



Modeling publications in SN SciGraph 2012-2019

June 2019

Michele Pasin

Lead Data Architect
Tech Product Owner

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



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

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.



Institution disambiguation,
95k+ organisation IDs,
openly available



4.3M projects in grant database, enrichment services and analytical application core



Tracking attention in news, social media and policy papers - as an immediate resonance



Global patent database - more than 100M patent records



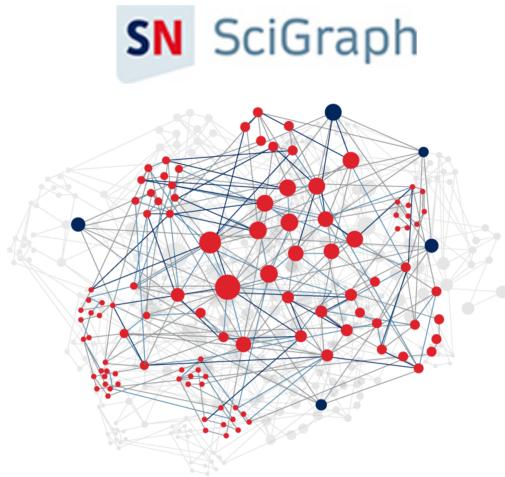
Research information management system - deep knowledge about institutional requirements for data and metrics



Serving publishers, access to +60M journal articles and books

Springer Nature SciGraph

A Linked Open Data platform for the scholarly domain

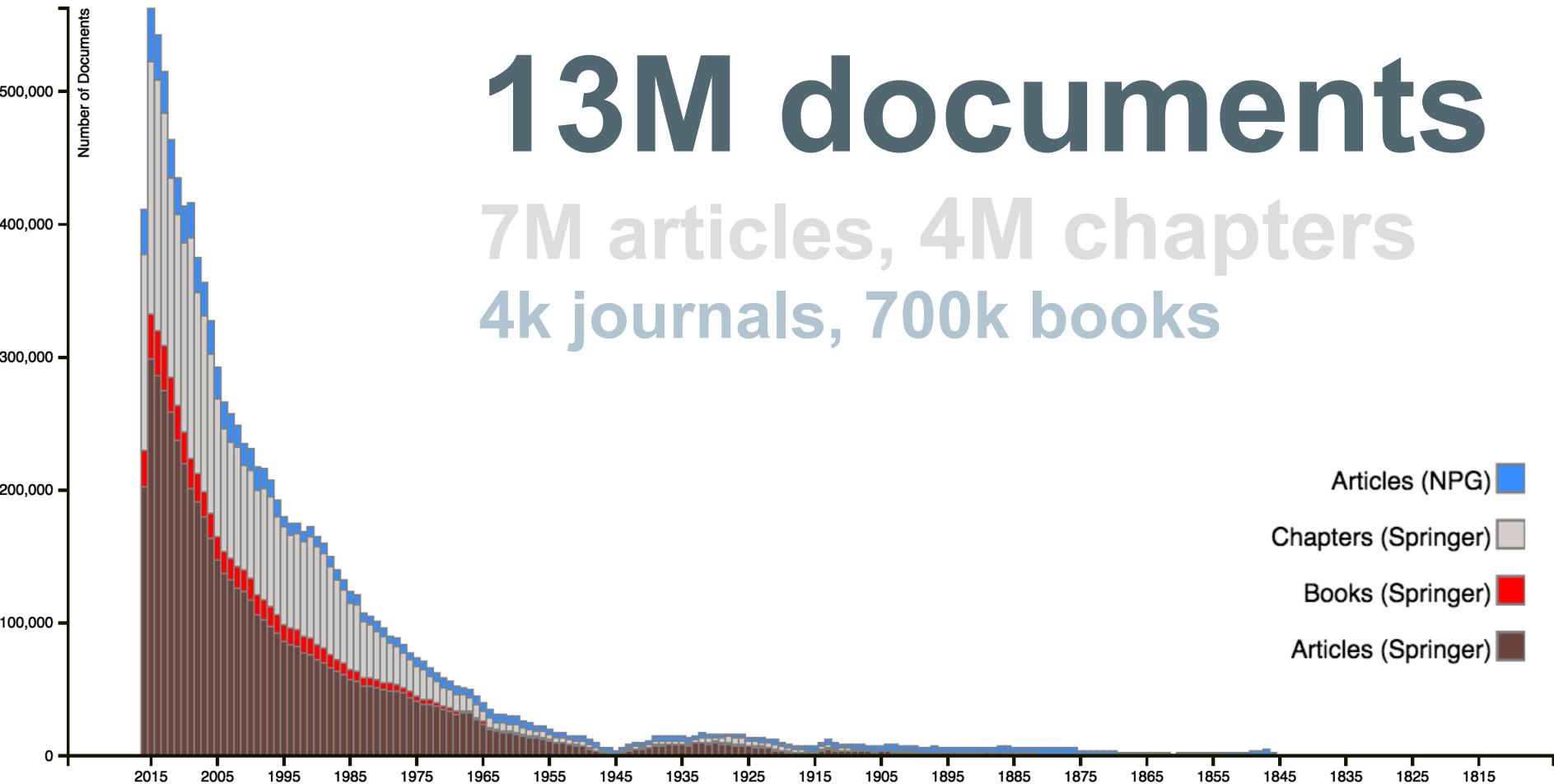


- > Collaborative effort between Springer Nature and Digital Science (mid 2016)
- > Increasing discoverability of content by using linked data and semantic technologies
- > Supporting internal use cases, but also contributing to an emerging web of linked scholarly data

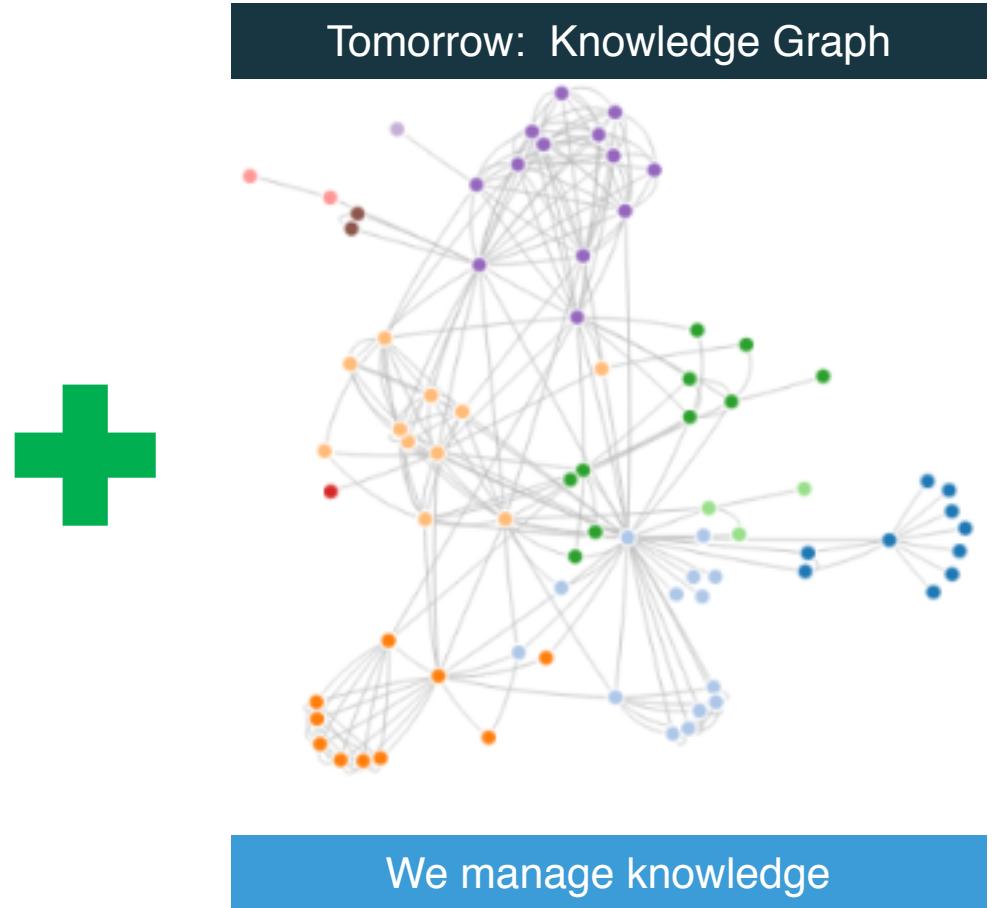
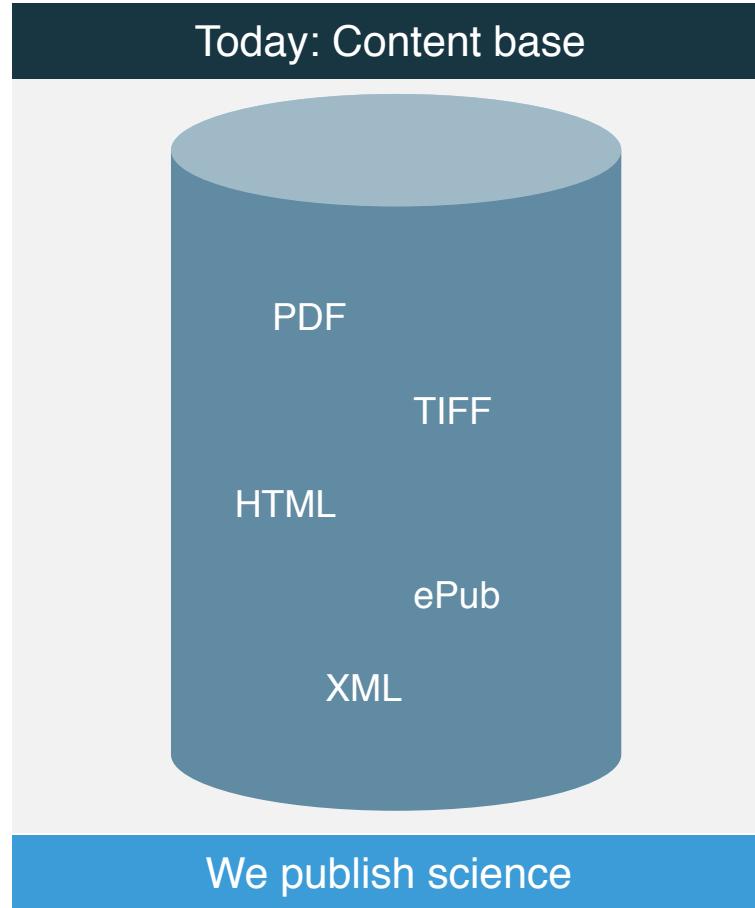
www.springernature.com/scigraph

Motivation: integrating Springer Nature archive

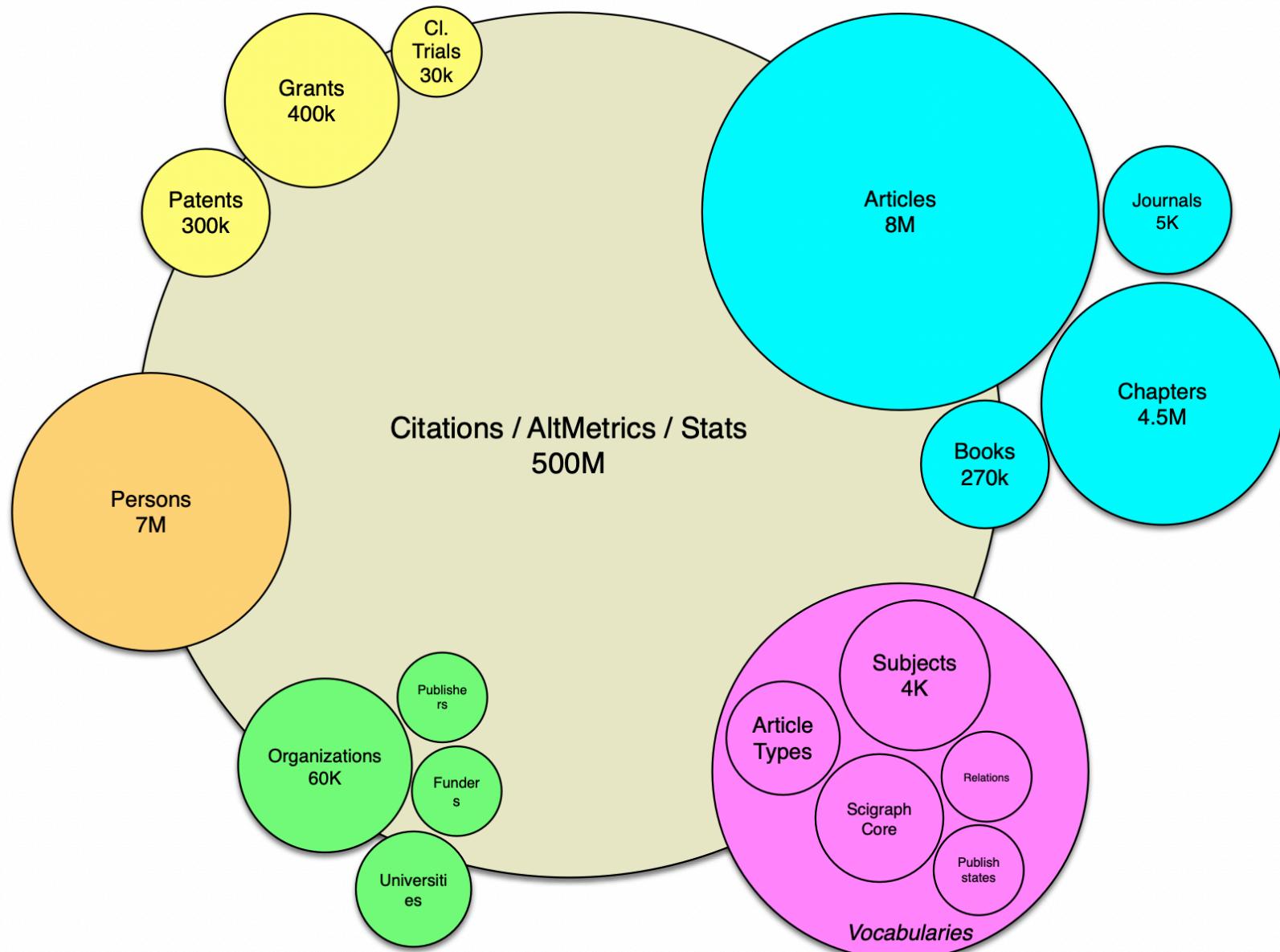
13M documents
7M articles, 4M chapters
4k journals, 700k books



Vision



SciGraph data landscape



Applications: analytics dashboards for editors, publishers etc..

Springer Nature SciGraph Analytics Dashboards Journals Institutions Countries Subject Areas

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 indexed in external databases.

Article - Total number: 841 Articles in Total

Article - Count from 2012: 201 Articles Published Last 5 Years

Article - Count by publication year

publicationYear	Count
2000	20
2001	25
2002	30
2003	20
2004	45
2005	40
2006	35
2007	55
2008	60
2009	100
2010	105
2011	55
2012	40
2013	50
2014	35
2015	30
2016	35
2017	30

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.abs.gov.au/ausstats/abs@.nsf/f06BB427AB9696C225CA2574180004463E>

Article - FieldOfResearch by code and de... Article - top 15 Fields of Research over time

Field of Research	Percentage of Count
PHYSICAL SCIENCES	~80%
NEUROSCIENCES	~10%
MEDICAL AND HEALTH SCIENCES	~5%
IMMUNOLOGY	~2%
GENETICS	~2%
CARDIOLOGY AND VASCULAR MEDICINE	~2%
BIOLOGICAL SCIENCES	~2%
BIOCHEMISTRY AND MOLECULAR BIOLOGY	~2%
TECHNOLOGY	~2%
STATISTICS	~2%
PLANT BIOLOGY	~2%
PHARMACOLOGY AND TOXICOLOGY	~2%
ONCOLOGY AND CARCINOLOGY	~2%
MEDICAL MICROBIOLOGY AND IMMUNOLOGY	~2%
MEDICAL BIOTECHNOLOGY AND BIOPROCESS ENGINEERING	~2%
MEDICAL BIOCHEMISTRY AND PHYSIOLOGY	~2%
MATHEMATICAL SCIENCES	~2%
INFORMATION AND COMPUTER SCIENCES	~2%
ENVIRONMENT	~2%

Percentage of publicationYear ranges

Applications: SciGraph open data publishing

SN SciGraph Data Explorer Getting Started Models ▾ Downloads License FAQ

You are here: Home

Springer Nature SciGraph Data Explorer

Search across one billion facts from the scholarly domain.

Search for...

Not sure where to start? Try searching for an organization, e.g. the ['Francis Crick Institute'](#), a topic, e.g. 'machine learning', or an author, e.g. '[Steven Pinker](#)'.

SciGraph is the Springer Nature Data Explorer, providing information from across the documents, people, places, science and scholarly domain.

Metadata for millions of entities on the site, as well as for download application.

See <https://www.springernature.com>

DNA nanotechnology SKOS SUBJECT SKOS SUBJECT SKOS CONCEPT

Overview Details

How to use: Click on a node to preview its contents. Double click to open its homepage.

See <http://scigraph.springnature.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

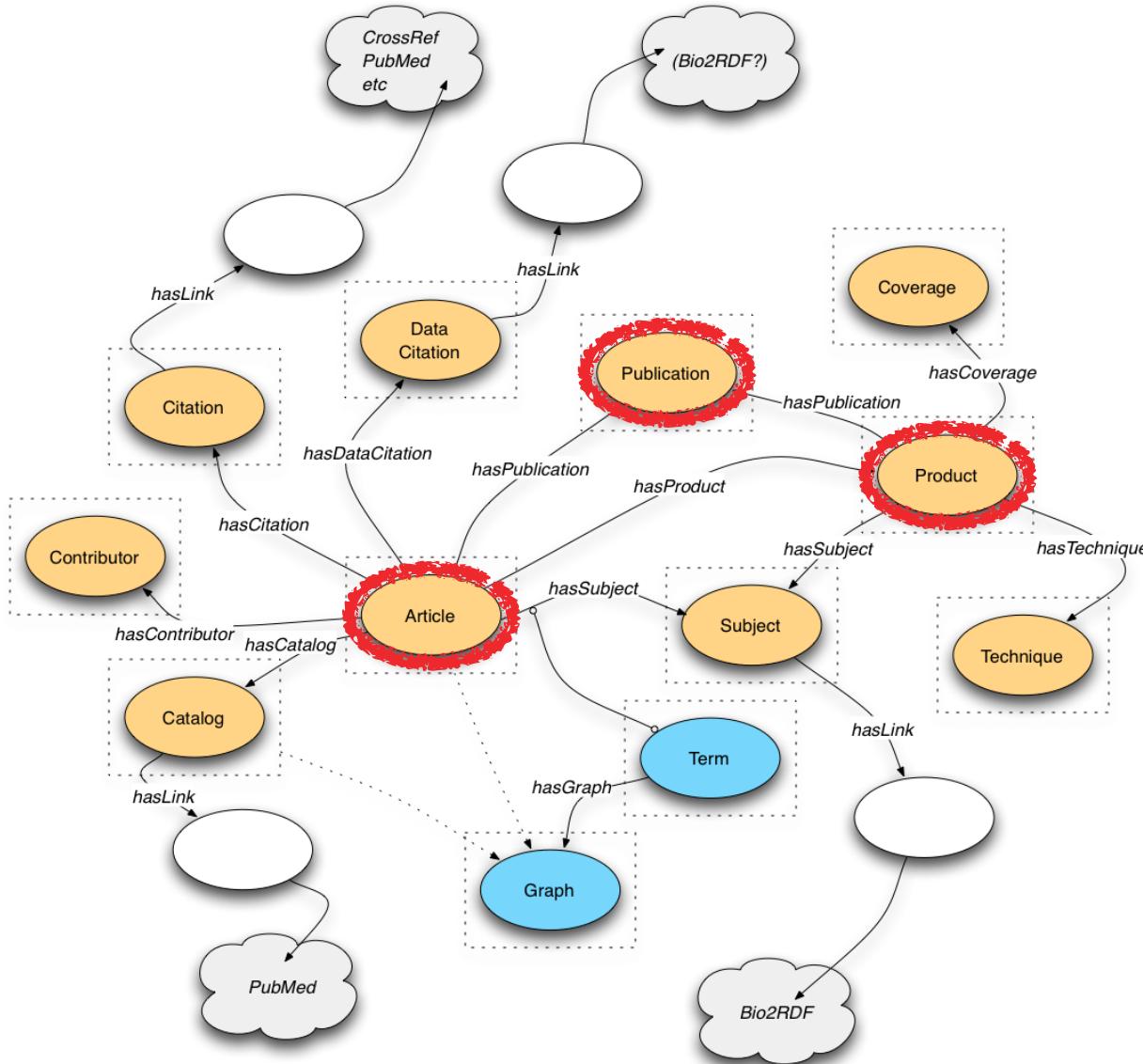
Summary of approaches

- 1. [2012-14] Ontologies Mix and Match**

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

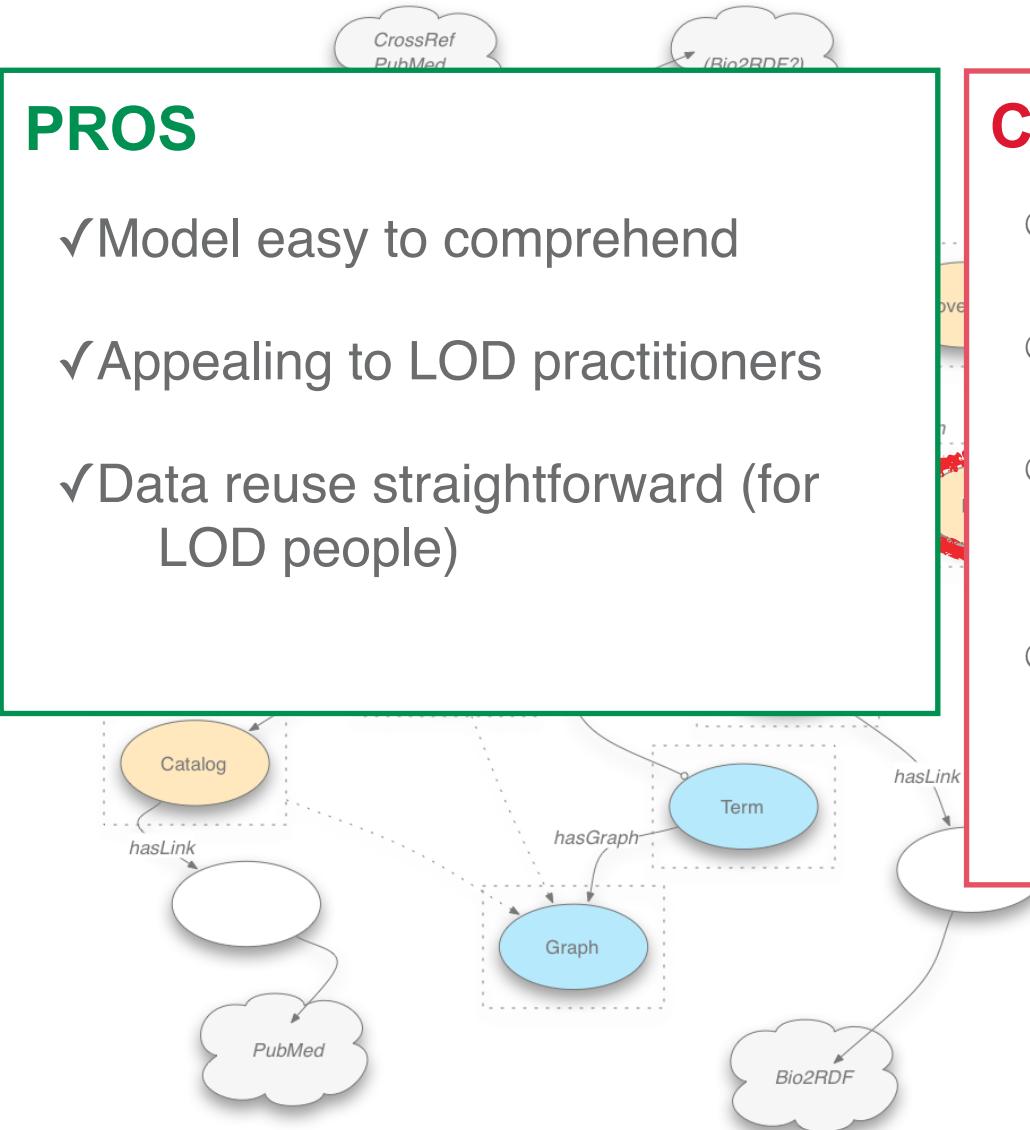
- 3. [2018-19] Building on Schema.org**

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, :familyName, :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

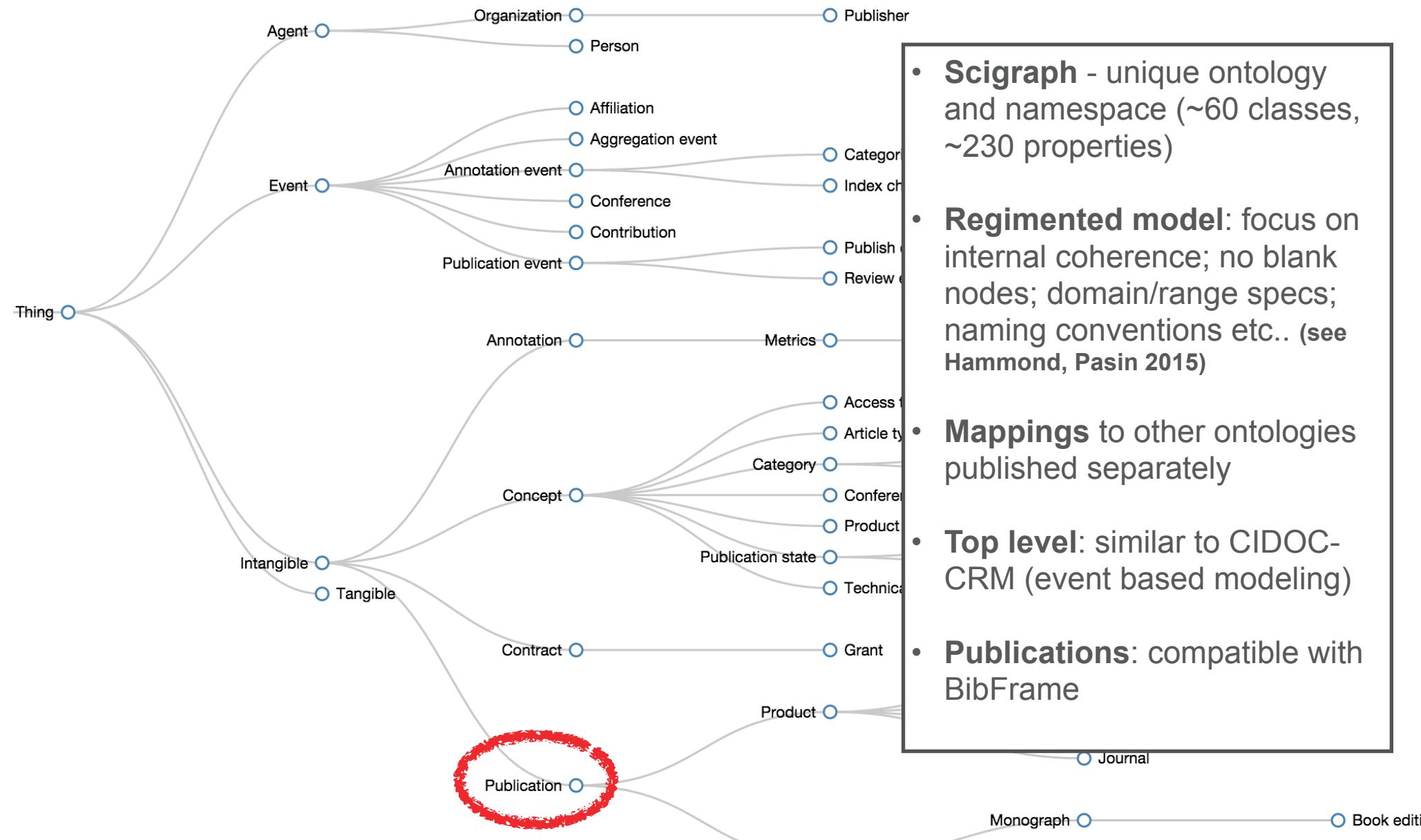
- Model rather simplistic
- Hard to maintain / extend / query
- Loose semantics, not suited for formal reasoning
- Not solid enough for internal uptake at Springer Nature

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

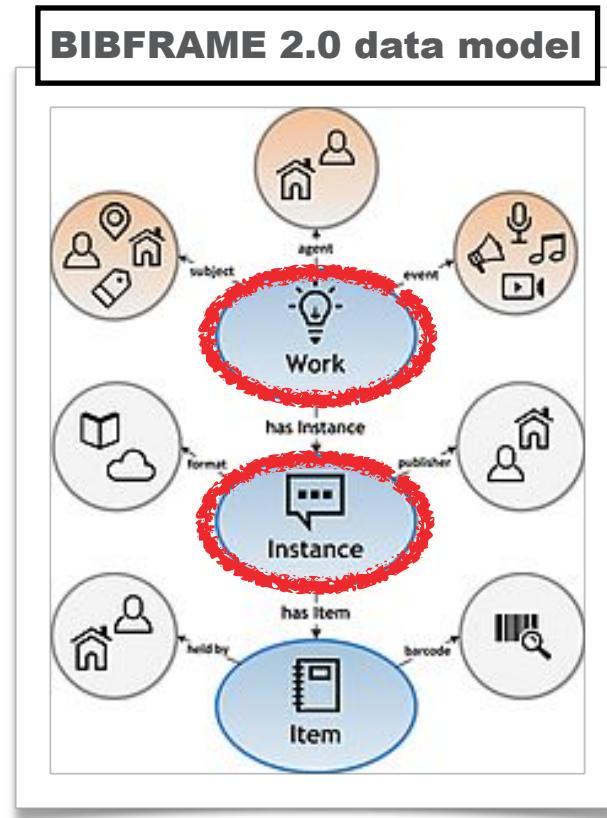
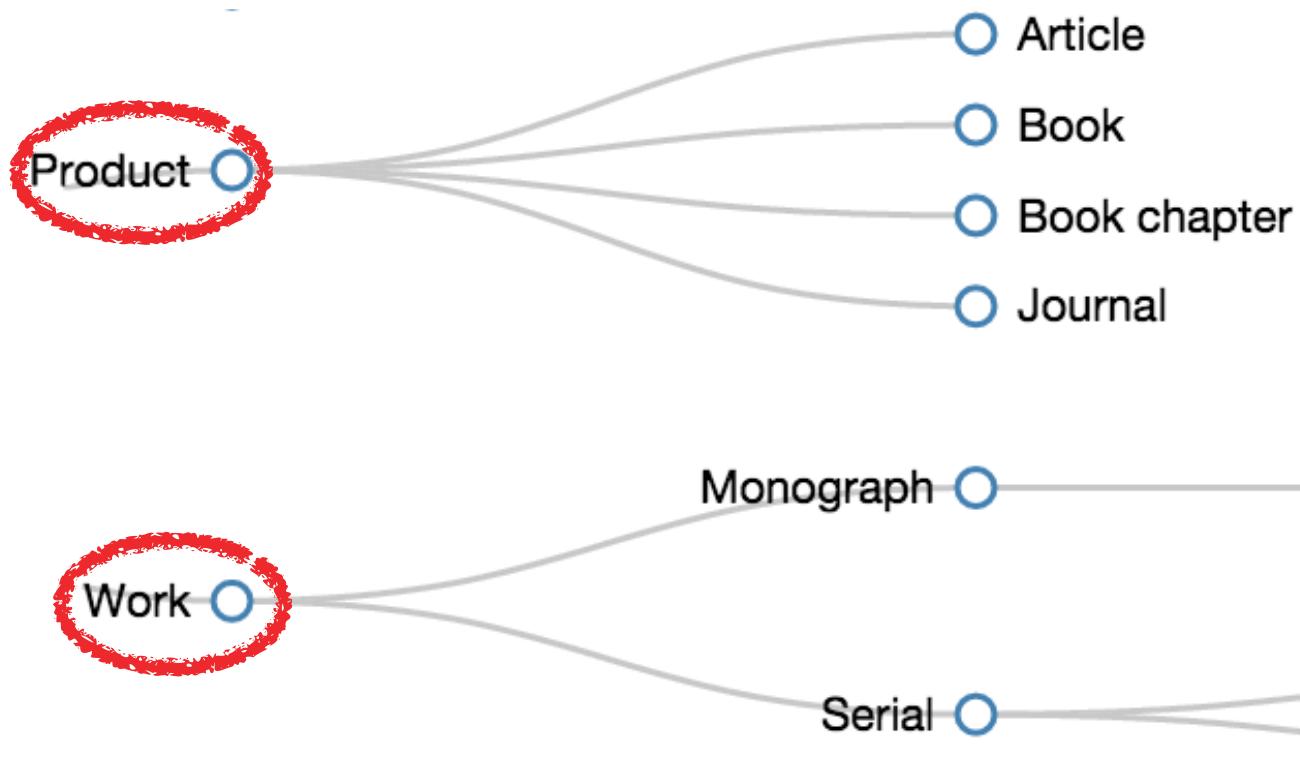
Summary of approaches

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

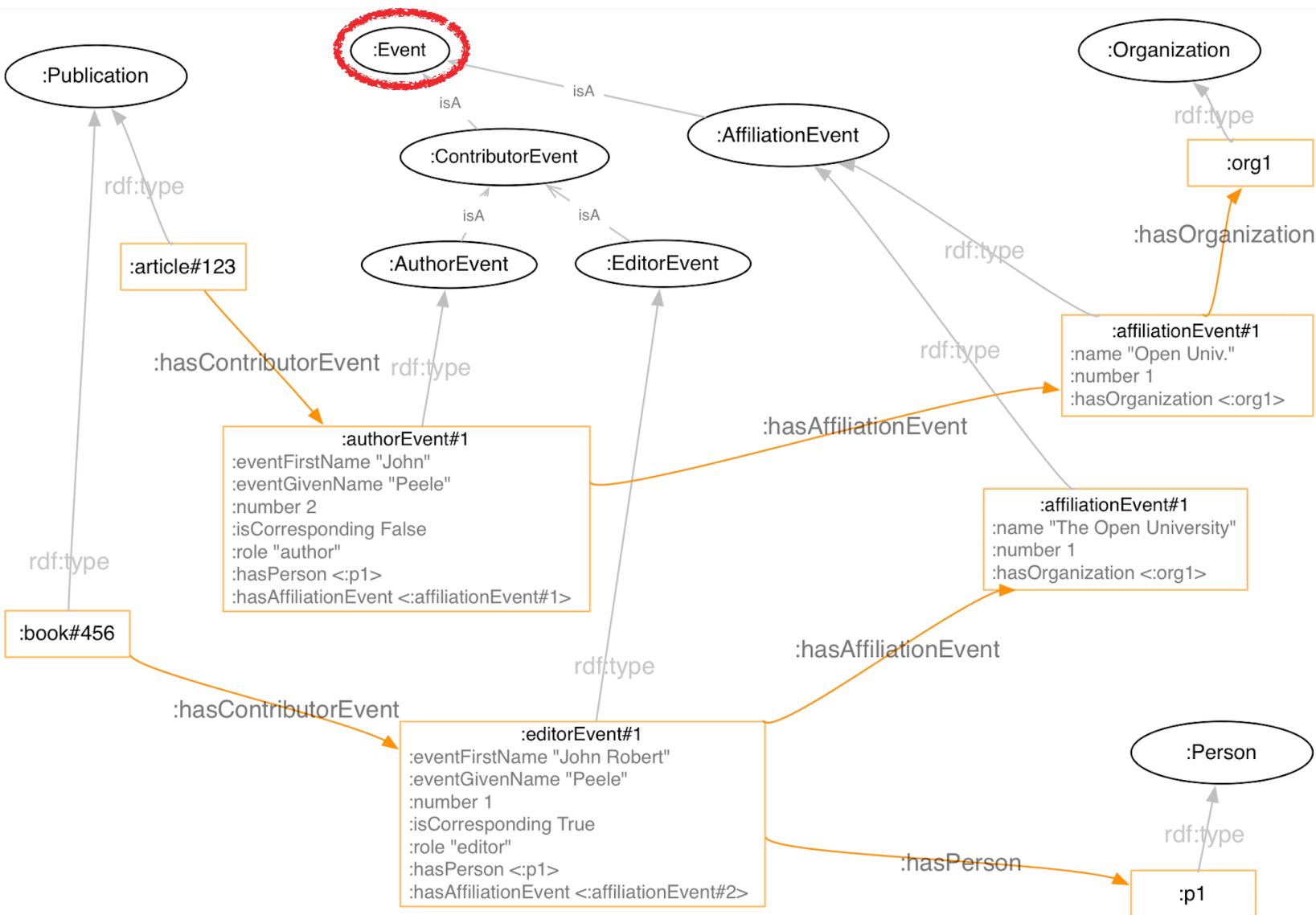
2015-2017: Bespoke ‘SciGraph’ Ontology



2015-2017: Publications in ‘SciGraph’ Ontology



2015-2017: Events in ‘SciGraph’ Ontology



2015-2017: Bespoke ‘SciGraph’ Ontology

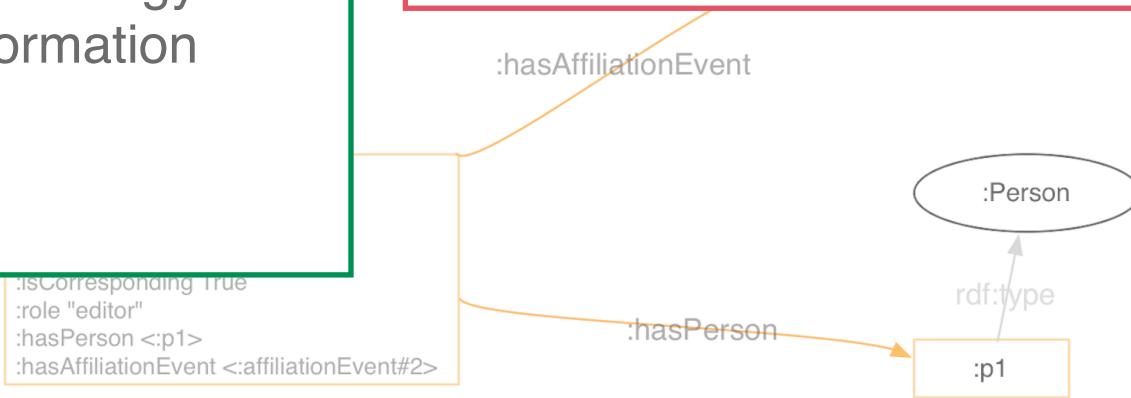
PROS

- ✓ 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

:isCorresponding True
:role "editor"
:hasPerson <:p1>
:hasAffiliationEvent <:affiliationEvent#2>



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

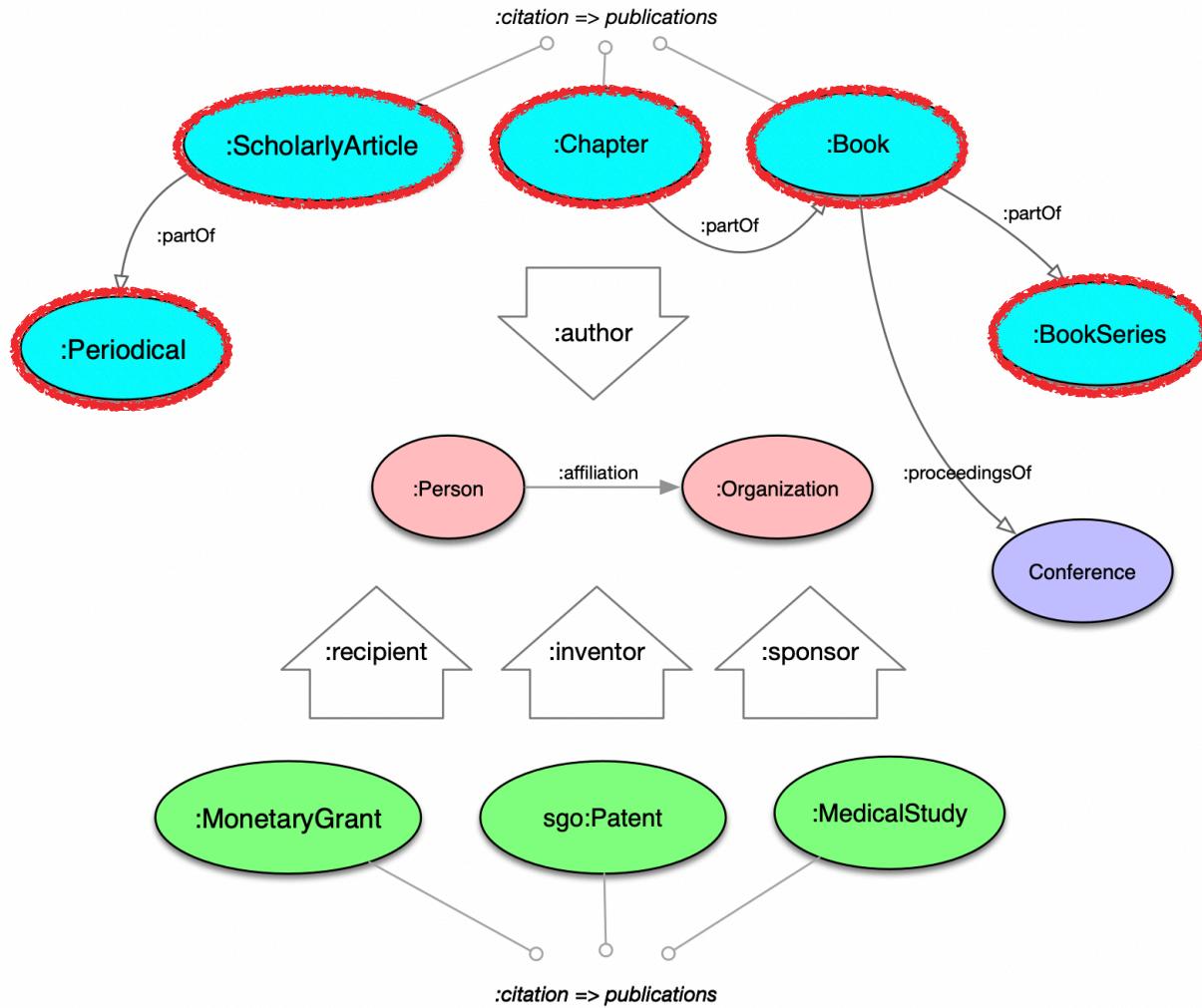
A screenshot of a search results page for "alan watts" on a search engine. The results include:

- Alan Watts - Wikipedia**
https://en.wikipedia.org/wiki/Alan_Watts ▾
Alan Wilson Watts (6 January 1915 – 16 November 1973) was a British-American philosopher who interpreted and popularised Eastern philosophy for a Western audience. Born in Chislehurst, England, he moved to the United States in 1938 and began Zen training in New York.
Notable work: *The Way of Zen* (1957) Institutions: American Academy of Asian Studies
School: Zen Buddhism; Hinduism; Pantheism; ... Alma mater: Seabury-Western Theological Se...
Works by Alan Watts · Jean Burden · *The Way of Zen* · Druid Heights
- Videos**
 - How To NEVER Be Worried Again - Alan Watts | A life-changing speech
 - Alan Watts - Can You Handle The Truth
 - What Happens When You Only Pursue Pleasure - Alan Watts
- Alan Watts**
American-British philosopher
Alan Wilson Watts was a British-American philosopher who interpreted and popularised Eastern philosophy for a Western audience. Born in Chislehurst, England, he moved to the United States in 1938 and began Zen training in New York. [Wikipedia](#)
Born: January 6, 1915, Chislehurst, United Kingdom
Died: November 16, 1973, Druid Heights
Influenced: Robert Anton Wilson, Seraphim Rose, Monica Furlong
Influenced by: Carl Jung, Jiddu Krishnamurti, Joseph Campbell, MORE



- **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 apes,",
  "genre": "research_article",
  "author": [
    {
      "id": "sg:person.01403111",
      "type": "Person",
      "name": "Koobi Fora Research Project"
    }
  ]
}
```

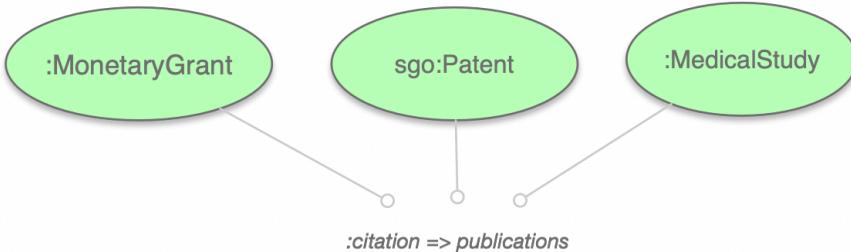
2018-2019: Building on Schema.org

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>