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…

OXFORD UNIVERSITY PRESS

- 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…

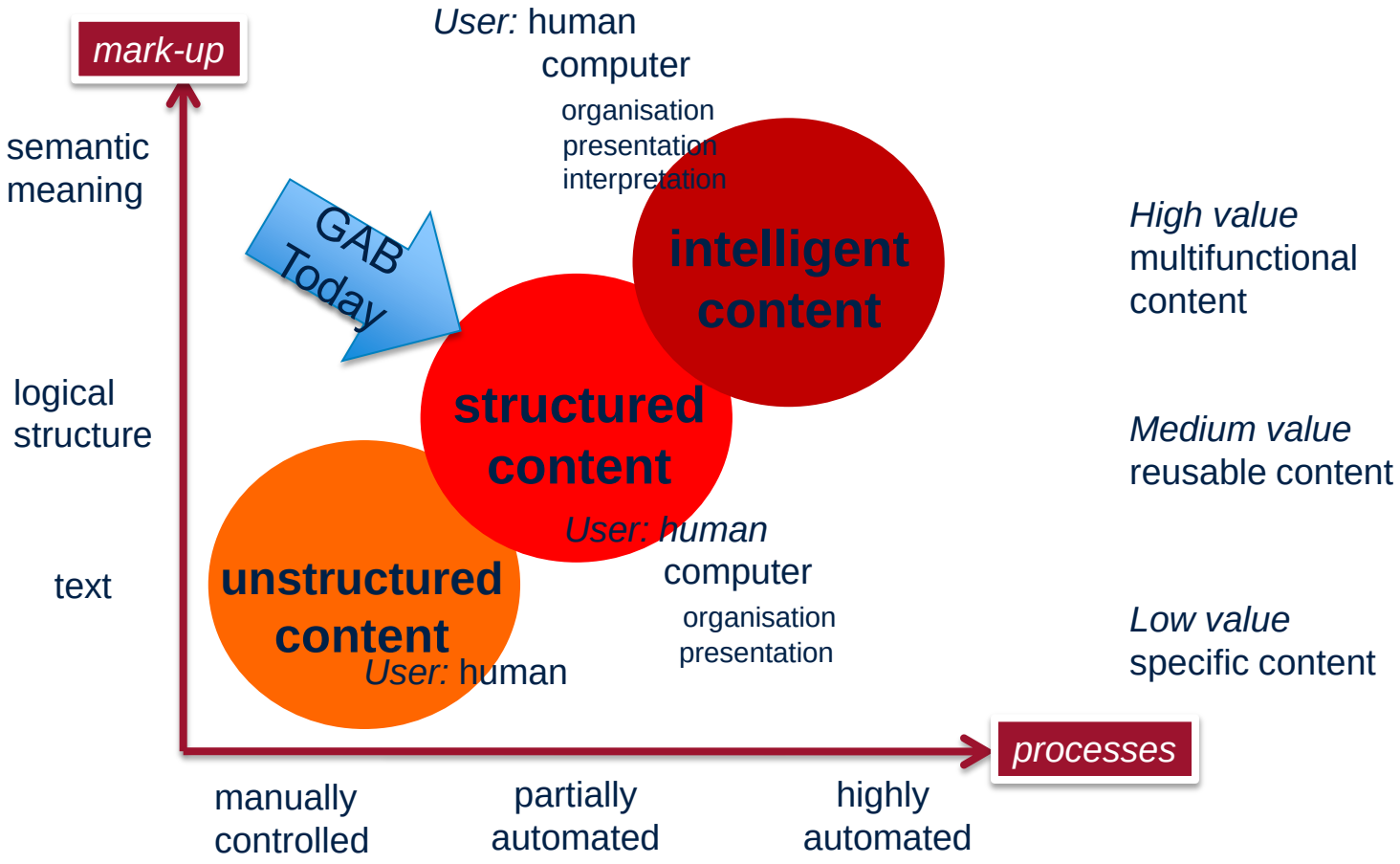
with limited amounts of this:

The challenges we face…

or this:

How GAB are facing the challenge

From structured content to intelligent content



How GAB are facing the challenge

Documents versus data

OXFORD UNIVERSITY PRESS

- 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

OXFORD UNIVERSITY PRESS

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

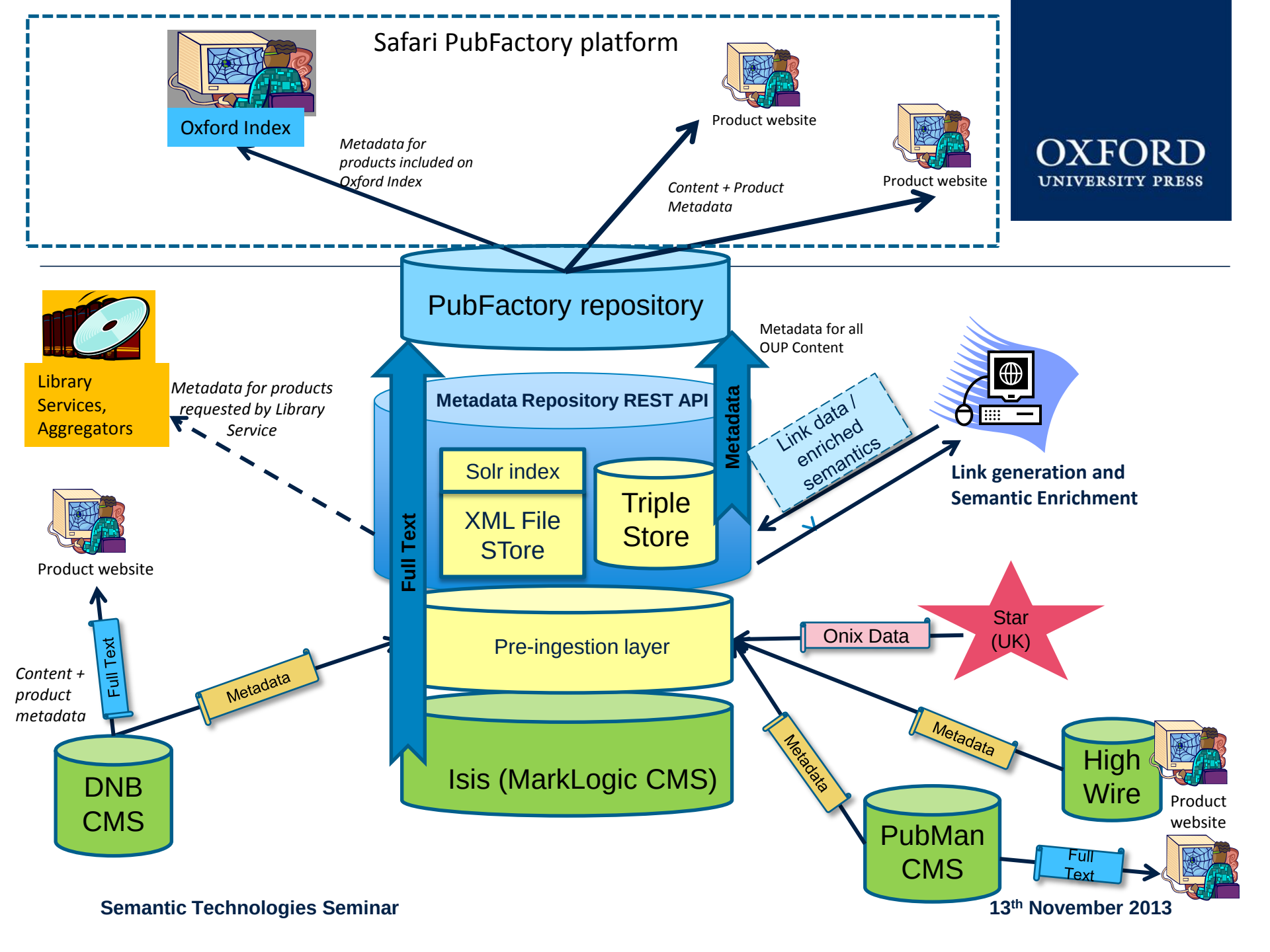
Move towards explicit meaning that can be easily understood

How GAB are facing the challenge

Metadata Repository

OXFORD UNIVERSITY PRESS

- 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

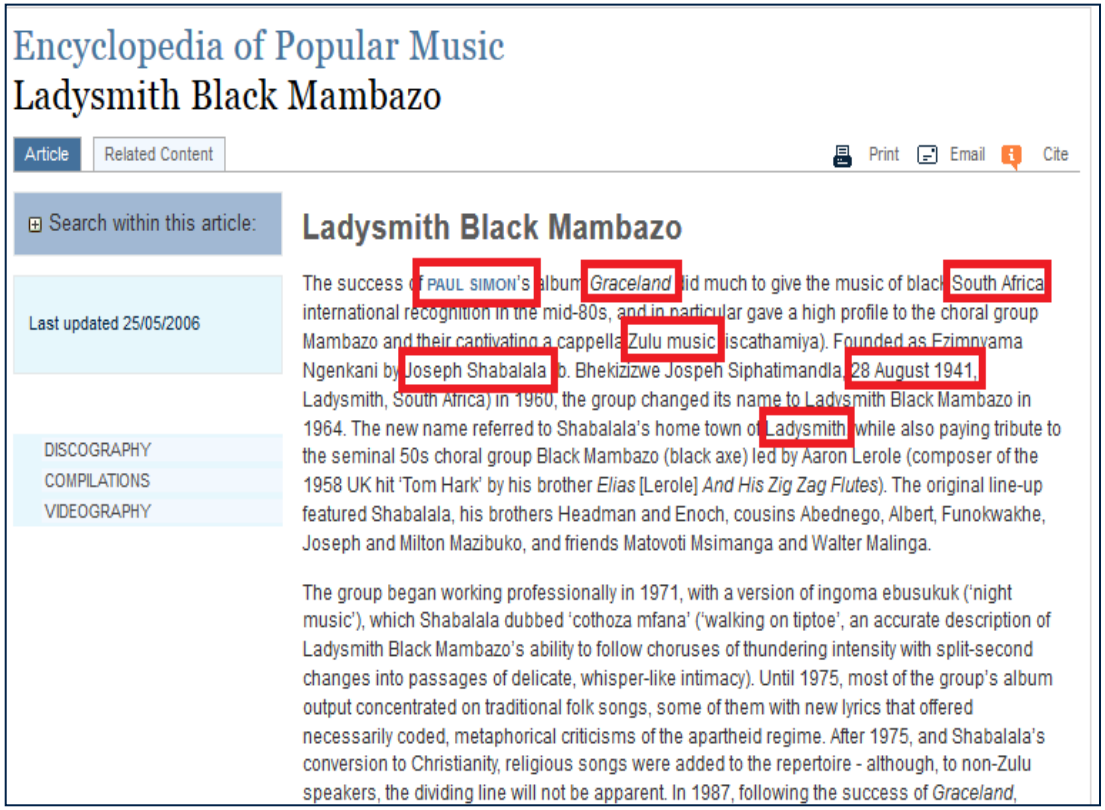

Safari PubFactory platform
Oxford Index
Metadata for products included on Oxford Index
Product website
Content + Product Metadata
Product website
OXFORD UNIVERSITY PRESS
PubFactory repository
Metadata for all OUP Content
Library Services, Aggregators
Metadata for products requested by Library Service
Metadata Repository REST API
Solr index
XML File STore
Triple Store
Full Text
Metadata
Link data / enriched semantics
Link generation and Semantic Enrichment
Product website
Content + product metadata
Full Text
Metadata
Pre-ingestion layer
Onix Data
Star (UK)
DNB CMS
Isis (MarkLogic CMS)
Metadata
Metadata
High Wire
Product website
PubMan CMS
Full Text

How do we add meaning to our content?

Content enrichment - “Semantic tagging”

OXFORD UNIVERSITY PRESS

- 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)



Encyclopedia of Popular Music

Ladysmith Black Mambazo

Article | Related Content | Print | Email | Cite

Search within this article:

Last updated 25/05/2006

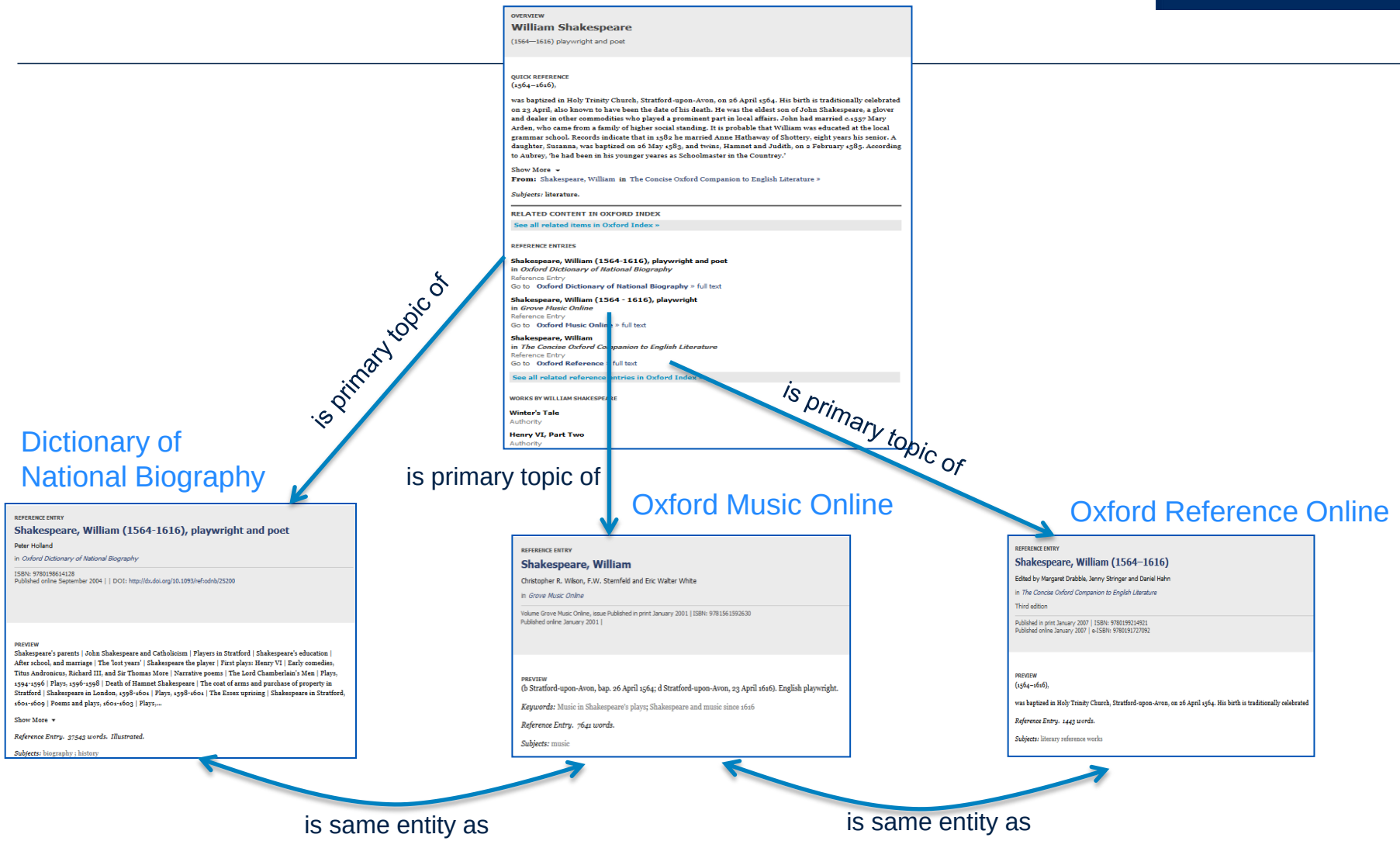
DISCOGRAPHY
COMPILATIONS
VIDEOGRAPHY

Ladysmith Black Mambazo

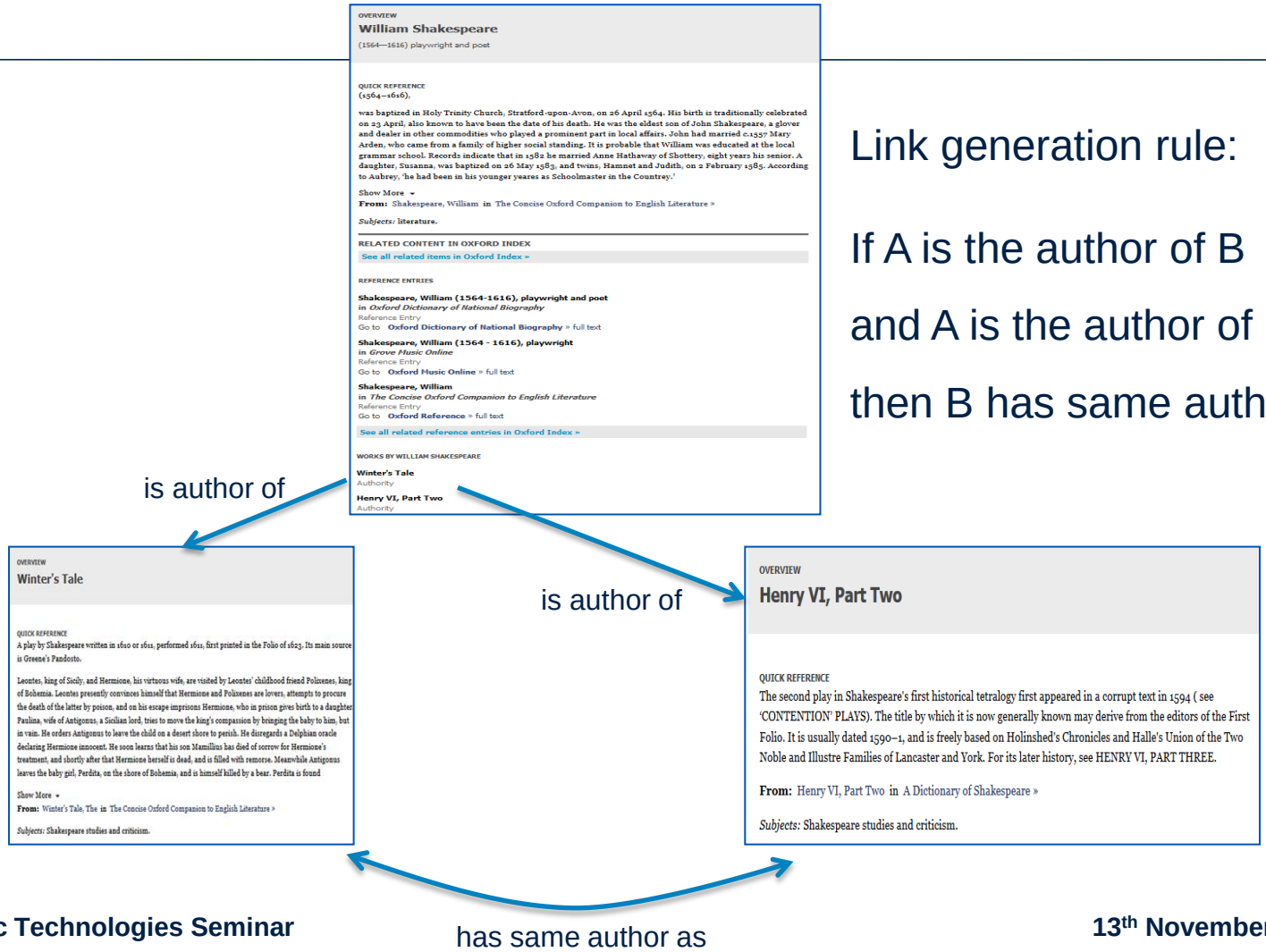
The success of PAUL SIMON's album *Graceland* did much to give the music of black South Africa international recognition in the mid-80s, and in particular gave a high profile to the choral group Mambazo and their captivating a cappella Zulu music (iscathamiya). Founded as Ezimnyama Ngenkani by Joseph Shabalala (b. Bhekizizwe Jospeh Siphatimandla, 28 August 1941, Ladysmith, South Africa) in 1960, the group changed its name to Ladysmith Black Mambazo in 1964. The new name referred to Shabalala's home town of Ladysmith, while also paying tribute to the seminal 50s choral group Black Mambazo (black axe) led by Aaron Lerole (composer of the 1958 UK hit 'Tom Hark' by his brother *Elias* [Lerole] *And His Zig Zag Flutes*). The original line-up featured Shabalala, his brothers Headman and Enoch, cousins Abednego, Albert, Funokwakhe, Joseph and Milton Mazibuko, and friends Matovoti Msimanga and Walter Malinga.

The group began working professionally in 1971, with a version of ingoma ebusukuk ('night music'), which Shabalala dubbed 'cothoza mfana' ('walking on tiptoe', an accurate description of Ladysmith Black Mambazo's ability to follow choruses of thundering intensity with split-second changes into passages of delicate, whisper-like intimacy). Until 1975, most of the group's album output concentrated on traditional folk songs, some of them with new lyrics that offered necessarily coded, metaphorical criticisms of the apartheid regime. After 1975, and Shabalala's conversion to Christianity, religious songs were added to the repertoire - although, to non-Zulu speakers, the dividing line will not be apparent. In 1987, following the success of *Graceland*,

Metadata Repository: Cross-product linking



Metadata Repository: Cross-product linking



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)

OXFORD UNIVERSITY PRESS

- 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