SN SciGraph



# Modeling publications in SN SciGraph 2012-2019

June 2019

Michele Pasin
Lead Data Architect
Tech Product Owner

DIGITAL science

SPRINGER NATURE

# Outline

1. A bit of history: how we got here
   - The SciGraph project: motivation, applications and data releases

2. Modeling the publications domain
   - Three phases

3. Conclusions

# SPRINGER NATURE


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

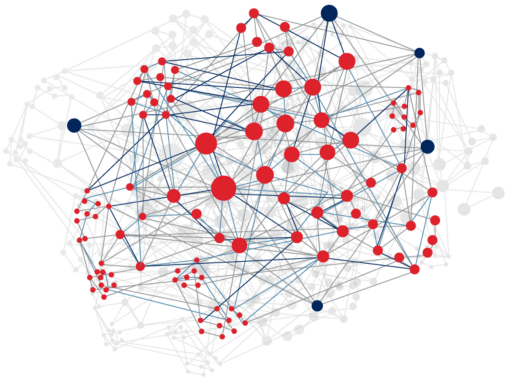
**Digital Science** is a **technology** company formed in 2010 that focuses on strategic investments into startup companies that support the **research lifecycle**. In 2018 it launched Dimensions, a scholarly search engine that is free to use.

SPRINGER NATURE

# Springer Nature SciGraph

A Linked Open Data platform for the scholarly domain

SN SciGraph



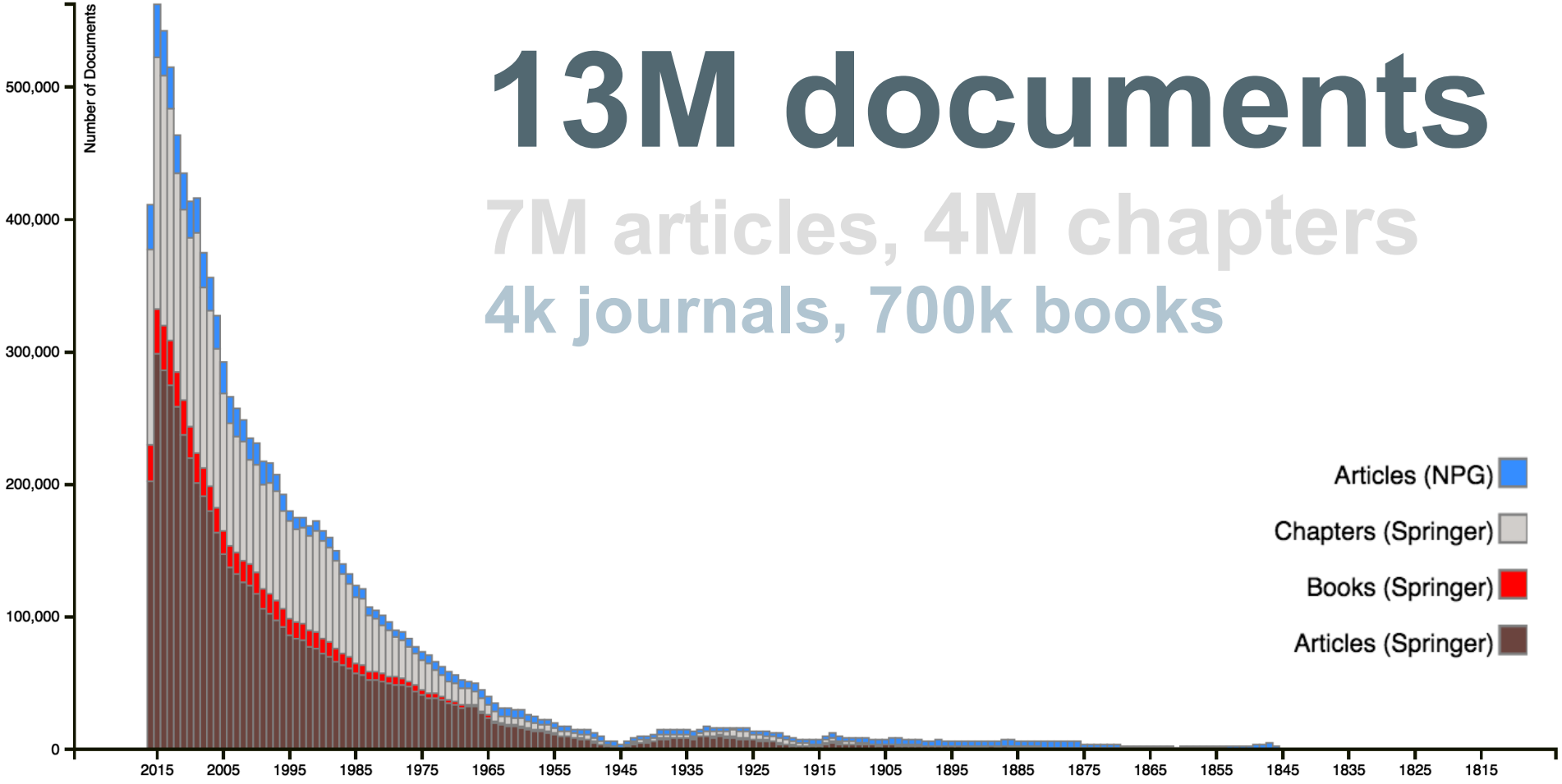
> Collaborative effort between Springer Nature and Digital Science (mid 2016)

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

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

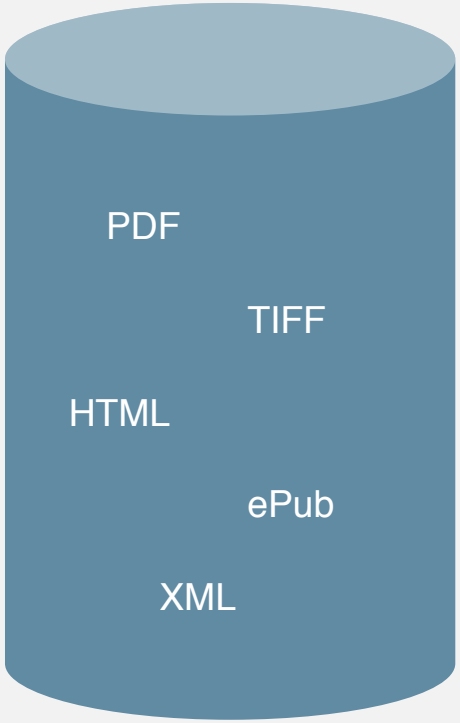
www.springernature.com/scigraph

# Motivation: integrating Springer Nature archive



# Vision

## Today: Content base

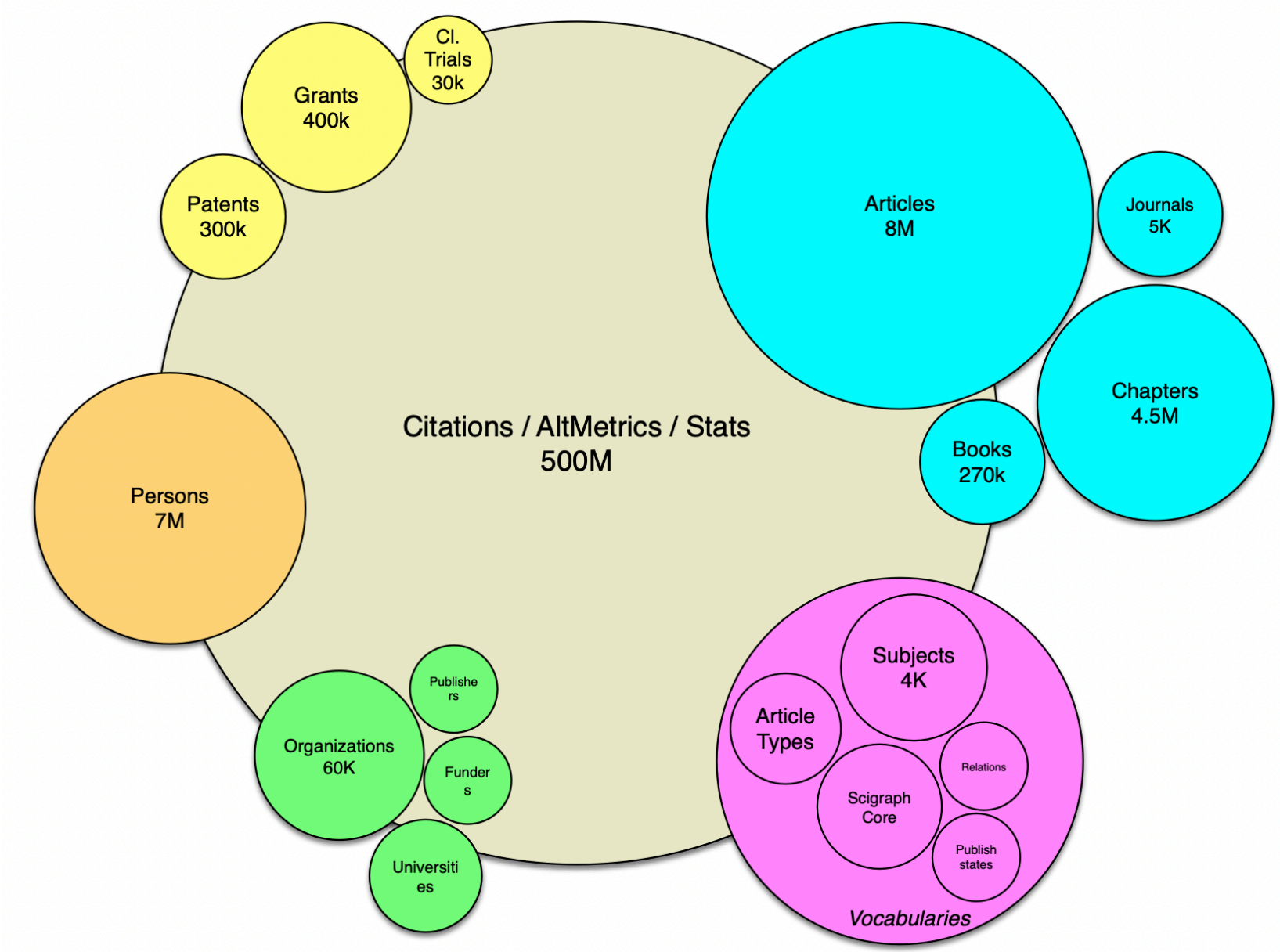



We publish science

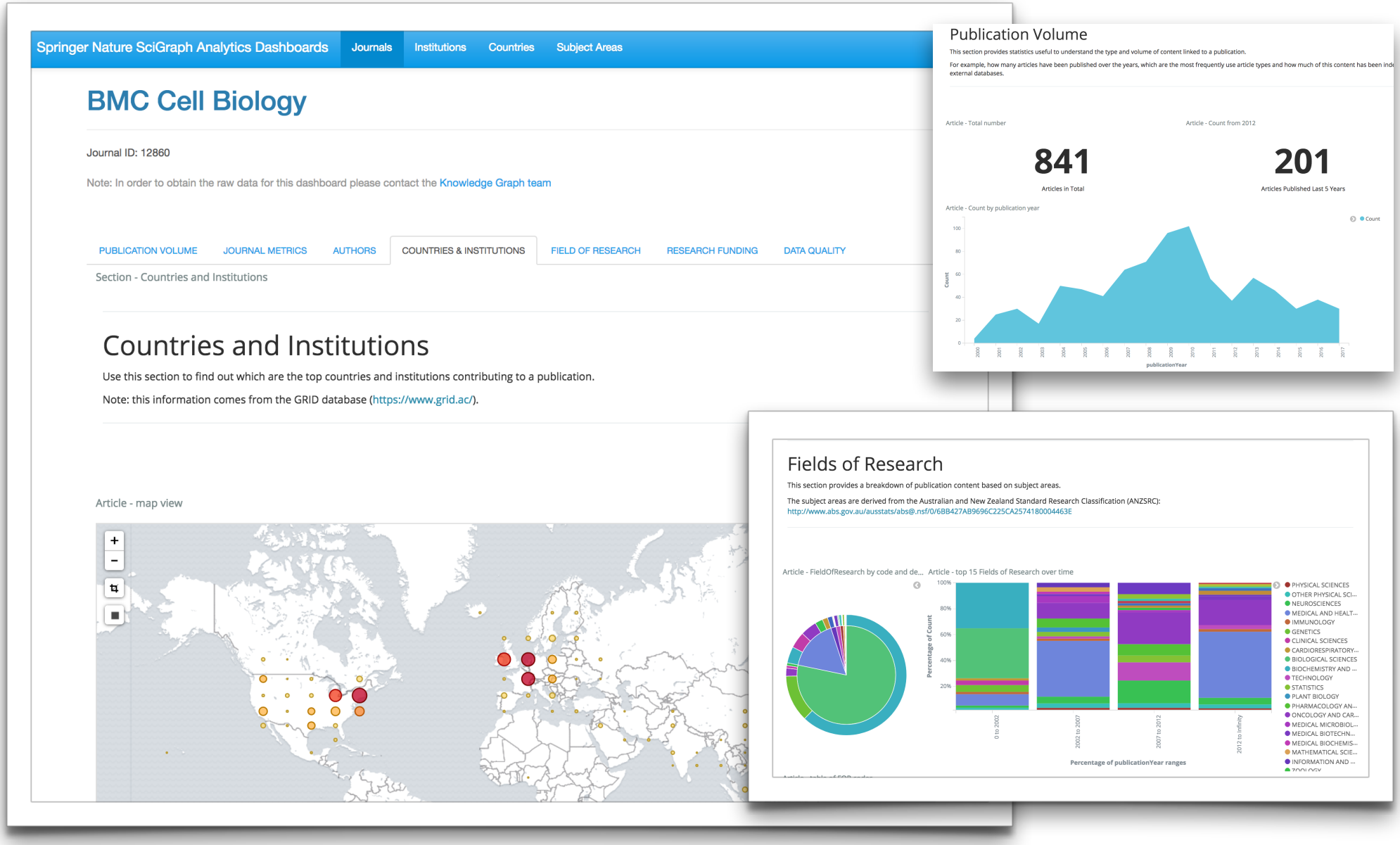
## Tomorrow: Knowledge Graph

We manage knowledge

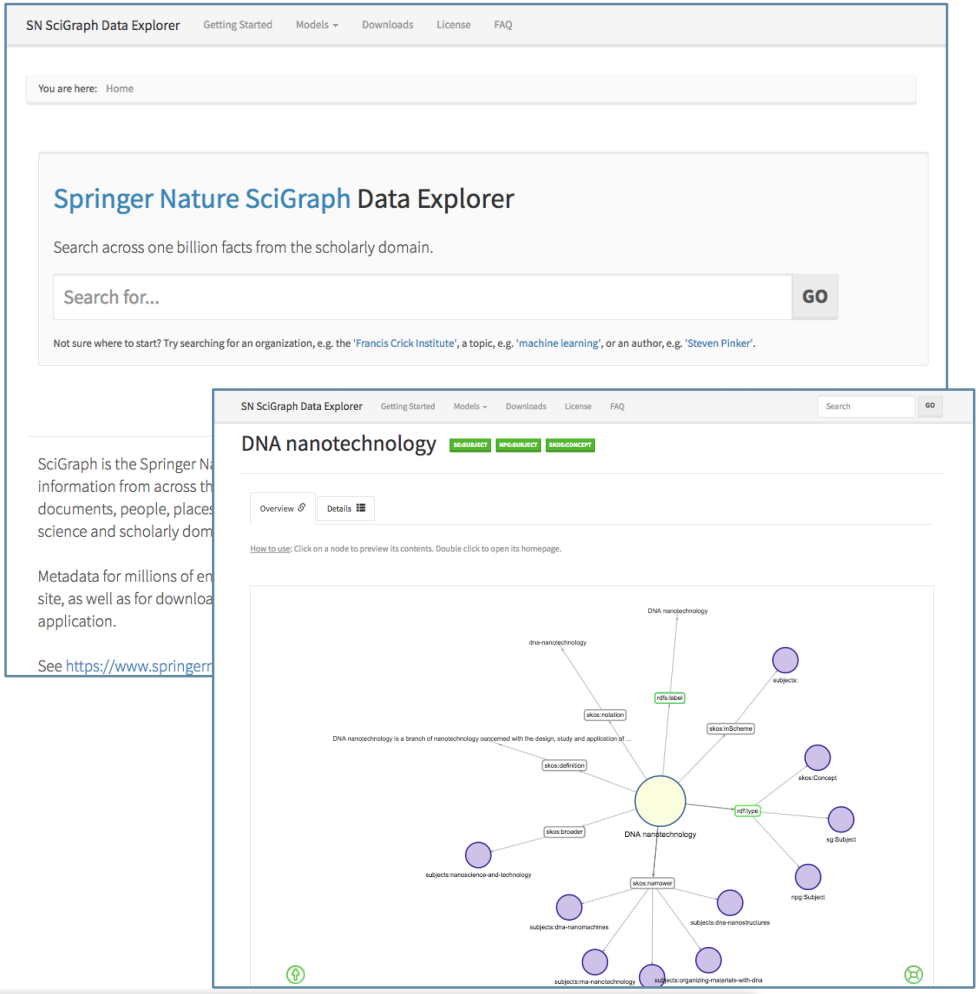
# SciGraph data landscape



# Applications: analytics dashboards for editors, publishers etc..



# Applications: SciGraph open data publishing




http://scigraph.springernature.com

## Releases

- Three major releases: 2017 (Feb & Nov), 2019 (Jan)
- Currently: daily updates for publications

## Features

- Linked Data Explorer (dereference & search)
- Bulk downloads (1 Billion+ triples)

## License

- Hybrid model
- CC-BY most metadata; CC-BY-NC abstracts and grants; CC-0 conferences data

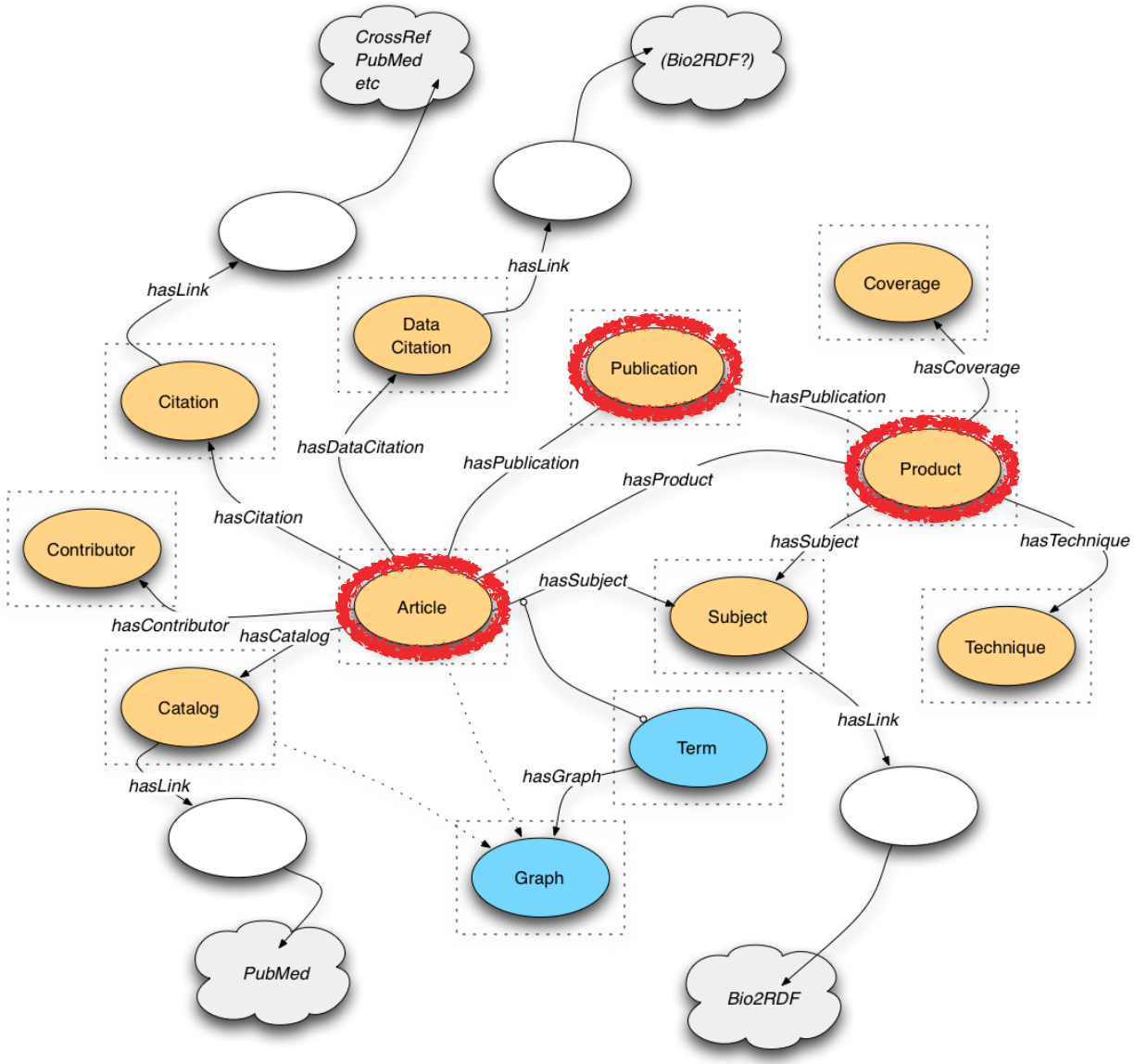
# Modeling Publications

## Challenges and Solutions

# Summary of approaches

1. **[2012-14] Ontologies Mix and Match**
2. [2015-17] Bespoke ‘SciGraph’ Ontology
3. [2018-19] Building on Schema.org

# 2012-2014: Ontologies Mix & Match



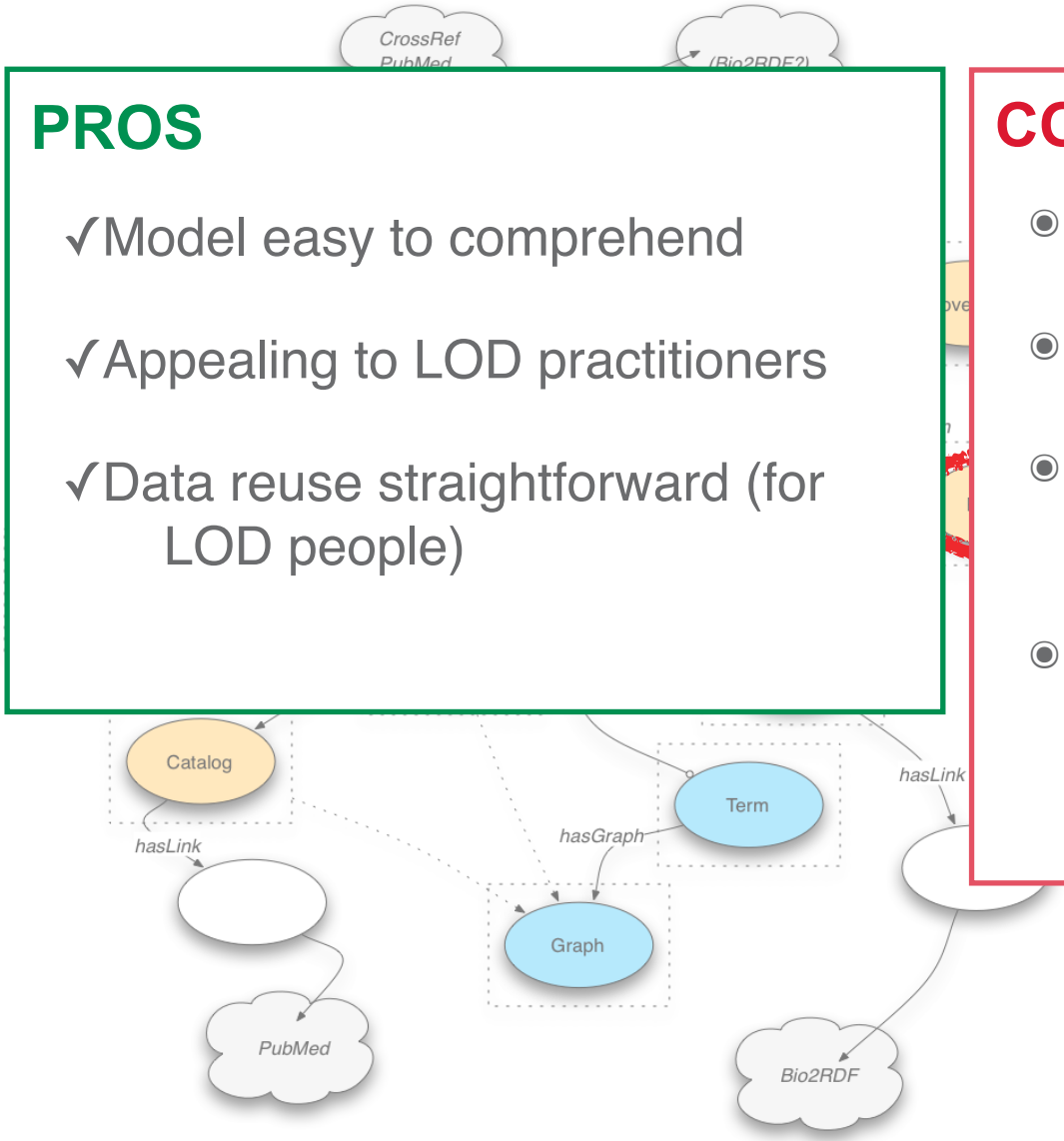

- **Common vocabularies** were used as much as possible

- **bibo** :issue, :pageStart, pageEnd, :Volume
- **dc** :identifier, :publisher, :title, :subject
- **prism** :copyright, :doi, :genre, :number, :publicationDate, :url, :volume
- **foaf** :name, :familiyName, :givenName
- **skos** :broader, :label, :Concept
- **npg** :hasProduct, :doihash, :hasCitation, :hasContributor, :hasPublication, :hasDataCitation etc..

- **RDF exported** only, not used internally

- **Flat publications model** main objects are represented, but no hierarchy

# 2012-2014: Ontologies Mix & Match




## PROS

- ✓Model easy to comprehend
- ✓Appealing to LOD practitioners
- ✓Data reuse straightforward (for LOD people)

## CONS

- Model rather simplistic
- Hard to maintain / extend / query
- Loose semantics, not suited for formal reasoning
- Not solid enough for internal uptake at Springer Nature

- **npg** :hasProduct, :doihash, :hasCitation, :hasContributor, :hasPublication, :hasDataCitation etc..

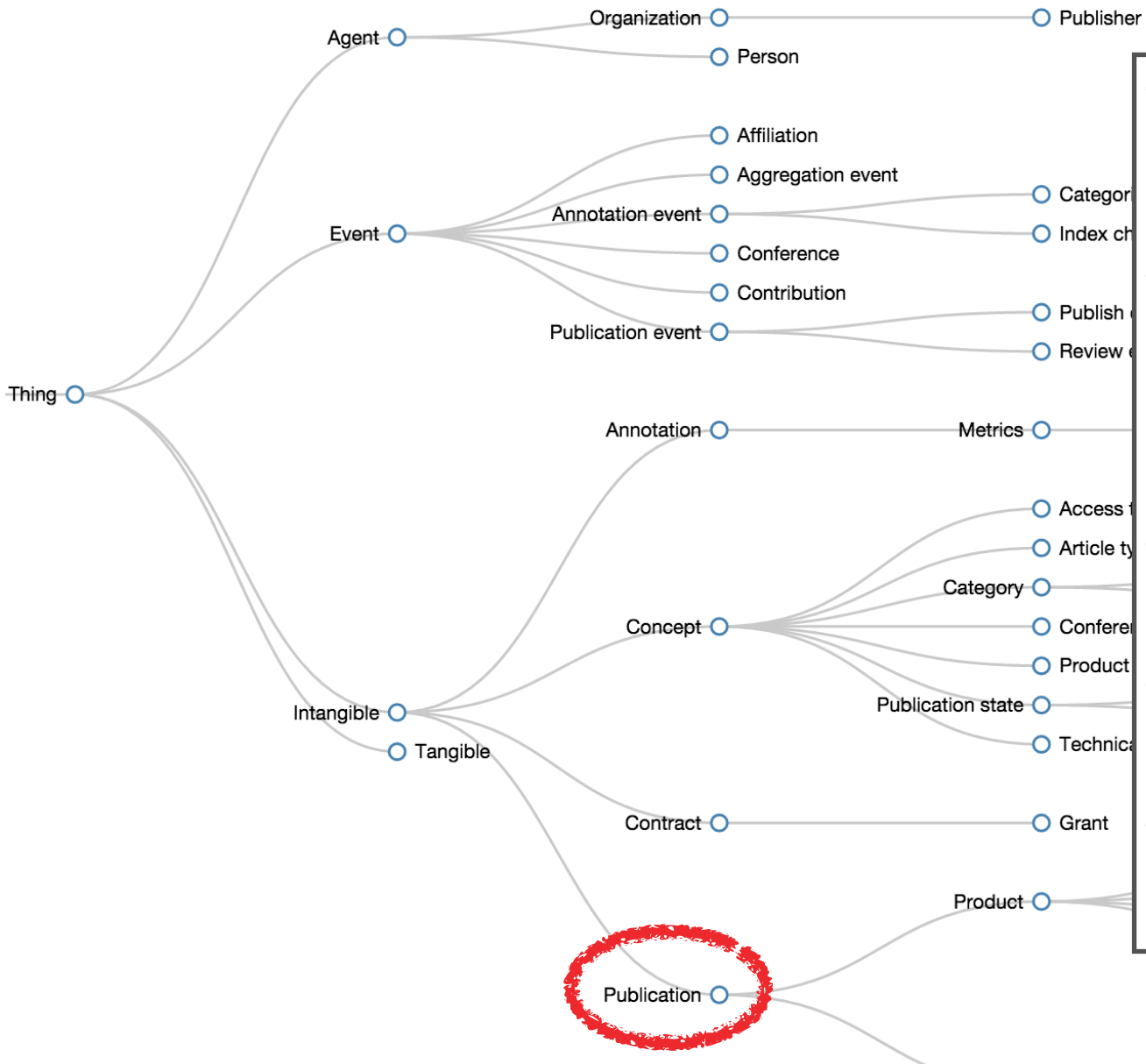
# Summary of approaches

1. [2012-14] Ontologies Mix and Match

2. **[2015-17] Bespoke ‘SciGraph’ Ontology**

3. [2018-19] Building on Schema.org

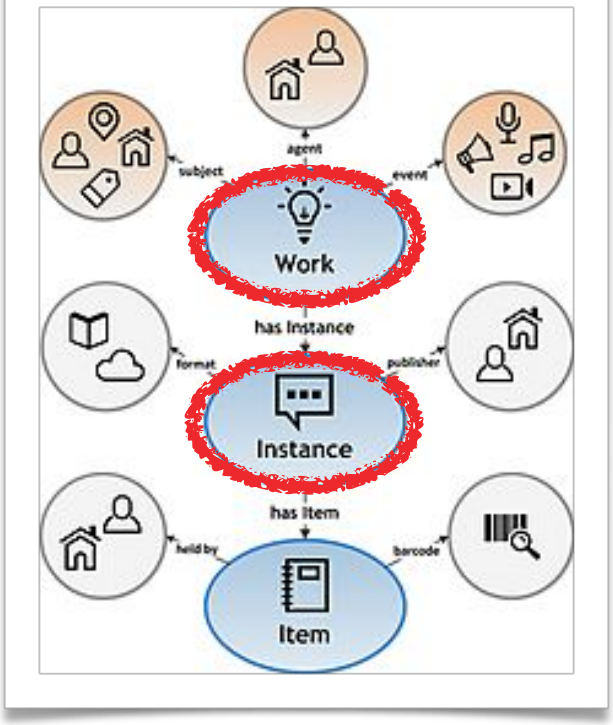
# 2015-2017: Bespoke ‘SciGraph’ Ontology




- **Scigraph** - unique ontology and namespace (~60 classes, ~230 properties)
- **Regimented model**: focus on internal coherence; no blank nodes; domain/range specs; naming conventions etc.. **(see Hammond, Pasin 2015)**
- **Mappings** to other ontologies published separately
- **Top level**: similar to CIDOC-CRM (event based modeling)
- **Publications**: compatible with BibFrame

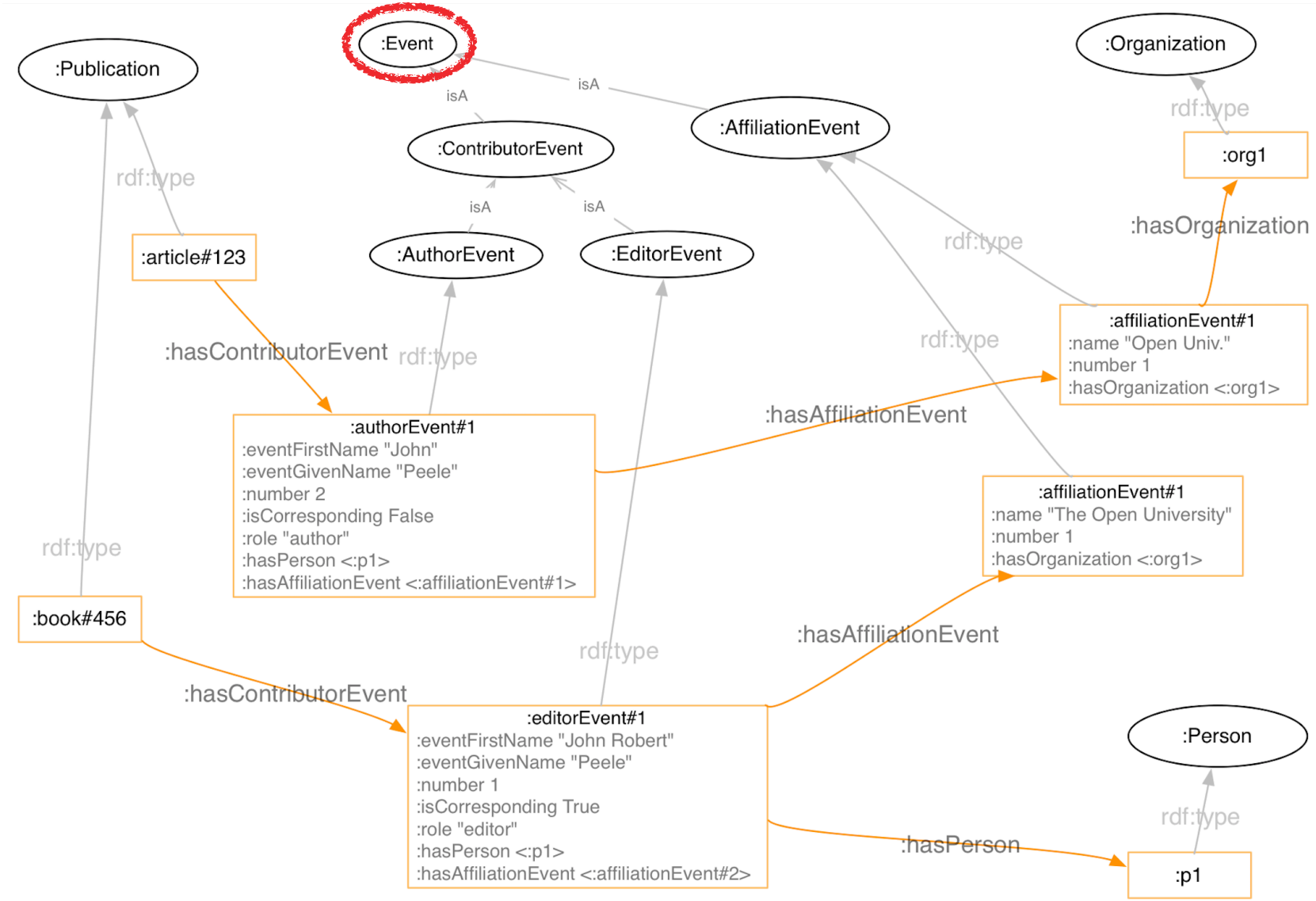
# 2015-2017: Publications in ‘SciGraph’ Ontology

**BIBFRAME 2.0 data model**




- Product
  - Article
  - Book
  - Book chapter
  - Journal
- Work
  - Monograph
    - Book edition
  - Serial
    - Book series
    - Journal brand

# 2015-2017: Events in ‘SciGraph’ Ontology



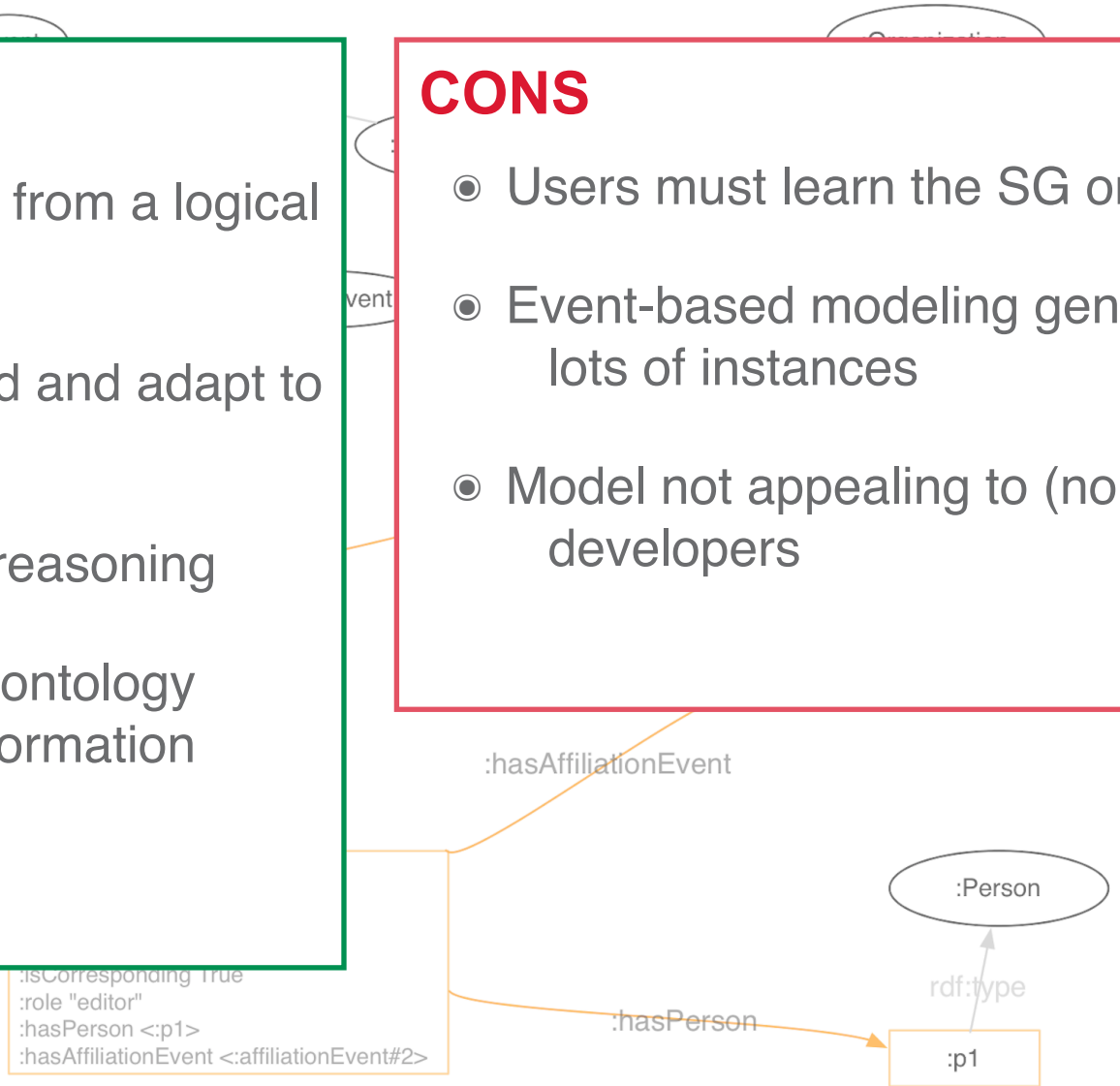
# 2015-2017: Bespoke ‘SciGraph’ Ontology

## PROS

- ✓Model very coherent from a logical perspective
- ✓Model easy to extend and adapt to project needs
- ✓Supports automatic reasoning
- ✓ Well understood by ontology specialists and information scientists

## CONS

- Users must learn the SG ontology
- Event-based modeling generates lots of instances
- Model not appealing to (non LOD) developers



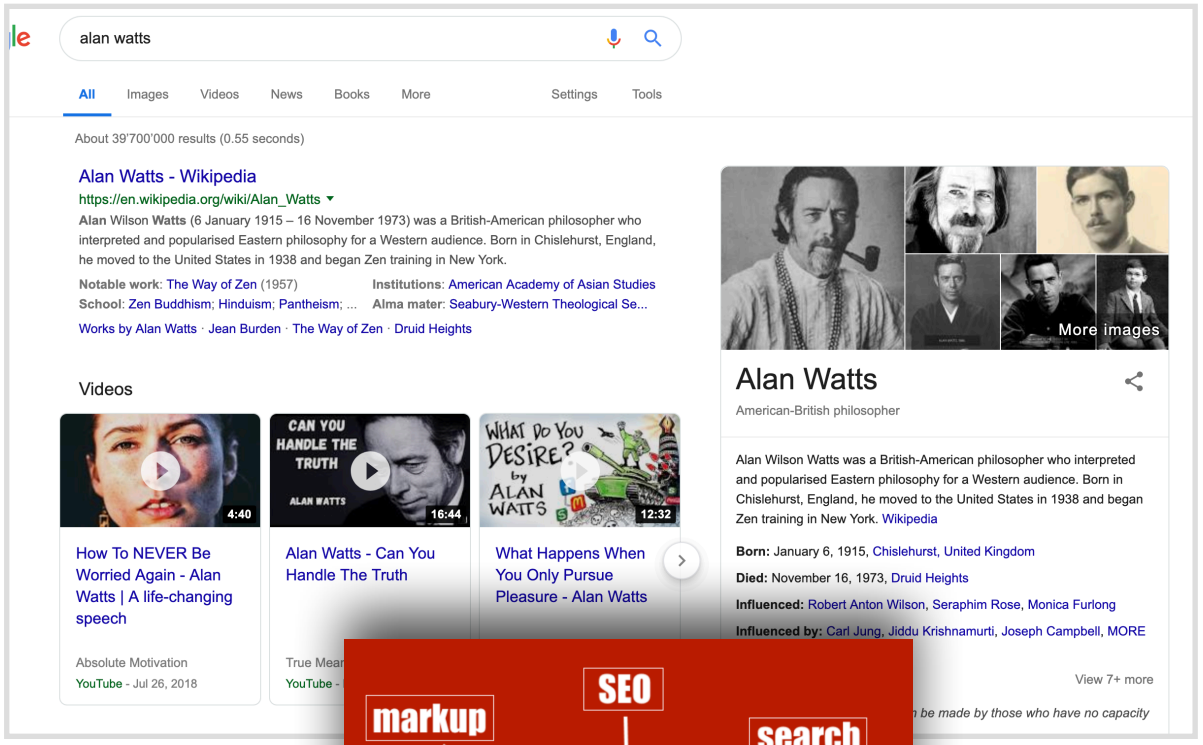
# Summary of approaches

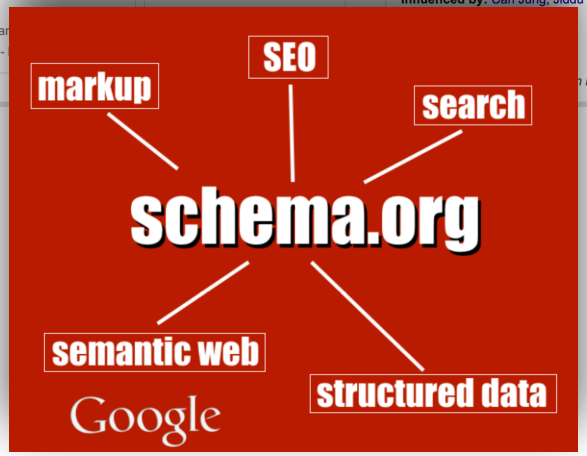
1. [2012-14] Ontologies Mix and Match

2. [2015-17] Bespoke ‘SciGraph’ Ontology

3. **[2018-19] Building on Schema.org**

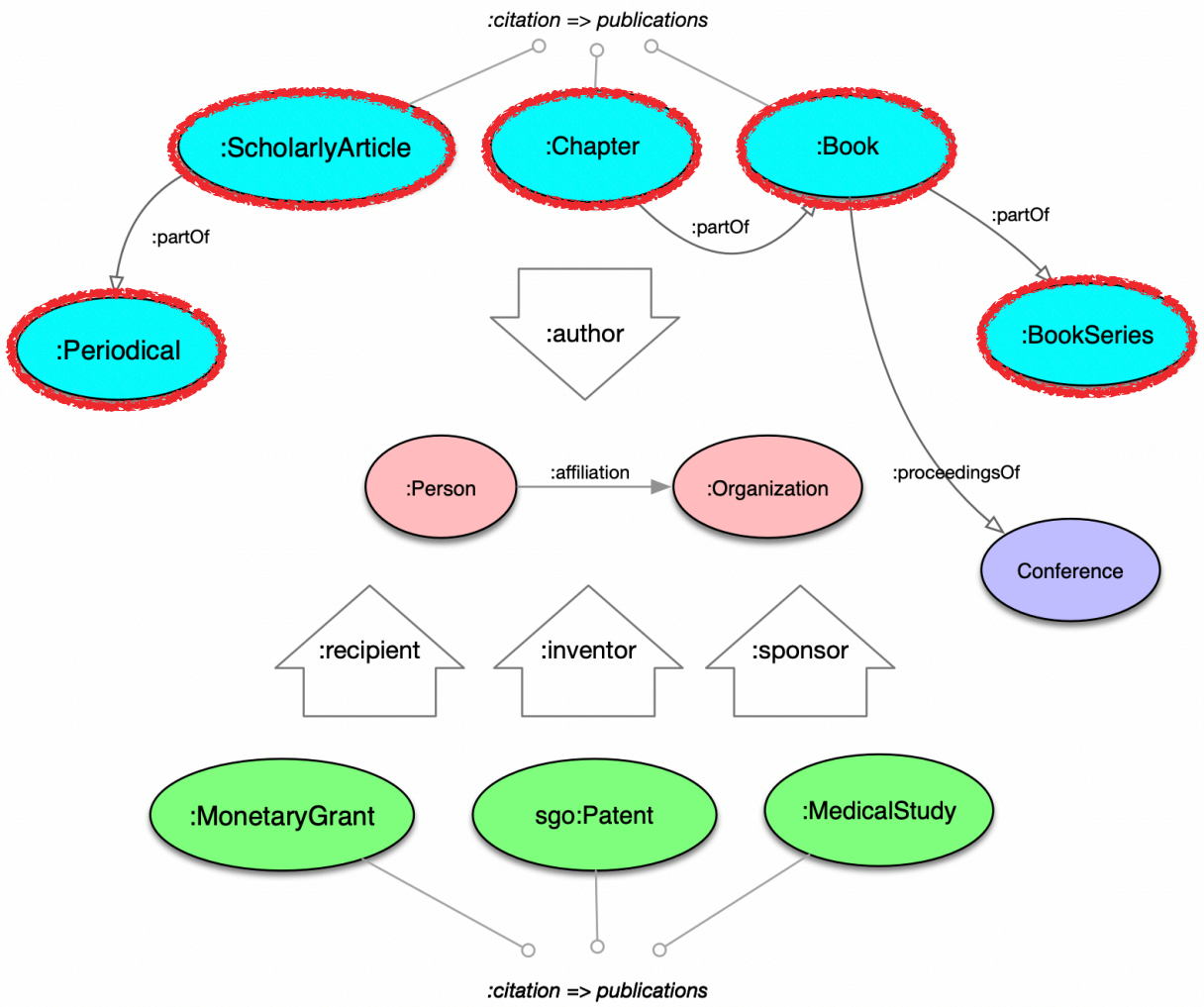
# 2018-2019: Building on Schema.org




- **Collaborative** community activity from major search engines (Bing, Google, Yahoo!, and Yandex - 2011)
- Focus on structured data on the Internet and search engine optimisation (discoverability)
- **JSON-LD** is the recommended representation format

# 2018-2019: Building on Schema.org




- **All Scigraph classes** could be modelled using **schema.org**
- **Some exceptions** still require Scigraph entities
- **JSON-LD** used as canonical format for all data
- **Publication** model rich enough (https://bib.schema.org/)

**Sample JSON-LD Markup** https://scigraph.springernature.com/explorer/datasets/articles/


```
{
    "@context": "https://springernature.github.io/scigraph/jsonld/sgcontext.json",
    "id": "sg:pub.10.1038/164322a0",
    "type": "ScholarlyArticle",
    "name": "A New Type of Fossil Man",
    "url": "http://www.nature.com/articles/164322a0",
    "sameAs": [
        "https://doi.org/10.1038/164322a0",
        "https://app.dimensions.ai/details/publication/pub.1000676771"
    ],
    "isPartOf": [
        {
            "type": "PublicationIssue",
            "issueNumber": "4164"
        },
        {
            "type": "PublicationVolume",
            "volumeNumber": "164"
        },
        {
            "id": "sg.journal.1018957",
            "type": "Periodical",
            "name": "Nature",
            "issn": ["0028-0836", "1476-4687"],
            "publisher": "Nature Publishing Group UK"
        }
    ],
    "datePublished": "1949-08",
    "pagination": "164322a0",
    "description": "IN the cave at Swartkrans which has now yielded the jaws and skulls of the huge a
    "genre": "research_article",
    "author": [
```

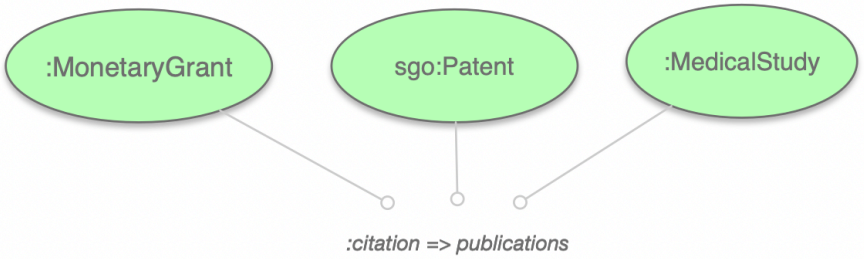
# 2018-2019: Building on Schema.org

## PROS

- ✓ Schema.org increasingly popular
- ✓ Schema.org is flexible and actively developed
- ✓ Suited for JSON-LD serialisation
- ✓ SEO friendly out-of-the-box

## CONS

- ◉ JSON-LD can have side-effects on other serialisations
- ◉ More work required for rigorous logical inference
- ◉ ?



# Conclusions

# #Conclusions: Usage and Data Models

* **Most users have very simple needs**
  * eg search for publications title+abstracts (few other metadata items are needed)
  * Hard core users (eg LOD specialists) are often more interested in the model and tech stack than actual reusing of the data
    * *=> see our paper at 'Semantic Science' workshop 2016*

* **Ontological hair-splitting is fun but not always to the point**
  * FRBR: how many of your users know the difference between an expression and a work?
  * Specialist communities: a lot of interest, but no practical uptake
    * Useful to focus on concrete use cases (justifying the effort)

# #Conclusions: Technology and Implementation

* **Models have a big impact on data volumes**
  * A good data model shouldn't make it harder to process/query your data *"a data model is a practical theory"*

* **Scalability & performance**
  * Semantic graph databases are often harder to scale
  * *=> If you have lots of data you'll probably end up using non-RDF technologies too eg Elasticsearch*

* **Data Serialization**
  * RDF still very unpopular with developers / data consumers
  * Good to go beyond the LOD community with new standards like JSON-LD

# Thanks

Email:
m.pasin@digital-science.com

Project Homepage:
http://www.springernature.com/scigraph