Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections

Georgi Kobilarov[2], Tom Scott[1], Yves Raimond[1], Silver Oliver[1], Chris Sizemore[1], Michael Smethurst[1], Christian Bizer[2], and Robert Lee[3]

[1] British Broadcasting Corporation, London, UK
firstname.lastname@bbc.co.uk
[2] Freie Universität Berlin, Berlin, Germany
firstname.lastname@fu-berlin.de
[3] Rattle Research, Sheffield, UK
robl@rattlecentral.com

Abstract. In this paper, we describe how the BBC is working to integrate data and linking documents across BBC domains by using Semantic Web technology, in particular Linked Data, MusicBrainz and DBpedia. We cover the work of BBC Programmes and BBC Music building Linked Data sites for all music and programmes related brands, and we describe existing projects, ongoing development, and further research we are doing in a joint collaboration between the BBC, Freie Universität Berlin and Rattle Research in order to use DBpedia as the controlled vocabulary and semantic backbone for the whole BBC.

1 Introduction

The Linking Open Data project[1] became one of the main showcases for successful community-driven adoption of Semantic Web technologies in the last year. It aims at developing best practices to opening up the "data gardens" on the Web, interlinking open data sets on the Web and enabling web developers to make use of that rich source of information. But the data made available in that process, the practices and technologies developed, are not only useful for open web data, they also provide benefits to end users and the enterprises at large.

In this paper, we describe how Linked Data technologies [1][2] were applied within the BBC, one of the world's largest broadcasters, and how DBpedia [3], often considered as the linking hub of the Linking Open Data project, and MusicBrainz are used in that process as both an interlinking vocabulary and a data provider. The paper is structured as follows: Frst, we describe the former status quo of multiple (largely unconnected) data sources and categorization systems within the BBC and describe the problems and the requirements for a transition from those unconnected data sources to a more interlinked ecosystems of data.

[1] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 723–737, 2009.

We then describe how we publish programmes data, how we link concepts with DBpedia and through that documents to each other and how that manifests in actual benefits for the users of the BBC main website at www.bbc.co.uk.

1.1 Problem Definition

The BBC publishes large amounts of content online – both text and audio & video. Historically this has focused largely on supporting broadcast brands and a series of domain specific microsites (e.g. food, gardening etc.) to the exclusion of wider integration with the rest of bbc.co.uk, let alone the rest of the web; albeit with some notable exceptions such as *news online*. That is, the focus has been on providing separate, largely stand alone, microsites designed to be accessed via HTML on the desktop.

This means that the BBC has been able to support the primary use cases within the context of the individual microsite but not those use cases that span programme brands or domains. For example, we can tell you who presents Top Gear, but not what else those people have presented. By developing self-contained microsites the BBC has produced some very popular services but it has also failed to reach its full potential because it hasn't been able to join up all of its resources. By failing to link up the content (on both a data and a user experience level) the content published has never been able to be greater than the sum of its parts. As a user it is very difficult to find everything the BBC has published about any given subject, nor can you easily navigate across BBC domains following a particular semantic thread. For example, you can't yet navigate from a page about a musician to a page with all the programmes that have played that artist.

Furthermore, while BBC Backstage[2] has made great progress in making BBC data available for third party developers we realized that there was much more that could be achieved if machine representations were made available in the same way as the HTML views.

As with many large publishers of Web content, the BBC has divided its different services by domain, e.g. food, music, news etc. Each service is maintained by a different team; as have the programme support sites, which have historically been commissioned independently of each other. This has made it difficult to coordinate interlinking between services and programme sites.

The development of bbc.co.uk/music/[beta] sought to address the interlinking of services around the music domain - music played on programmes, at events or in sessions. Likewise bbc.co.uk/programmes aims to provide a central programme support service which the rest of the business can use.

However, neither music nor programmes solved the problem of cross domain linking, nor do they address the issue of disambiguation between multiple controlled vocabularies - the fact that Madonna is both an artist (in MusicBrainz), and an actor and a person (in Wikipedia). And possibly more significantly nor does it address the other domains the BBC has an interest in - news, food, books,

[2] http://backstage.bbc.co.uk

sport, natural history etc. To address these issues the BBC, Freie Universität Berlin and Rattle Research are investigating using DBpedia to provide a common "controlled" vocabulary and equivalency service, which in turn is used to add "topic badges" to existing, legacy web pages.

1.2 Objectives

As we've seen in the previous section, there are a large number of data sources within the BBC that are either unconnected or only cross-connected for specific use cases. This scenario is not an uncommon one in large enterprises. While - in theory - when designing a data infrastructure it should be possible to design a system that connects and interlinks all domains – the reality however is almost always different. Organic growth of organizations and supporting systems leads to a diverse ecosystem and trying to simply start again from scratch is clearly an untenable option.

Enterprises the size of the BBC have invested significant resources, over a number of years, into their existing web offerings and redesigning and redeploying all of them at once is impossible not only from a deployment point of view, but also from an organizational one. Therefore an approach needs to be taken that results in better connections and interlinking of existing systems – providing a soft transition and reducing the impact on existing systems – while at the same time, where possible, developing new services to maximize the interlinking of domains. The Semantic Web, especially Linked Data technologies, can offer such a soft transition.

Our objectives then are fourfold:

1. to develop a new service that supports the branding of our Radio stations, TV channels and programmes (bbc.co.uk/programmes) while at the same time ensuring that users and third party developers were able to traverse the graph of BBC data though to other data sources elsewhere in the Linked Data cloud – following contextually related links across all BBC content or content elsewhere on the web.
2. to develop a new music offering (bbc.co.uk/music/beta) that builds on existing open web standards and is fully integrated with the emerging programme support service.
3. to retrofit simple navigational elements (i.e. Topic Badges and term extraction) to existing, legacy pages, and new pages built with legacy systems to support contextual, semantic navigation.
4. to provide a common set of web scale identifiers to help classify all BBC online content (and external URLs) and to help create equivalency between multiple vocabularies.

In meeting these objectives it is our hope that bbc.co.uk becomes a more coherent, useful place by providing contextual, semantic links connecting content across different domains and providing meaningful navigation paths for our users.

So the main part of this paper will first describe how BBC Programmes and Music were developed in order to publish Linked Data, both for use internal and

for external developers. We will then describe legacy systems, in particular a content categorization system called CIS and how we interlinked concepts from CIS with DBpedia as the controlled vocabulary for integrating the different BBC domains. Then we turn to documents, which still are the main part of the content published by the BBC, and how we interlink those through the work of using DBpedia as vocabulary for concepts. And finally we will present how our users benefit from better, more coherent, user journeys across www.bbc.co.uk.

2 Publishing and Linking BBC Programmes and Music

The previous piecemeal approach to programme support afforded an opportunity to develop a new service bbc.co.uk/programmes (which launched in Summer 2007) with the objective of providing one URI per programme - for every programme the BBC broadcasts - allowing other teams within the BBC to incorporate those pages into new and existing programme support sites, TV Channel and Radio Station sites.

Likewise there was an opportunity to redevelop the music site along the same principles - but in this case a URI for every artist (and eventually a URI for every track) the BBC plays. However, although there are commonalities between programmes and music there are also differences in the development approach that are worth noting.

BBC programmes is underpinned with a proprietary database – PIPs – which seeks to contain the definitive record of all public facing BBC programme metadata. The public representation of this data is, in part, exposed at bbc.co.uk/iplayer and also at bbc.co.uk/programme. The later representation is published in accordance with the programmes ontology[3].

Unlike BBC programmes the BBC does not "own" the music domain and as such it is important that the BBC adopts existing web identifiers to ease the development effort for both BBC software engineers and third party engineers. MusicBrainz, the community maintained music metadata service, provides such web identifiers.

However, artists also "exist" in other domains – for example Madonna is a singer and an actress (hence is also part of a movies domain) and also a person. The BBC therefore needs a mechanism to create equivalence between two or more identifiers from different domains.

Both BBC Music and Programmes provide persistent web identifiers and a set of corresponding representations for BBC programmes and music artists, bringing BBC data to the Semantic Web. These services were designed according to the Linked Data principles coined in [1]:

- Web identifiers are used to denote things: entities within the scope of the Programmes Ontology (brands, series, episodes, versions, services and broadcasts) and of the Music Ontology [4] (music artists);
- These web identifiers have multiple representations, including:

[3] http://purl.org/ontology/po/

- An XHTML representation, designed for human interaction;
- An RDF representation, exposing structured BBC data;
- These representations hold links to further web identifiers, allowing to discover more structured data. For example, the representation of an artist in BBC Music holds an *owl:sameAs* link to the corresponding artist in DBpedia.

We were also convinced that the value in programme websites is not in the implicit metadata of the domain model but rather in the way this domain model overlapped and intersected with other domains. As ever, the links are more important than the nodes because that's where the context lives:

`programmes:segment <features> music:track`,
`programmes:segment <features> food:recipe` etc.

In this way we could weave new "user journeys" into and out of /programmes, into and out of bbc.co.uk. But to achieve that, we needed to bring more of our legacy content and systems into the picture.

3 Cross-Linking Legacy Content and Legacy Systems

We have seen how the BBC programmes domain has successfully modeled programme specific relationships. This, though progress, represents a comparatively small amount of the BBC content and it is desirable to link to further BBC domains. One way of doing this was through an *about*-relationship between programmes and people, places and subjects. This data was created with a legacy auto-categorization system called CIS.

CIS holds a hierarchy of terms in five main top-level classes: proper names, subjects, brands, time periods and places. The system was originally used for annotating regional news stories in a content management system. This meant that the vocabulary was primarily focused around British regional content. This included many local bands, people and events. In addition to this at its core is a more useful general subject vocabulary and list of locations that would be of use to the annotation of BBC programming. Based on this more useful parts of its vocabulary (around about 40% of the total concepts) and the expertise already existing in the BBC it was chosen as an initial solution for annotating BBC programmes.

CIS was put in place as an automated system to categorize programmes based on their textual description. For example, the programme synopsis "a look ahead to the Beijing Olympics, including a preview of British boxing hopes" would be categorized with "Beijing" (place), "British" and "Boxing" (subjects). The term "Beijing" can then become the link between this programme and other programmes about Bejing. Addtionally this could be used to link to BBC news stories (if news use similar identifiers) related to Beijing. This approach could help the interlinking of different BBC services, whilst keeping developing them independently. The key to this is ensuring there are mappings between the various vocabularies in use.

However, even though this approach highlights the need for a common vocabulary, shared across the different BBC services and acting as a set of links

between them, a single categorization system is difficult to maintain, and it is difficult to cover every single entity that might be of interest. Moreover, no relations relating terms together are available within CIS. For example, it would be impossible using a CIS-based framework to access the relationship between Beijing and the Beijing Olympics. In order to provide a satisfying user experience, we need richer relationships between these different terms. Furthermore CIS terms will only ever be an internal identifier and so will never help automating the linking of BBC resources to non-BBC data.

In addition, the objects identified with /programmes and /music are also to be found within other domains. Ideally we need a mechanism by which we map between equivalent terms.

So we decided to look for a common set of web identifiers for the BBC. The DBpedia project[4], which is bringing information extracted from Wikipedia to the Semantic Web, had at that time already developed into the de-facto interlinking hub for the Linking Open Data project, so it was the obvious choice to rely on DBpedia identifiers and this way at the same time joining the Linked Data Web. And while DBpedia doesn't only provide Linked Data URIs for a broad range of concepts, but also structured data about those concepts and their relationships, that data can be used to power (semi-)automatic interlinking algorithms, serve as a source of data to be used and displayed in BBC applications, and provide the rich relationships between terms we need. So DBpedia becomes the controlled vocabulary to connect our BBC domains such as Music, News, Topics and Programmes as shown in Figure 1. Although considering Wikipedia as a "controlled" vocabulary is arguable due to changes being made to Wikipedia article URIs, Hepp et al. [5] note that these changes occur less often than expected, and DBpedia can provide mechanisms to capture those changes through redirect resolution.

BBC domains contain both concepts and structured data as well as documents and other content. In order to achieve our cross-domain linking of concepts and content, we will in the next section describe our attempts to first interlink concepts and then content about concepts with DBpedia.

4 Linking CIS Concepts to DBpedia

In order to quickly bridge our existing CIS content with the DBpedia-centric ecosystem, we developed a system to automatically interlink CIS concepts with DBpedia. In the following sections, we will describe our approach, give an overview of the algorithm we have implemented and provide a preliminary evaluation of the interlinking results.

4.1 Interlinking Approach

CIS contains 150,000 terms overall, covering four different domains: BBC brands, locations, people and general subjects. Each domain has its one term hierarchy categorizing the terms expressed in SKOS [9].

[4] http://dbpedia.org

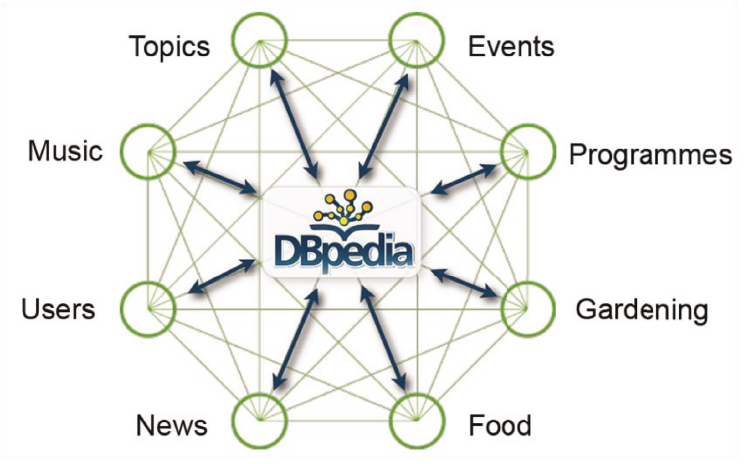

Fig. 1. Linking BBC Domains

For the interlinking only the CIS taxonomy itself could be used, for various reasons there was no access to documents tagged with CIS terms. So the interlinking algorithm had to work only with the information we could get from CIS, including labels and categorization information. The main challenge was to find the most probable matches based on label lookup of CIS terms in DBpedia and to disambiguate those matches using classification information.

This problem is a well known one in the research area of ontology matching [6][7]. However, we needed to developed a solution that is tailored to our very specific use case, taking into account the very limited information on CIS terms that were available for the matching implementation, as well as some of the specific characteristics of DBpedia.

The core idea of our approach can be described as similarity clustering based on the "context" of concepts. While the term "apple" is in itself ambiguous, given the context of the terms "microsoft" and "google", the meaning of "apple" referring to Apple Inc. becomes clear. This assumption forms the basis for our CIS interlinking. We use the available classification information in CIS to build similarity clusters (concepts in the same category or parenthesis-based class) that help us disambiguate the meaning of a given term. In DBpedia, we can then calculate a similarity metric for the tuples of meanings based on the node distances in the DBpedia categorization and classification graph.

So the algorithm we developed is divided into two parts: DBpedia label lookup and context-based result disambiguation. The overall automated interlinking approach is therefore similar to the one described in [8], but differs in the way the context identification is performed and the results are ranked.

4.2 DBpedia Label Lookup

The core of the linking process is a weighted label lookup [10]. DBpedia contains 2.5 million concepts, and most input strings from CIS match to several concepts:

"shakespeare" matches to over 50 DBpedia resources. In order to find the most likely matches for a given term, the system uses a weighted label lookup, using Wikipedia inter-article-links as weight indicator. PageRank works based on the assumption that the more important a web page is, the more hyperlinks on the web will point to that page. We found out that in Wikipedia the number of inter-article-links pointing to an article can be similarly treated as indicator for the articles' overall relevance. The Wikipedia article about William Shakespeare has over 6000 links pointing to it, while the article about Nicholas Shakespeare only has 18. In order to handle synonyms and abbreviations (such as "EU" referring to the European Union) we included Wikipedia redirects into the index, calculating their weight as `links(redirect)*log2(weight(article))`.

DBpedia Lookup builds on a custom Lucene index which combines our relevance metric with Lucene's string-similarity-based ranking, and is available as web-service[5].

4.3 Context-Based Disambiguation

In order to disambiguate the possible matches, we identify similarity contexts of CIS terms by clustering matches and finding according contexts in DBpedia. We used the CIS categorization hierarchy and parenthesis texts as clusters. The concept "Mary (1985 sitcom)" falls into the clusters "television" (the category), "1985" and "sitcom" (parenthesis texts). The algorithm creates these clusters for all CIS terms and identify matching DBpedia categories, classes and templates for each cluster based on the multiple possible DBpedia matches.

Those identified contexts are then used to disambiguate matches for every CIS concept. In addition, we added restrictions based on the CIS domain: stemming is used for subjects, not for people's names, locations must use a geo-coordinate template, and for brands the most relevant templates in DBpedia were identified manually.

With the previous example of "Mary (1985 sitcom)" falling into the clusters together with other sitcoms and television shows, we were able to reject the top ranked label-based result "Mary (Holy Mother)" for the search term "Mary" due to its DBpedia class, reject "Something about Mary" and "Mary Tylor Moore Show" and accept the match `dbpedia:Mary_(1985_TV_series)` based on the DBpedia category "1980s American television series" (see figures 2 and 3).

4.4 Evaluation

The goal of this first automated interlinking approach was to only create links with high confidence values (i.e. reducing the number of false positives or false links) while accepting that we will have missed many potentially correct links (i.e. having many false negatives). Table 1 shows our interlinking results.

One reason why we have only interlinked 20% - 30% of the brands, locations and names datasets is that many concepts don't have their own Wikipedia article. TV Episodes for example are often merged into one single list-article, which

[5] http://lookup.dbpedia.org/api/search.asmx

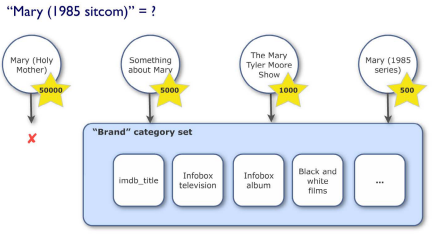

Fig. 2. Context Identification

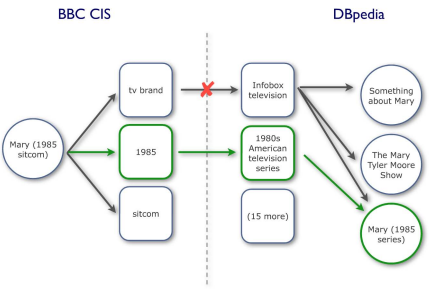

Fig. 3. Context-based Disambiguation

Table 1. CIS / DBpedia Interlinking Results

	Total	Linked	Precision (test set)	Recall (test set)
Brand	6,630	1,267 (19%)	86%	41%
Location	55,943	11,316 (20%)	99%	77%
Name	73,442	22,341 (30%)	92%	67%
Subject	11,231	6,822 (61%)	92%	75%

would be inappropriate to use as URI for every episode. And many people don't exist in Wikipedia due to low notability. We manually created a test set of 540 CIS-DBpedia links for important concepts of each domain (100 brands, 150 locations, 130 names, 160 subjects), in order to check the precision and recall of our algorithm. While we have included CIS concepts with no according DBpedia concept into the test sets to test for false positives, their percentage in the test sets does not represent their real distribution in the main data sets. We want to also note that this work has been done before the release of DBpedia 3.2, which provides a much cleaner dataset and a hierarchical ontology. We are convinced that we would now be able to achieve much less false negative results, but haven't evaluated that yet.

5 Linking Documents to Concepts

In the previous sections we focused on how we have published and interlinked existing structured data and concepts. Now we turn to documents. The BBC has got a lot of textual documents, such as editorial web sites and news articles. Interlinking the CIS vocabulary with DBpedia brings a part of the available documents into our linked data ecosystem, but there are much more. In this section, we describe how we use a named entity extraction system called Muddy Boots to process BBC News articles, recognize entities in those articles, and use DBpedia identifier for those entities in order to bring BBC News into the picture. Muddy Boots also provides us with the text processing capabilities we need for our semi-automated Content Link Tool, described later.

5.1 Named Entity Recognition: Muddy Boots

During the past year, Named Entity Recognition (NER) became one of the main focus of so called Semantic Web startups and companies. Products such as OpenCalais[6], Twine[7], and Zemanta[8] built on NER in order to extract concepts (mostly people, locations and companies) from textual input. OpenCalais and Zemanta provide APIs to enable bloggers and businesses to use their NER systems. But those systems did use their own entity identifiers, so we looked for a system that actually reuses existing web identifiers, i.e. Wikipedia/DBpedia URIs. (OpenCalais and Zemanta recently announced to link their entity IDs to DBpedia URIs.)

The BBC worked with Rattle Research to create a system that used DBpedia URIs as a controlled vocabulary to identify entities [11] [12] in their news archive, called Muddy Boots. The Muddy Boots system parses the story body from a given BBC news URI and then uses a NER system in combination with the Yahoo Term Extraction API[9] to extract the main entities from the story. These entities are just text and have no semantic meaning or classification attached to them at this point. They are matched to possible resources in DBpedia using a simple fuzzy logic algorithm that compares DBpedia resource titles to the extracted entities. In the event of multiple DBpedia resources being identified, the system allocates each of these to the extracted term as a possible valid match.

The system follows any DBpedia `redirects` and `disambiguates` predicate and compiles a list of possible DBpedia resources for each term and ranks each identified DBpedia resource, using a contextual disambiguation. For a story about 'Apple Inc', there will be terms such as 'Steve Jobs', 'Iphone' and 'Macbook'. By using the complete list of extracted terms from the story and applying a similarity algorithm across the resources identified to rank them in terms of their similarity to a document consisting of just the extracted terms, the system uses the highest ranking resource as the disambiguation for the extracted term.

This allows Muddy Boots to create a mapping between a list of extracted terms to their probable counterparts in DBpedia. The final step is to identify the resources that correspond to 'people' and 'companies', by examining the predicates for each resource and creating a scoring algorithm that assigns points to a categorisation (either 'people' or 'company') based on the predicates that are present, such as 'birthdate' or 'birthplace' for a person. The categorisation with the highest score is assigned to the resource and an overall 'confidence' metric is assigned using a weighted algorithm that combines results from the similarity metric, the categorisation score and the ambiguity of any given term.

The Muddy Boots prototype produced a system that uses DBpedia as a controlled vocabulary to unambiguously identify the main actors in a piece of content. The system has been further developed to utilise the new DBpedia ontology and only attempts to categorise resources that do not have a classification

[6] http://www.opencalais.com/
[7] http://www.twine.com
[8] http://www.zemanta.com
[9] http://developer.yahoo.com/search/content/V1/termExtraction.html

within this. This allows the system to identify an increased number of entities in a wider variety of categories. The system currently powers the Content Link Tool resource recommendation system, described in the next section, and is also available as a set of publically available web services (http://www.muddy.it).

5.2 Content Link Tool

No matter how good auto-categorisation methods are there will always be the need for editorial intervention. This meant the BBC would need a tool with which to add or remove DBpedia identifiers from any given BBC URL. The DBpedia concepts and associated URIs could be then stored centrally to be made available for any service that required BBC URLs about a DBpedia concept.

Previous experience with semi-automated tagging had taught us that the user interface of any annotation tool was critical to the creation of high quality metadata. The tool would also need to incorporate high quality automated suggestions to make the annotation process as painless as possible. The Muddy Boots system was chosen as the source of automatic suggestions based on the DBpedia dataset. In addition users could search for terms that had not been suggested. The search box incorporates auto-completion of keywords based on DBpedia resource titles and synonyms using the DBpedia Lookup web service for auto-completion, and displayed abstracts make it easy for BBC editors to choose from the suggested concepts. Selected terms are simply added to the list of concepts for that URL. Figure 4 shows a screenshot of the Content Link Tool.

The cultural aspect of editorial annotation was also considered and lessons have been learnt from the past. Any added or removed concepts will be immediately reflected in the related links on the content page. This means there is a user facing editorial aspect to the annotation, concepts end up as links on the page as opposed to hidden in the HTML. Another change has been the move away from the language of tagging to the language of linking, where as editors will use the 'Content Link Tool' as opposed to the 'Content Tag Tool'. These are attempts to ensure content creators take annotation seriously and see it as an integral part of the content creation process.

6 Using Concept Links to Create User Journeys: Topic Pages and Navigation Badges

The lack of linked structured data discussed in the beginning made it difficult to present a coherent BBC website. Scalable cross site navigation on the whole is dependent on the interlinking of well modeled data via agreed identifiers. Obviously this is more difficult with the unstructured content, which, in the case of the BBC, is the majority of it.

One solution to this problem is the creation of aggregation pages of unstructured and structured content. These pages pull together the modeled world of BBC programmes (CIS identifiers mapped to DBpedia) and the unstructured world of BBC News articles. These types of aggregation pages (topic pages)

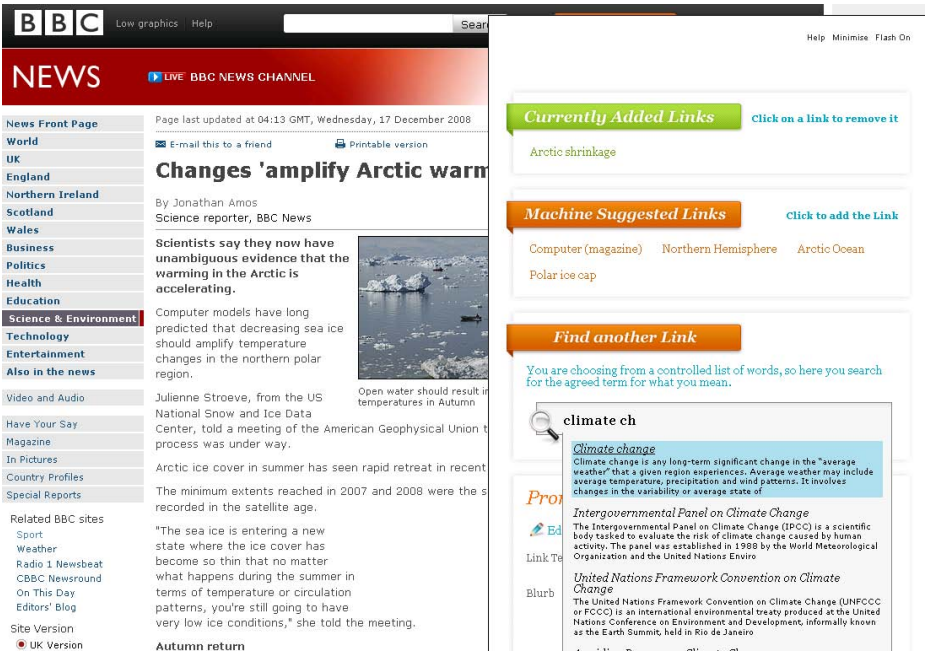

Fig. 4. Tagging and Linking BBC News Articles

have become popular with sites like the New York Times and CNN primarily because of their ability to focus search engines on a given keyword. For the BBC aggregation pages have the additional benefit as navigational nodes to facilitate journeys across unlinked content domains.

When deciding on a vocabulary to drive the BBC aggregation pages a decision was made to use DBpedia. The DBpedia vocabulary offered a number of advantages including:

- Interoperability with other BBC domain specific data such as MusicBrainz.
- It is cheaper to maintain than existing internally maintained vocabularies.
- DBpedia has additional data available, for example: short descriptions, temporal and geo-location data to enrich aggregation pages.
- DBpedia offers rich associative and hierarchical relationships.
- DBpedia's descriptive text can be used as training material for auto-categorisation systems.

Using DBpedia as our vocabulary meant we could join the domain modeled structured data that are using Linking Open Data identifiers (like MusicBrainz) with the unstructured content aggregated using auto-categorisation systems based on the DBpedia dataset (in this case an implementation of Autonomy). In addition any machine annotation can be editorially managed using the Content Link Tool.

These aggregation pages allow a user to go from a given person, place or subject to any BBC content area. But this is only half the story because once a user has entered an area of BBC content there are few links through to other related content. An example might be going from the 'carrots' aggregation page to a particular recipe (BBC Food) but then not being able to go to a related health story (BBC Health). What is needed is a link on the content page back to the related aggregation page linking up the user journey. Providing this link is the role of the navigation badge.

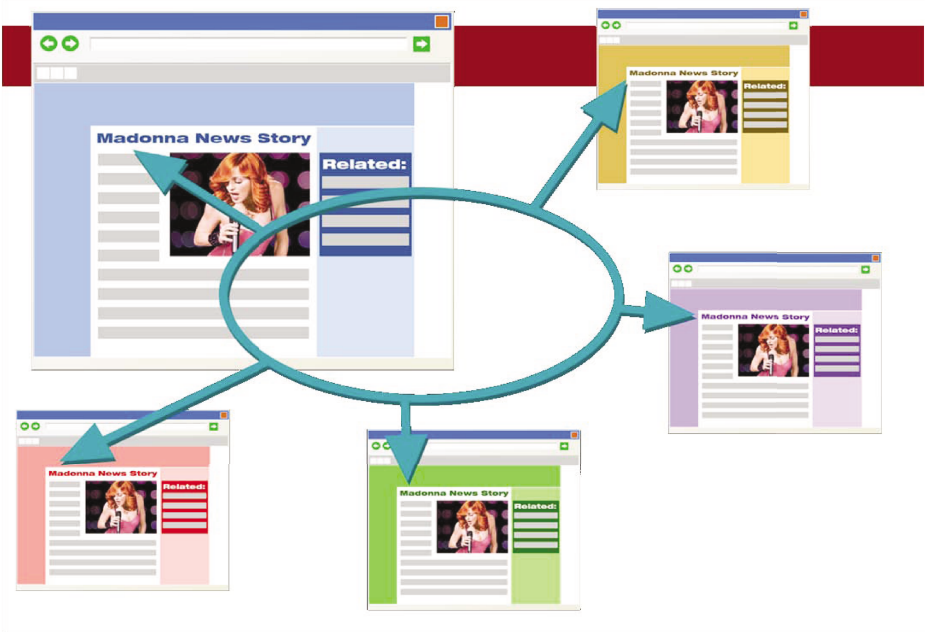

Fig. 5. Linking Documents via Navigation Badges

A navigation badge is dynamically created based on entity extraction of the content on the page. The identified concepts are then compared to existing aggregation pages and where an aggregation page exists it is linked to. This is done using Muddy Boots via the Content Link Tool. In addition to the automatic application of links to aggregation pages editorial staff will be able to add and remove links as appropriate. The navigation badge 'knows' which navigation page to link to because both services are based on DBpedia identifiers. Figure 5 illustrates how different stories about Madonna are interlinked via navigation badges, in the figure shown as "related"-boxes.

The result is a coherent and scalable user journey via a combination of aggregation pages and navigation badges. It is DBpedia in this case that is providing the semantic backbone with which to create the nodes on these journeys and ensure all services are using the same identifiers to express a given concept.

7 Conclusion and Outlook

In this paper, we described how Linked Data technologies were applied by the BBC. We mentioned two BBC services, BBC Music and BBC Programmes, publishing data in two different domains. We then described a categorisation system, CIS, allowing us to interlink data items in these different services because of the successful interlinking of CIS categories with corresponding DBpedia web identifiers, in order to get access to richer relationships between concepts. Finally, we demonstrated how these links between data items can benefit our user facing web sites, through topic pages and navigation badges.

There are three main next development steps we want to outline, as those need to be addressed to continue our path of better connecting BBC content with the help of DBpedia:

DBpedia Live. The DBpedia dataset is currently updated only every 2-3 months based on the releases of Wikipedia database dumps. In order to make DBpedia a real-time RDF view of Wikipedia, the DBpedia framework will be adjusted to support live extraction and updates of the dataset. Live extraction will provide instant feedback for users adding information and make that information available for use immediately. BBC News editors can at the time they write an article about an event create according DBpedia concepts (such as a previously unknown person being involved in some event) and instantly access them via the mechanisms described in this paper.

More domains. Along the same lines as programmes and music we are looking to add additional domains, specifically, events (e.g. music festivals) and natural history (species, habitats) content. This will allow the BBC to more accurately reflect the nature of events such as the Proms - cross linking with those programmes covering the event and those artists featured at the event. The work to create and publish a nature history domain at bbc.co.uk will allow the cross linking of programmes, news stories etc. through natural history concepts.

Equivalency Engine. Finally, our work linking CIS and DBpedia was just a starting point. There are more ID systems and concept stores in use at the BBC, and we want to build a generic equivalency engine based on the methods described above (and similar to link discovery tools being developed in the meantime[13][14]) in order to interlink those remaining systems. We think that such an equivalency engine would not only help the BBC interlink their internal datasets with each other and with DBpedia, but would also be beneficial for outside developers, enterprises and the larger web.

In conclusion we believe that the BBC and their users can largely benefit from the better connected ecosystem of content we are creating, and we hope that more content providers will join us on that road of interlinking content with the Linking Open Data project and in particular with DBpedia and MusicBrainz to create more meaningful navigation paths not only within websites but across the whole web.

Acknowledgments

We are grateful to all the people who made this work possible, in particular to Patrick Sinclair, Nicholas Humfrey, Derek Harvie, Matthew Wood, Frances McNamara, Andrew Shearer and Sophie Walpole from the BBC and Christian Becker from Freie Universität Berlin.

References

1. Berners-Lee, T.: Design Issues: Linked Data (2006), http://www.w3.org/DesignIssues/LinkedData.html
2. Bizer, C., Cyganiak, R., Heath, T.: How to publish Linked Data on the Web (2007), http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia - A Nucles for a Web of Data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
4. Giasson, F., Raimond, Y.: Music Ontology Specification (2007), http://musicontology.com
5. Hepp, M., Bachlechner, D., Siorpaes, K.: Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements. In: Proceedings of the Workshop on Semantic Wikis at the ESWC 2006 (2006)
6. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
7. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. Journal on Data Semantics IV (2005)
8. Raimond, Y., Sutton, C., Sandler, M.: Automatic Interlinking of Music Datasets on the Semantic Web. In: Proceedings of the Linked Data on the Web (LDOW 2008) Workshop at 17th International World Wide Web Confernce (2008)
9. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS Core: Simple Knowledge Organisation for the Web. In: Proceedings of the International Conference on Dublin Core 2005 (2005)
10. Zaragoza, H., et al.: Ranking Very Many Typed Entities on Wikipedia. In: Proceedings of the Sixteenth ACM International Conference on Information and Knowledge Management (2007)
11. Dakka, W., Cucerzan, S.: Augmenting Wikipedia with Named Entity Tags. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing (2008)
12. Bunescu, R., Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proceedings of the 11th Conference of the EACL (2006)
13. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk – A Link Discovery Framework for the Web of Data. In: Proceeding of the 2nd Workshop about Linked Data on the Web (2009)
14. Hassanzadeh, O., et al.: A Declarative Framework for Semantic Link Discovery over Relational Data. Poster at 18th World Wide Web Conference (2009)