The Web as a Content Management System

Patrick Sinclair, Nicholas Humfrey, Yves Raimond, Michael Smethurst, Tom Scott

BBC Audio and Music Interactive
Henry Wood House, 3-6 Langham Place,
W1B 3DF, London, UK
+442077654640

{firstname.lastname}@bbc.co.uk

ABSTRACT

In this paper, we describe the BBC Music Beta, providing a comprehensive guide to music content across the BBC. We publish a persistent web identifier for each resource in our music domain, which serves as an aggregation point for all information about it. We describe a promising approach in building web sites, by re-using structured data available elsewhere on the Web --- the Web becomes our Content Management System. We therefore ensure that the BBC Music Beta is a truly Semantic Web site, re-using data from a variety of places and publishing its data in a variety of formats.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Data sharing

General Terms

Management, Design, Human Factors.

Keywords

Semantic web, Linked Data, Music, API

1. INTRODUCTION

The aim of the BBC Music Beta [3] is to provide a comprehensive guide to music content across the BBC, linking information about an artist to those BBC programmes that have played them. The first step is to provide detailed information about artists who appear on BBC programmes, have had an album reviewed on the BBC Music site or have been covered on BBC News.

Rather than maintain our own source of music information, we are using existing, open repositories: MusicBrainz [4], an open source community-maintained database of music information, and Wikipedia [8].

This paper describes how by using these repositories we are using the web as a content management system, effectively enabling anyone to indirectly contribute to the BBC Music Site, while also helping to create links between existing URIs. We also describe how MusicBrainz has allowed us to automatically link one of our richest assets, BBC News stories, to individual artists using the hyperlinks, and how this is in turn encouraging a richer web.

2. BBC MUSIC BETA

The BBC Music Beta is a major rewrite of the existing site, in redeveloping the service we set out with a clear objective to design and building the service around the primary objects within the music domain and integrate those with the other BBC domains our audience is interested in, namely programmes, events and users.

The primary music objects are: artists, releases (and their reviews) and labels, for each of these we are working to provide persistent URIs. To ensure that we minimise the number of web identifiers we have reused those provided for Musicbrainz. Our hope in doing this is that we make it easier for others to link to and reuse the data we are publishing.

Once these persistent web identifiers for the different types of resources in our domain are set up, the most difficult step towards a Semantic Web site is achieved. Different representations for each of these resources can suit different types of user agents, from traditional web browsers to more sophisticated user agents making use of structured web data. Our HTML documents are designed for the earlier, whereas our RDF documents are designed for the latter. We automatically deliver the representation suiting a particular user agent's need using HTTP content negotiation.

For example, for an artist on the BBC Music Beta, we consider the following web resources:

The artist itself, where `:guid` represents it's MusicBrainz identifier (e.g. `a3cb23fc-acd3-4ce0-8f36-1e5aa6a18432`):
`http://www.bbc.co.uk/music/artists/:guid#artist`

A document about the artist:
`http://www.bbc.co.uk/music/artists/:guid`

Several content-negotiated versions of that document, e.g.
`http://www.bbc.co.uk/music/artists/:guid.html` and
`http://www.bbc.co.uk/music/artists/:guid.rdf`

These web identifiers ensure we follow the principles outlined in [9], whilst not requiring any HTTP redirects, which would significantly increase the traffic on our website. Both the HTML and the RDF documents provide links to further web identifiers, e.g. programmes in which an artist has been featured, a corresponding artist in DBpedia [1] or an identifier for the "music artist" concept in the Music Ontology [6].

In addition to those primary pages we also identified a large number of additional resources -- fragments of a page if you will -- that are also assigned their own URI. Each document, each page, is then composed by transcluding the relevant set of resources into the document. For example, the URL for an artist at: `http://www.bbc.co.uk/music/artists/:guid`

is in turn composed of the following resources:

```
http://www.bbc.co.uk/music/artists/:guid/news
http://www.bbc.co.uk/music/artists/:guid/reviews
http://www.bbc.co.uk/music/artists/:guid/links
http://www.bbc.co.uk/music/artists/:guid/labels
```


WWW 2009, April 20-24, 2009, Madrid, Spain.

This approach has the advantage that different representation can transclude a different set of resources i.e. the URIs remain the same while the set that makes up any given representation may vary (as will the document format). For example the mobile (XHTML MP) representation excludes the wikipedia biography whereas the desktop (XHTML) does include this information. We of course recognise that this approach may be considered controversial since the approach is not allowed within the HTTP spec. However, we believe that it is an acceptable compromise since it does improve the user experience while retaining one URI per concept.

2.1 Web as a Content Management System

On the BBC Music Beta, there are three sources of information: MusicBrainz, Wikipedia and the BBC. MusicBrainz is used as the backbone of the site, providing data such as artists' releases, relationships with other artists and links to external websites. Wikipedia is used for artists' biographies. The BBC provides additional material, such as an image, album reviews, details about which programmes have played that artist and links to featured content elsewhere on the BBC site.

To obtain data from MusicBrainz, we are using their replication mechanism [5]. This consists of SQL change event packets delivered hourly over FTP, which we download and apply to our own local MusicBrainz database. We then use the Wikipedia article link for each artist provided by MusicBrainz to fetch the Wikipedia article synopsis that forms the artist biography. To keep the Wikipedia text up to date, we have developed a system that tracks the Wikimedia recent changes IRC channel [7] that is updated whenever an article is created, edited or deleted. When we track an edit to an article linked to from MusicBrainz, the system downloads that article and stores it in our local database, allowing us to keep the Wikipedia biographies up to date in real time.

The use of MusicBrainz and Wikipedia to provide the underlying data for the site has allowed us to cover a much wider range of artists than would otherwise be possible - it is beyond our resources to maintain a biography for every artist heard on the BBC. It is also ensures that the data is kept up to date and doesn't go stale. For instance, when an artist dies their profile is updated within a few hours by the community and reflected on our site.

The BBC is also an active member of both the MusicBrainz and Wikipedia communities, at the time of writing the BBC's music team has contributed over 2,800 edits to MusicBrainz. Many of these edits include content about new or specialist music but a significant number of edits includes updating links to other sites. The addition of these links is important because it helps the web aggregate more resources around those concepts. Thus the contributions of BBC staff means that not only does the BBC benefit, so does the quality and quantity of data at those services, but possibly more importantly the web at large benefits because it is more coherent, more linked and so more easy to navigate.

2.2 Automatically Linking Artists and News

On many of the news stories published on BBC News journalists add related Internet links. If a story covers a music artist, it might link out to their home page, their MySpace site or even a Wikipedia article. In MusicBrainz, artists can have several URLs associated to them. By simply cross-referencing each link on a news story with the URLs in MusicBrainz, when we find a match we can confidently say that the news story relates to the artist associated with that URL.

For example, a news story covering Madonna links to her homepage at `http://www.madonna.com/`. When we look up the `http://www.madonna.com/` in MusicBrainz we find it associated with the MusicBrainz artist Madonna, allowing us to link to that news item from Madonna's artist profile. By tracking an RSS feed and checking the links for each story we can generate a news feed with RSS for any artist.

This simple technique is completely automatic and will perform across any of the 400,000 artists in MusicBrainz. As it is based on matching links added by BBC editors we can be very confident that a news item will be associated with the correct artist. We also believe that it adds value to the web, as BBC editors will be encouraged to add useful links from news stories so that these can be aggregated on artist profile pages. Editors from BBC Music will play a more active part in maintaining artist links in MusicBrainz instead of manually associating artists with news items, improving the data quality in MusicBrainz.

We would like to extend this technique on other music news sites besides the BBC but it requires that these sites link out to external sites, not just to their own pages. As discussed by Tim O'Reilly [2], there is a current trend for sites to prefer linking to themselves. Perhaps our approach to linking news stories with artists will encourage such sites to start linking out again.

3. SUMMARY

BBC Music Beta uses open source, community maintained MusicBrainz and Wikipedia projects to publish comprehensive information about the music heard on the BBC. We are exploiting techniques powered by the web links stored in MusicBrainz to aggregate content around artists, encouraging contributions by BBC staff to MusicBrainz and Wikipedia. The reuse of the existing MusicBrainz identifiers for our persistent URIs makes it easier for others to link to and reuse the data we are publishing.

4. ACKNOWLEDGMENTS

Our thanks to everyone at Audio and Music Interactive, in particular the Music Discovery and Music Interactive teams.

5. REFERENCES

[1] S. Auer and J. Lehmann, “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content”, The Semantic Web: Research and Applications, pages 503-517, 2007, http://www.eswc2007.org/pdf/eswc07-auer.pdf

[2] Tim O’Reilly, “Is Linking to Yourself the Future of the Web?”,http://radar.oreilly.com/2008/08/is-linking-to-yourself-the-future-of-the-web.html

[3] BBC Music Beta: http://www.bbc.co.uk/music/beta

[4] MusicBrainz: http://www.musicbrainz.org

[5] MusicBrainz replication mechanism: http://musicbrainz.org/doc/ReplicationMechanics

[6] Music Ontology: http://musicontology.com

[7] Wikmedia Recent Changes IRC channel: http://meta.wikimedia.org/wiki/IRC_channels

[8] Wikipedia: http://www.wikipedia.org

[9] W3C, “Architecture of the World Wide Web, Volume One”,http://www.w3.org/TR/webarch