Linked Data Experiences at Springer Nature

Michele Pasin
Lead Data Architect
Knowledge Graph Team



SPRINGER NATURE

Outline

- Who we are
- Why semantic technologies
- Our work so far
- The Scigraph project
- Looking ahead

SPRINGER NATURE

Who We Are

SPRINGER NATURE



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

SPRINGER NATURE

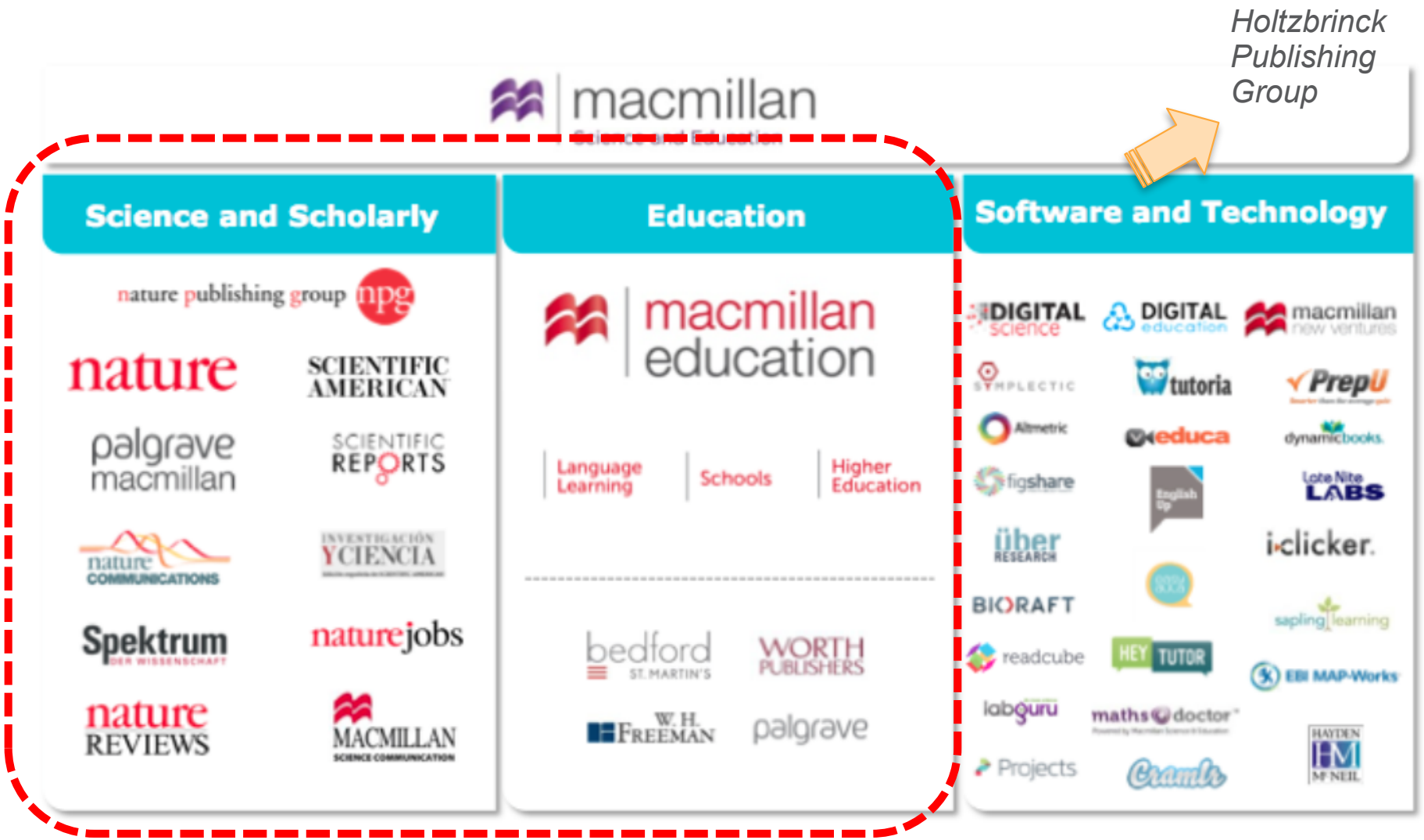


13k employees in over 50 countries, EUR 1.5 billion turnover

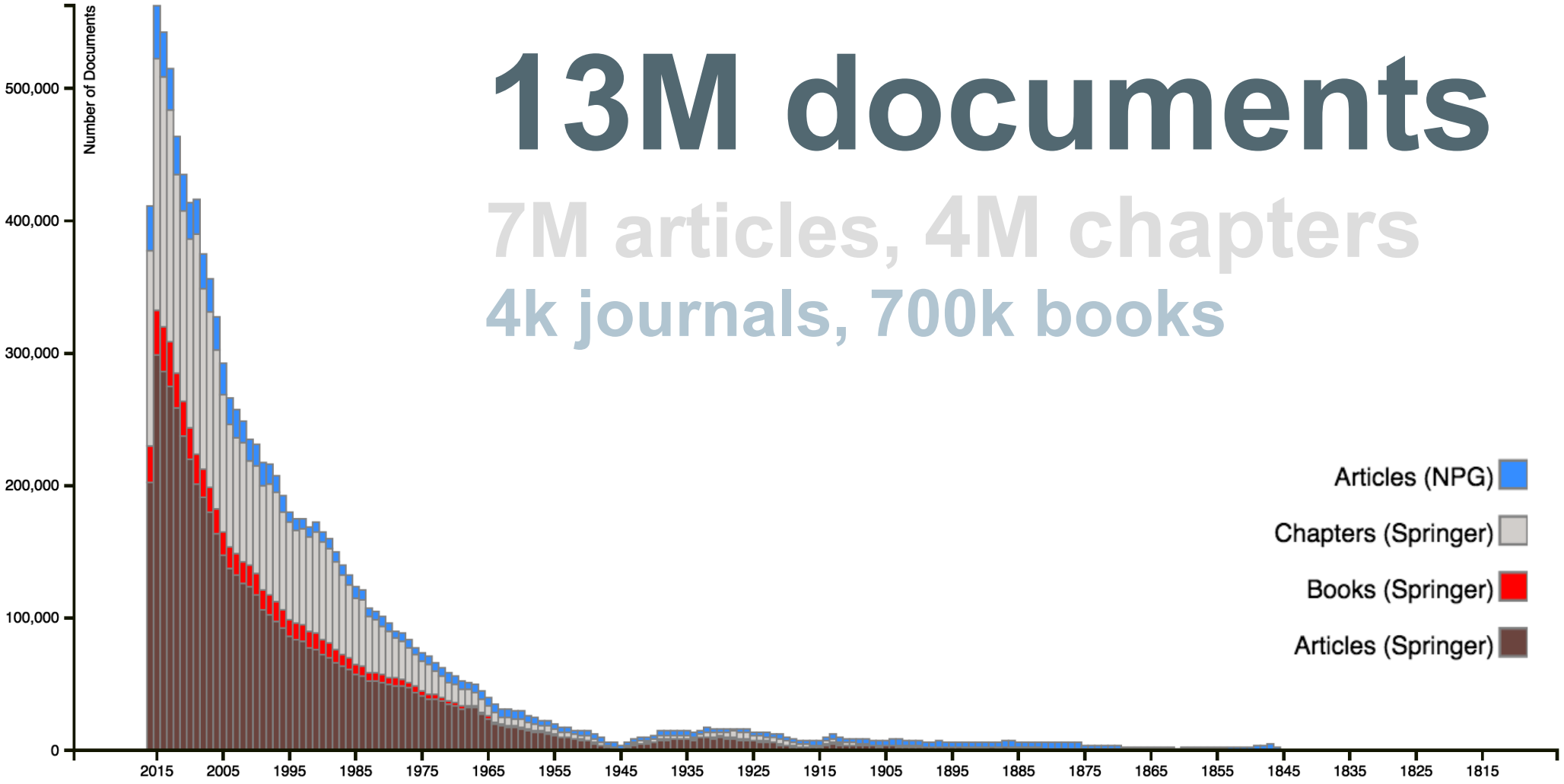
[Pre-Merger] Springer Science + Business Media brands



[Pre-Merger] Macmillan Science & Education brands



We publish a lot of science! (since 1815)



..and generate a lot of traffic

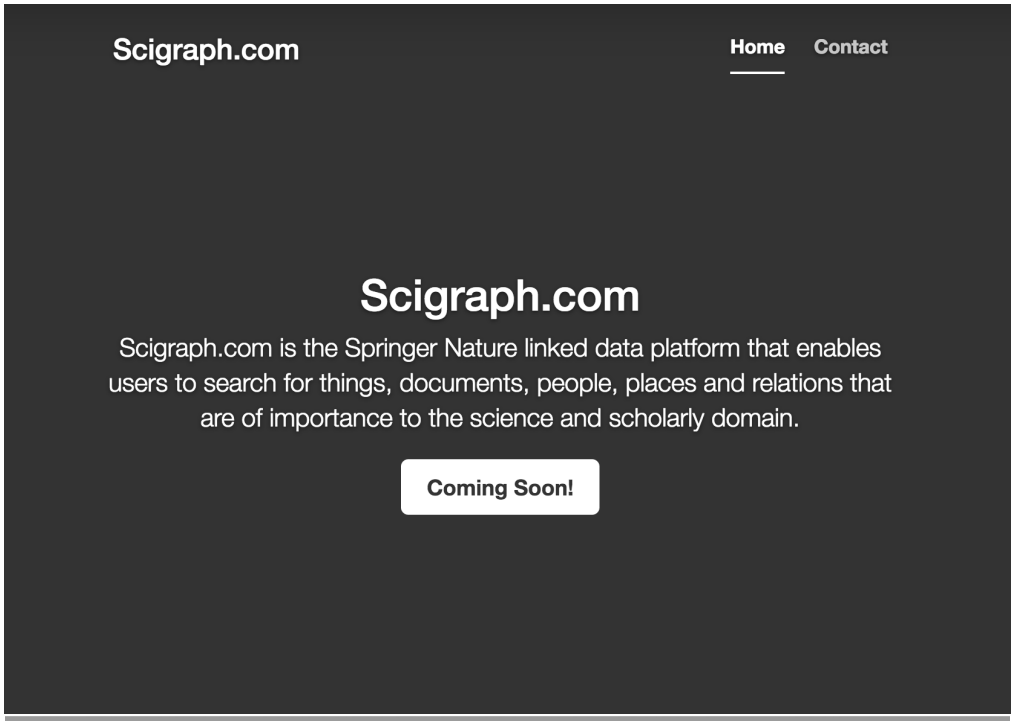
11.5M monthly visitors
(nature.com)

260M visits per year
600M downloads per year
(link.springer.com)

Scigraph.com

Scigraph.com is the Springer Nature linked data platform that enables users to search for things, documents, people, places and relations that are of importance to the science and scholarly domain.

Coming Soon!





> Collaborative effort between Springer Nature and Digital Science

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

> Not just publications data but a wealth of other related information

Why Semantic Technologies

Why is Semantics Important To Us?

Change Drivers

- Digital first workflow
- User-centric design
- Unified Springer Nature domain

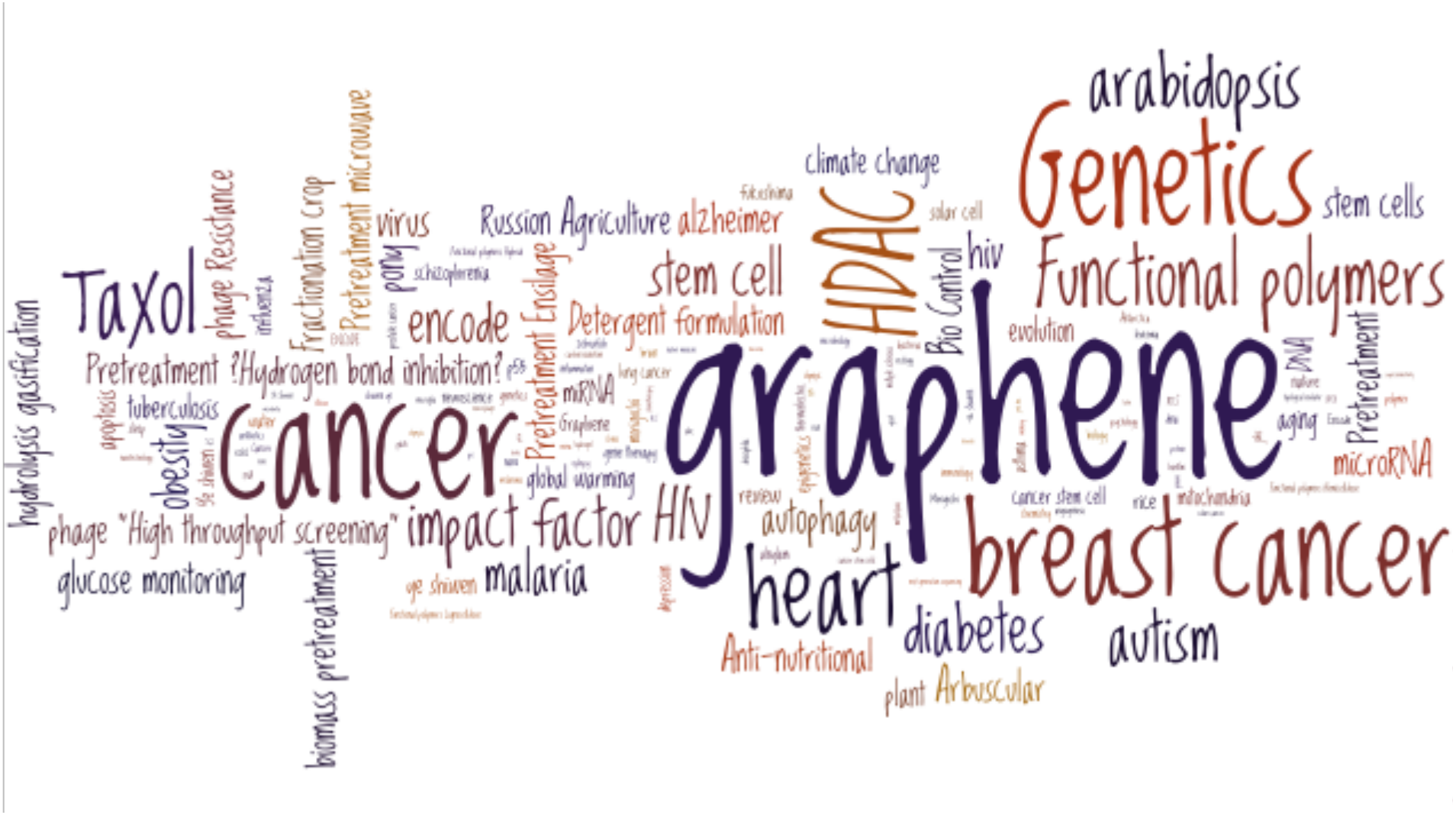
Challenges: Data Silos

- Data is fragmented
- Data gets duplicated
- Data is hardcoded into applications

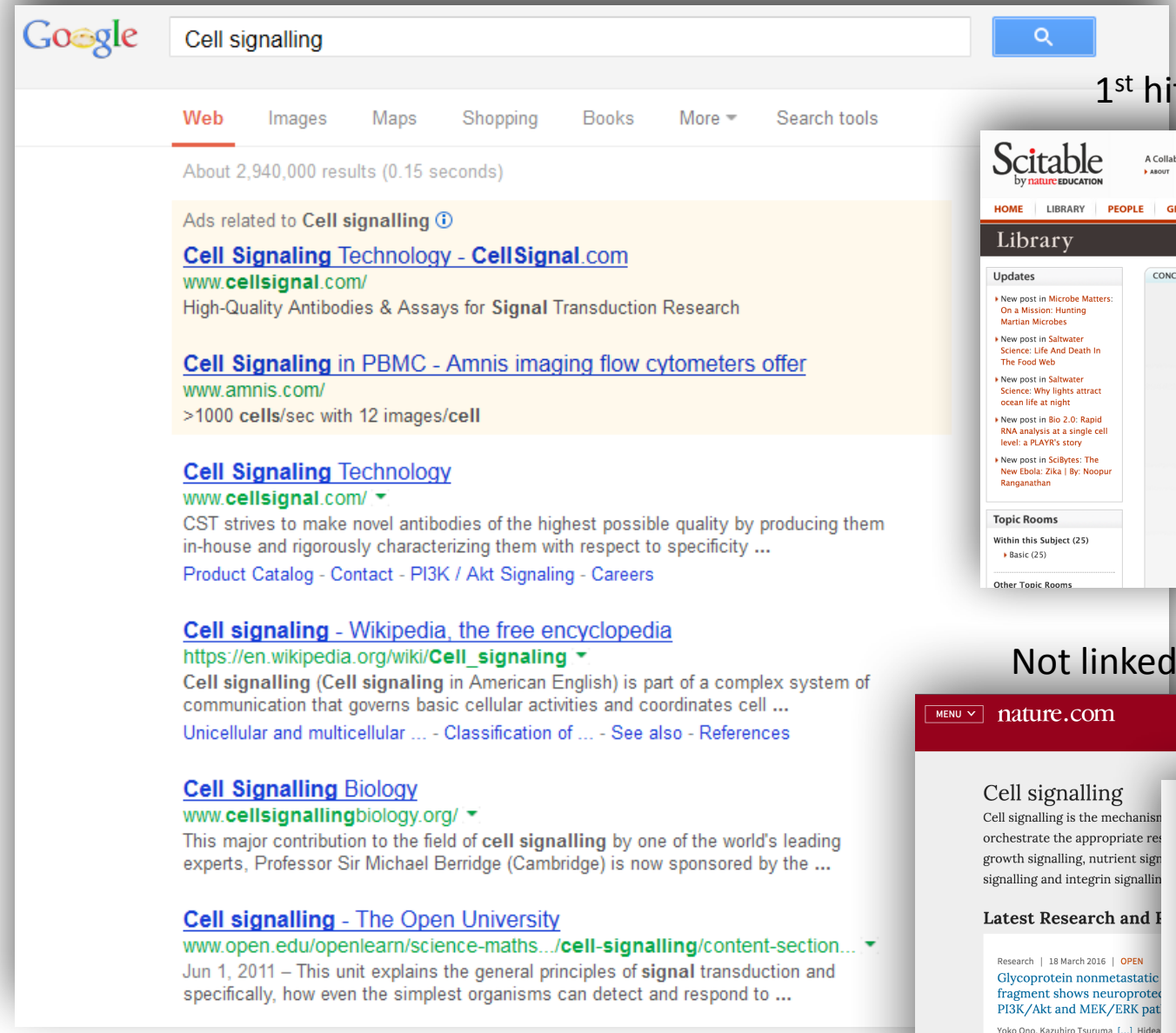
For example: our sites are currently organised around articles, journals and issues…


nature journal
nature jobs
nature magazine
nature communications
nature neuroscience
nature medicine
nature genetics
nature materials
nature chemistry
nature cell biology
nature biotechnology
nature immunology
nature physics
nature methods
nature photonics
nature nanotechnology
nature chemical biology
nature protocols
nature reviews
nature.com
naturejobs
scientific reports
nature publishing group
nature job
nature news
nature reviews
nature.com
nature climate change
nature reviews
nature reviews
nature reviews

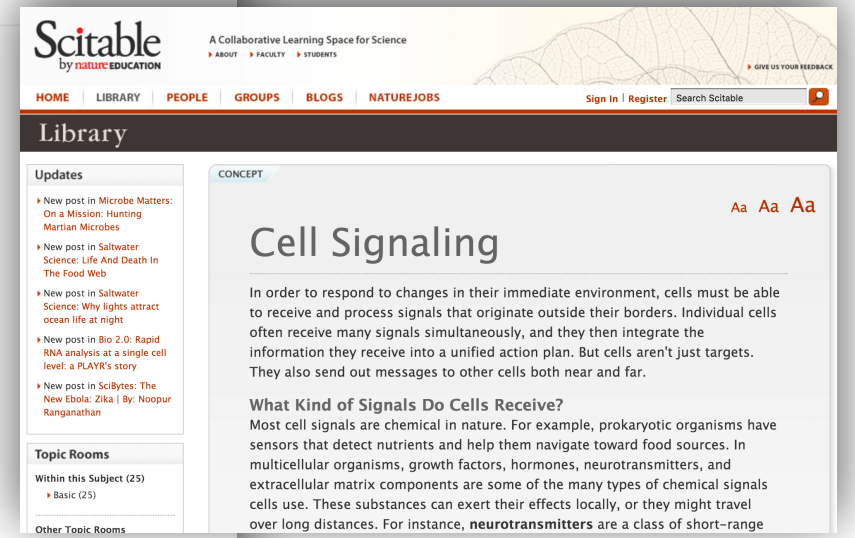
However, scientists are interested in answering questions about real world things…



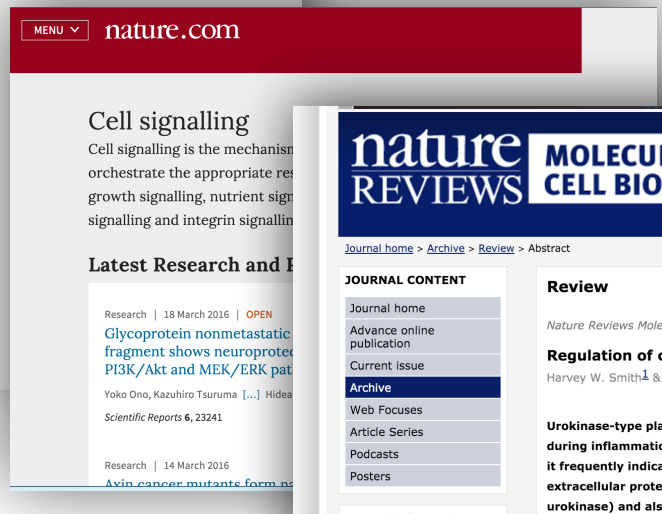
Search engines do not know we have content about these things…

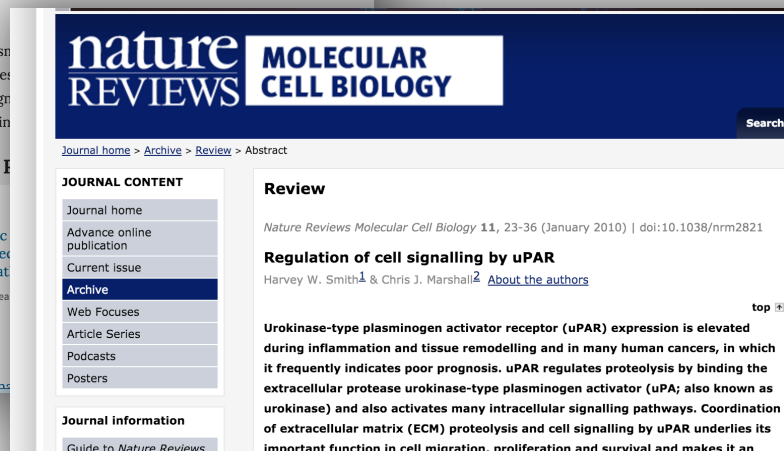


1st hit from nature.com…



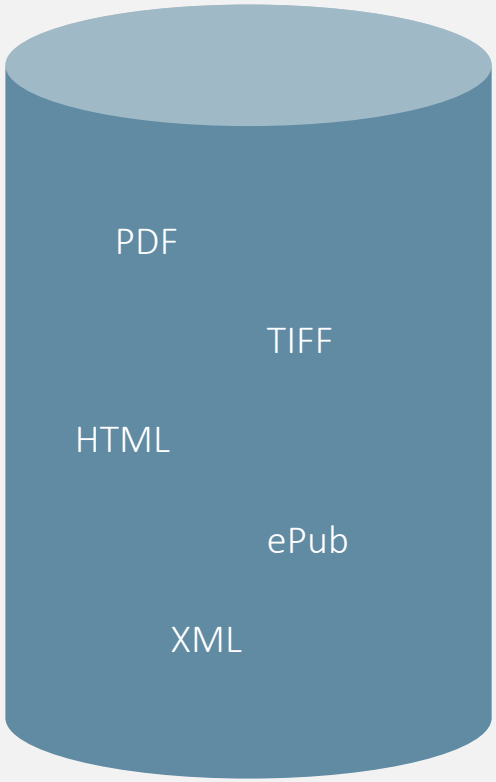
Not linked to/from..





Vision

Today: Content base



We publish science

Tomorrow: Knowledge Graph



We manage knowledge

The Knowledge Graph is about collecting information about objects in the real world

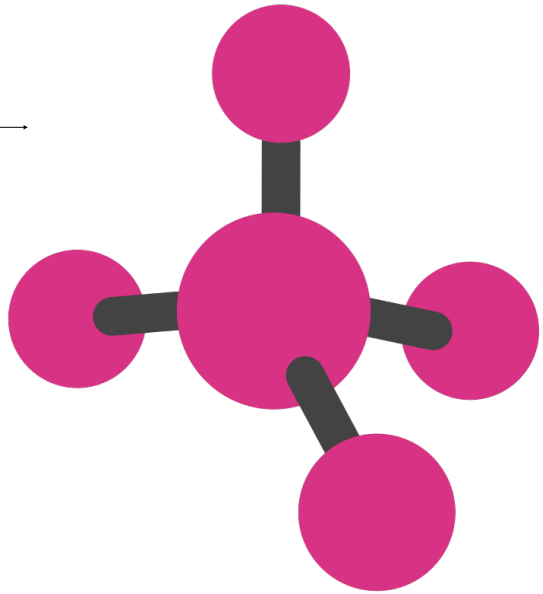
…so that we can do a better job of providing users with what they're looking for

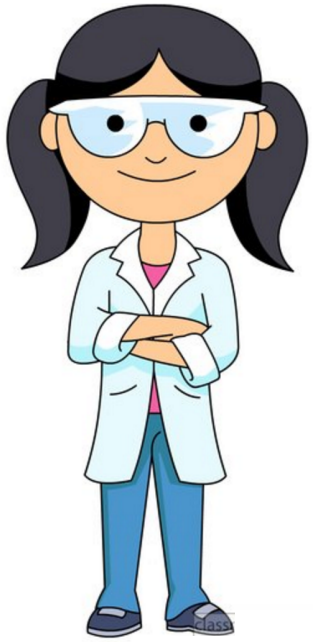
Google

Three areas of knowledge we care about



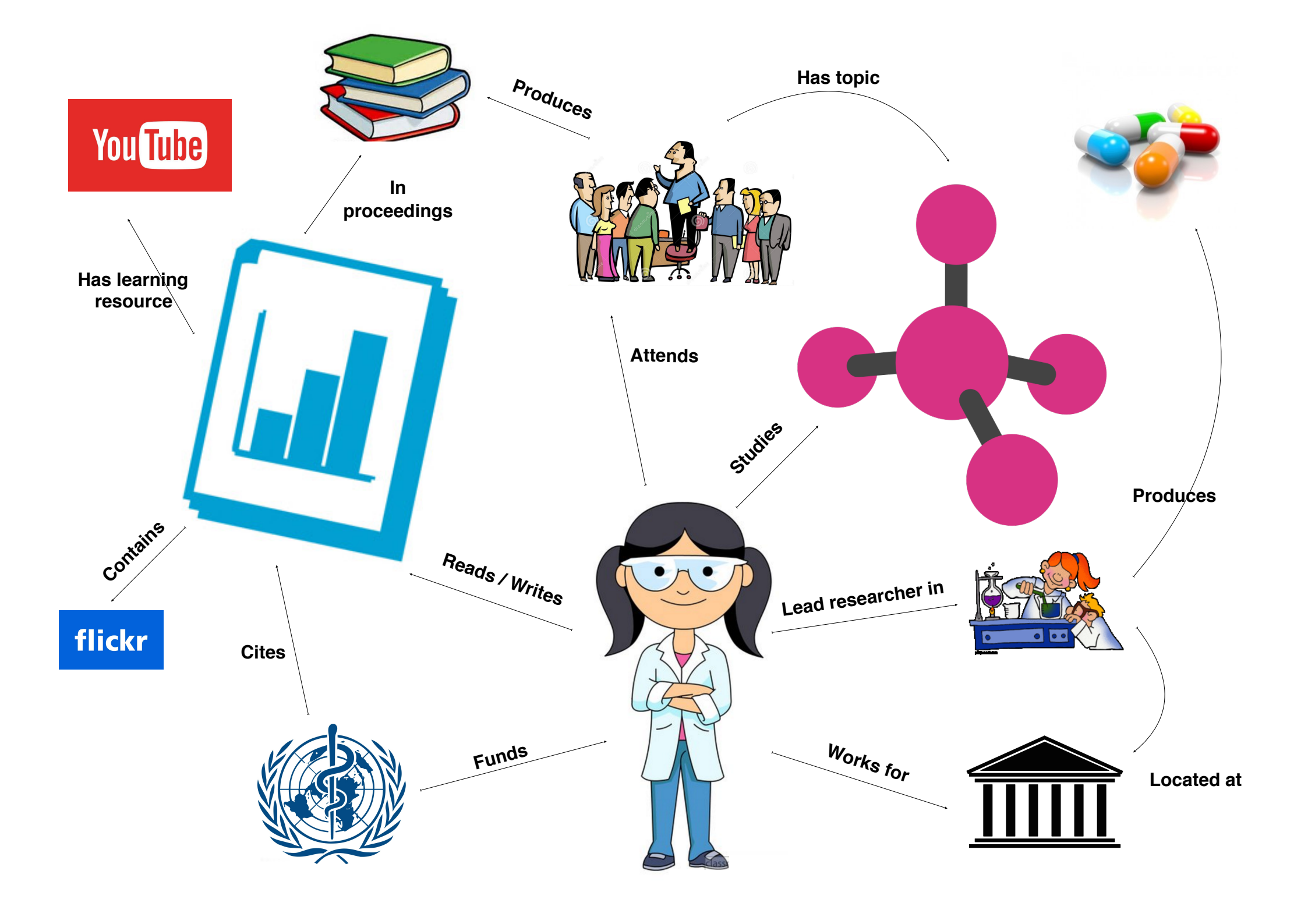
is about





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

Opportunities: Tools & Services Along the Publishing Life Cycle



Our Work So Far

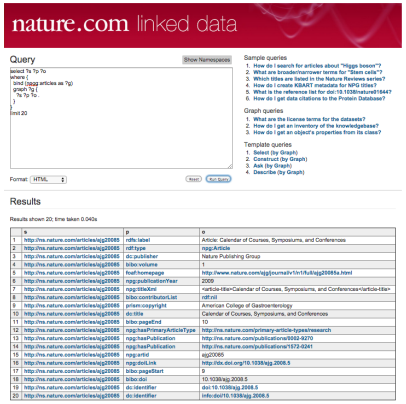
Our Work So Far

2012

Linnaeus Project

2013

NPG Linked Data Platform



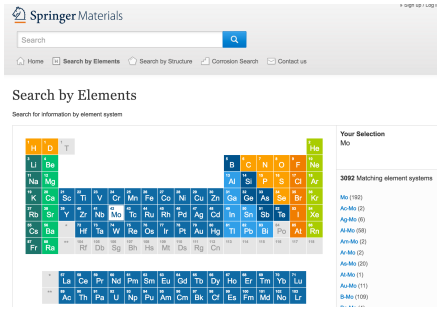
Content Hub

CURI Semantic Annotation Project

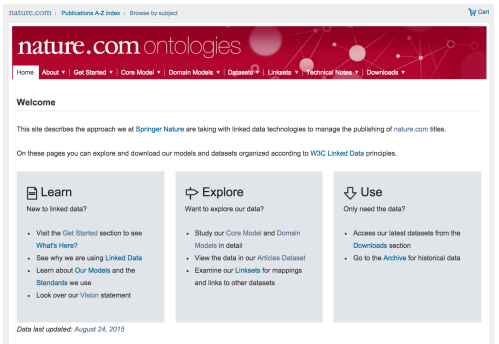
Scigraph prototype

2014

Springer Materials

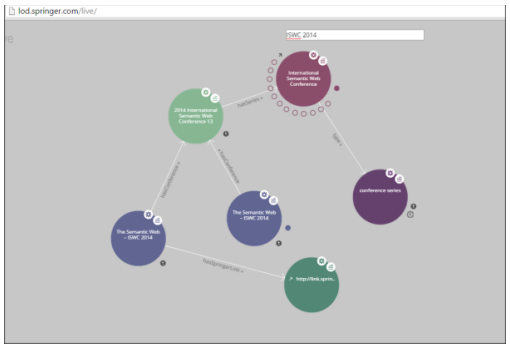


Nature Ontologies Portal



2015

Springer Conferences

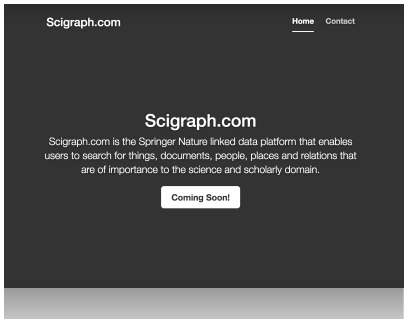


Nero Project

Springer Protocols

2016

Scigraph



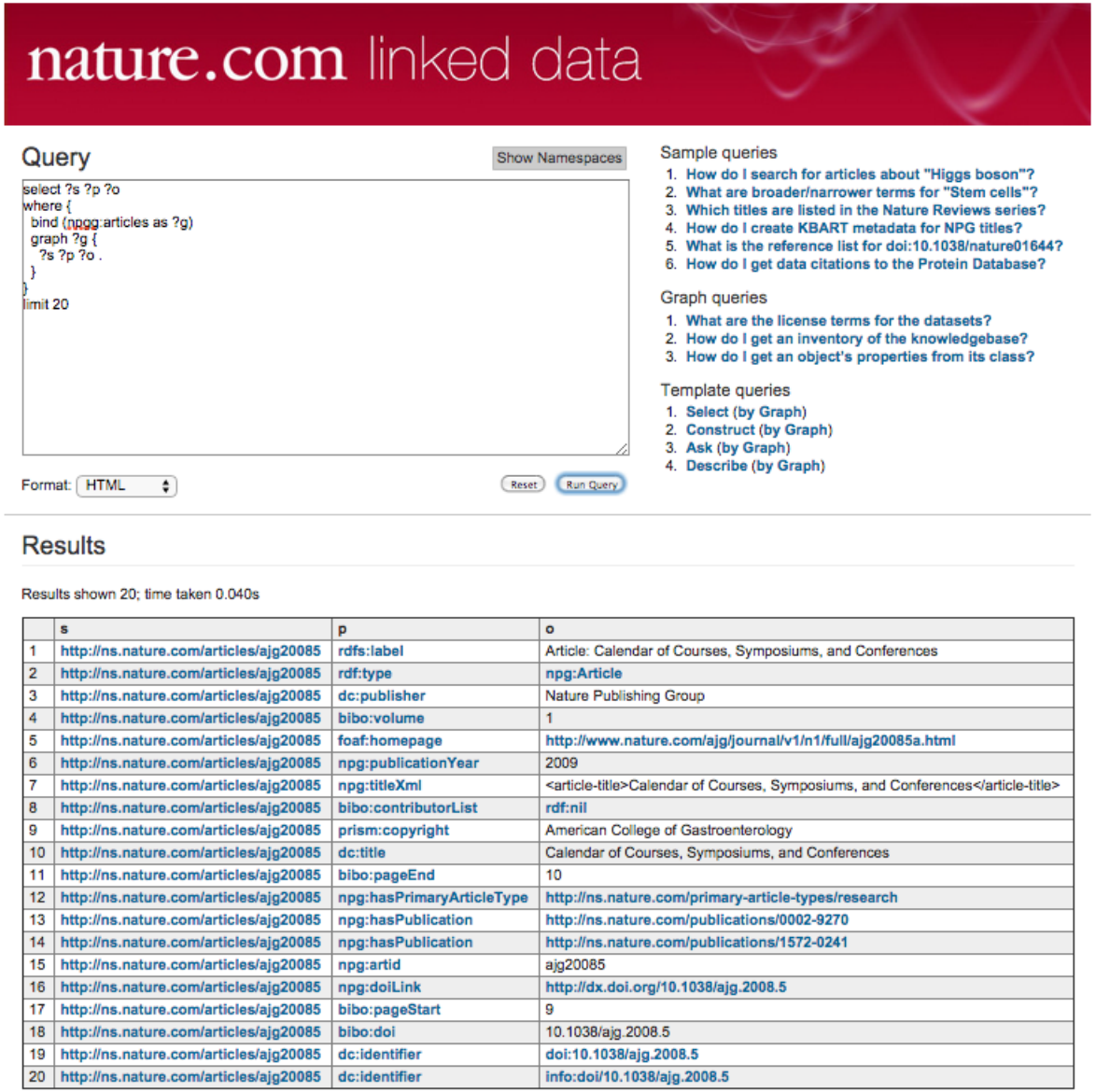
NPG Linked Data Platform (2012)

Deliverables (2012–2014)

- Prototype for external use
- SPARQL query service
- Two RDF dataset releases in 2012
  - April 2012 (22m triples)
  - July 2012 (270m triples)
- Live updates to query endpoint

Led to (2014–)

- Focus on internal use-cases
- Publish ontology pages
- Periodic data snapshots



	s	p	o
1	http://ns.nature.com/articles/ajg20085	rdfs:label	Article: Calendar of Courses, Symposiums, and Conferences
2	http://ns.nature.com/articles/ajg20085	rdf:type	npg:Article
3	http://ns.nature.com/articles/ajg20085	dc:publisher	Nature Publishing Group
4	http://ns.nature.com/articles/ajg20085	bibo:volume	1
5	http://ns.nature.com/articles/ajg20085	foaf:homepage	http://www.nature.com/ajg/journal/v1/n1/full/ajg20085a.html
6	http://ns.nature.com/articles/ajg20085	npg:publicationYear	2009
7	http://ns.nature.com/articles/ajg20085	npg:titleXml	<article-title>Calendar of Courses, Symposiums, and Conferences</article-title>
8	http://ns.nature.com/articles/ajg20085	bibo:contributorList	rdf:nil
9	http://ns.nature.com/articles/ajg20085	prism:copyright	American College of Gastroenterology
10	http://ns.nature.com/articles/ajg20085	dc:title	Calendar of Courses, Symposiums, and Conferences
11	http://ns.nature.com/articles/ajg20085	bibo:pageEnd	10
12	http://ns.nature.com/articles/ajg20085	npg:hasPrimaryArticleType	http://ns.nature.com/primary-article-types/research
13	http://ns.nature.com/articles/ajg20085	npg:hasPublication	http://ns.nature.com/publications/0002-9270
14	http://ns.nature.com/articles/ajg20085	npg:hasPublication	http://ns.nature.com/publications/1572-0241
15	http://ns.nature.com/articles/ajg20085	npg:artid	ajg20085
16	http://ns.nature.com/articles/ajg20085	npg:doiLink	http://dx.doi.org/10.1038/ajg.2008.5
17	http://ns.nature.com/articles/ajg20085	bibo:pageStart	9
18	http://ns.nature.com/articles/ajg20085	bibo:doi	10.1038/ajg.2008.5
19	http://ns.nature.com/articles/ajg20085	dc:identifier	doi:10.1038/ajg.2008.5
20	http://ns.nature.com/articles/ajg20085	dc:identifier	info:doi/10.1038/ajg.2008.5

NPG Content Hub (2014): Hybrid Architecture

Features

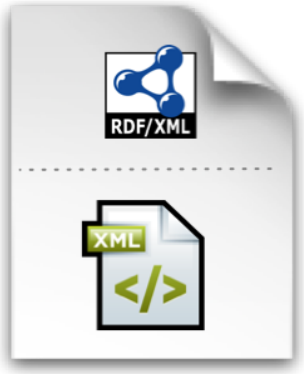
- Hybrid RDF + XML architecture
  - MarkLogic for XML, RDF/XML
  - Triplestore (TDB) for RDF validation
- Repo's for binary assets

Layout

- Semantic RDF/XML includes in XML
- RDF objects serialized in list order
- Application XML for subject hierarchy

Indexes

- Indexes over all elements
- Range indexes for datatypes (e.g. dates)

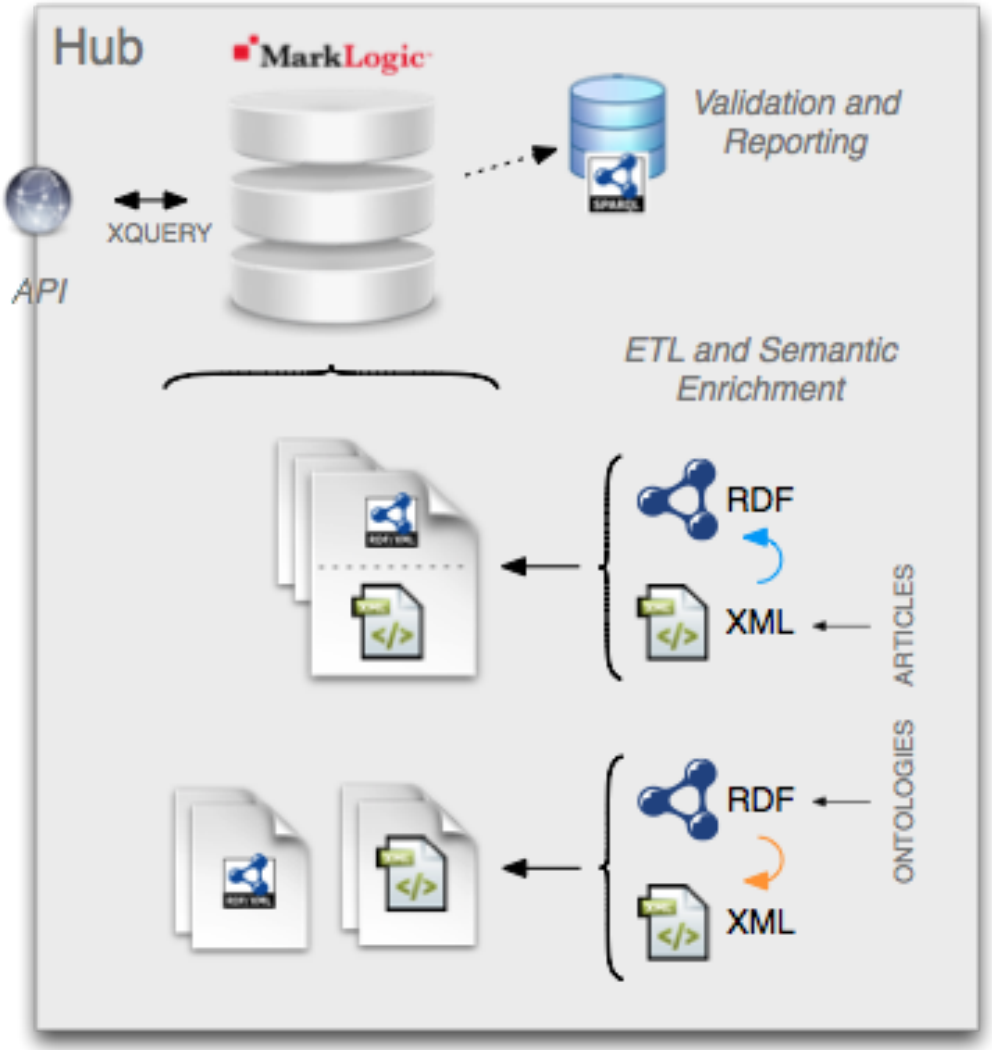



```
<semantic-metadata xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
 <graph uri="http://ns.nature.com/graphs/articles">
  <rdf:RDF>
   <npg:Article rdf:about="http://ns.nature.com/articles/maestrorj2
    ...
    <bibo:doi>10.1038/maestrorj.2014.4</bibo:doi>
    <bibo:number>140004</bibo:number>
    <bibo:volume>1</bibo:volume>
    <dc:title>A new climate for grazing livestock</dc:title>
    ...
   </npg:Article>
  </rdf:RDF>
 </graph>
 <graph uri="http://ns.nature.com/graphs/journals">
  <rdf:RDF>
   <npg:Journal rdf:about="http://ns.nature.com/journals/maestro
    <dc:title>Maestro RJ</dc:title>
    <foaf:homepage  rdf:resource="http://test-www.nature.com/ma
   </npg:Journal>
  </rdf:RDF>
 </graph>
 ...
</semantic-metadata>
```

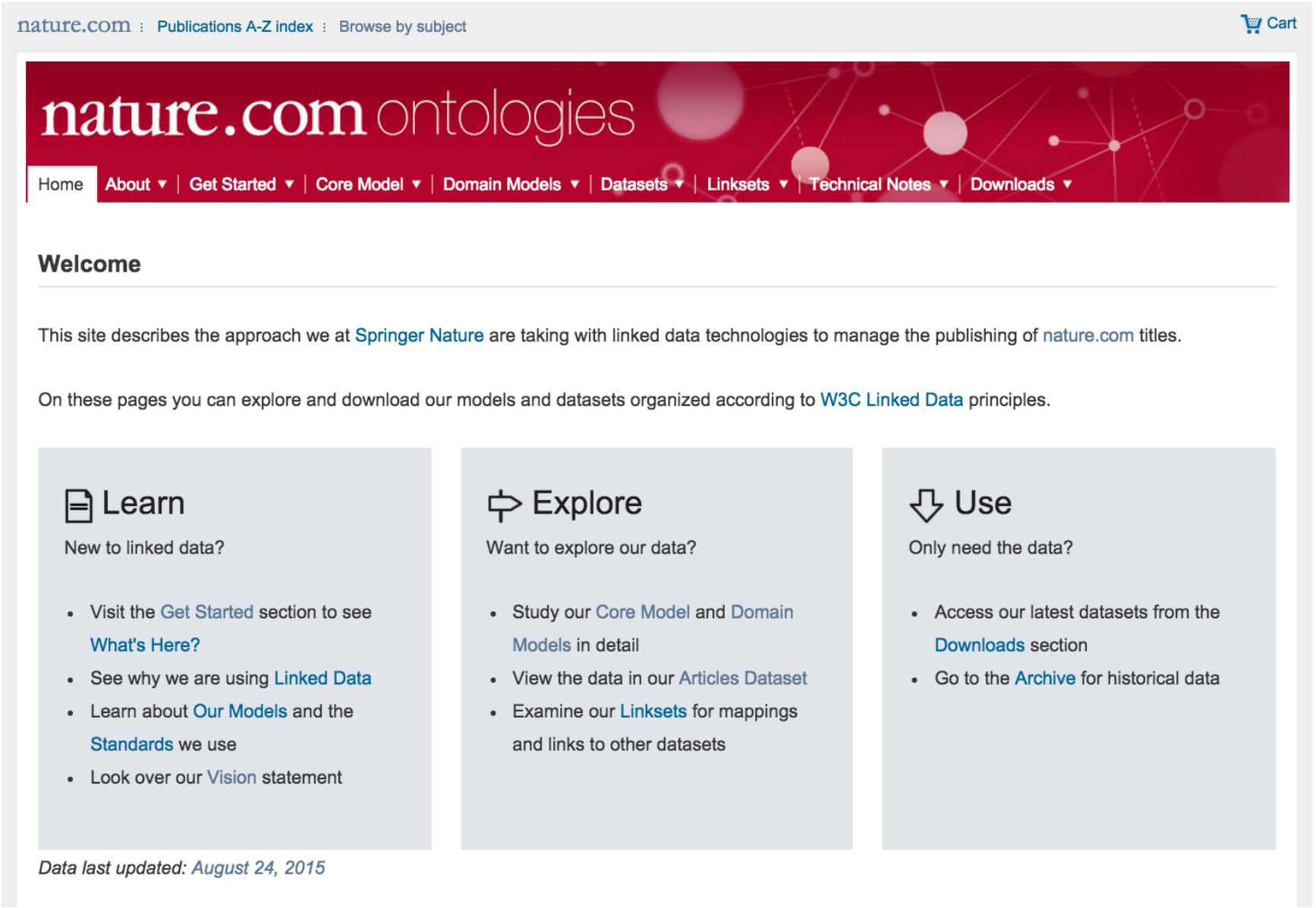
Subject Pages (2014)



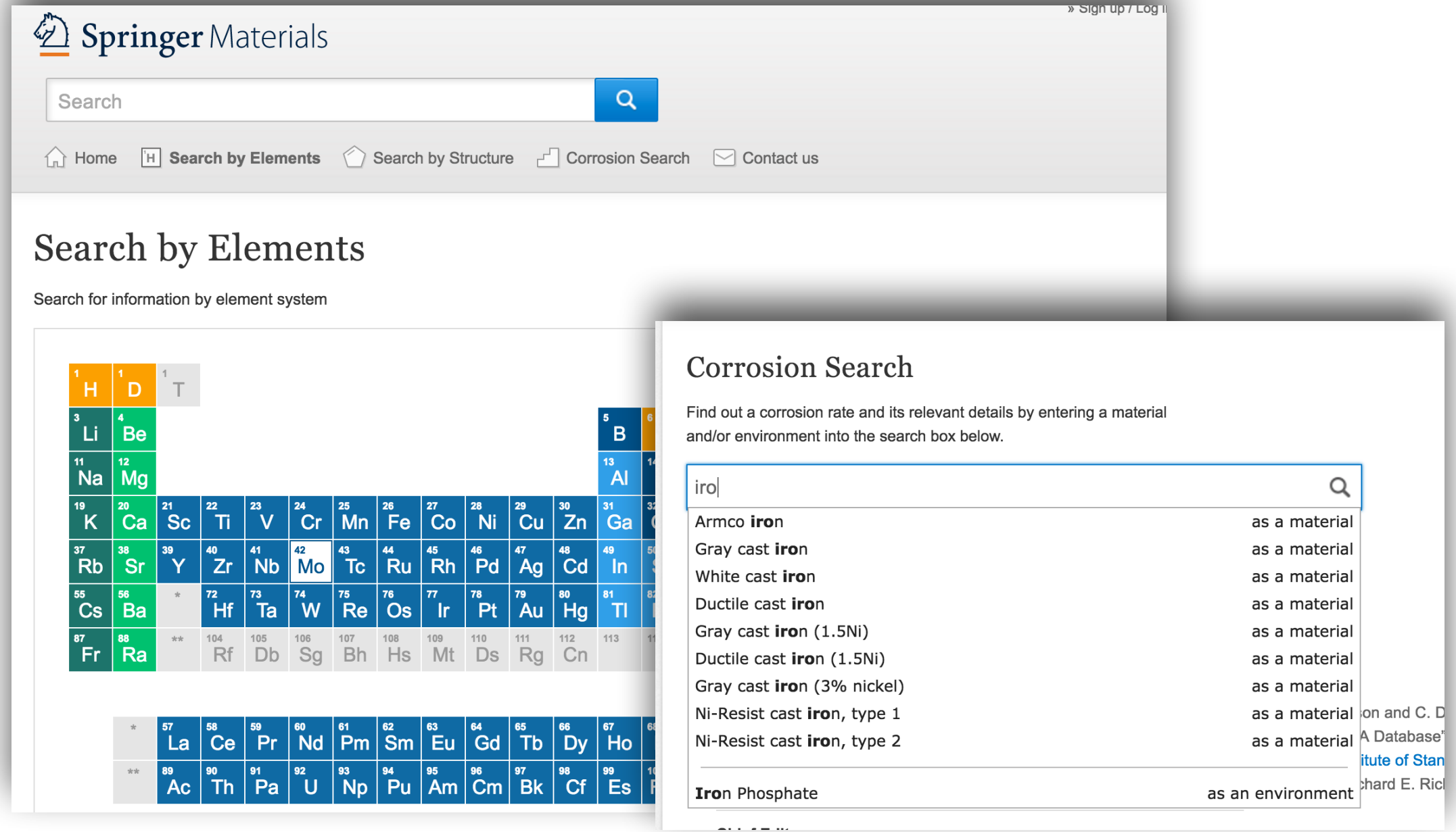




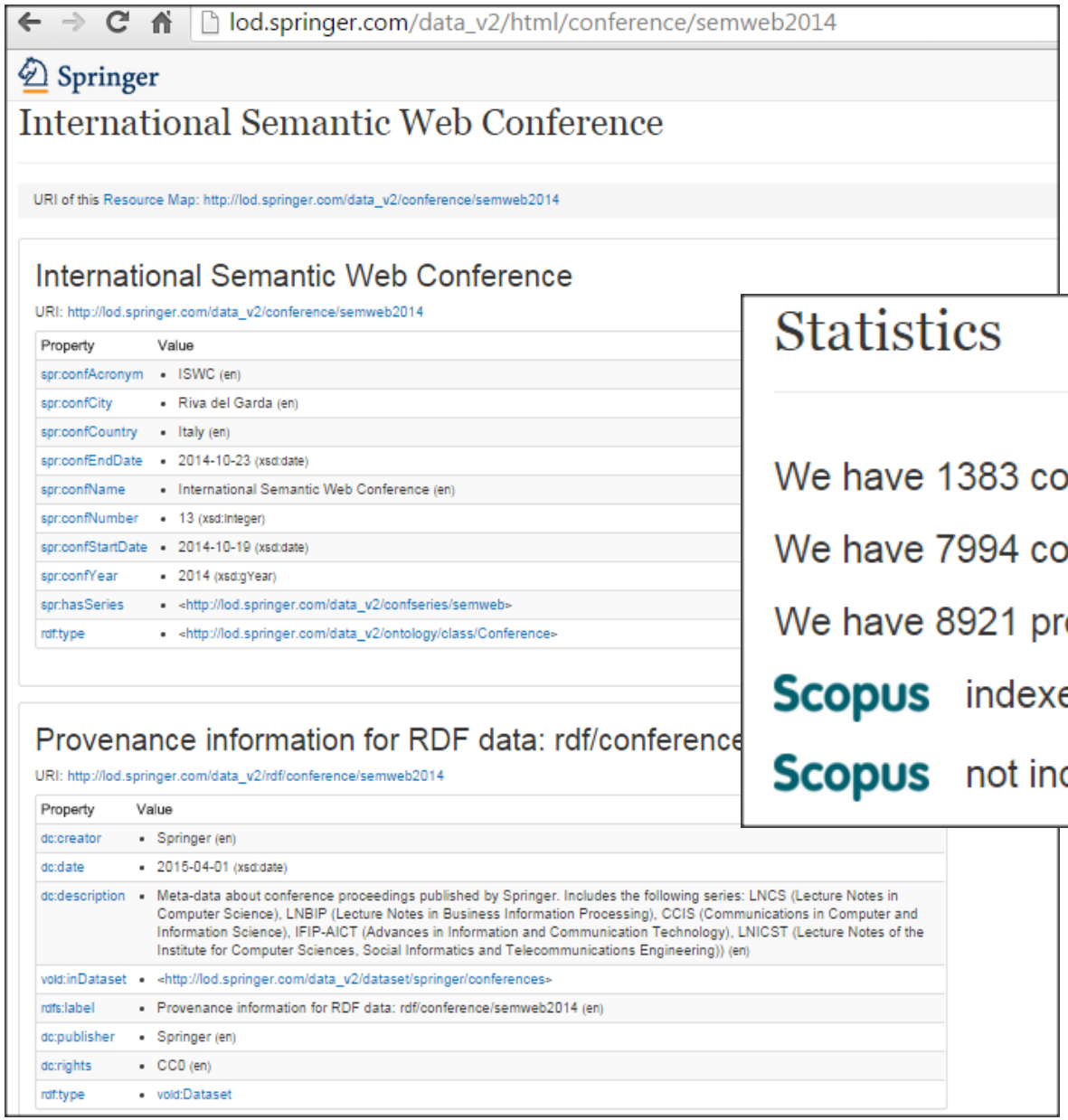
NPG Ontologies Portal (2015): Data Publishing



Springer Materials (2014)



Springer Conferences Portal (2015)



Statistics

We have 1383 conference series

We have 7994 conferences

We have 8921 proceedings volumes

Scopus indexed: 7555 (download as CSV)

Scopus not indexed: 1366 (download as CSV)

Scigraph Project (2016): main objectives

Data Integration

> Consolidation of existing LD efforts via a single domain mode

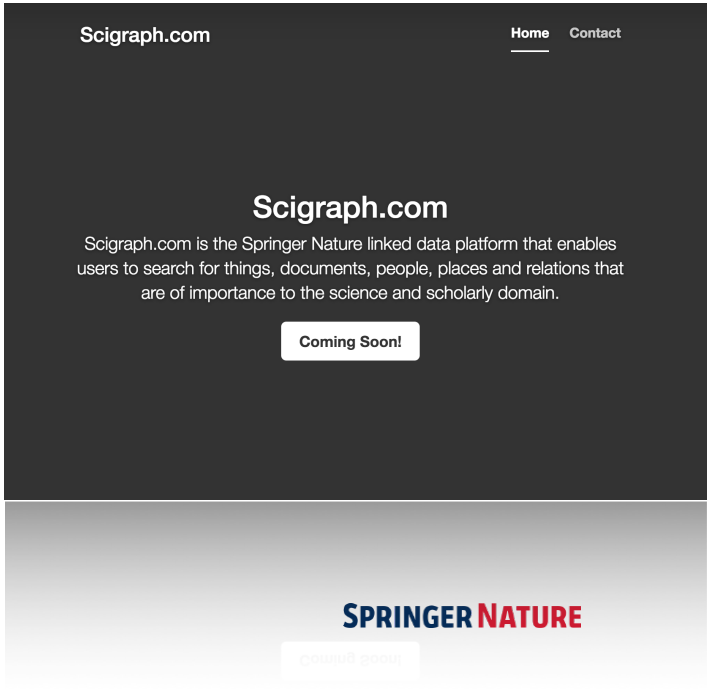
> Ingestion and normalisation of third party datasets

Discoverability

> Better end user applications [B2C]

> Metadata delivery & validation [B2B]

> Data publishing [B2developers]



Scigraph

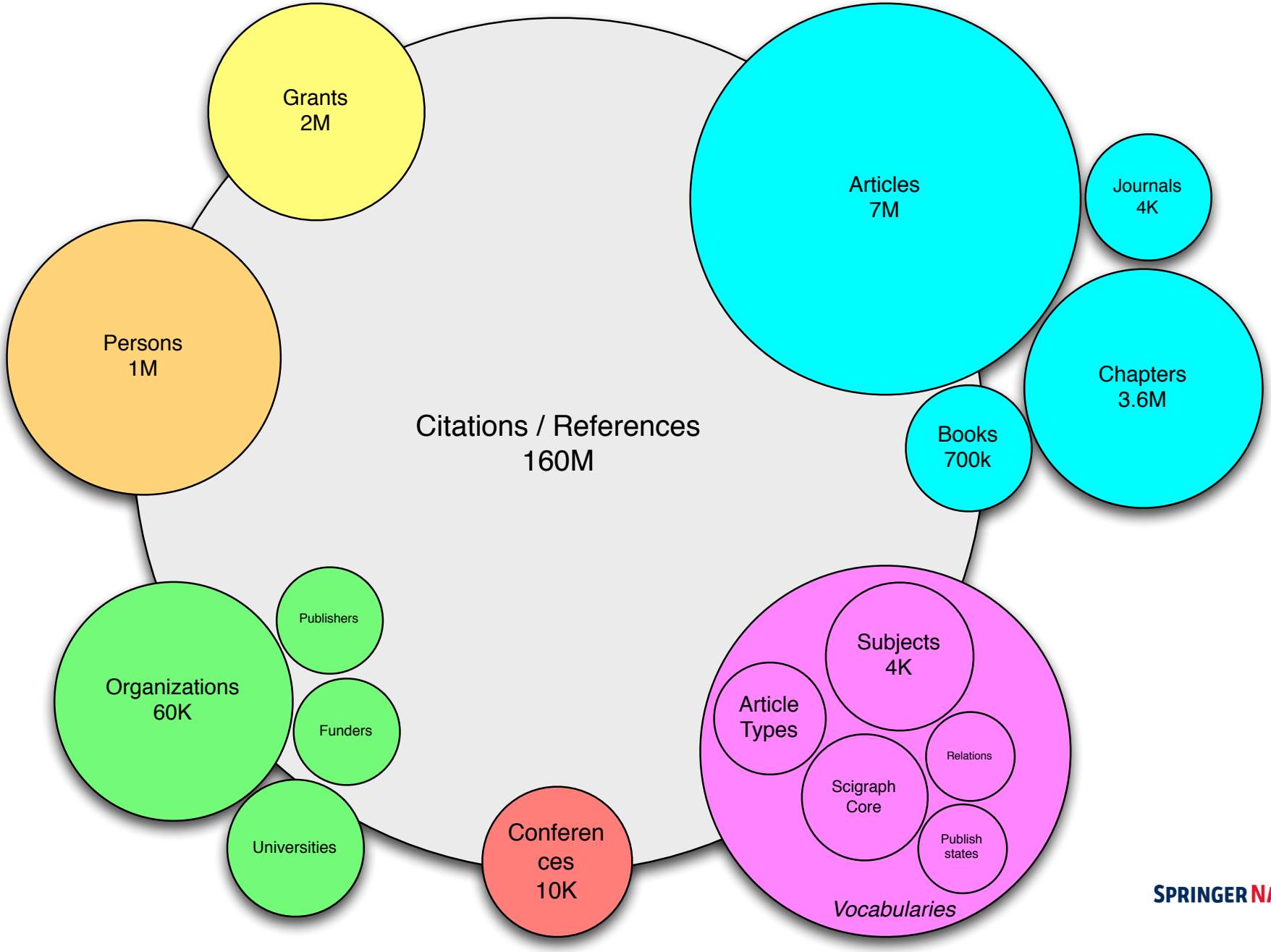
what's in it

> data architecture, taxonomies, ontologies

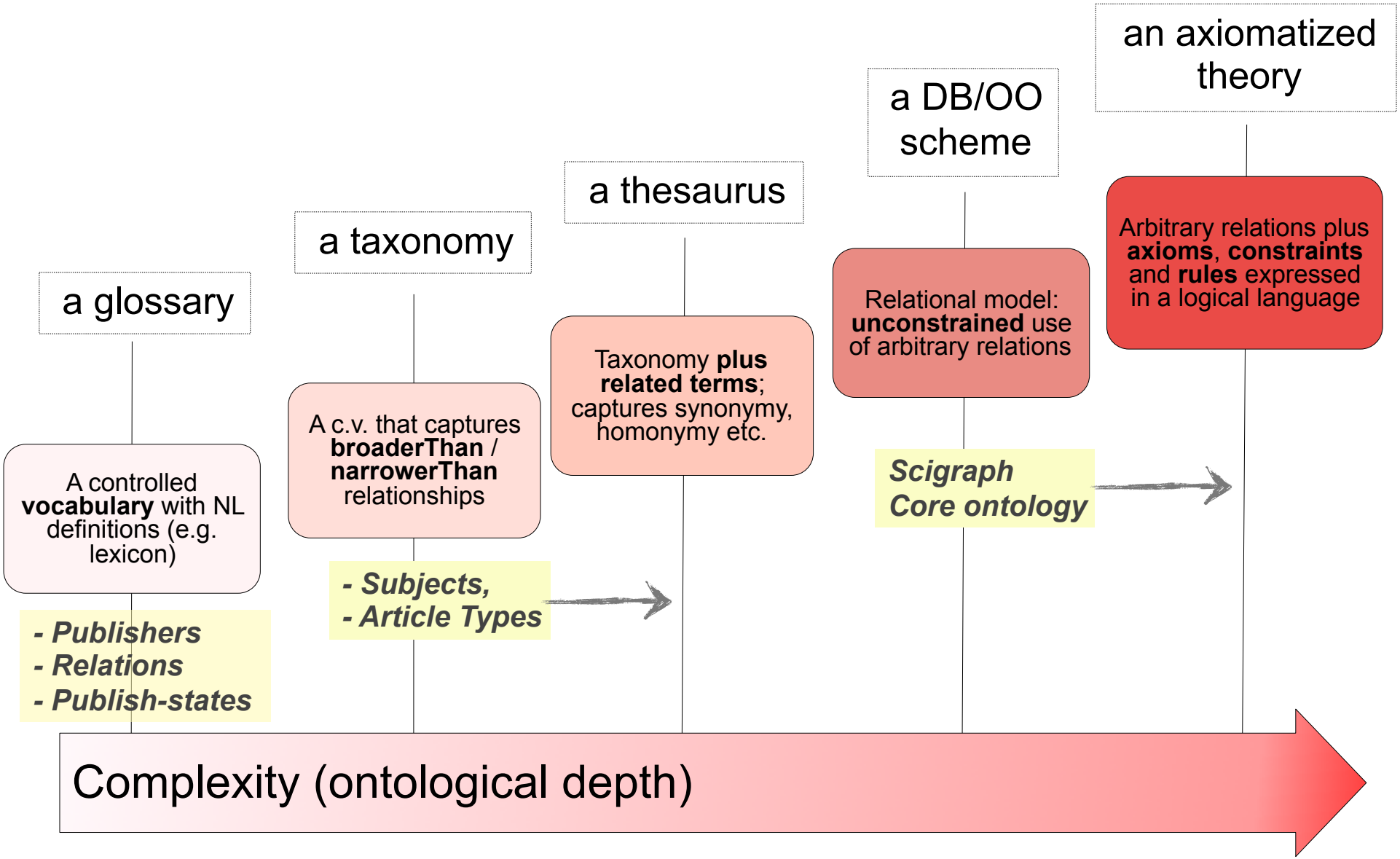
how it works

> ETL, naming, validation, identity

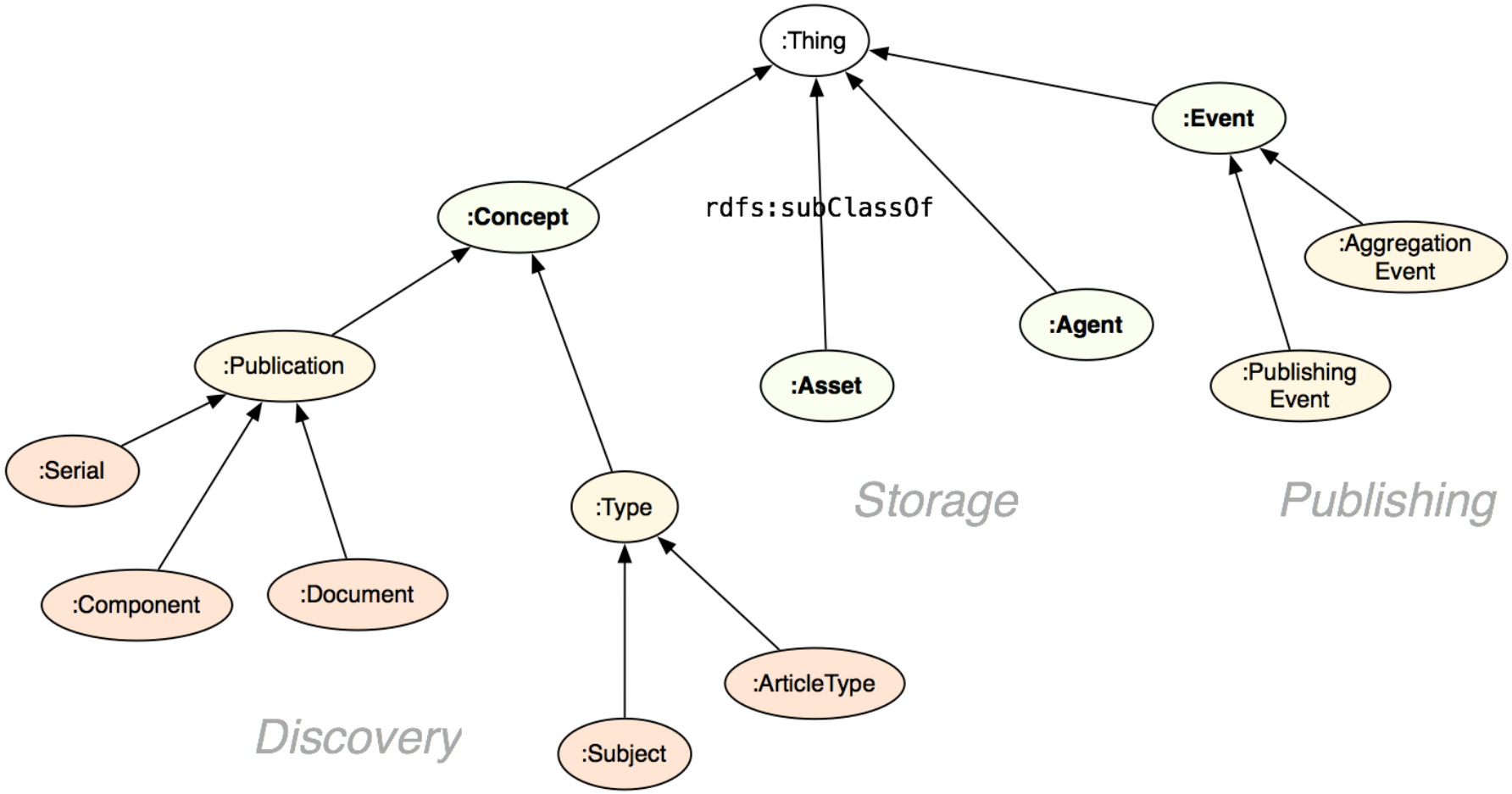
Data Landscape



Ontologies and Taxonomies: overview



The Core Ontology

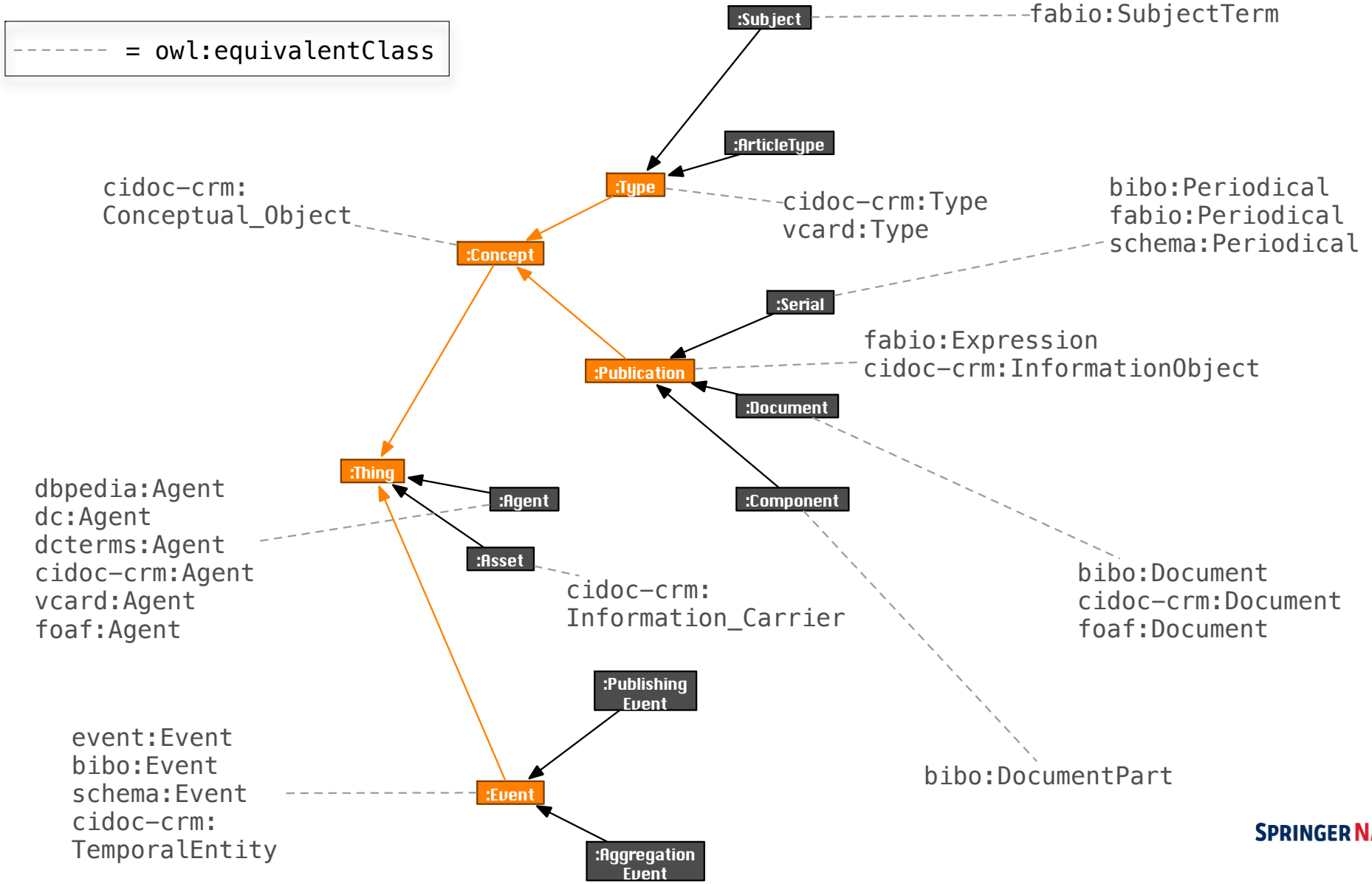


- **Language**: OWL 2, Profile: *ALCHI(D)*
- **Entities**: ~73 classes, ~250 properties
- **Principles**: Incremental Formalization/ Enterprise Integration / Model Coherence

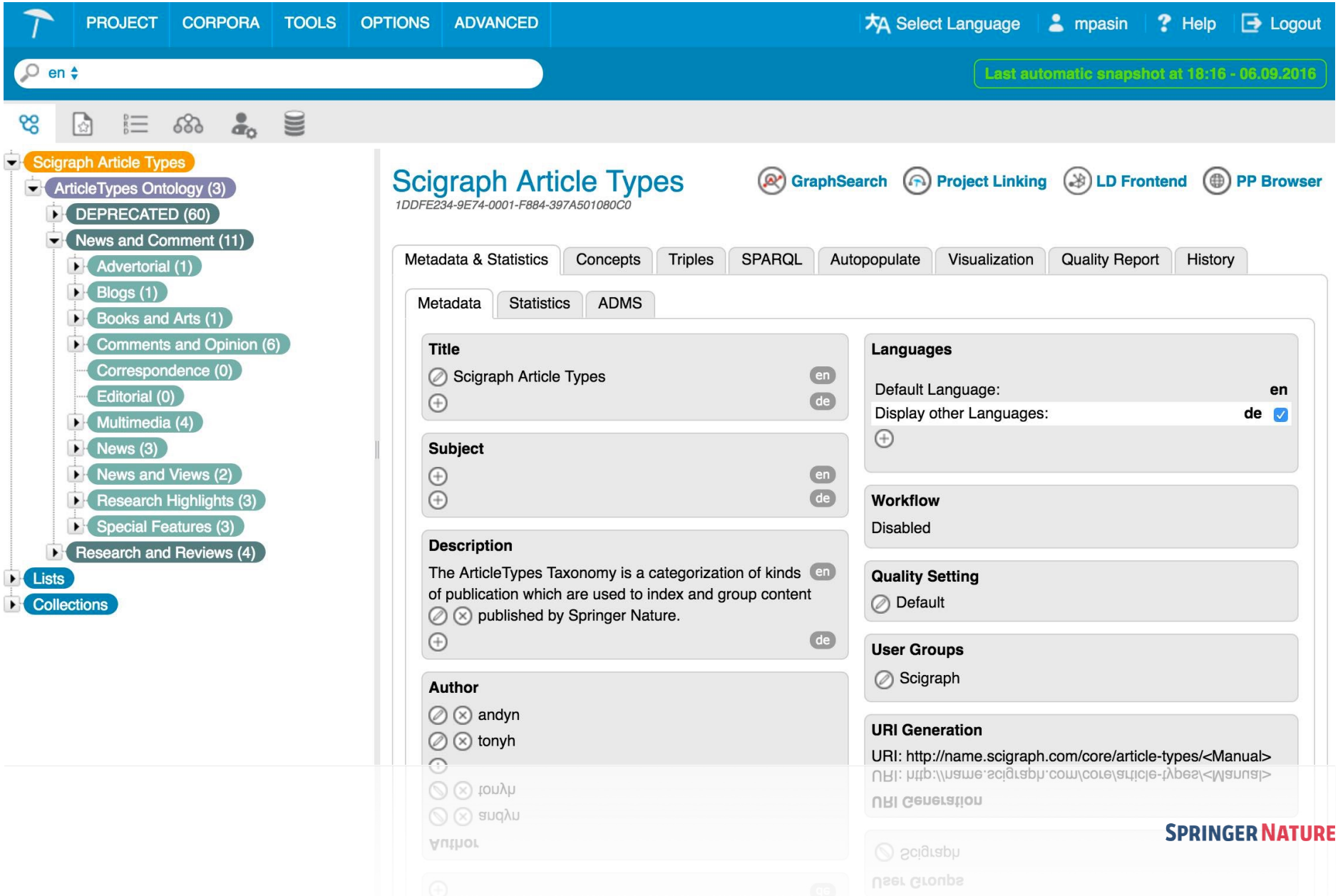
SPRINGER NATURE

http://www.nature.com/ontologies/core/

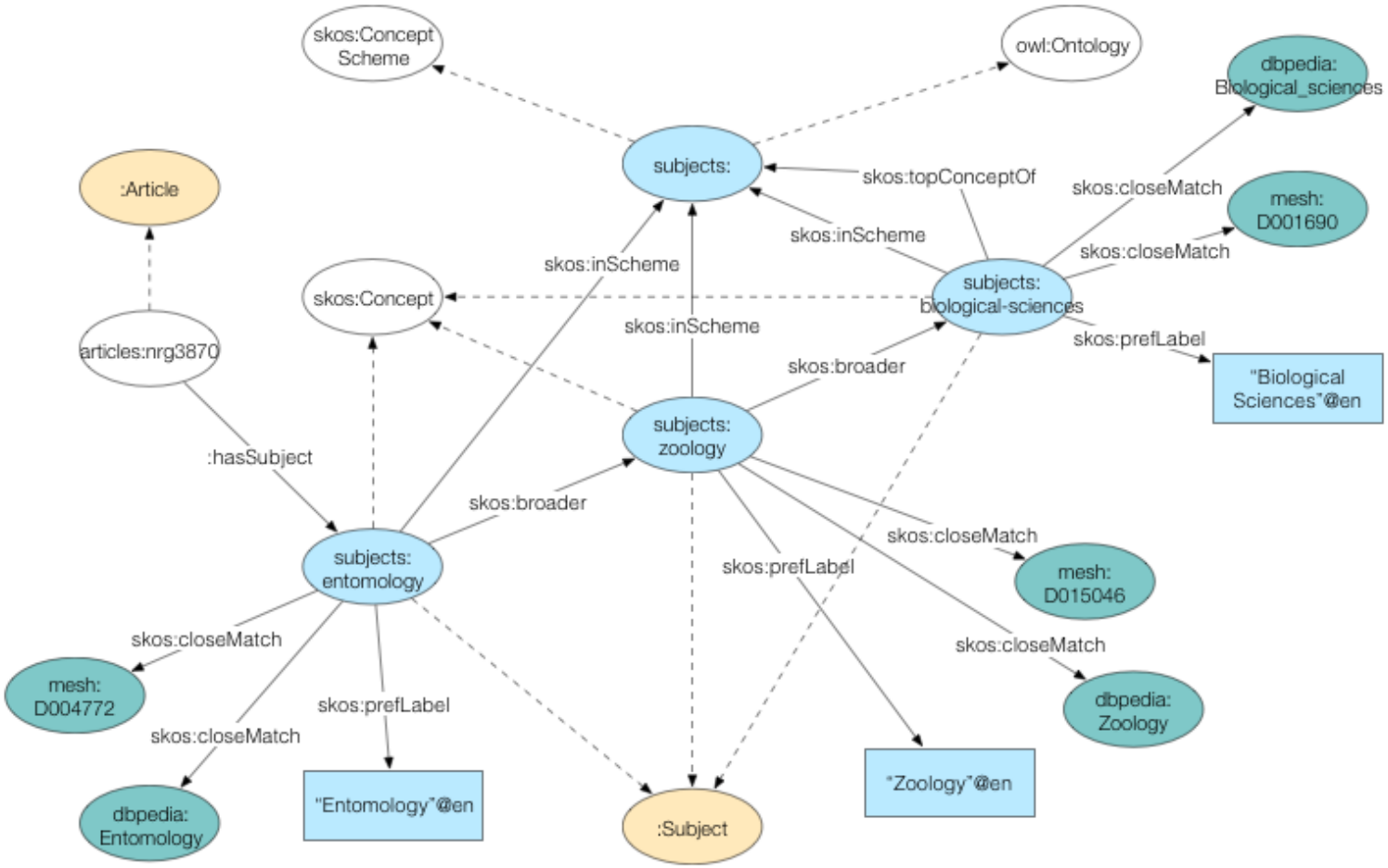
The Core Ontology: mappings



SKOS taxonomies: Poolparty integration

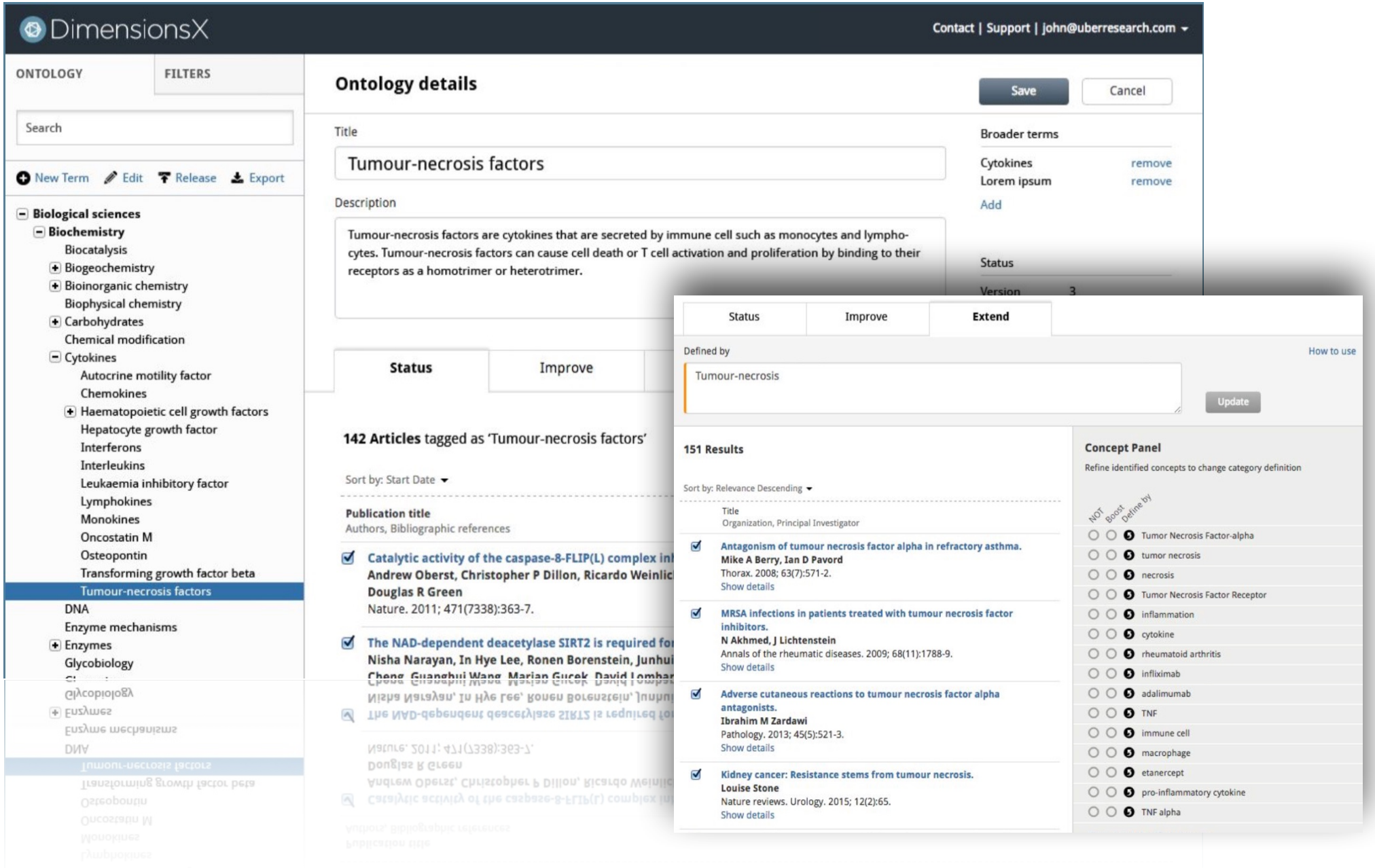


SKOS taxonomies: Subjects



- **Structure**: SKOS, ~2500 concepts, multi hierarchical tree, 6 branches, 7 levels of depth
- **Mappings**: 100% of terms, using *skos:broadMatch* or *skos:closeMatch,* (Dbpedia and MESH)
- **Document tagging:** mostly manual, different workflows, often costly and inconsistent

Semi-Automatic tagging with Dimensions (from *UberResearch*)



Scigraph

what's in it

> data architecture, taxonomies, ontologies

how it works

> ETL, naming, validation, identity

Naming Architecture: federated model

> **Dereference** and 303 redirects:
  - http://name.scigraph.com/{things}/
  - http://data.scigraph.com/{things}/

> Two **patterns**: schemas and instances
  - http://name.scigraph.com/ontologies/{domain}/
  - http://name.scigraph.com/{domain}/{things}/

> **Prefixes** for schemas and instances
  - @prefix sg: <http://name.scigraph.com/ontologies/core/> .

> Entity **names** follow a robust convention
  - camel-case for naming terms, with an initial uppercase for classes and an initial lowercase for properties.

> Named graphs used to track **provenance**

Prefix	Namespace
sgg:	http://name.scigraph.com/core/graphs/
sgo:	http://name.scigraph.com/ontologies/
sg:	http://name.scigraph.com/ontologies/core/

- sg: = sgo:core/
- sg:Thing = sgo:core/Thing
- demo:Thing = sgo:demo/Thing

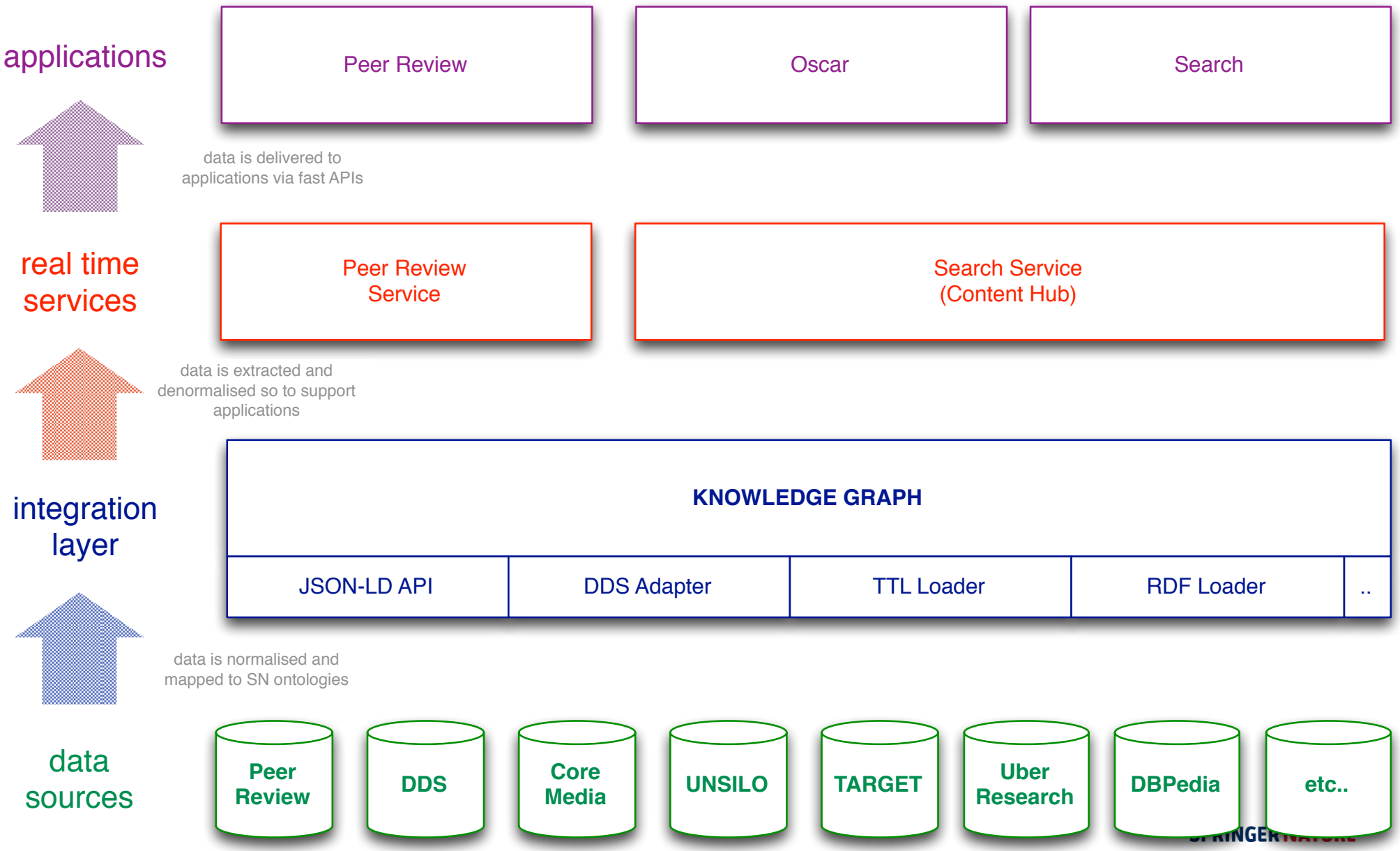
```
Example (Classes):

> sg:Book    # class
> sg:BookChapter    # class

Example (Properties):

> sg:hasBookChapter    # object property
> sg:isOpenAccess    # datatype property
> sg:xmlUriTemplate    # datatype property
```

Scigraph - Data Flow



ETL Architecture: main features [in evolution]

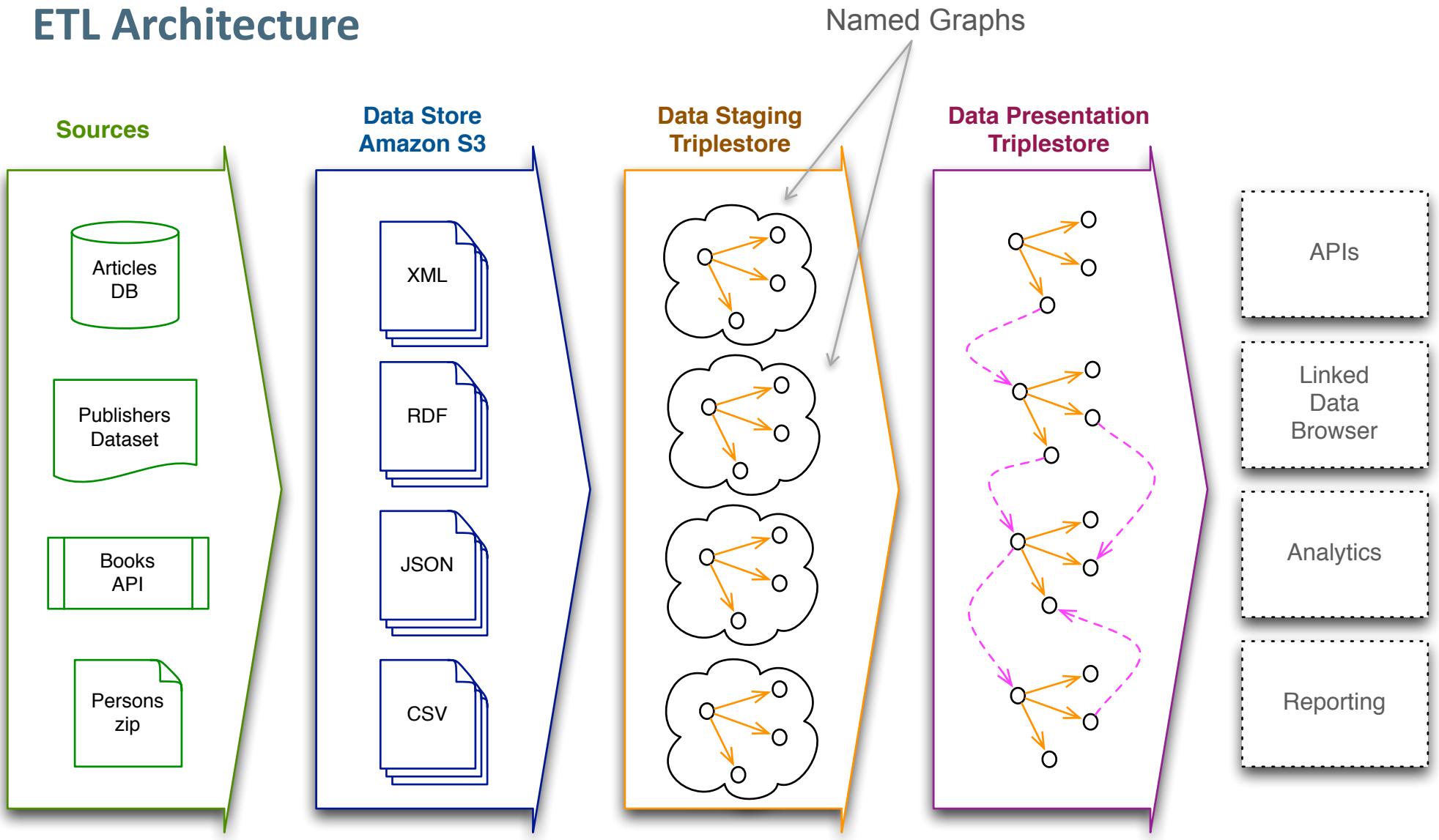
Tech stack

> 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...)

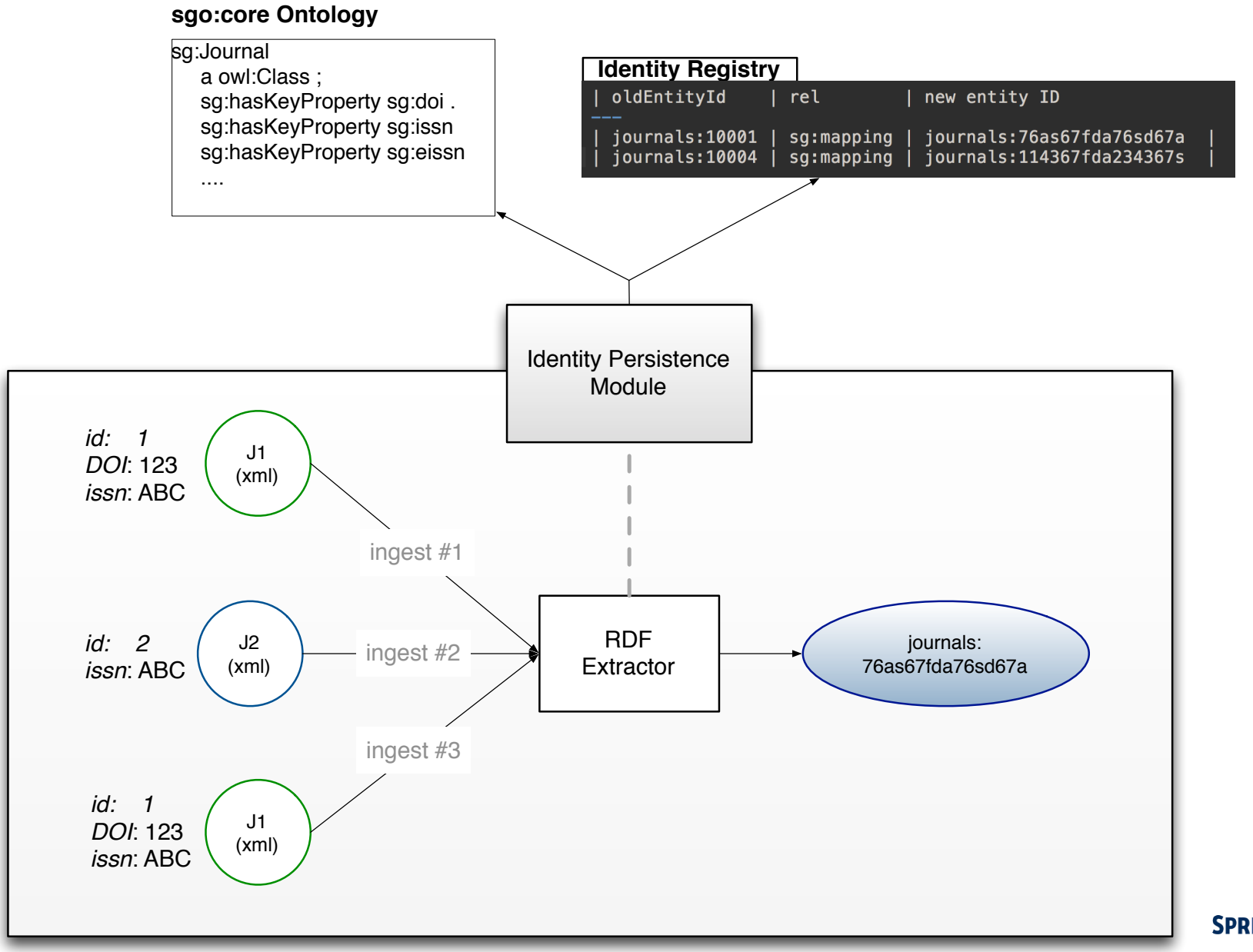
→

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

→

- Integration (union graph)
- Inference

Identity Persistence



Data Validation: from SPIN to SHACL

> **SPIN** SPARQL syntax (2011, TopQuadrant)

> Example: "*if a Journal instance has no short title, raise an Exception*"

> Main drawback: hard to maintain and to read by non specialists

```
construct {
  npgg:journals npg:hasConstraintViolation [
    a spin:ConstraintViolation ;
    npg:severityLevel "Warning" ;
    rdfs:label ?message ;
    spin:rule [ a sp:Construct ; sp:text ?query ; ] ;
  ] .
}
where {
  { select (count(?s) as ?count)
    where {
      ?s a npg:Journal .
      filter ( not exists { ?s bibo:shortTitle ?h . } ) }
  }
  bind (concat("! Found ", str(?count), " journals with no short title") as ?message)
  bind ("""
construct {
  npgg:journals npg:hasConstraintViolation [
    a spin:ConstraintViolation ;
    spin:violationRoot ?s ; ... ] .
} where { ... }
""" as ?query)
}
```

Data Validation: from SPIN to SHACL

> **SHACL** - Shapes Constraint Language (2016, TopQuadrant)

> **Example**: "*all article instances should have a valid DOI*"

> **Example**: "*all grants instances should have max 1 start year and end year*"

> **Approach**: polish data before entering the triplestore, use triplestore inference primarily for integration

```
@prefix shapes: <http://name.scigraph.com/core/shapes/> .

shapes:ArticleShape
    a sh:Shape ;
    sh:targetClass sg:Article ;
    # sh:closed true ;
    sh:ignoredProperties ( rdf:type ) ;
    sh:property [
        sh:predicate sg:doi ;
        sh:datatype xsd:string ; sh:maxCount 1 ;
        sh:pattern "^10\\.\\d{4,5}\\/\\S+$" ;
    ] ;
    # ...
    .

shapes:GrantModelShape
    a sh:Shape ;
    sh:targetClass sg:Grant ;
    # sh:closed true ;
    sh:property [
        sh:predicate sg:endYear ; sh:datatype xsd:gYear ;
        sh:maxCount 1 ;
    ] ;
    sh:property [
        sh:predicate sg:startYear ; sh:datatype xsd:gYear ;
        sh:maxCount 1 ;
    ] ;
    .
```

Next Steps

Looking Ahead

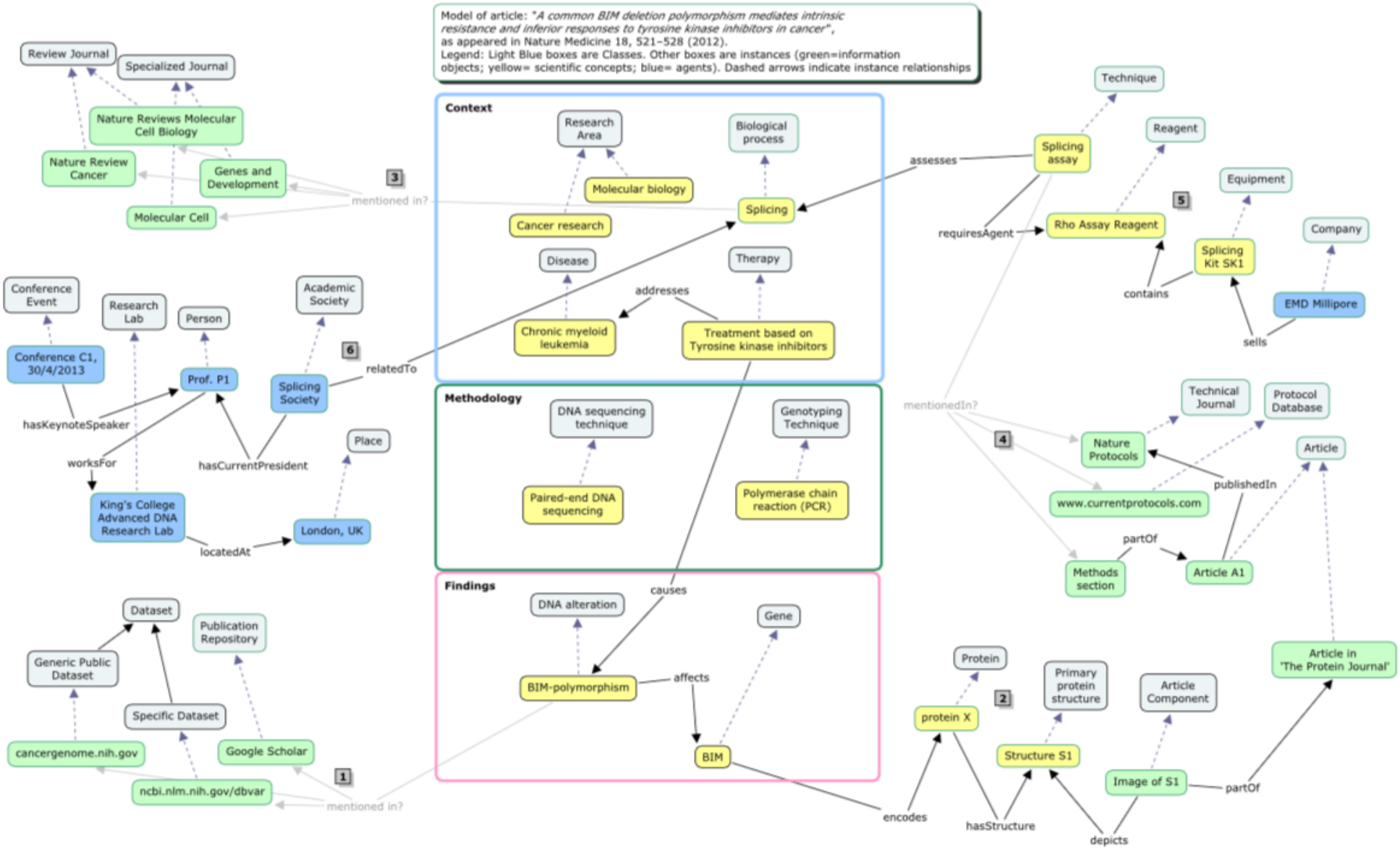
Summary

- Scigraph is our latest LD platform - public version live in late 2016
- SW tech allows for scalable enterprise-level metadata management
- It is crucial to distinguish between data Integration VS (real time) data delivery
- Still a work in progress… suggestions or feedback very welcome!

Ongoing Work

- Ontology: federated model, more advanced inferencing capabilities
- Build internal/external APIs (JSON-LD) by integrating also NoSQL
- Tools for analytics, reporting, visualisation, interactive exploration of the graph
- Entities extraction: scientific entities, places, people, events etc..
- We're looking to collaborate… Crossref, W3C, building a Linked Science Web

Future: a scientific article X-ray?



The Knowledge Graph team

CORE TEAM

* Markus Kaindl: Product Owner
* Ben Kirkley: Project Manager

* Michele Pasin: Lead Data Architect
* Tony Hammond: Data Architect
* Matias Piipari: Lead Engineer
* Hilverd Reker: Software Engineer
* Artur Konczak: Software Engineer

* <blankNode>: Data Scientist
* <blankNode>: Data Engineer

PROGRAM SPONSOR

* Henning Schoenenberger: Director Data & Metadata

DIGITAL SCIENCE

* Martin Szomszor: Data Scientist
* Richard Koks: Data Scientist
* Mario Diwersy: CTO, Uber Research

Thanks

michele.pasin@nature.com