# Media Meets Semantic Web
# How the BBC uses DBpedia and Linked Data to Make Connections

Andreas Müller

Technische Universität Berlin, 10623, Germany,
andreas.mueller.4@campus.tu-berlin.de,
WWW home page: http://www.user.tu-berlin.de/hpdesigner_20

**Abstract.** This paper describes, how the BBC managed to better interlink different BBC domains by introducing DBpedia as a common vocabulary for every domain. Given the existing legacy systems, the BBC was already using, it is shown, how the new Semantic Web technology was integrated and used, to interlink documents and providing a better usability and user experience, allowing the user to browse different BBC domains by following a semantic thread and getting cross-domain information.

**Keywords:** linked data, semantic web, bbc, dbpedia

## 1 Introduction

The British Broadcasting Corporation (BBC) is the oldest and still one of the largest Broadcasting Companies in the world. Given the fact, that the BBC is producing online content since 1994 [2], they have a huge amount of online media content today in text, audio and video format. To make the data accessible, it was categorized and organized in different domains, i.e. news, sport, weather etc. Figure 1 and 2 present these domains the way they were displayed on the BBC website in the years 2009 and 2017. Each domain became a separate microsite with its own content, vocabulary and datasets.

This separation of content created a clear structure where the user instantly knew where to navigate to when he searched for content of a particular domain. But it also came with a big disadvantage: It was neither possible to find everything, the BBC has published to a given subject nor to navigate between different BBC domains following a semantic thread (i.e. on a page about a musician was no possibility to see all programmes that played this artist). The reason for this lack of interoperability was the missing interlinking between the different microsites. Without this interlinking, the real potential of the available data was not used.

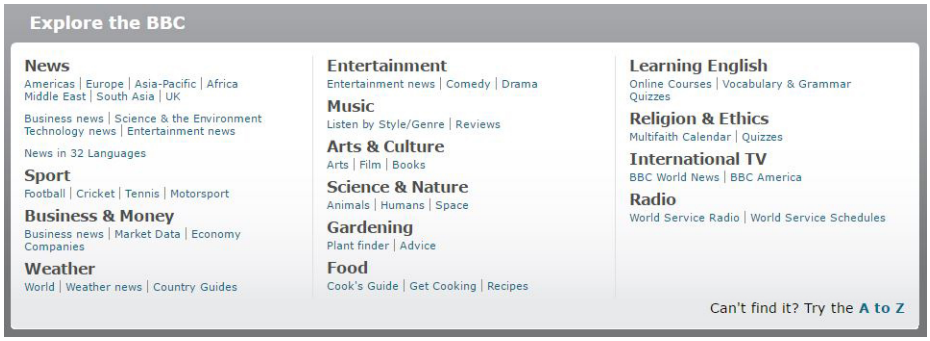


**Fig. 1.** BBC microsites in June, 2009

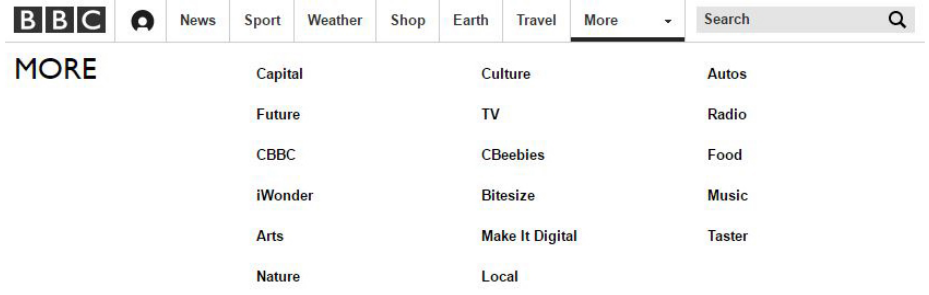


**Fig. 2.** BBC microsites in June, 2017

To make the BBC website more coherent and more useful, G.Kobilarov et al. proposed a solution [1] with the following objectives:

1. Build better connections and interlinking of existing systems
2. Reducing impact on existing systems while adding new services to maximize interlinking of domains

The next section will give a short overview of the existing legacy systems of the BBC website.

## 2 Background: Legacy Systems

One of the systems, the BBC was already using to automate content relations, was a legacy auto-categorization system called *CIS*. With this system it was possible, to categorize programmes by their textual description, that contained meta data like brands, locations, people and subjects. Despite being used for the programmes domain only, it was not possible to cover every single entity that might have been of interest with CIS. This system also held no information about

relations between different terms; i.e. the terms "Bejing" and "Bejing Olympics" obviously are related by location, but in the CIS system these were just two independet terms. Additionally the system only relied on internal identifiers, so linking to non-BBC data was not possible either.

G.Kobilarov et al. pointed out, that CIS could only be used to interlink between different domains if there were mappings between the various vocabularies of each domain. In this case, it would be possible to further developing them independently [1].

## 3 Solution: Integration of DBpedia

To solve the problem of the need for a common set of web identifiers for all domains, a well-known, crowd-sourced extraction framework was chosen: *DBpedia*[1].

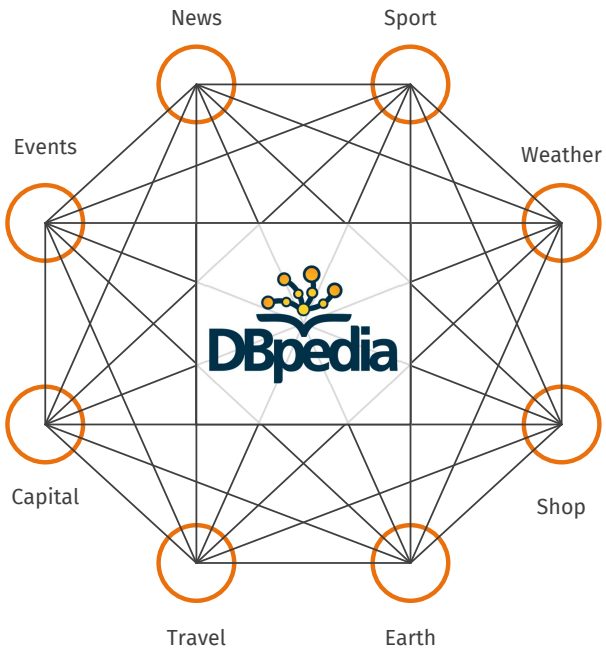


**Fig. 3.** BBC microsites connected by DBpedia

[1] The first public DBpedia datasets were released 2007. Its purpose is to extract structured content from the data of Wikipedia using Semantic Web and Linked Data technologies [4].

To interlink all domains (as to be seen in figure 3), DBpedia serves as a common vocabulary. This works in two steps. At first, a DBpedia label loopup is performed for a given CIS term. Therefore the most likely matches to a given term and possible DBpedia resources are found and ranked by their relevance[2]. After that, the best match is found by context-based disambiguation. All relevent matches are disambiguated by clustering them and finding an according context in DBpedia. The term "apple" i.e. is simply a fruit for itself, but in the context of "Microsoft" and "Google" it becomes "Apple Inc.".

With the use of the same vocabulary, the interlinking of text documents (i.e. news or documentation articles) became possible. Therefore the text body of a BBC document URI was parsed to extract the main entities using *Named Entity Recognition* (NER)[3]. These main entities then are matched by an algorithm to possible DBpedia resources and ranked by contextual disambiguation, like described in the previous paragraph. This creates a mapping of extracted terms to possible counterparts in DBpedia. Finally the DBpedia resources are filtered in a way that only resources that correspond to "people" or "companies" are kept. This is achieved by the already existing predicates of the terms provided by CIS. This process of interlinking documents is realized by a system called *Muddy Boots* and is presented in figure 4.

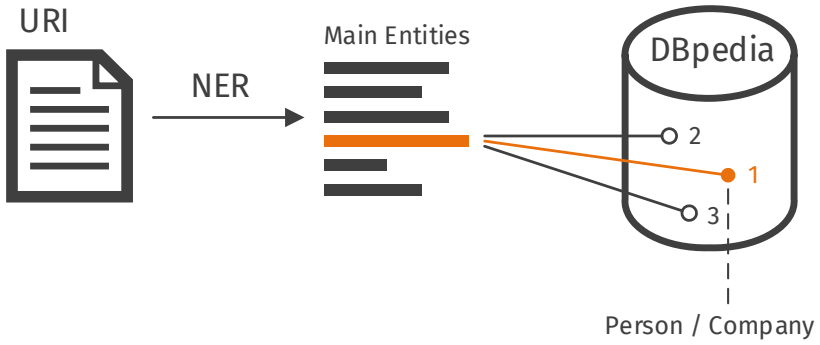


**Fig. 4.** Muddy Boots: Identify main actors in a piece of text content

## 4 Evaluation

According to G.Kobilarov et al. [1] ...

[2] The relevance here is calculated by the number of backlinks.
[3] TODO

## 5 Related Work

...

## 6 Future Work & conclusions

...

## References

1. G.Kobilarov, T.Scott,Y .Raimond, S.Oliver, C.Sizemore, M.Smethurst, C.Bizer and R.Lee.: *Media meets semantic web - How the bbc uses dbpedia and linked data to make connections.* Lecture Notes in Computer Science (including subseries Lecture Notes in Articial Intelligence and Lecture Notes in Bioinformatics) 5554LNCS:723737, 2009.
2. The Sunday Times: *THE BBC is launching an on-line service.* 17 April 1994. Quoted in Connor, Alan (25 December 2007). "The WWW Info-Rainforest". *BBC Internet Blog.* BBC.
3. Yves Raimond, Tristan Ferne, Michael Smethurst, Gareth Adams: *The BBC World Service Archive prototype.* Web Semantics: Science, Services and Agents on the World Wide Web Semantic Web Challenge 2013:2-9, 2014
4. J.Lehmann, R.Isele, M.Jakob, A.Jentzsch, D.Kontokostas, P.N.Mendes, S.Hellmann, M.Morsey, P.van Kleef, S.Auer, C.Bizer: *DBpedia A large-scale, multilingual knowledge base extracted from Wikipedia.* 10.3233/SW-140134 Semantic Web, vol. 6, no. 2, pp. 167-195, 2015