What’s New in Semantic Enrichment
4 Million Content Items, 120 Disciplines, and 1 Metadata Repository

Jess Lawson
Head of Content Architecture, GAB-IT
It’s all in the Title…

- Why semantic enrichment: 4 million content items (and counting)…
- What are the challenges: 4 million content items and 120 subject disciplines…
- How are we facing them: 1 metadata repository
The case for semantic enrichment in GAB
Describing what your content is about enables...

- More accurate data integration (e.g. mashups, integrating internal silos)
- Reuse and repurposing (e.g. microsites or other custom websites)
- Link generation based on an understanding of what each content unit (chapter, article, dictionary definition) is actually about.
- Semantic search (e.g. Google Hummingbird & Knowledge Graph) – focuses on the meaning behind the query and content

Intelligent and sustainable content
The challenges we face...
From this:
The challenges we face... to this:
The challenges we face... with limited amounts of this:
The challenges we face...
or this:
How GAB are facing the challenge
From structured content to intelligent content

Mark-up

Semantic meaning

Logical structure

Text

User: human

computer

organisation

presentation

interpretation

GAB Today

Structured content

User: human

partially automated

High value
multifunctional content

Medium value
reusable content

Low value
specific content

Unstructured content

User: human

manually controlled

Processes

High value
intelligent content

Low value
specific content

Medium value
reusable content

High value
multifunctional content

How GAB are facing the challenge
From structured content to intelligent content
How GAB are facing the challenge
Documents versus data

• Currently GAB publishes documents created from XML
  – HTML
  – eBook
  – print
• We structure our content as documents, as separate files, with a
  sequential order of information, in display order
• We are moving towards data
  – Data that can be understood by anyone
  – Data can used in software applications, but not necessarily
    directly published as text
  – Discoverability of our data
• RDF data model captures meaning and relationships
  independently of what is displayed
How GAB are facing the challenge
Adding meaning to our data

Using what we’ve already got!

- Implicit structures (headings, text order, cross-references)
- Book indexes
- Keywords and subject taxonomy categorisation
- Biographical metadata (life dates, occupations, family groups)
- Oxford Index Authorities (bespoke multi-domain ontology)
- Dictionary entries and their metadata

Move towards explicit meaning that can be easily understood
How GAB are facing the challenge
Metadata Repository

- **Aim:** To have an overview of all GAB’s content
  - Uses metadata, since content in multiple silos
  - Metadata: data about data for each chapter/article
  - One common XML schema => OxMetaML
  - Architecture uses Solr-indexed XML file store (c.f. PIM/title by title) plus triple store

- **Using metadata as documents (XML)**
  - Published on the Oxford Index for discoverability

- **Using metadata as data (RDF)**
  - Understanding of its meaning allows link generation
  - E.g. this OSO chapter discusses the person who has this ODNB biography
How do we add meaning to our content?
Content enrichment - “Semantic tagging”

- Uses text mining:
  - Split into words/phrases
  - Tag different parts of speech
  - Coreference (identify terms that refer to the same object)
  - Named entity recognition (find people, organisations, place names etc)
Metadata Repository: Cross-product linking

Dictionary of National Biography

Oxford Music Online

Oxford Reference Online

is primary topic of

is same entity as

is same entity as
Link generation rule:
If A is the author of B and A is the author of C, then B has same author as C.
And finally…

SEO using RDFa (RDF in attributes)

- Embedding RDF metadata in HTML web pages
- Improves click-through rate (30% reported by BestBuy) as search results more eye-catching
- BBC reported 20% increase in search rankings
- Adding RDFa to the Safari platform and Oxford Index