ADA – Automated Data Architecture

Creating user journeys through content using Linked Data

Jo Kent

BBC Radio & Music Multiplatform, UK
jo.kent@bbc.co.uk

Abstract. The BBC has a wealth of permanently available programmes across a wide range of subjects with very low usage. We wanted to create a route into these programmes which balanced the need for curated, high quality journeys between programmes and the limited resource available for that curation effort. I will demonstrate ADA, a system created to create consistent, meaningful high-quality links between programmes with limited user input.

1 Introduction

There is a need for content providers to create consistent, high quality onward journeys to available content. Across the industry solutions used range from the heavily internally manually curated approach of Netflix[1], to the user-driven algorithmically determined approach of Spotify[2].

In this demo I will demonstrate ADA (Automated Data Architecture) which uses minimal manual curation and linked data to provide high quality serendipitous onward journeys.

2 Understanding the problem space

2.1 Assigning metadata

The BBC has at least 34,000 permanently available speech radio programmes to which the traffic is low. There is no easy path into all available content. Some programmes have archive navigation, but these are isolated, specialised and heavily curated, and with decreasing team sizes, even these are not sustainable in the long term.
Our news and sport teams have long used linked data to dynamically populate article pages, which are set up using a strict, pre-existing ontology[3]. This constrains the browser to a rigid structure which may not match their world view. In any case, such an ontology does not exist across all programmes, the subject matter is too diverse.

[1] Netflix manual tagging process: http://www.techradar.com/news/television/netflix-wants-to-pay-you-to-watch-shows-here-s-why-1256098

[2] Spotify's algorithm explained: http://qz.com/571007/the-magic-that-makes-spotifys-discover-weekly-playlists-so-damn-good/

[3] Sport ontology: http://www.bbc.co.uk/ontologies/sport

Crowdsourcing metadata creation has been used by our R&D department on the World Service archive, this achieved at best 30.3% precision (36.7% recall)[4] largely because every person may have a different perception of the subject matter or the meaning of a term[1]. Also it was inconsistent: people only added tags to programmes which interested them, so some programmes have lots of tags and some none at all.

A team of researchers would provide better quality, more consistent data[2], but at a permanently high cost, in staff time. This is unfeasible with fewer staff available.

A fully automated system would not be able to deliver consistent quality standards for an audience facing offer: the automated interpretation of a homograph can give an erroneous or even offensive connection (e.g. Georgia the country as opposed to the American state). Any loss of data quality can cause a loss of trust in our content[3].

A middle ground needs to be found between these levels of automation, without compromising quality. We cannot expect producers to classify consistently, but they do know the precise subject of their programme. Therefore we need a system which only requires them to enter that subject, (e.g. a programme on autism can just be tagged with autism) without the need to classify the concept. Without this classification therefore, we need a system that will automatically supply the links.

2.2 Classification systems

The ideal classification system would need to be recognisable and therefore trusted by our audiences and also be flexible and maintainable over time, as perceptions change[4].

Maintenance of our own ontology requires a significant staff time overhead, but the use of eternally maintained ontologies means we cannot control when changes happen, and still have to adapt when they do. Given the diversity of subject matter (subjects include the A470 (a road in Wales), Munch's "The Scream", virtue, Canada geese, existentialism and the Battle of Bosworth Field) the task of creating an ontology to cover and group every possible subject would be unfeasibly large. To make it manageable, we would have to make arbitrary choices about classification to make the multidimensional world fit in a two dimensional hierarchical structure. This is increasingly viewed as an outmoded and dictatorial organisational method, compared to open ontologies and collaborative folksonomies[5], and any arbitrary divisions of data are no longer semantic distinctions but simply an organisational tool.

3 Unlocking the power of linked data to provide automated onward journeys

The most promising linked open data sources were the Wikipedia/Dbpedia and Wikidata datasets. We found that there was no consistent hierarchical navigation or grouping information applied to the datasets. Wikidata has classification such as Li-

[4] http://www.bbc.co.uk/rd/blog/2014/08/data-generated-by-the-world-service-archive-experiment-draft

brary of Congress and Dewey Decimal mappings, but these are inconsistently applied[5]. It also offers classes and subclasses[6], again inconsistently applied, which often simply cut off without reaching the top class of 'Thing'. Dbpedia has classes[7], but again these are only applied to a fraction of instances[8]. Dbpedia has categories for every subject, which offer a skos:broader[9] journey to other categories, however, due to the way it is structured, often the category that was two hops broader was the initial category we started with, which meant that we had simply introduced more categories without any additional clarity.[10]

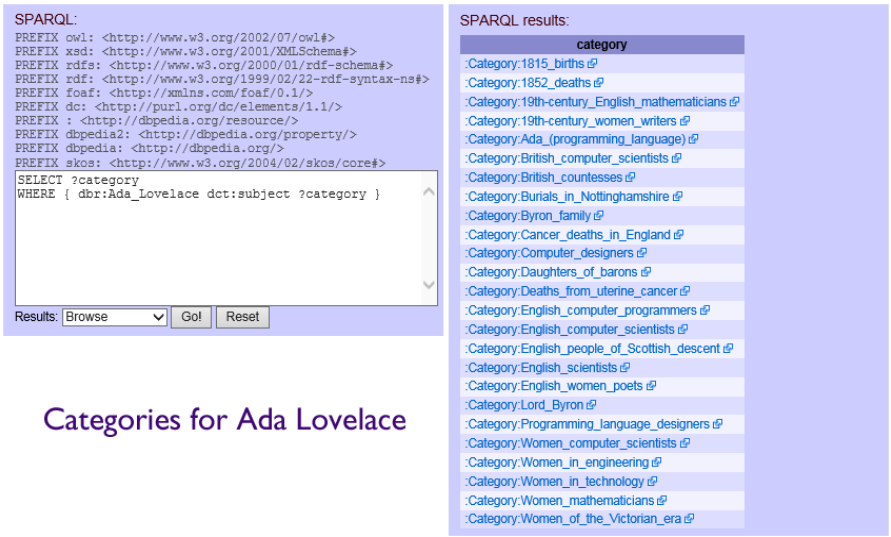

Figure 1 - Categories in dbpedia

Having found no usable hierarchical or grouping information we looked again at categories in Wikipedia/dbpedia. These have been added by Wikipedia editors, each adding the facts they felt were most salient. Anyone can remove them if they disagree, so they are effectively crowdsourced and peer reviewed. This means they have the recognisable relevance that people will respond to, while being of a high quality.

Asking producers to simply identify the subject for their programme means we can assure the quality of the initial reference, and automatically link to all of the categories (an average of seven per subject), which are matched to others to create user journeys that we could not create using manual curation without hours of research. At best a curatorial team might have added tags to Ada Lovelace like 'computer scientist' or 'mathematician' but here we have links to such diverse groups as 'programming language designers' and 'British countesses'. These small, precise categories

---

5 Fewer than 3,500 Wikidata entities have Dewey Classifications attached

6 https://www.wikidata.org/wiki/Property:P279

7 http://mappings.dbpedia.org/server/ontology/classes/

8 Only 182 of 720 (25%) sampled had types applied

9 http://www.w3.org/2009/08/skos-reference/skos.html#broader

10 Using the 3.9 dataset, this issue seems to have been improved in the 1/4/2016 release

give a serendipitous feel to the journey and allow the users to learn more about the subjects even as they are navigating between programmes.

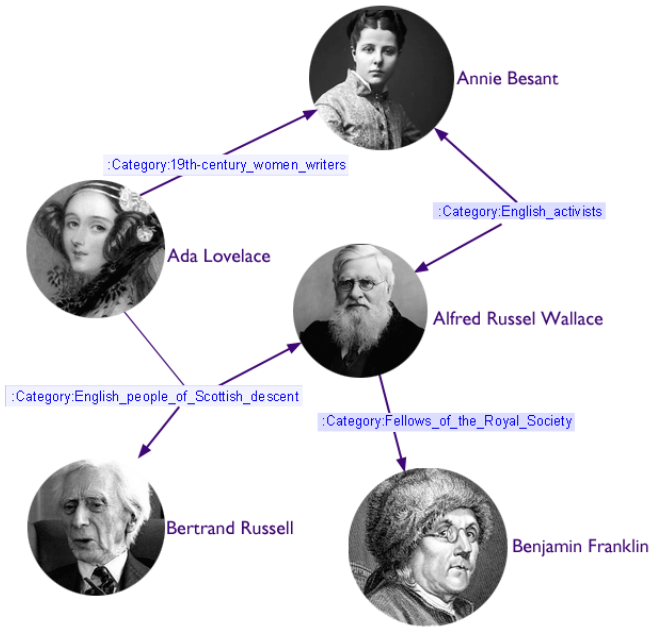

Figure 2 - Category links between people

By discarding the notion of a hierarchy and instead presenting a graph, the journeys are not constrained in to a single worldview. We know from Lobel and Sadler's work on homophily (i.e., love of the same) that "In a relatively sparse network, diverse preferences present a clear barrier to information transmission. In contrast, in a dense network, preference diversity is beneficial."[6] So we can see that people respond better to a wide range of links that may not match their world view than to a narrow one. Therefore providing a broad range of linking categories (which have been selected by their peers as relevant) to each subject will present links the user will instinctively have a positive response to.

4 Evaluation

Beginning with our initial sample of 610 programmes, we extracted over 1000 categories, of which 554 were linked to more than one programme, some of them to as many as 12. We only use categories which link to two programmes or more because only those offer an onward journey. We keep the non-matching categories in the ADA triple store so that they can be used as soon as a new programme with a matching category is added. Some maintenance categories such as "World Digital Library

related"[11] or "Articles with inconsistent citation formats"[12] were added to a blacklist as these are not useful user journeys. We were then able to examine the quality of the journeys offered. A programme on Roman Satire yielded links to 14 other programmes through five different categories; a greater and more detailed level of linking than in our bespoke archives.

We launched a beta[13] to gauge the audience reaction to the new navigation, and the response has been overwhelmingly positive with a rating of 4.15 (out of 5) stars on BBC Taster[14] and 164 (out of 250) positive verbatim responses through the demo feedback link. Once we rolled out ADA to all of our programmes, we plan to roll this out to other departments in the BBC to bring in news articles and educational literature and then to partner agencies (particularly cultural heritage organisations and learning institutions) to create learning journeys across all of our content by subject, rather than content type.

5 The demo

In the demo I'll be showing the beta and the API calls that power it. Visitors will be able to see and experiment with the semantic onward user journeys made possible by the use of linked data.

References

1. Oluwaseyi Feyisetan1, Markus Luczak-Roesch, Elena Simperl, Ramine Tinati, and Nigel Shadbolt. Towards hybrid NER: a study of content and crowdsourcing-related performance factor, In *ESWC*, 2015
2. Marianne Lykke; Haakon Lund; Mette Skov: Metadata in CHAOS: how researchers tag and annotate radio broadcasts. In *Knowledge Organization – making a difference (ISKO UK biennial conference)*, 2015
3. Frances Johnson, Laura Sbaffi and Jenny Rowley. Assessing trustworthiness in digital information. In *International Data and Information Management Conference (IDIMC)*, 2014.
4. Drahomira Cupar, Diachronic semantics: changes of meaning of words over time and the consequences for keeping classification systems up to date. In *Knowledge Organization – making a difference (ISKO UK biennial conference)*, 2015
5. Joseph Busch, Web-based Content Organization and the Transformation of Traditional Classification Systems. . In *Knowledge Organization – making a difference (ISKO UK biennial conference)*, 2015
6. Ilan Lobel and Evan Sadler. Preferences, Homophily, and Social Learning. In *Operations Research*, 2014

[11] https://en.wikipedia.org/wiki/Category:World_Digital_Library_related

[12] https://en.wikipedia.org/wiki/Category:Articles_with_inconsistent_citation_formats

[13] http://www.bbc.co.uk/inourtimeprototype

[14] Feedback on BBC Taster overall rating 4.15 stars: http://www.bbc.co.uk/taster/projects/in-our-time