress data values within your metadata via Linked Data.

Other libraries:

- British Library (**BL**)
- French National Library (**BNF**)
- German National Library (**DNB**)
- National Library of Spain (**BNE**)
- National Library of Sweden (**LIBRIS**)
- Hungarian National Library (**NSL**)

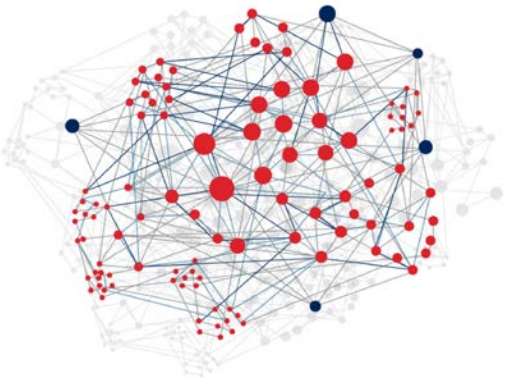
SPRINGER NATURE

just released

Springer Nature SciGraph

A Linked Open Data platform for the scholarly domain

SN SciGraph



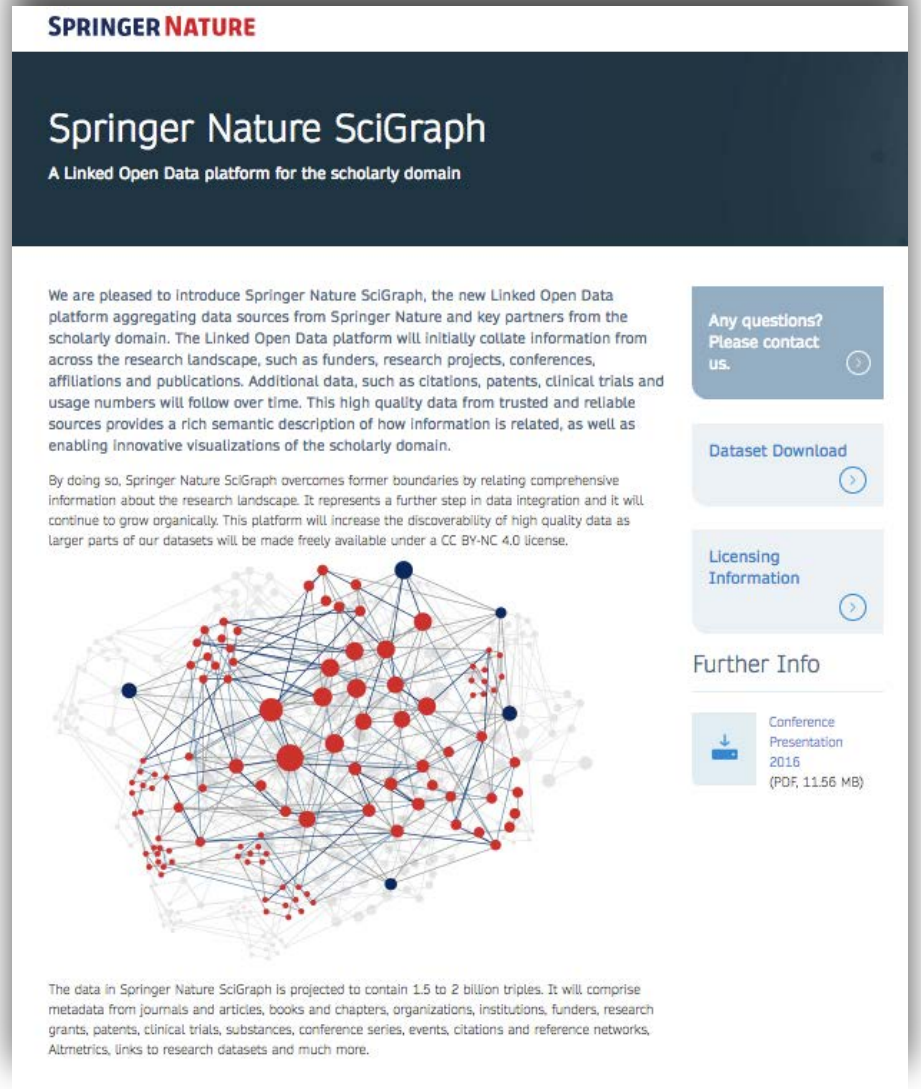
> Collaborative effort between Springer Nature and Digital Science

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

> Integrating data from a variety of sources using Linked Data technology

> Not just publications but a wealth of other related data

Linked Open Data Publishing: making science more accessible



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)

Feb. 2017 Data Release

At a glance:

- 150 M triples / 32 GB download size
- CC-BY-NC License

Metadata about:

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

Springer Nature SciGraph: Feedback on Twitter after PR

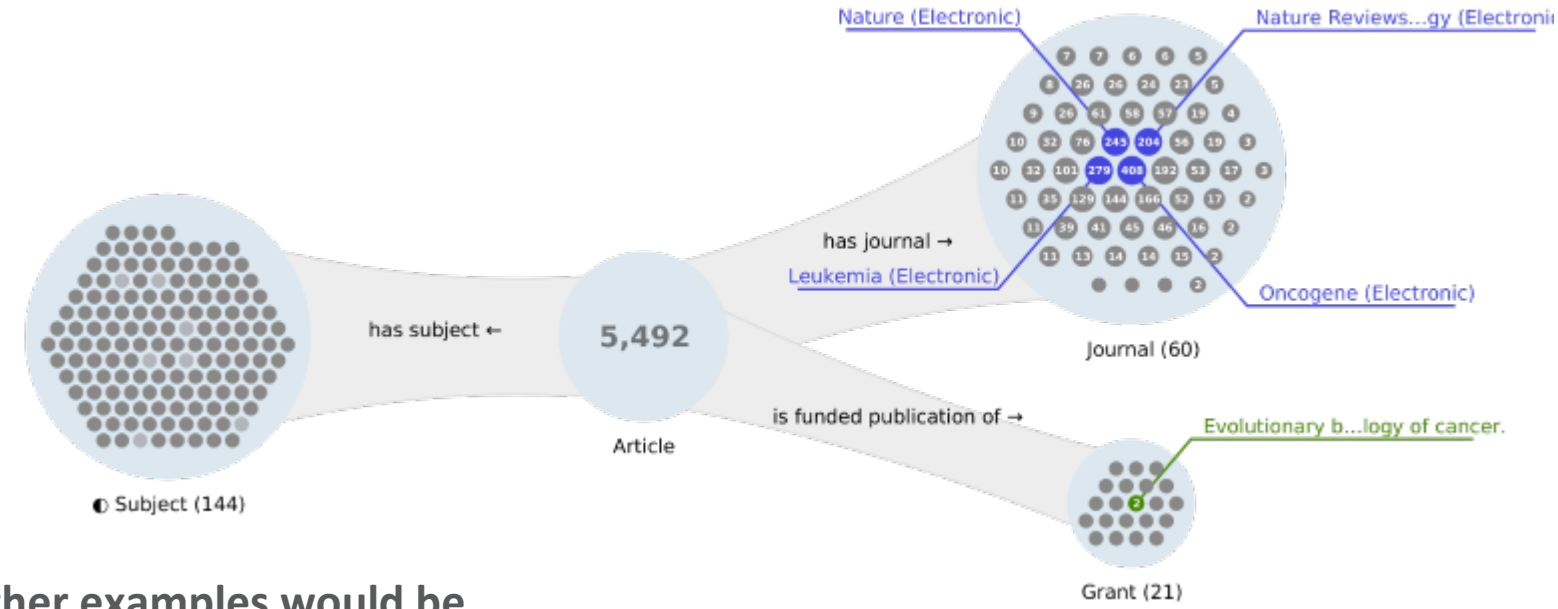
- **Selected tweets:**
  - "@SpringerNature Thank you for SciGraph integrated research data for helping the knowledge revolution #Semantic #Reason #ThinkingWeb"
  - "@SpringerNature @NatureCellBio SciGraph excellent new tool"
  - "Interesting, but only has info from one publisher. Is it extensible to other publishers? Single-publisher solutions are no solution!"
  - "Springer Nature SciGraph Dataset ネイチャー誌5年分の記事メタデータ1.55億トリプルをCC-BY-NCで提供。年内にさらに追加と。モデル説明あ"
  - "I hope this will go far beyond the mere understanding of 'which conferences are taking place in subject areas with rising funding'"
  - „Springer Nature SciGraph: Supporting open science and the wider understanding of research"

Workshop with DBpedia (April 2017)

- DBpedia is the Linked Open Data transformation of facts extracted from Wikipedia
- We are uploading our data sets with links to Wikipedia on the DBpedia GitHub repository in order to generate backlinks from DBpedia and increase traffic to our content platforms and usage of our SN SciGraph LOD.
- Currently in the process of hiring an intern that works both for DBpedia and SciGraph

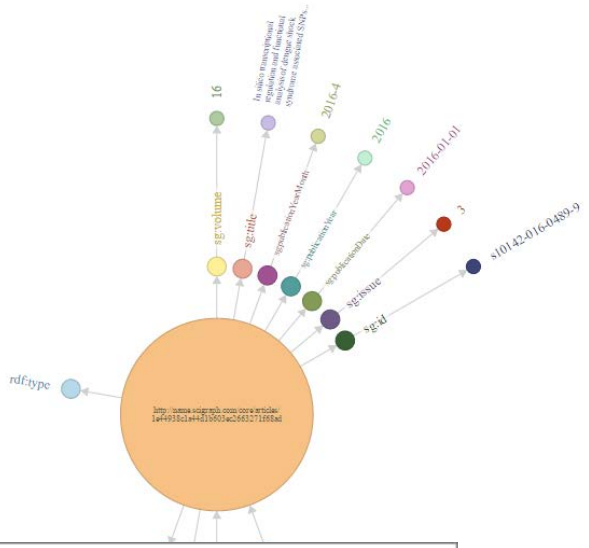
Examples of users working with SN SciGraph data

- **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.

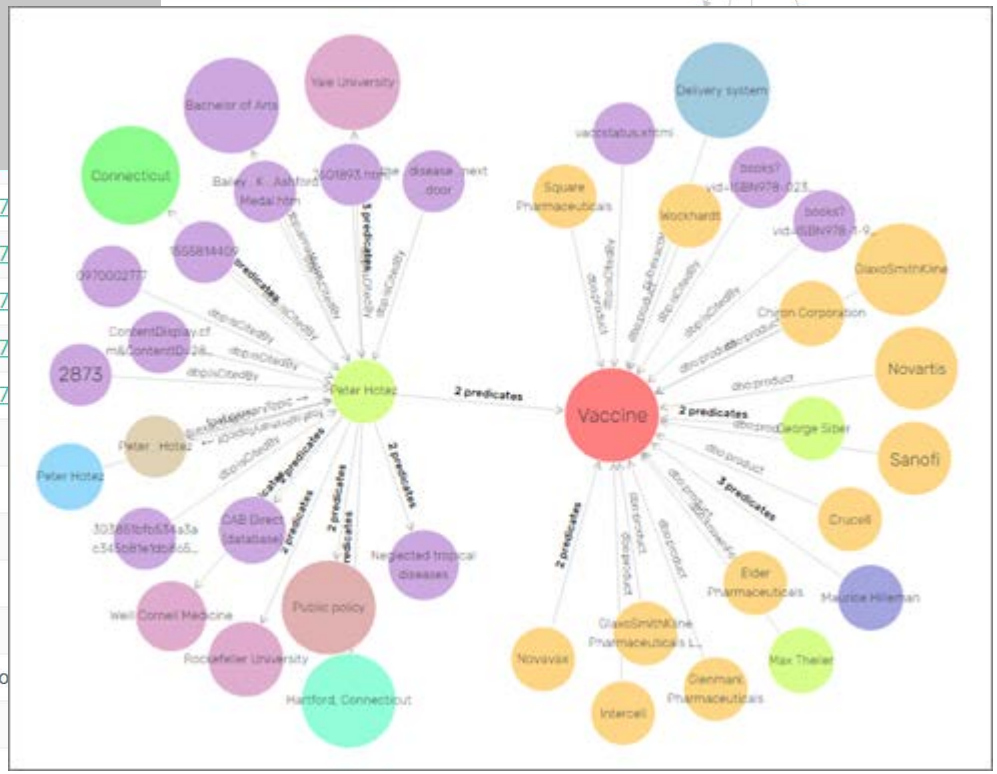


- **Other examples would be**
  - ResearchGraph (Australia)
  - Science Media Center (Cologne)

Springer Nature SciGraph: LOD Browsers



Property	Value
sg:doi	10.1007/s101
sg:doiLink	<http://dx.doi.
sg:hasAbstract	<http://name.s
sg:hasContribution	<http://name.scigraph.com/core/contributions/1e44938c1a44d1b603ec266327
sg:hasContribution	<http://name.scigraph.com/core/contributions/1e44938c1a44d1b603ec266327
sg:hasContribution	<http://name.scigraph.com/core/contributions/1e44938c1a44d1b603ec266327
sg:hasContribution	<http://name.scigraph.com/core/contributions/1e44938c1a44d1b603ec266327
sg:hasContribution	<http://name.scigraph.com/core/contributions/1e44938c1a44d1b603ec266327
sg:id	s10142-016-0489-9
sg:issue	3
sg:publicationDate	2016-01-01
sg:publicationYear	2016
sg:publicationYearMonth	2016-4
sg:title	In silico transcriptional regulation and functional analysis of dengue shock syndro
sg:volume	16
rdf:type	sg:Article



Status
- SN SciGraph Hack Day
- Analytics Dashboards

#2

Status
- SN SciGraph Hack Day
- Analytics Dashboards

#2.1

Key Partners

Digital Science is a primary data provider and development partner on the project - also helping to kick-start the project with senior staff joining our team.

derivo is supporting the project with their expertise in knowledge modeling. They also provide an instance of SemSpect for visually exploring SN SciGraph data: http://scigraph.semspect.de

We are using Ontotext's triple store GraphDB as scalable semantic graph database. Their support during the initiation of the project and afterwards has been excellent.

InfoChem provides named entity recognition in the chemical domain, annotating relevant substances, chemical compounds & molecules.

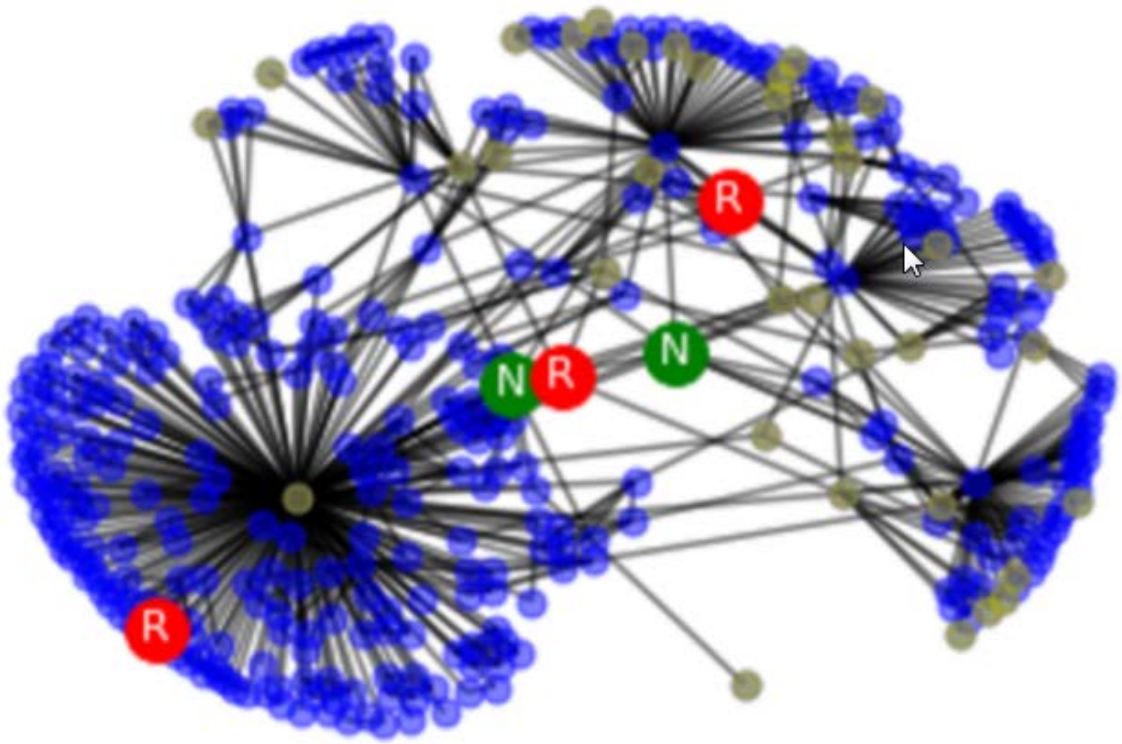
We will be storing the main concepts of our publications as extracted by Unsilo in our graph to cluster documents around certain topic areas.


a runaway
SciGraph
ACK DAY June 23rd
SPRINGER NATURE
SN SciGra

Springer Nature SciGraph: Hack Day June 23rd in London

What have participants worked on?

- "Helping you get into the right journal"
- "Connecting Articles and Data Repositories"
- "Peer Reviewer assignment system"



More project details of this productive day can be found in these slides.

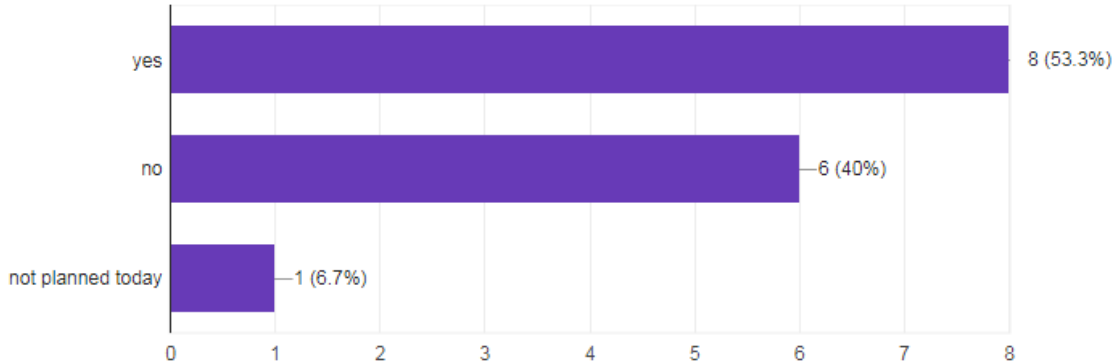
Springer Nature SciGraph: Hack Day June 23rd in London

What were the key insights?

- One day was enough time; the event was well structured (but some attendees of course wanted more time for hacking)
- Our data model is very much understandable; there is room for improvement around data quality, data coverage and data APIs – all of which were considered adequate
- A clear mandate that citations would be most useful to users – this is aligned with our goal to include citations in the graph next
- Finally, more than 50% would want to use the data commercially – (un)fortunately, our current CC-BY-NC license doesn't allow that
- It is really affirming that 87% would like to remain in contact with us and be informed about upcoming events.

License: are you interested in using the data commercially?

15 responses



You can read more details about the day in [my blog post on Hive](#).

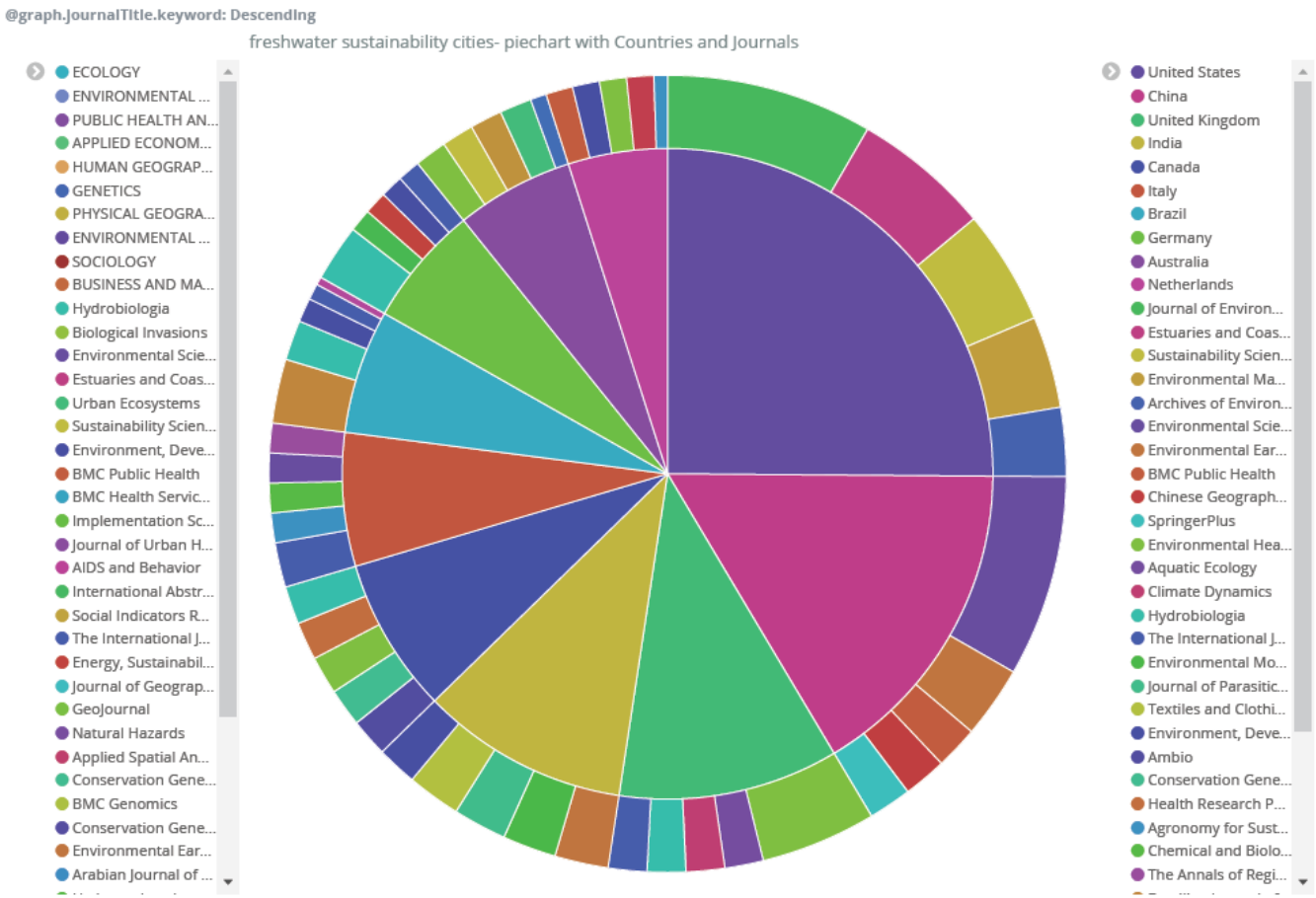
g mentions

Status
- SN SciGraph Hack Day
- Analytics Dashboards

#2.2

Springer Nature SciGraph: What are we currently working on?

- **Analytics dashboard**
  - Make data available and queryable within Elasticsearch
  - Use Kibana to create visualizations on top of the data



SciGraph: questions from sales, marketing and editorial

- What's the **Open Access adoption ratio**, i.e. how many articles are open access in this journal and what was the trend over the last 5 years?
- How many articles of this journal were **indexed in Web of Science / Scopus** in the last 3 years?
- What is the **funding pattern and amount** for the top 10 funded institutes in a given country?
- Which institution shows the **most growth in article output** over the past 3 years?
- Who has **interacted with Springer Nature** at a certain institution?
- Which **conferences** take place in subject areas that show an increase of funding over the last 5 years in a given geographical area?
- 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?

- 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

- 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 in this area?

- i.e. "where you can see an institution's relationships with Springer Nature'
- How r
- How many a
- simple
  - SN editor
  - SN reviewers
  - and grants by fields
  - output and citations
  - ey are publishing in
  - rs from that institute
  - ed with that institute
- ourse 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 w development of research in fie

Article - FieldC

uthors with ORCID
uthors without ORCID

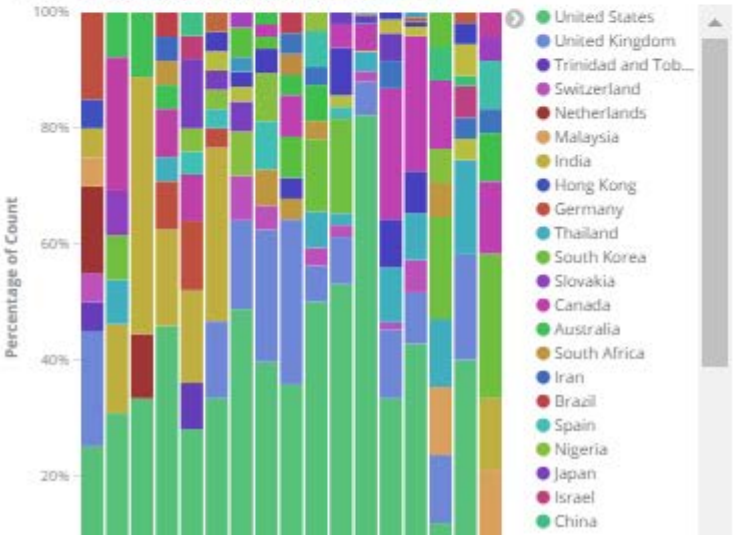
UBLIC HEALTH A...
LANT BIOLOGY
HARMACOLOGY ...
EDICAL AND HE...
OMPLEMENTARY...
LINICAL SCIENCES
IOLOGICAL SCIE...
IOCHEMISTRY A...
PECIALIST STUDI...
SYCHOLOGY AN...
SYCHOLOGY
HYSICAL CHEMIS...
AEDIATRICS AND...
THER MEDICAL A...
NCOLOGY AND ...
UTRITION AND D...
EUROSCIENCES
ICROBIOLOGY
EDICAL PHYSIOL...

Article - by country

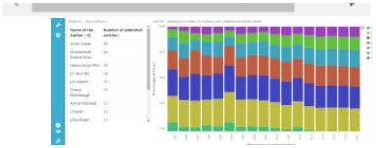
contribution.affiliation.countryName: Descending	Count
United States	562
China	281
South Korea	237
India	134
Malaysia	128
Germany	108
United Kingdom	95
Australia	94
Canada	92
Taiwan	90

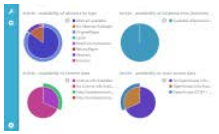
Export: Raw Formatted

Article - top ten countries distribution as a timeline



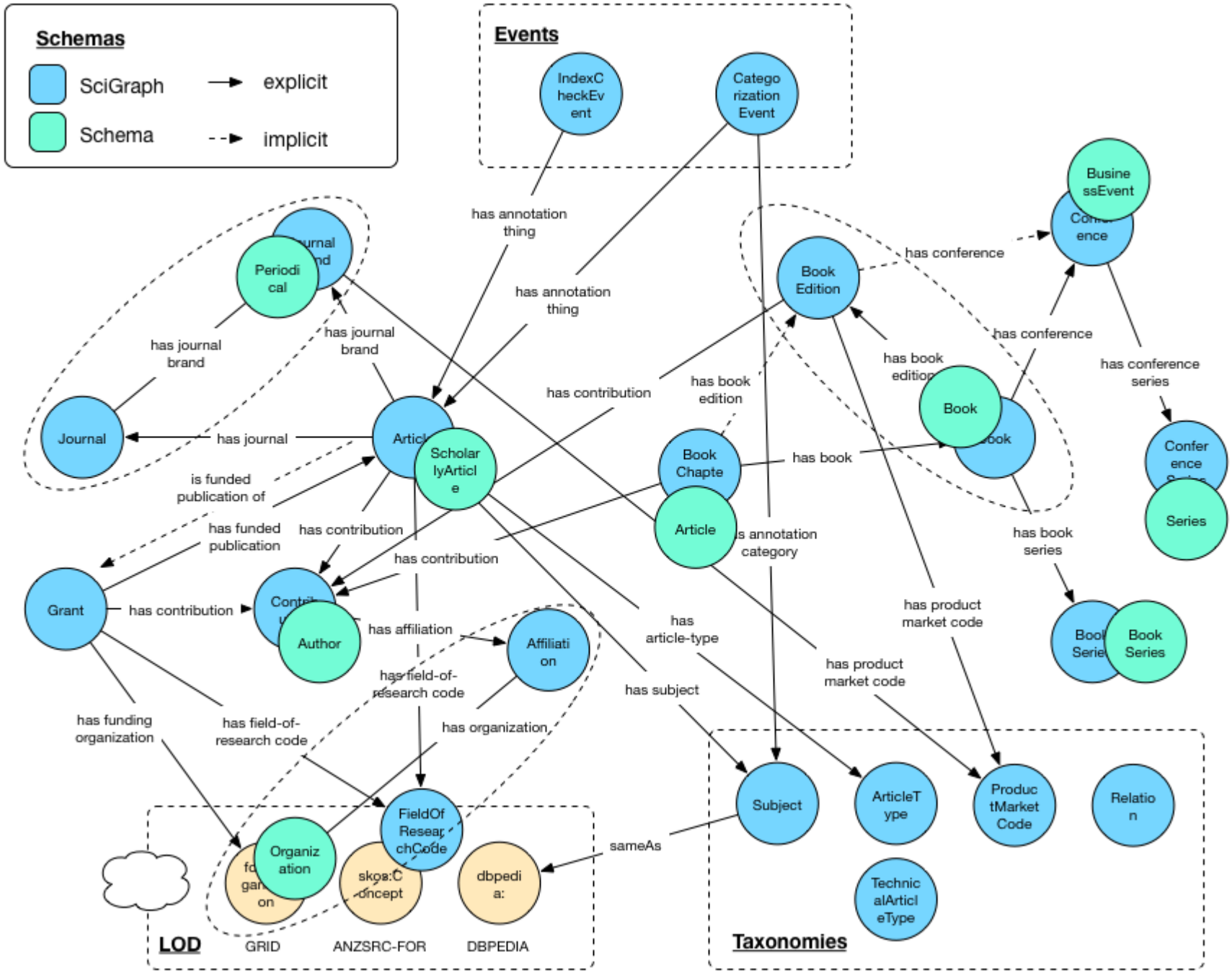
Article - table



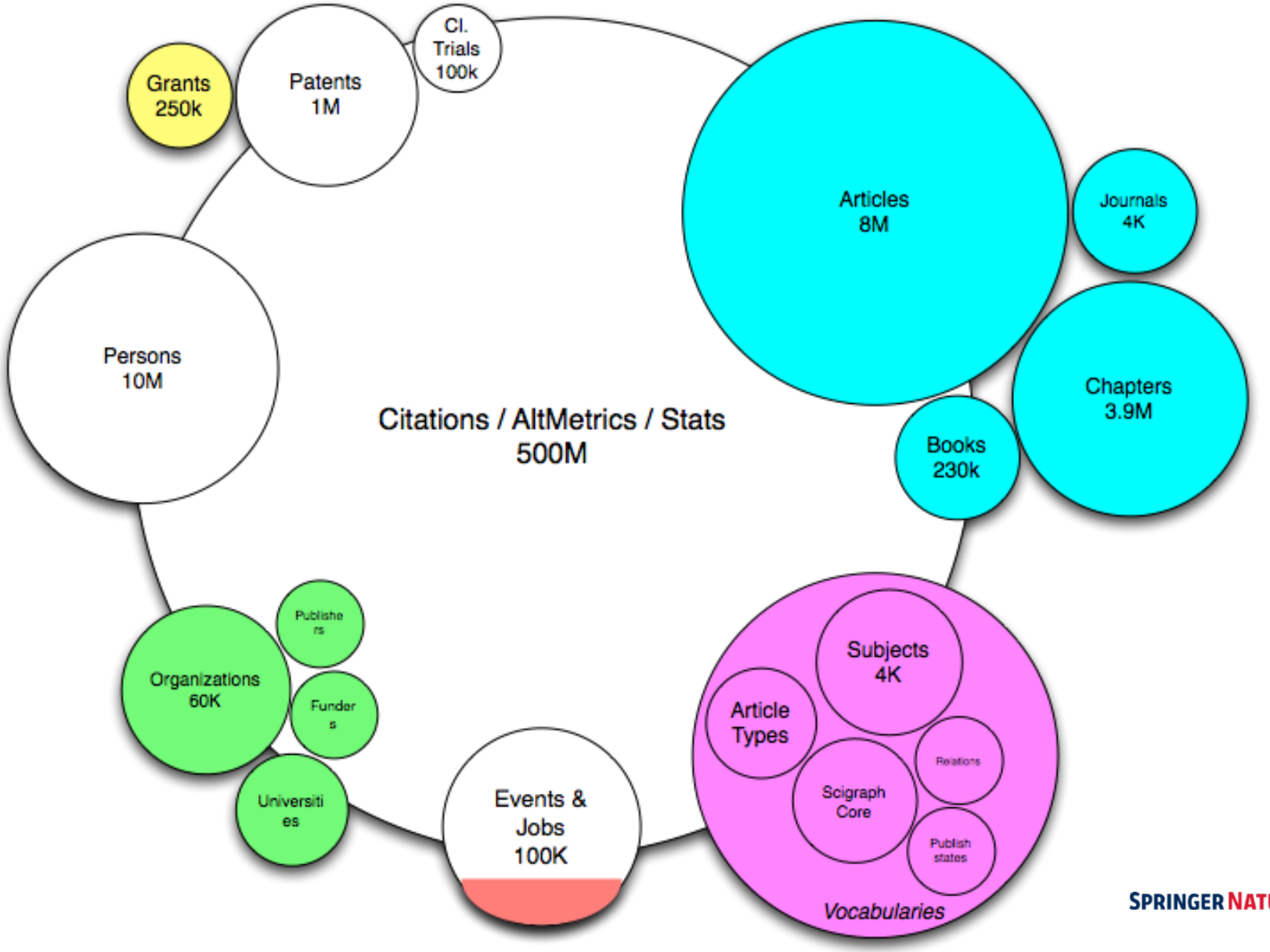


Data: Roadmap EOY and beyond

#3

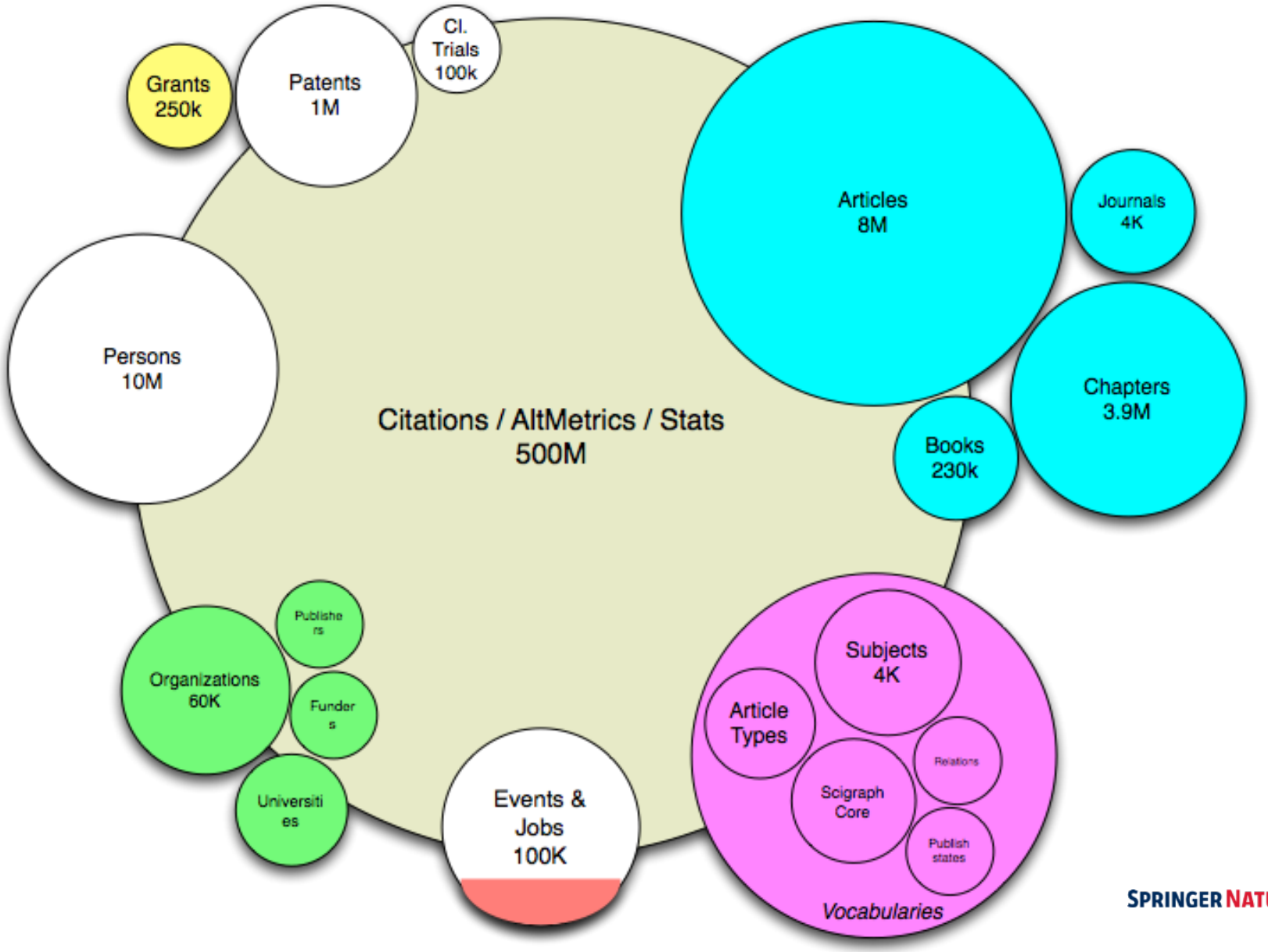

Schemas
SciGraph
explicit
Schema
implicit
Events
IndexCheckEvent
Categorization Event
has annotation thing
has annotation thing
Periodical
has journal brand
has journal brand
Journal
has journal
Article
ScholarlyArticle
is funded publication of
has funded publication
has contribution
has contribution
has contribution
has contribution
Grant
Author
has affiliation
Affiliation
has field-of-research code
has organization
has funding organization
has field-of-research code
FieldOfResearchCode
Organization
skos:Concept
dbpedia:
sameAs
LOD
GRID
ANZSRC-FOR
DBPEDIA
Book Edition
has conference
BusinessEvent
Conference
has conference
has conference series
has book edition
has book edition
Book
Book
has book
Book Chapter
Article
has annotation category
has book series
Conference Series
Series
Book Series
Book Series
has product market code
has product market code
has article-type
has subject
Subject
ArticleType
ProductMarketCode
Relation
TechnicalArticleType
Taxonomies

Springer Nature SciGraph: Data Landscape - now (Q2)



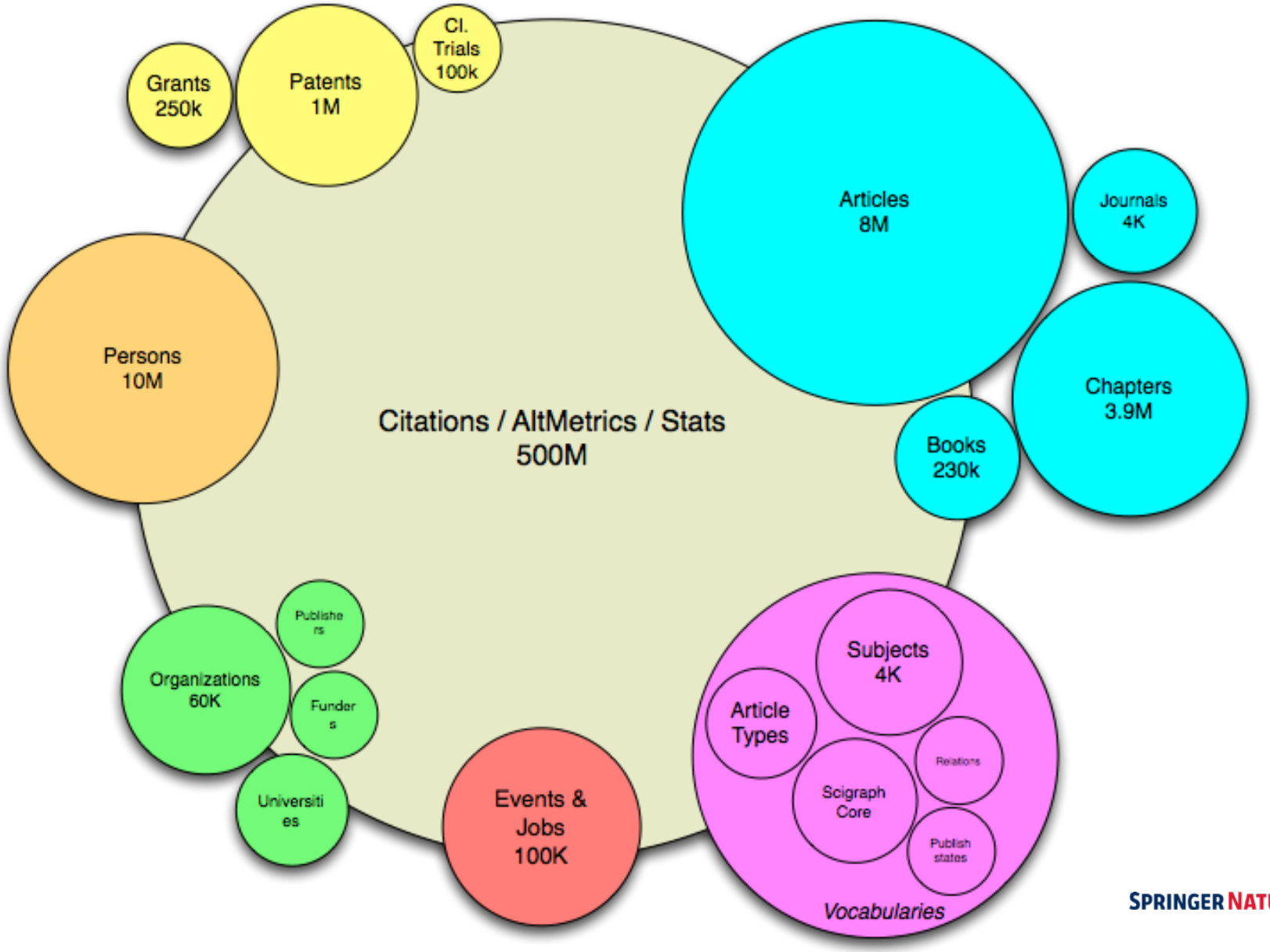
SPRINGER NATURE

Springer Nature SciGraph: Data Landscape - Q3



SPRINGER NATURE

Springer Nature SciGraph: Data Landscape - Q4



SPRINGER NATURE

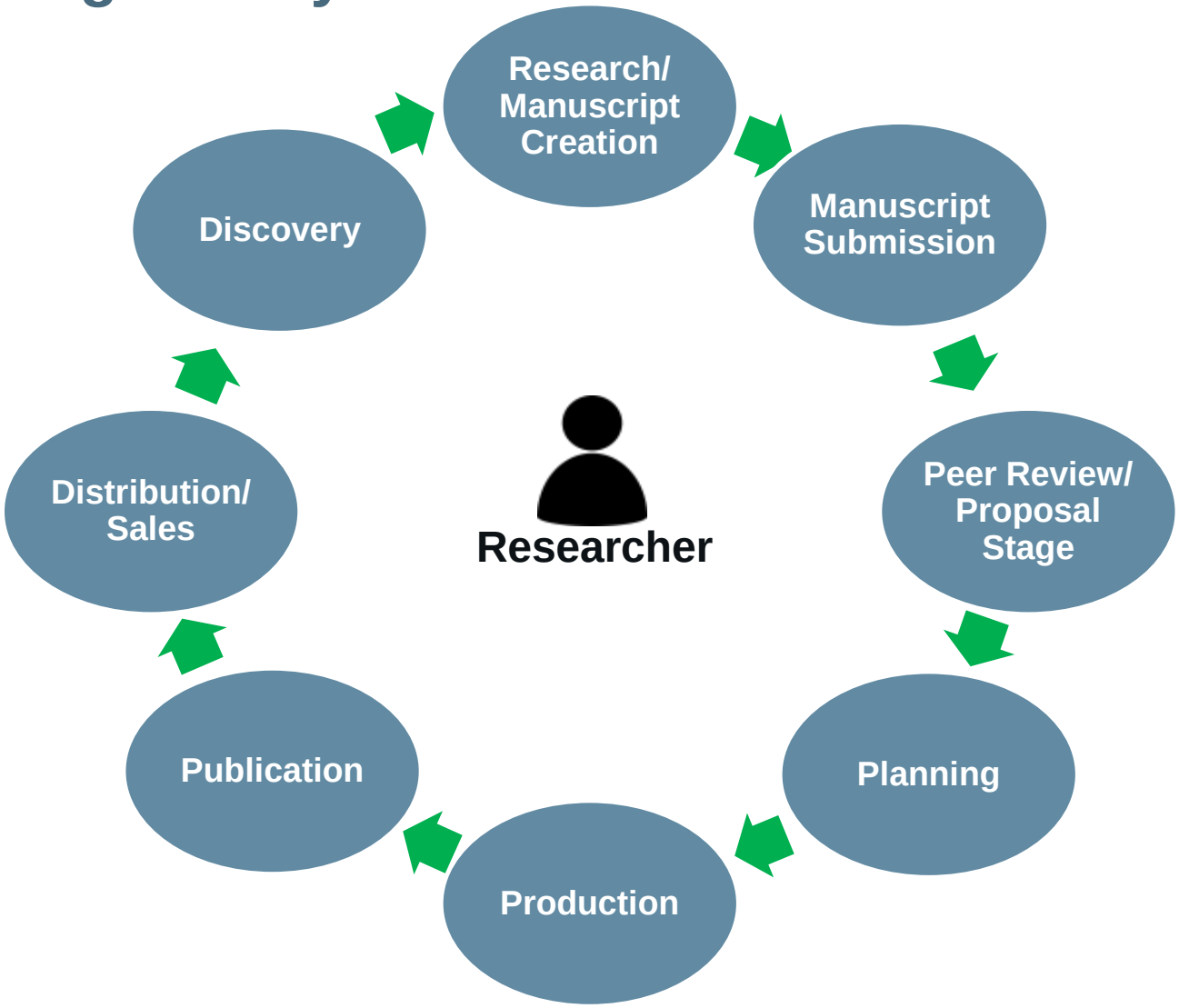
Data Roadmap 2017

1	**Journals + Articles data**
2	**Institutions (GRID)**
3	**Books + Chapters data**
4	**Field of Research categories (FOR)**
5	**Conferences**
6	Disambiguated authors
7	Citations / References
8	**Research grants** and OA funding information
9	Download + reader numbers
10	Concepts + chemical substances
11	Patents + clinical trials
12	Links to research datasets

Data Roadmap: General approach

1. Scale	2. Size	3. Scope
1.5 billion facts (triples) at the end of 2017	Extend focus from Springer Nature publications to the entire research publishing landscape	Shift focus from published documents to submission, planning and production

Publishing Life Cycle




ANY
QUESTIONS
?

Thank you

Email : markus.kaindl@springernature.com

Senior Manager Semantic Data
& Product Owner SN SciGraph

>> How to get in touch:

- Portal
  http://www.springernature.com/scigraph
- E-Mail
  scigraph@springernature.com
- Twitter
  #scigraph