LINKED DATA EXPERIENCE AT MACMILLAN

Building discovery services for scientific and scholarly content on top of a semantic data model

Tony Hammond
Michele Pasin

22 October 2014



macmillan
Science and Education

Transforming Learning and Discovery

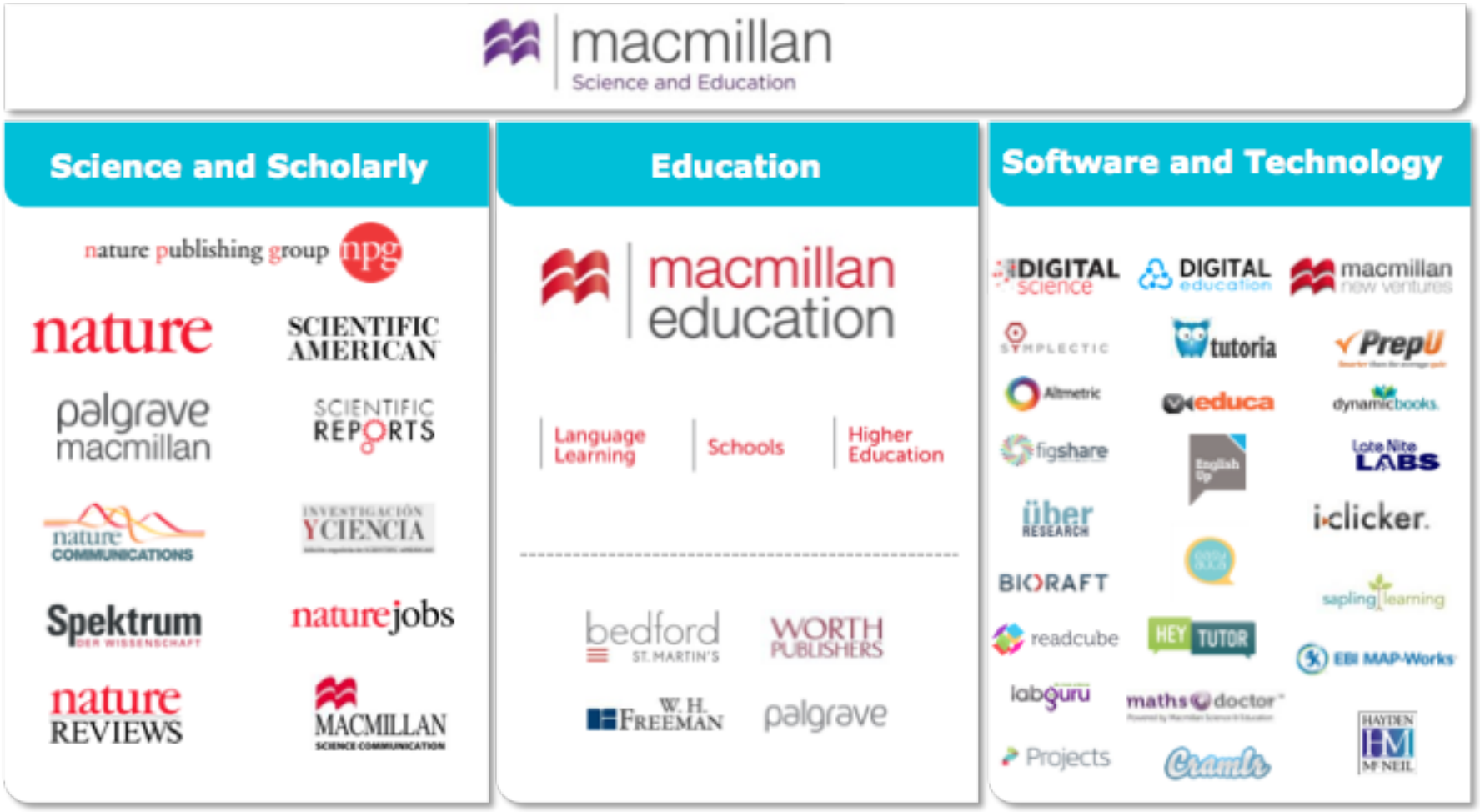
Background

About Macmillan and what we are doing

1

Macmillan Science and Education

Group brands and businesses

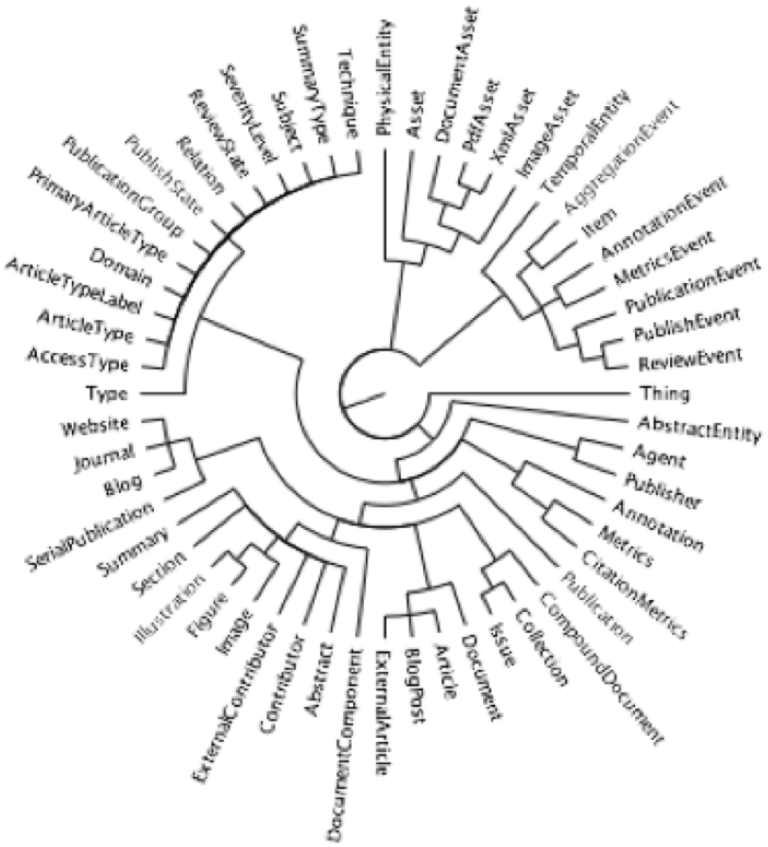


MS&E Current trends

Developing a richer graph of objects

Change Drivers

- Digital first workflow
  - print becomes secondary
  - support for multiple workflows
- User-centric design
  - things, not data
  - focus on user experience
- Deeply integrated datasets
  - standard naming convention
  - common metadata model
  - flexible schema management
  - rich dataset descriptions



NPG Linked Data Platform (2012)

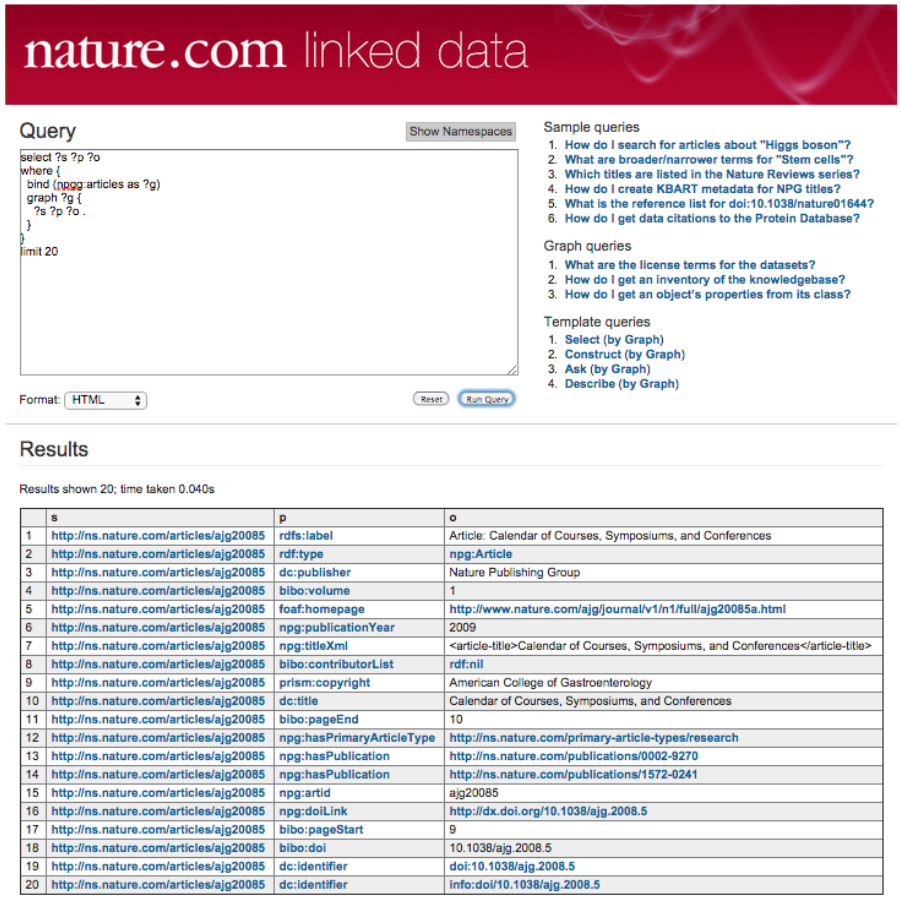
data.nature.com

Deliverables (2012–2014)

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

Current Work (2014–)

- Focus on internal use-cases
- Publish ontology pages
- Periodic data snapshots (no endpoint)



	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 Core Ontology (2014)

Things: assets, documents, events, types

Features

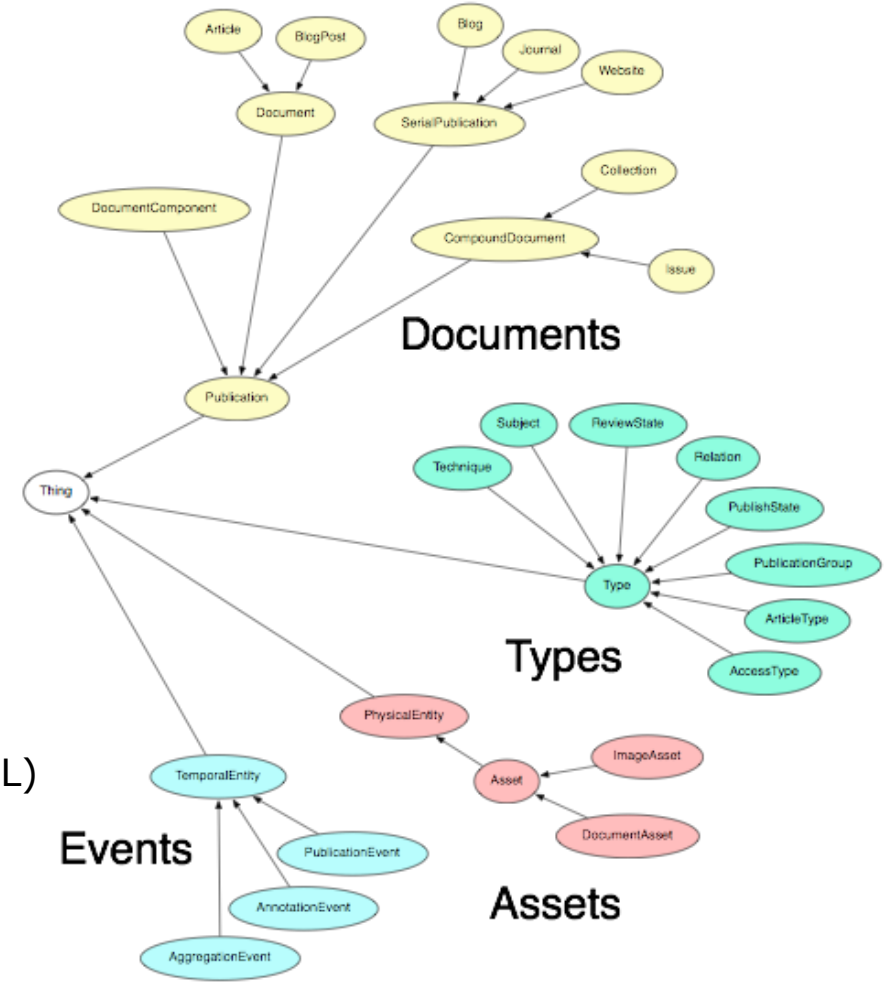
- Classes: ~65
- Properties: ~200
- Named graphs (per class)

Namespaces

- npg: => http://ns.nature.com/terms/
- npgg: => http://ns.nature.com/graphs/

Approach

- Minimal commitment to external vocabs
- Incremental formalization (RDF, RDFS, OWL-DL)
- Shared metamodel vs. automatic inference

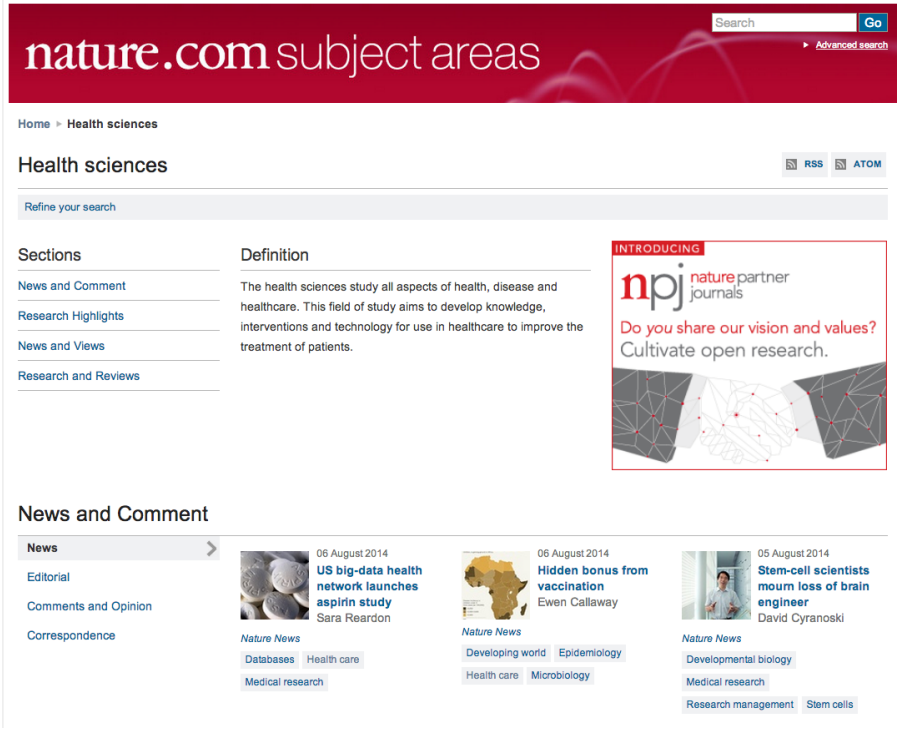


NPG Subject Pages (2014)

Topical access to content

Features

- Based on SKOS taxonomy
  - >2750 scientific terms
  - content inherited via SKOS tree
- Completely automated
  - one webpage per subject term
  - structure based on article type
  - secondary pages for specific types
- Various formats e.g. eAlerts, feeds, etc.
  - allows people to ‘follow’ a subject
- Customized related content
  - ads, jobs, events, etc.



Data Storage and Query

Achieving speed by means of a hybrid architecture

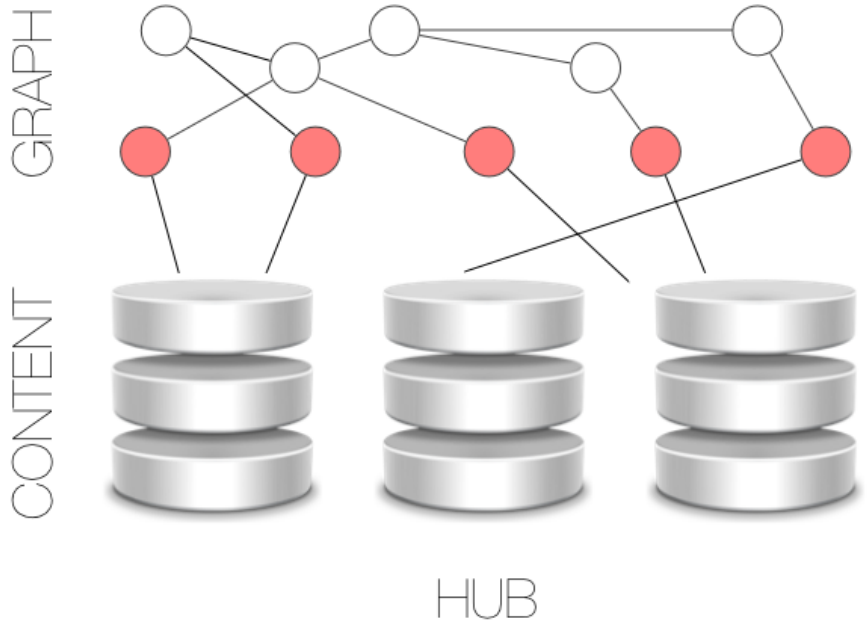
2

Content Hub

Managed content warehouse for data discovery

Capabilities

- Discovery – Graph
- Storage – Content Repos



Features

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

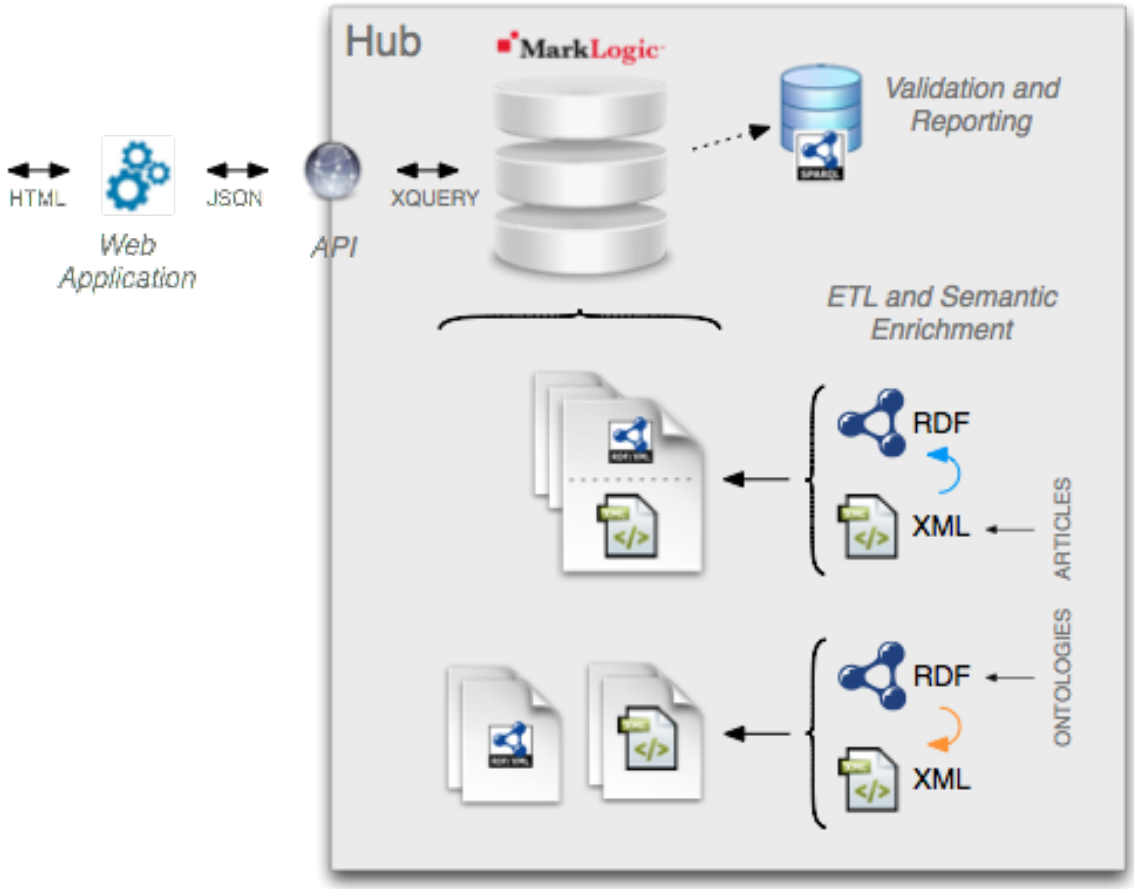
Datasets

- Documents (large; >1m)
- Ontologies (small; <10k)

System Architecture

Hub content





Content Discovery – Principles

Readying the API for applications

Generations

- 1st – Generic linked data API (RDF/*)
- 2nd – Specific page model API (JSON)

Concerns

- Speed (20ms single object; 200ms filtered object)
- Simplicity (data construction)
- Stability (backup, clustering, security, transactions)

Principles

- Chunky not chatty, all data in a single response
- Data as consumed, rather than as stored
- Support common use cases in simple, obvious ways
- Ensure a guaranteed, consistent speed of response for more complex queries
- Build on foundation of standard, pragmatic REST (collections, items)

Content Discovery – Optimization

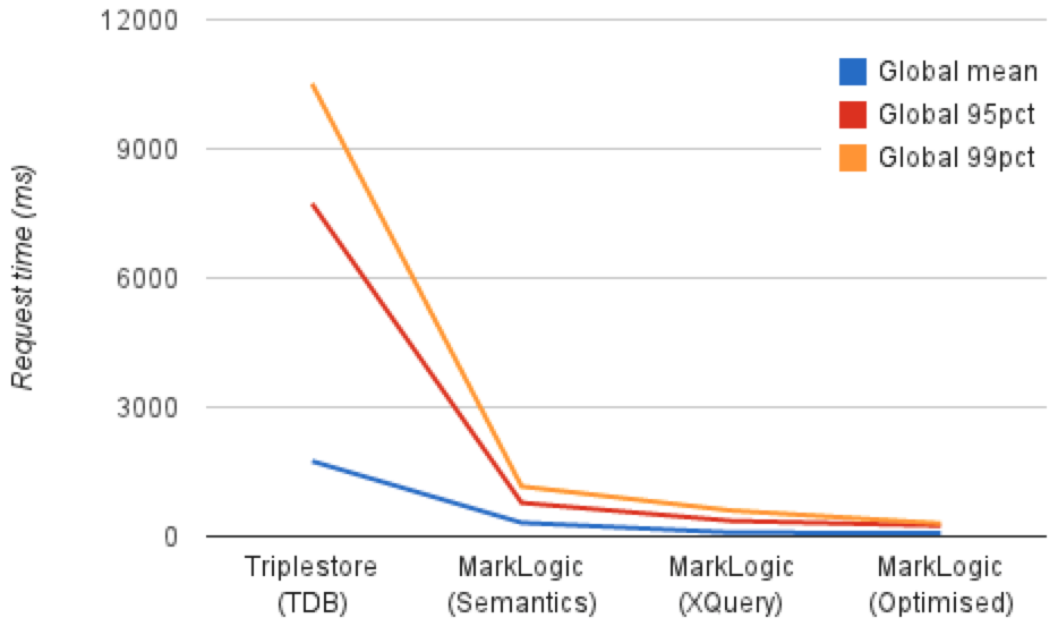
Tuning the API for performance

Approaches

- TDB + Fuseki – SPARQL
- MarkLogic Semantics – SPARQL
- MarkLogic – XQuery
- MarkLogic (Optimized) – XQuery

Techniques

- Partitioning – RDF/XML objects
- Streaming – serialization
- Hashing – dictionary lookup
- Cacheing – *Varnish*



Content Storage – Layout and Indexing

Readying the data for page delivery

Challenges

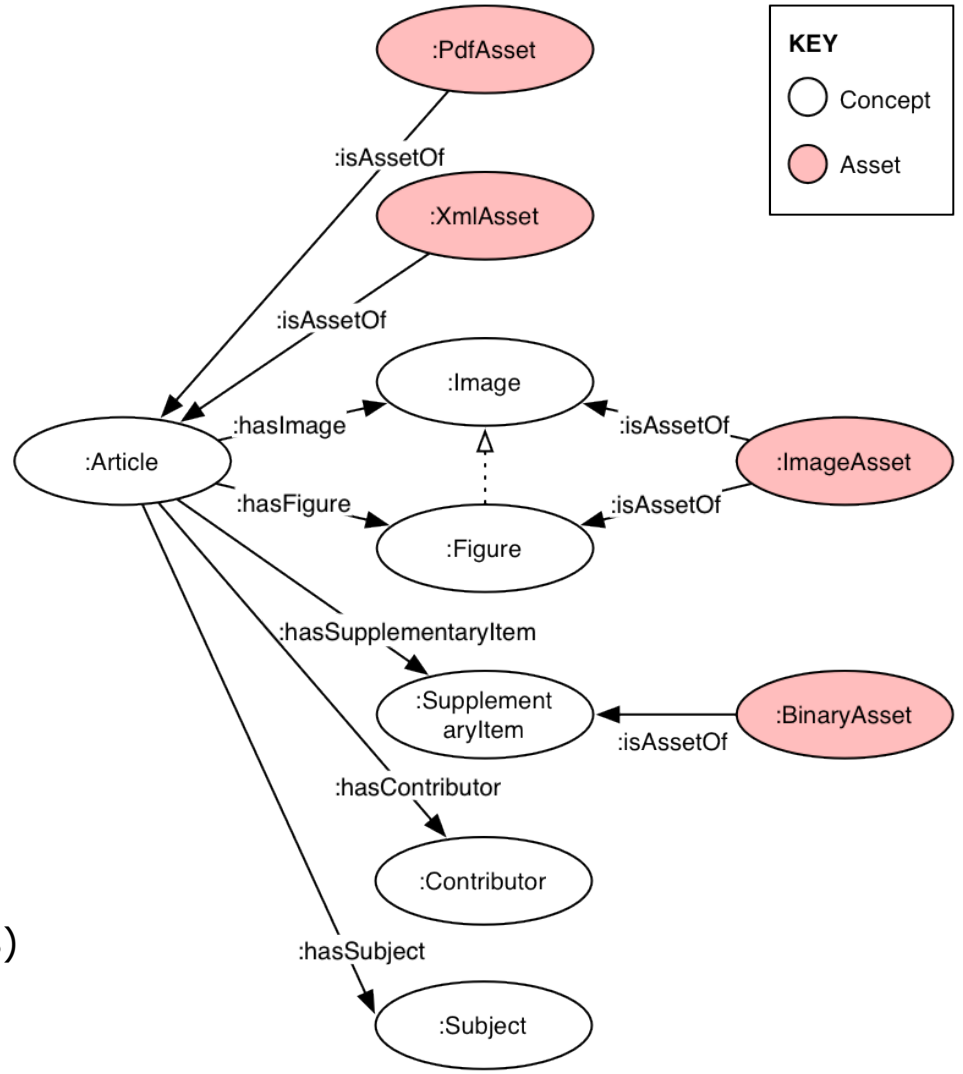
- Sort orders
- RDF Lists
- Facetting, counting

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)

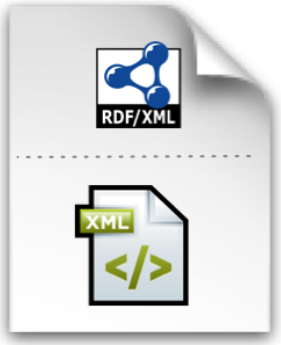


Content Storage – Example

Semantic metadata

Techniques

- XML header for semantic metadata
- All article data is localized
- Maintain named graphs via <graph/> elements
- RDF/XML-ABBREV
- Simple XML :: JSON mapping




```
<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>
```

In Conclusion

A few lessons learned

Summary

- An RDF metamodel allows for scalable enterprise-level data organization
- It is crucial to adequately distinguish between internal and external use cases
- A hybrid architecture proved to be an efficient internal solution for content delivery

Future Work

- Grow the ontology so that it matches product requirements more closely
- Allow for more advanced automatic inferencing
- Provide richer query options both via the API and SPARQL endpoints
- Maintain and expand the vision of a shared semantic model as a core enterprise asset

Thank you

For more information
please contact

TONY HAMMOND
Data Architect, Content Data Services
tony.hammond@macmillan.com

MICHELE PASIN
Information Architect, Product Office
michele.pasin@macmillan.com



Transforming Learning and Discovery