Use of Semantic Web technologies on the BBC Web Sites

Yves Raimond, Tom Scott, Silver Oliver, Patrick Sinclair and Michael Smethurst

Abstract The BBC publishes large amounts of content online, as text, audio and video. As the amount of content grows, we need to make it easy for users to locate items of interest and to draw coherent journeys across them. In this chapter, we describe our use of Semantic Web technologies for achieving this goal. We focus in particular on three BBC Web sites: BBC Programmes, BBC Music and BBC Wildlife Finder, and how those Web sites effectively use the wider Web as their Content Management System.

1 Introduction

The BBC is the largest broadcasting corporation in the world. Central to its mission is to enrich peoples lives with programmes that inform, educate and entertain. It is a public service broadcaster, established by a Royal Charter and funded, in part, by the licence fee that is paid by UK households. The BBC uses the income from the licence fee to provide public services including 8 national TV channels plus regional programming, 10 national radio stations, 40 local radio stations and an extensive website, `http://www.bbc.co.uk`.

1.1 Linking microsites for cross-domain navigation

The BBC publishes large amounts of content online, as text, audio and video. Historically the website has focused largely on supporting broadcast brands (e.g. Top Gear) and a series of domain-specific sites (e.g. news, food, gardening, etc.). That is, the focus has been on providing separate, standalone HTML sites designed to be

BBC, UK, e-mail: `firstname.lastname@bbc.co.uk`

accessed with a desktop Web browser. These sites can be very successful, but tend not to link together, and so are less useful when people have interests that span programme brands or domains. For example, we can tell you who presents Top Gear, but not what else those people have presented. 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, until recently you weren't able to navigate from a page about a programme to a page about an artist played in that programme.

This lack of cross linking has also limited the type of user interaction the BBC is able to offer, for example, it is a complex piece of work to recontextualise content designed for one purpose (e.g. a programme web site) for another purpose or to extract the underlying data and visualize it in a new or different way. This has been because of a lack of integration at a data level and a lack of semantically meaningful predicates making it difficult to repurpose and represent data within a different context.

1.2 Making data available to developers

The BBC, since 2005 through its Backstage project[1], has made 'feeds' (mainly RSS) available for third party developers to build non-commercial mash-ups. However, these feeds suffer from the same or similar issues to the microsites namely they lack interlinking. That is, it is possible to get a feed of latest news stories but it is not easy to segment that data into news stories about 'Lions'. Nor is it possible to query the data to extract the specific data required.

1.3 Making use of the wider Web

Developing internal Content Management Systems is expensive, both in terms of editorial staff required to add and curate data into them, and in terms of development and integration costs. A tremendous amount of community-curated data is available on the Web, which can be used to make our sites richer, either by providing a navigation backbone (e.g. Musicbrainz[2] for BBC Music) or by enhancing our pages with relevant information (e.g. Wikipedia[3] for BBC Music). Also, by involving our editorial staff in those community-curated datasets, we make sure the community at large benefit from our use of the data.

[1] `http://backstage.bbc.co.uk/`

[2] see `http://musicbrainz.org`

[3] see `http://wikipedia.org`

2 Programme support on the Web

When commissioning hand-crafted programme web sites for specific broadcast brands, only a small subset of programme can be covered. Hence, until recently, only the major BBC brands had a web presence on the BBC web site. Even between programmes that had a corresponding web site commissioned, the disparity in terms of programme support was high. Some programmes would have a very detailed web site, with for example information about cast and crew, about the fictional universe in which the programme takes place, etc. Some other programmes would just have a single web page with upcoming broadcast dates.

As the BBC broadcasts between 1,000 and 1,500 programmes a day, this meant that historically the long tail of programming didn't get any web presence. Hand-crafted web sites are also harder to maintain and they therefore often got forgotten and left unmaintained, or even removed. This meant that when referring to a particular programme from other content on the BBC web site, no persistent link could be used.

As new platforms become ubiquitous (mobile, game consoles, etc.), so the BBC web sites also needed to provide a coherent offering across those platforms. However, without a single, common source of integrated data and an efficient publishing mechanism this increase in platforms could result in a parallel and unsustainable increase in effort.

Hand-crafting programme web sites is inefficient - there is a limited amount of code reuse between sites but it is not only expensive in terms of actual expenditure, but also in terms of opportunity costs. The time spent writing HTML files is lost, and you can't spend it on developing new features or otherwise improving the site for its users.

2.1 BBC Programmes

BBC Programmes[4] launched in Summer 2007 to address these issues. It provides a persistent web identifier for every programme the BBC broadcasts. Each web identifier has multiple content-negotiated representations, ensuring that a coherent offering is proposed across multiple devices (e.g. desktop and mobile) and that the data used to generate our pages is re-usable in different formats (RDF/XML, JSON and plain XML) to enable building enhanced programme support applications. Other teams within the BBC can incorporate those programme pages into new and existing programme support sites, TV Channel and Radio Station sites, and cross programme genre sites such as food, music and natural history.

[4] `http://www.bbc.co.uk/programmes`

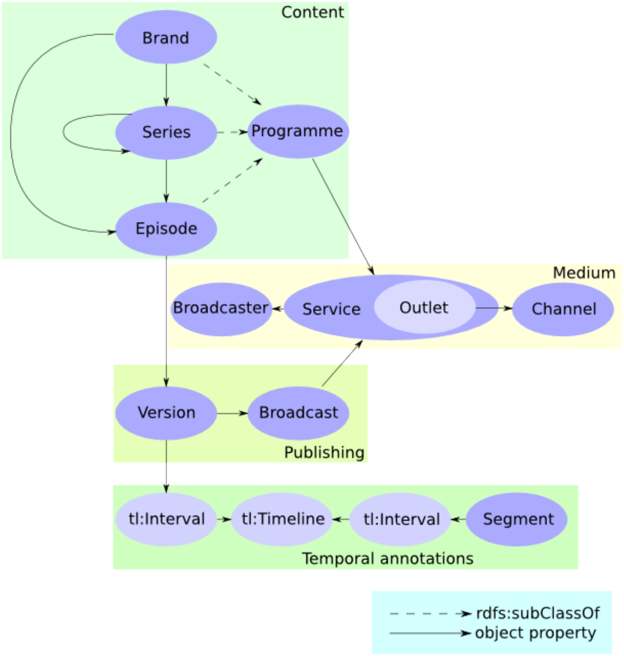


Fig. 1 Main terms within the Programmes Ontology and their relationships

2.2 The Programmes Ontology

In November 2007, we launched the Programmes Ontology [11]. The reason for publishing this ontology was three-fold. Firstly, it exposes the data model driving our web site as a formal OWL [5] ontology. As BBC Programmes was built using a Domain Driven Design methodology[5], this ontology can be seen as a 'map' of the different items we publish a web identifier for and of the links between these items. Secondly, it allows us to anchor our data feeds within a domain model. The RDF/XML feeds we provide, as well as the RDFa markup embedded within our HTML pages, refer to terms defined within this ontology. Thirdly, this ontology aims at assisting other organisations or individuals to publish programmes data on the Web. In the following, we give a brief overview of the different terms defined within our Programmes Ontology. A diagram of these terms and their relationships is given in Figure 1.

[5] For more details on the methodology used to build BBC Programmes, we refer the reader to `http://www.bbc.co.uk/blogs/radiolabs/2009/01/how_we_make_websites.shtml`, last accessed April 2010

2.2.1 Main terms

We consider a **Programme** as being the core of our domain model. A programme is an editorial entity, which can either be an **Episode** (e.g. ‘Top Gear, first episode of the first series’), a **Series** (e.g. ‘Top Gear, first series) or a **Brand** (e.g. ‘Top Gear’). All these programmes have multiple **Versions**, where a version is an actual piece of media content, either audio or audio and video. A single episode may have multiple versions. For example, an episode can have an original, unedited, version, a shortened one, a signed one, etc. Versions can have **Broadcasts**, each of them being on a particular **Service** and at a particular time, and they can have **Availabilities** — they can be made available through our iPlayer catchup service for a period of time on a particular set of devices.

2.2.2 Tagging programmes

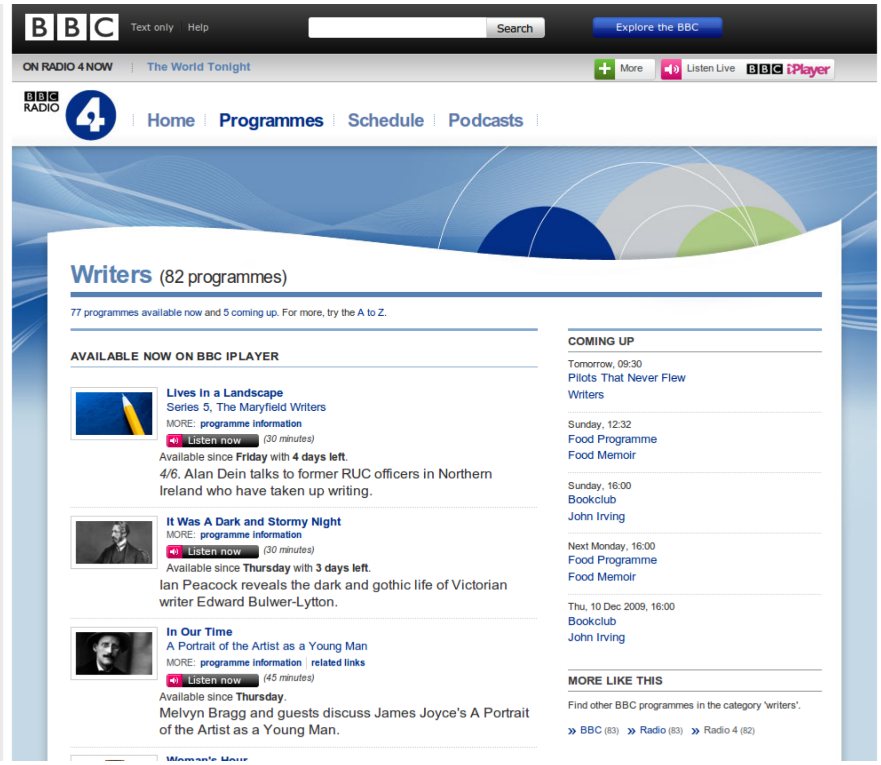
In order to generate simple aggregations of programmes, our domain model has a simple **category** predicate allowing us to relate a programme to a particular item in a SKOS categorisation scheme [8]. The Programmes Ontology defines two different categorisations scheme: genres (e.g. ‘drama’) and formats (e.g. ‘animation’). We also use the **subject** predicate defined within Dublin Core[6] to relate programmes to related subjects (e.g. ‘birds’), people (e.g. ‘William Shakespeare’), places (e.g. ‘Manchester’) and organisations (e.g. ‘BBC’). Using such categorisations and subject classifications, we can generate pages such as `http://www.bbc.co.uk/programmes/genres/drama/historical`, aggregating programmes in a particular category (here, in the ‘historical’ sub-genre of the ‘drama’ genre). Similarly, we can provide an aggregated view of all BBC Radio 4 programmes associated with the subject ‘writer’, as depicted in Figure 2. We are currently moving to use DBpedia [1] web identifiers as tags [7], so that we can aggregate richer information about them (e.g. geolocation of places and relationships between artists). By exploiting this ancillary data, we can provide richer ways of navigating our content.

2.2.3 A flexible segmentation model

When describing a programme, it is critical to be able to describe its actual content. Therefore, a good segmentation model which can accomodate track listings in a music show, points of interest within a programme, or editorially relevant sub-sections of a programme, is critical. The Programmes Ontology defines such a model, making use of the Event and Timeline ontologies[7] created within the scope of the Music Ontology [10]. A version of a programme has a temporal extent, which is defined

[6] see `http://dublincore.org/documents/dcmi-terms/`, last accessed May 2010

[7] See `http://purl.org/NET/c4dm/event.owl` and `http://purl.org/NET/c4dm/timeline.owl`



Writers (82 programmes)

77 programmes available now and 5 coming up. For more, try the A to Z.

AVAILABLE NOW ON BBC IPLAYER

Lives in a Landscape
Series 5, The Maryfield Writers
MORE: programme information
(30 minutes)
Available since **Friday** with **4 days left**.
4/6. Alan Dein talks to former RUC officers in Northern Ireland who have taken up writing.

It Was A Dark and Stormy Night
MORE: programme information
(30 minutes)
Available since **Thursday** with **3 days left**.
Ian Peacock reveals the dark and gothic life of Victorian writer Edward Bulwer-Lytton.

In Our Time
A Portrait of the Artist as a Young Man
MORE: programme information | related links
(45 minutes)
Available since **Thursday**.
Melvyn Bragg and guests discuss James Joyce's A Portrait of the Artist as a Young Man.

Woman's Hour

COMING UP

Tomorrow, 09:30
Pilots That Never Flew
Writers

Sunday, 12:32
Food Programme
Food Memoir

Sunday, 16:00
Bookclub
John Irving

Next Monday, 16:00
Food Programme
Food Memoir

Thu, 10 Dec 2009, 16:00
Bookclub
John Irving

MORE LIKE THIS

Find other BBC programmes in the category 'writers'.

» BBC (83) » Radio (83) » Radio 4 (82)

Fig. 2 An aggregation of BBC Radio 4 programmes associated with the subject ‘writer’

on a **Timeline**. On the same version timeline, we can anchor **Segments** — classifications of particular temporal sections of a version. Most links to other ontologies are done at the segment level, as we might want to describe e.g. the recipe being described or the track being played. For example, it is at the segment level that we link to the BBC Music web site described in section 3. We can also classify segments using the same mechanism as described above, to associate a segment with a particular place, subject, person or organisation.

2.3 Web identifiers for broadcast radio and television sites

Human-readability is often deemed important when creating web identifiers. In the case of programmes this could mean that identifiers could be created from programmes titles. However, when a programme title changes, the corresponding web identifier would also change, which would make external links to that programme break. Programme titles can also clash — it can happen that two distinct programmes share the same title, e.g. many episodes don’t have a distinct title, for example long running weekly shows such as the ‘Today programme’[8]. We could imagine creating web identifiers from other literal attributes, such as broadcast dates. However, those identifiers become ambiguous when a programme gets repeated. Those identifiers would also break for off-schedule content — programmes only available through on-demand services such as the BBC iPlayer.

Any web identifier that assumes some structure of the object it is representing is likely to break when that structure changes[9].

In order to keep a level of indirection helping us to deal with such changes, we use opaque unique identifiers such as `b00cccvg` to construct our web identifiers. Given an opaque identifier, we need to consider several web identifiers for a single programme. We need to identify the actual programme (e.g. a particular episode of ‘Doctor Who’), and a page about this programme, as we want to state different things about both of them — the creation date of the page will not the be the same as the creation date of the programme, for example. We adopt the following scheme:

- `/programmes/b00cccvg#programme` – the actual programme;
- `/programmes/b00cccvg` – a document about that programme;
- `/programmes/b00cccvg.html` – an XHTML page about that programme;
- `/programmes/b00cccvg.mp` – an XHTML Mobile Profile page about that programme;
- `/programmes/b00cccvg.rdf` – an RDF/XML document about that programme

We also need to identify the associated versions, segments, broadcasts and availability windows. We use a similar mechanism for those, by generating unique identifiers and constructing web identifiers from them.

From `/programmes/b00cccvg` to one specific representation (e.g. XHTML or RDF/XML), we use content negotiation [12]. The representation that is most appropriate for the user agent will be sent back, along with a `Content-Location` HTTP header pointing to the canonical web identifier for that particular representation.

The use of content negotiation and the use of the fragment identifiers firstly reduces the number of requests the server needs to process compared to other methods for publishing Linked Data, such as the redirection-based method described in [12]; but more significantly it ensures that there is one web identifier for a resource. We only want users or automated user agents to see and work with the programme web identifier or the generic document web identifier. So that if a user bookmarks a web identifier on a desktop machine they can access that bookmark on a mobile and get an appropriate mobile representation. Similarly, an automated user agent aggregating BBC Programmes data needs information in a more structured format than an HTML document, so it will access an appropriate structured representation, e.g. RDF/XML.

[8] `http://www.bbc.co.uk/programmes/b006qj9z`

[9] See the web identifier opacity section in [6]

3 BBC Music

The aim of the BBC Music website[10] 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. BBC Music follows the same principles as BBC Programmes, and provides a persistent web identifier for primary objects within the music domain, and integrate those with the other BBC domains our audience is interested in, namely programmes, events and users. These primary music objects are: artists, releases and their reviews, and editorial genres.

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 information, such as audio snippets for tracks, images, album reviews, details about which programme have played which artist and links to related content elsewhere on the BBC site.

3.1 BBC Music as Linked Data

We are publishing Linked Data for most of the resources on BBC Music using a variety of different ontologies and vocabularies. The Linked Data community has developed several vocabularies around the music domain that we have been able to reuse. For example, we use the music ontology [10] for describing artists and release information, the Reviews Ontology [2] for describing album reviews and SKOS [8] for defining the BBC music genres.

3.2 Web identifiers for BBC Music

As with BBC Programmes, we decided to use opaque identifiers for constructing BBC Music web identifiers to improve their persistence. MusicBrainz uses a globally unique identifier (GUID) scheme for its resources. When it came to BBC Music, instead of coming up with our own identifiers we reused the MusicBrainz artist GUIDs:

- `/music/artists/:musicbrainz_artist_guid#artist` – the actual artist;
- `/music/artists/:musicbrainz_artist_guid` – a document about that artist;
- `/music/artists/:musicbrainz_artist_guid.html` – an XHTML page about that artist;

[10] `http://www.bbc.co.uk/music`

- `/music/artists/:musicbrainz_artist_guid.rdf` – an RDF/XML document about that artist

For album reviews, we have minted our own URL keys (e.g "b5rj") and use the following scheme:

- `/music/reviews/:url_key#review` – the actual review;
- `/music/reviews/:url_key` – a document about that review;
- `/music/reviews/:url_key.html` – an XHTML page about that review;
- `/music/reviews/:url_key.rdf` – an RDF/XML document about that review

We also have similar scheme for other resources such as reviewers and BBC content promoted through the site[11].

3.3 The Web as a content management system

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 that would otherwise be possible. It is beyond our resources to maintain a biography for every artist heard on the BBC. It also ensures 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 this change is reflected on our site.

BBC Music takes the approach that the Web itself is its content management system. BBC editors directly contribute to Musicbrainz and Wikipedia, and BBC Music will show an aggregated view of this information, put in a BBC context.

3.4 Using the BBC Programmes and the BBC Music Linked Data

The BBC Programmes Linked Data described in section 2.1 links to the BBC Music data. A programme that features an artist will be linked to that artist within BBC Music, using the segmentation model described above. Moreover, BBC Music artists are linked to corresponding resources within DBpedia. A number of prototype applications demonstrating the use of such links have been built, both within the BBC and outside.

3.4.1 Programmes and locations

When aggregating DBpedia information about BBC artists, we can access related geographical location (current location, place of birth, place of death, etc.). Using

[11] respectively at `/music/reviewers` and `/music/promotions`

this information we can display programmes on maps, according to the locations of the artists played in those programmes. We can also build geographical programme look-up services[12] which, given a place, give a list of programmes featuring an artist related to that place.

3.4.2 Artist recommendations

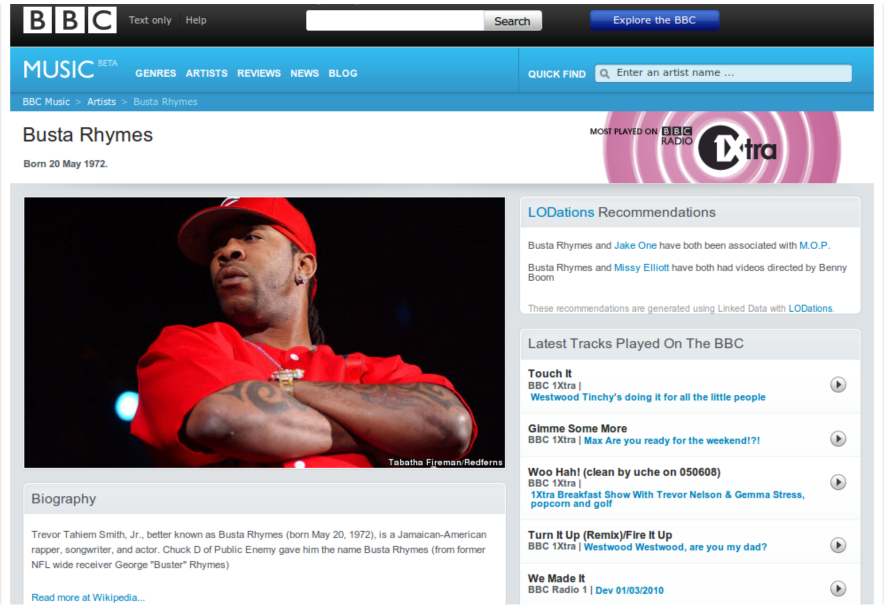
As mentioned in [9], Linked Data can be used to generate music recommendations. From BBC Music artists, a number of music-related datasets can be reached. By following links leading from one artist to another, we can derive connections between artists (e.g. ‘this artist has had his first music video directed by the same person as that other artist’) that can be used to drive recommendations. The path leading from one artist to another can then be used to explain why a particular recommendation has been generated. This is a fundamental shift from most current music recommender systems which, given an artist, return an ordered list of related artist without any clue for the user as to how these recommendations were generated. Although recent work has been done in trying to make music recommendation more transparent, such as the Aura [4] recommendation engine from Sun Microsystems Labs, the generated explanations are limited by the use of simple textual tags, which discards explanations derived from potential relationships between such tags.

Two prototypes have been built by the BBC to ilustrate such music recommendations generated from Linked Data. LODations[13] provides a collaborative way to specify editorially relevant connections between artists. New musical connections, such as ‘if two bands were formed in June 1976 in Manchester, then they are musically related’ can be specified, and music recommendations along with their explanations are derived from these connections. An example of LODations recommendations is depicted in Figure 3. The ‘MusicBore’[14] derives connections between artists in a similar way and use them to generate an original radio programme. An automated DJ, built using an off-the-shelf text-to-speech software, uses these connections to explain how the next artist in the tracklist relates to the previous one (e.g. ‘they were both born in Detroit in the mid-1960s’).

[12] an example of such a service, using Ordnance Survey, BBC and DBpedia data, is available at `http://www.johngoodwin.me.uk/boundaries/meshup.html`, last accessed July 2009

[13] see `http://lodations.heroku.com/`, last accessed April 2010

[14] see `http://www.bbc.co.uk/blogs/radiolabs/2009/07/the_music_bore.shtml`, last accessed April 2010



Busta Rhymes

Born 20 May 1972.

LODations Recommendations

Busta Rhymes and Jake One have both been associated with M.O.P.

Busta Rhymes and Missy Elliott have both had videos directed by Benny Boom

These recommendations are generated using Linked Data with LODations.

Latest Tracks Played On The BBC

Touch It
BBC 1Xtra |
Westwood Tinchy's doing it for all the little people

Gimme Some More
BBC 1Xtra | Max Are you ready for the weekend!?!

Woo Hah! (clean by uche on 050608)
BBC 1Xtra |
1Xtra Breakfast Show With Trevor Nelson & Gemma Stress, popcorn and golf

Turn It Up (Remix)/Fire It Up
BBC 1Xtra | Westwood Westwood, are you my dad?

We Made It
BBC Radio 1 | Dev 01/03/2010

Biography

Trevor Tahiem Smith, Jr., better known as Busta Rhymes (born May 20, 1972), is a Jamaican-American rapper, songwriter, and actor. Chuck D of Public Enemy gave him the name Busta Rhymes (from former NFL wide receiver George "Buster" Rhymes)

Read more at Wikipedia...

Fig. 3 Artist page for Busta Rhymes, along with recommendations generated from Linked Data

4 BBC Wildlife Finder

BBC Wildlife Finder[15], provides a web identifier for every species (and other biological ranks), habitat and adaptation the BBC has an interest in. The information presented is aggregated from data within the BBC and across the Web, including: Wikipedia, the WWF's Wildfinder, the IUCN's Red List of Threatened Species, the Zoological Society of Londons EDGE of Existence programme and Animal Diversity Web. BBC Wildlife Finder repurposes this external data and puts it in a BBC context adding to it with programme clips extracted from the BBC's Natural History Unit archive and links to programme episodes and BBC news articles.

An underlying principle behind the design of Wildlife Finder is the notion that people care more about real world concepts than abstract web pages, by providing web identifiers and associate documents about things that people care about, think about and talk about, it is more likely that the site will be intuitive and more likely to be discovered via search. As a result the Wildlife Finder provides web identifiers for real world natural history objects — animals, plants their habitats and adaptations. Each of these resources are then linked to adjacent concepts within the ontology, for example, a web idenfier for Lions links to web identifiers for the "tropical grassland" (because lions live there) and for "pack hunting" because that's one of the lion's behaviours; this is in addition to BBC programmes and news stories about lions.

[15] `http://www.bbc.co.uk/wildlifefinder/`

Wildlife programmes (clips and episodes) are transcluded and linked to from the Wildlife Finder web site. The programmes are identified by 'tagging' the clip or episode with the appropriate DBpedia web identifier. Programmes are identified as "coming soon", "catch up" or "archived" by checking the BBC Programmes RDF described in section 2.1 to extract the relevant broadcast or catch-up details. RDF was found to be a convenient approach when it came to integrating two separate but related domains within the BBC.

4.1 The Wildlife Ontology

As noted above, BBC Wildlife Finder was designed following similar principles to BBC Music and BBC Programmes - that is modelling the site around real world concepts, in this case that means the animals, plants, habitats and adaptations the BBC films. We recently published the Wildlife Ontology [3], describing how those different concepts we are interested in relate to each other. Our objective in publishing the ontology is similar as what we described in section 2.2 for the Programmes Ontology. It should be noted that although the Wildlife Ontology was designed with the BBC Wildlife Finder application in mind it should be applicable to a wide range of biological data publishing use cases and to that end care has been taken to try and ensure interoperability with more specialised ontologies used in scientific domains such as taxonomy, ecology, environmental science, and bioinformatics.

4.1.1 Main terms

Biologists group organisms based on their current understanding of a species evolution. This has resulted in a hierarchical grouping or taxonomy of species. However, in addition to the absolute hierarchy the relative hierarchy when compared to other species can also be of interest and as a result each of these groups, or ranks, have historically been given their own names. Ranks are useful because they help you know how far down in the tree of life you are (e.g. a Class is further 'down' than Phylum). Certain ranks also allow you to usefully lump groups of organisms together. For example, the rank of "Class" (Class Aves, Class Mammalia, Class Insecta etc.) is convenient and useful to biologists.

Of all the biological ranks that of "species" is the most significant to us. Although there are a number of different definition for species the most common is that of a group of organisms capable of interbreeding and producing fertile offspring of both genders. The classification for all other ranks (e.g. Genus and Class) are not as strictly codified, and hence different authorities often produce different classifications.

In addition to the Linnean taxonomy the wildlife ontology also describes a species's habitat, grouped into terrestrial, aquatic (freshwater) and marine habitats;

their adaptations to their environment, their conservation status (as defined by the IUCN[16]) and the habitats and species found within an ecozone[17].

The core to the Wildlife Ontology, however, is the species. Species, in addition, to being for us the most significant biological rank also tend to be the point of interest with regard to other areas of research, for example, both conservation and distribution studies tends to focus on the species. It is therefore the species which links to other classes (habitat, adaptations, conservation status etc.) within the Wildlife Finder application. However, it is worth noting that this level of linking is not enforced within the ontology and it is acceptable to link habitats to other taxonomic ranks.

4.1.2 Species as Classes vs Species as Instances

One perennial problem associated with modelling biological taxonomies using RDF is whether to attempt to model individual species as Classes, or whether to simply model species as instances of a generic Species class. The latter approach is simpler and avoids creating a huge ontology that attempts to model all biological organisms. Existing ontologies have taken different approaches to resolving this issue, some choosing one style, others another. At present there doesn't seem to be a consensus. With this in mind, the Wildlife Ontology adopts the simpler of the two approaches, i.e. modelling species as instances of a Species class, as this maximises interoperability with many of the existing Linked Data sources, particularly DBpedia, which adopt similar approaches.

4.1.3 Web identifiers - using DBpedia as a controlled vocabulary

As noted above DBpedia provides a controlled vocabulary to help link programme clips and episodes, and news stories to web identifiers within Wildlife Finder. All resources within Wildlife Finder are constructed using the corresponding Wikipedia web identifier slug. A URI slug is the fragment of a URI that uniquely identifies a resource within a domain, for example, in the case of Wikipedia the URI slug for the entry Stoat: `http://en.wikipedia.org/wiki/Stoat` is “Stoat”.

By using identifiers that are already widely used across the Web it means that:

1. The BBC can effectively outsource a significant proportion of the effort required to maintain a controlled vocabulary to the Web;
2. It makes it easier for third party developers to integrate with BBC content because of a shared definition of a resource;
3. The BBC can contribute its knowledge to the Web by linking data to those common identifiers and by creating new identifiers where necessary.

[16] `http://www.iucnredlist.org/`

[17] See `http://simple.wikipedia.org/wiki/Ecozone`

The added advantage in using Wikipedia is the addition of a large evidence set. The Wikipedia article text defines the meaning and use of the identifier and hence allows the BBC to confirm which identifier to use in which context.

DBpedia URIs are used to categorise programme episodes and clips and, news stories. The cagorisation is used to assert that the programme, clip or news story is ”about” the concept identified by the DBpedia URI i.e. it is about an animal, plant, habitat or adaptation. This categorisation is thus used to identify news stories and programmes before transcluding the relevant information on Wildlife Finder pages. The canonical web identifier for a clip, news story etc. remains with BBC Programmes (see section 2.1), news site etc. The clip is therefore discoverable both in a BBC Programmes context and within a Wildlife Finder context (a clip can be about both a species and an adaptation).

4.2 Web identifiers

As mentioned above, in addition to using DBpedia as a controlled vocabulary to tag content, Wildlife Finder also reuses Wikipedia URI slugs to construct its web identifiers. The high level web identifier scheme is therefore as follows:

For biological taxa:

- `/nature/rank/:wikipedia_slug#:rank` – the actual organism;
- `/nature/rank/:wikipedia_slug` – a document about that organism;
- `/nature/rank/:wikipedia_slug.html` – an XHTML page about that organism;
- `/nature/rank/:wikipedia_slug.rdf` – an RDF/XML document about that organism

For habitats:

- `/nature/habitats/:wikipedia_slug#habitat` – the actual habitat;
- `/nature/habitats/:wikipedia_slug` – a document about that habitat;
- `/nature/habitats/:wikipedia_slug.html` – an XHTML page about that habitat;
- `/nature/habitats/:wikipedia_slug.rdf` – an RDF/XML document about that habitat

For adaptations:

- `/nature/adaptations/:wikipedia_slug#:adaptation` – the actual adaptation;
- `/nature/adaptations/:wikipedia_slug` – a document about that adaptation;
- `/nature/adaptations/:wikipedia_slug.html` – an XHTML page about that adaptation;

- `/nature/adaptations/:wikipedia_slug.rdf` – an RDF/XML document about that adaptation

The only exception to this web identifier scheme is with collections (editorially curated aggregations of BBC content):

- `/nature/collections/:pid#:adaptation` – the actual collection;
- `/nature/collections/:pid` – a document about that collection;
- `/nature/collections/:pid.html` – an XHTML page about that collection;
- `/nature/collections/:pid.rdf` – an RDF/XML document about that collection

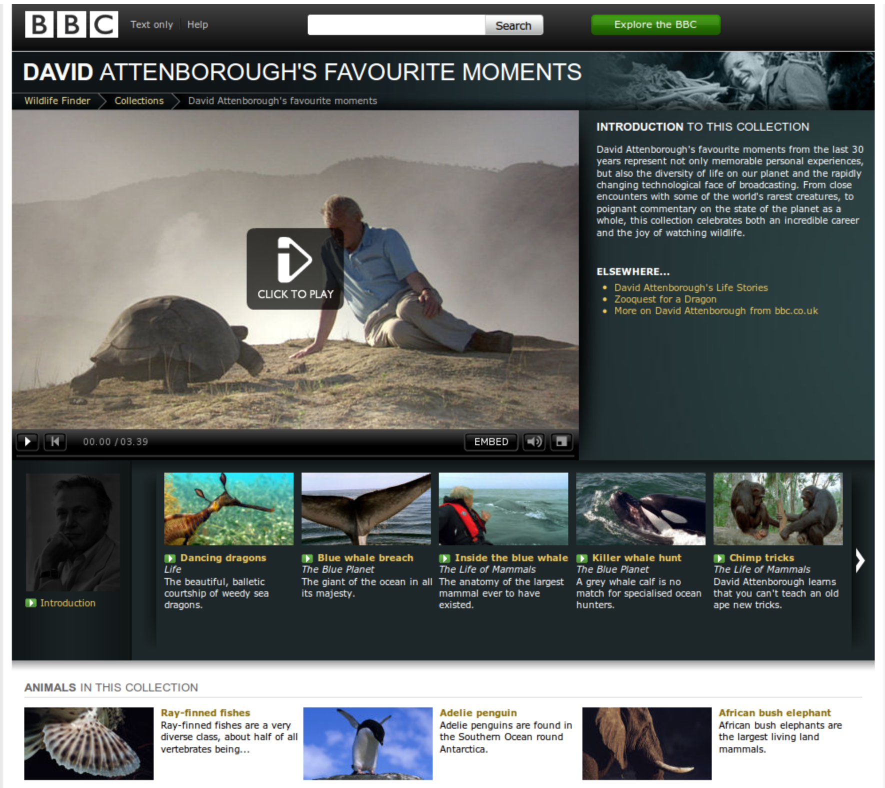
Programme identifiers similar to the ones used in BBC Programmes (PIDs) where chosen to identify collections because, like a programme, a collection is an editorially curated entity created by the BBC. The provenance of a collection means that not only will Wikipedia not have an entry but also that its ownership resides with the BBC not with the Web at large, since it is the BBC who chose the clips and edited the introductory video. It is therefore appropriate that the BBC provide the identifier for these objects.

From `/nature/rank/:wikipedia_slug` etc. to one specific representation (e.g. XHTML or RDF/XML), we use content negotiation, exactly as described in section 2.3. The format that is most appropriate for the user agent will be sent back, along with a `Content-Location` HTTP header pointing to the canonical URL for that particular format.

4.3 The Web as a Content Management System

The data that makes up a page on Wildlife Finder is taken from a range of range of sources, both inside and outside the BBC. Some of this data is in effect read-only that is the BBC nor its audience is at liberty to modify it. This includes, for example, the information about the conservation status of a species and is used, as supplied, by the IUCN. Other sources of data notably Wikipedia can be edited by both the BBC and its audience - this has the effect of making Wikipedia part of the BBC's content management 'system'.

The use of Wikipedia on Wildlife Finder means that the BBC and its audience benefits from access to (generally) high quality content about things in the natural world and additional, contextual links to that content. People can continue their journey and discover information elsewhere on the Web. In addition, the rest of the Web also benefits because where the BBC is able to improve the content on Wikipedia, it does so not on bbc.co.uk but on Wikipedia directly and in doing so users of Wikipedia also benefit.



DAVID ATTENBOROUGH'S FAVOURITE MOMENTS

Wildlife Finder > Collections > David Attenborough's favourite moments

INTRODUCTION TO THIS COLLECTION

David Attenborough's favourite moments from the last 30 years represent not only memorable personal experiences, but also the diversity of life on our planet and the rapidly changing technological face of broadcasting. From close encounters with some of the world's rarest creatures, to poignant commentary on the state of the planet as a whole, this collection celebrates both an incredible career and the joy of watching wildlife.

ELSEWHERE...

- David Attenborough's Life Stories
- Zooquest for a Dragon
- More on David Attenborough from bbc.co.uk

Dancing dragons
Life
The beautiful, balletic courtship of weedy sea dragons.

Blue whale breach
The Blue Planet
The giant of the ocean in all its majesty.

Inside the blue whale
The Life of Mammals
The anatomy of the largest mammal ever to have existed.

Killer whale hunt
The Blue Planet
A grey whale calf is no match for specialised ocean hunters.

Chimp tricks
The Life of Mammals
David Attenborough learns that you can't teach an old ape new tricks.

ANIMALS IN THIS COLLECTION

Ray-finned fishes
Ray-finned fishes are a very diverse class, about half of all vertebrates being...

Adelie penguin
Adelie penguins are found in the Southern Ocean round Antarctica.

African bush elephant
African bush elephants are the largest living land mammals.

Fig. 4 A collection of David Attenborough's favourite moments from the last 30 years

4.4 The importance of curation

It is not always possible nor desirable to automate all aspects of page building. It is sometimes advantageous to curate specific collections of content - for example a collection of David Attenborough's favourite moments from the last 30 years[18], as depicted in Figure 4, or a collection highlighting one of the worlds endangered animals, the tiger[19].

While in theory both these collections could be separately modelled and codified within the ontology it wouldn't be practical to do so - there would be little additional benefit in modelling one off collections. Instead it is sufficient to define a collection as a group of editorially selected and sequenced clips, habitats, adaptations and taxa introduced through a bespoke clip i.e. one edited specifically for the collection.

Collections add a personalised context to the content within Wildlife Finder — they are not simply a set of aggregated clips, they have been selected, sequences and

[18] see `http://www.bbc.co.uk/nature/collections/p0048522`

[19] see `http://www.bbc.co.uk/nature/collections/p0063wt7`

used to tell a specific story. This added layer adds a level of trust to the content and helps guide the users of the site through particular aspects of the site. Without this layer the site would remain largely encyclopedic, providing information to those that know what they are looking for but not introducing people to the content and acting as a guide through that content. With the addition of collections the BBC can guide people through some of that information, to present it in a different light and add a new context to the raw information; and in doing so the collections not only provide the audience with an easier route into the site but can also facilite a conversation by positioning the content within a specific light.

5 Journalism

BBC Journalism incorporates News, Sport, Travel and the Weather. The majority of this content (News and Sport) consist of stories published out of a content production system. Once published these stories are manually managed on to a small number of topical indexes. For years this has been sufficient but recent business requirements have created an impetus for more sophisticated publishing and navigation strategies. For example:

- Automating the creation of lower profile indexes;
- More sophisticated information architectures with many more topical indexes;
- Merging data (sports statistics for example) with stories to create a more coherent product;
- Linking to other BBC sites and external sites.

The starting point has been sporting events like the Winter Olympics and World Cup. These are easier to model and populate with data because you know when they will happen, who will participate and where the event will happen. For example the model for the Winter Olympics includes athletes, sports disciplines and sports venues. For the 2010 Winter Olympics we published a page for every sports discipline. These pages consisted of stories, statistics, other BBC content and links to external sites.

The process of creating these pages was to first model the event. Working with domain experts we established the key concepts and relationships important to the event. This was then turned into a formal OWL ontology. Where possible bits of existing ontologies were reused and the Winter Olympics model was based largely on the Event Ontology mentioned in section 2.2. This will make it easier for others to query and access the data in the future as well as reducing the number of design decisions required.

5.1 Populating and using the ontology

The modelling of the Winter Olympics was relatively low cost in terms of design time but populating the ontology would involve significant effort. Because of this we looked to consume data freely available on the web. The primary source is Wikipedia but there were gaps that we would need to fill with data the BBC would author. To facilitate the integration of the Wikipedia data and BBC data we are using the Uberblic service[20]. This means live updates from Wikipedia can be combined along with data created in a local MediaWiki[21] instance. The data can then be consolidated in real time and the resulting RDF used by services producing the Winter Olympics site.

The Uberblic service provides DBpedia identifiers for those entities that are in Wikipedia. This ensures Winter Olympics data can interoperate with other BBC systems that work with DBpedia identifiers. For those entities not in Wikipedia new BBC identifiers were created. This means we are not restricted to entities existing in Wikipedia alone. This approach is taken, as opposed to creating new Wikipedia pages, because the concepts are considered to be unlikely to be of significant cultural interest to justify the existence of a page in Wikipedia (for example Canadian Winter Olympics Team at the Winter Olympics 2010).

The populated ontology was primarily used to provide data for an auto-categorisation system. The presence of various entities in a document could be used, once compared to concepts and relationships in the ontology, to help disambiguate the entities extracted. This service suggested concepts to journalists to tag stories with. Once done, tagged stories were then dynamically included on Winter Olympics pages.

5.2 Future developments

In future projects greater use could be made of the ontology and the common web identifiers. Sharing common identifiers across the BBC ensures the aggregated pages contain as much variety of BBC content as possible. For example the use of DBpedia identifiers by the BBC Programmes service and sporting events like the Winter Olympics allow for the transclusion of BBC programmes in to the sport event pages. In addition if there was a requirement to add a rich navigation structure linking the aggregated pages we have the option of using the relationships in the ontology. For example the relationships in DBpedia could be used to generate links between Winter Olympic athletes and the sports disciplines they participate in.

An easily achieved use of Linked Data for news organisations would be to make their topic aggregation pages available as Linked Data. This is something the New

[20] see `http://uberblic.org`

[21] see `http://www.mediawiki.org`

York Times have already done[22] and there are now established design patterns for publishing lists of stories associated with a particular topic. This will offer a clear and simple path for other news organisation to follow the New York Times lead. Once we have a number of organisations publishing stories as Linked Data it will be increasingly easy to link from a BBC topic index to the most recent stories on other news sites. The key here will be the use of common web identifiers. We can already see in the BBC, New York Times and Reuters the use of DBpedia as a way to commonly identify topics and entities. The publishing of established controlled vocabularies (like IPTC) as Linked Data and mapping them to common web identfiers will also be critical to lowering the barrier of entry to those organisations already using these vocabularies.

Going beyond the syndication of lists of stories the next stage for a news organisation would be to take advantage of the data freely available as Linked Data. Data sets like DBpedia, CIA Fact Book and the recently released UK Government data[23] could all be used to add context and navigation to otherwise dry aggregation pages. For example an aggregation page for an MP could be enriched with a personal profile, data about their previous election results and the policies they have voted on. In addition links between an MP aggregation page and a constituency aggregation page could be provided for free from the Linked Data sources. This is a technique already used by Wildlife Finder and BBC Music, as described above, and Linked Data makes this process considerably easier than collecting distributed data sets in different formats (for example CSV files and custom database dumps).

Not only will this improve the user experience but could also significantly improve the process of story creation for the journalists. If common Government identifiers are provider for a politician and we tag our assets with the same identifier then when a journalist is researching a story about a politician it becomes increasingly easy to pull data to them. By identifying the politician a journalist is interested in, it would be trivial to pull together all assets created by their news organisation as well as Linked Data about the politician from trusted web sources. This goes beyond the document retrieval of Google as it could merge useful data and documents to provide context regarding the politicians career, popularity with voters and impact on their constituency.

The question of trust regarding Linked Data is very important for journalism. Where sources like Wikipedia may not be acceptable by editorial standards it puts even more emphasis on trusted sources. For this reason the recent publishing of UK Government Linked Data sets is critically important to journalism, providing trusted identifiers, labels and relationships for things like politicians, schools and UK locations.

[22] see `http://data.nytimes.com`

[23] see `http://data.gov.uk`

6 Conclusion

Creating web identifiers for every item the BBC has an interest in, and considering those as aggregations of BBC content about that item, allows us to enable very rich cross-domain user journeys. This means BBC content can be discovered by users in many different ways, and content teams within the organisation have a focal point around which to organise their content.

Re-using data from existing online sources such as Musicbrainz or Wikipedia means that the community at large benefits from our use of the data, as our editorial staff is directly contributing to those sources. It is also more efficient than maintaining an in-house Content Management System, which would require development and integration costs, and which would be very difficult to bootstrap, curate and maintain up-to-date.

The RDF representations of these web identifiers allow developers to use our data to build applications. The two issues, providing cross-domain navigation and machine-readable representations, are tightly interleaved. Giving access to machine-readable representations that hold links to further such representations, crossing domain boundaries, means that much richer applications can be built on top of our data, including new BBC products. In addition the system gives us a flexibility and a maintainability benefit: our web site becomes our API. Considering our feeds as an integral part of building a web site also means that they are very cheap to generate, even when built in a best efforts way: they are just a different view of our data.

The approach has also proved to be an efficient one – allowing different development teams to concentrate on different domains while at the same time benefiting from the activities of the other teams. The small pieces loosely joined approach, which is manifest in any Linked Data project, significantly reduces the need to coordinate teams while at the same time allowing each team to benefit from the activities of others.

References

1. S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the International Semantic Web Conference*, Busan, Korea, November 11-15 2007.
2. Danny Ayers. Review vocabulary. Working draft.
3. Leigh Dodds and Tom Scott. Wildlife ontology. Online ontology, February 2010. Available at `http://purl.org/ontology/wo/`. Last accessed April 2010.
4. S. Green, P. Lamere, J. Alexander, and F. Maillet. Generating transparent, steerable recommendations from textual descriptions of items. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, 2009.
5. I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, 1:7–26, 2003.

6. Ian Jacobs and Norman Walsh. Architecture of the World Wide Web, volume one. W3C Recommendation, December 2004. `http://www.w3.org/TR/webarch/`. Last accessed July 2008.
7. Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Chris Bizer, and Robert Lee. Media meets semantic web - how the BBC uses DBpedia and linked data to make connections. In *Proceedings of the European Semantic Web Conference In-Use track*, 2009.
8. Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. W3C Recommendation, August 2009. Available at `http://www.w3.org/TR/skos-reference/`. Last accessed June 2010.
9. Yves Raimond. *A distributed music information system*. PhD thesis, Department of Electronic Engineering, Queen Mary, University of London, 2008.
10. Yves Raimond, Samer Abdallah, Mark Sandler, and Frederick Giasson. The music ontology. In *Proceedings of the International Conference on Music Information Retrieval*, pages 417–422, September 2007.
11. Yves Raimond, Patrick Sinclair, Nicholas J. Humfrey, and Michael Smethurst. Programmes ontology. Online ontology, September 2009. Available at `http://purl.org/ontology/po/`. Last accessed April 2010.
12. Leo Sauermann and Richard Cyganiak. Cool uris for the semantic web. W3C Interest Group Note, March 2008. `http://www.w3.org/TR/cooluris/`. Last accessed June 2008.