Opening up the BBC's data to the Web

A position paper for the April 2013 "*Open Data on the Web*" workshop

Sofia Angeletou, Michael Smethurst and Jeremy Tarling
With input from Oliver Bartlett, Tristan Ferne and Yves Raimond.
Edited by Olivier Thereaux

A short history of Open Data at the BBC

The BBC is one of the largest public service broadcasters in the world, with a mission to enrich people's lives with programmes and services that inform, educate and entertain. We have been using the web since 1994, not only to advertise and catalogue our programmes, but also as a leading news source and a platform for our audience to learn about history, the natural world, and many other topics.

It has been almost 10 years since we began work on PIPs (programme information pages), an effort to create a Web page for every radio programme we broadcast.

Thus began our approach of using one page (one URL) per thing and one thing per page (URL). It felt like the beginning of a new era. We called it the age of point-at-things.

Shortly after, in 2006, work began on /programmes, a replacement for PIPs covering both radio and TV. Around the same time we bought — in bulk — copies of Eric Evan's "Domain Driven Design" which influenced the way we designed and built websites to expose more of the domain model to users. Building on the recent Backstage initiative, we also added data views (JSON, XML, YAML, etc.) to /programmes.

The following years saw us work on a number of similar efforts for nature, food, and music. When building /music, we published data about our music news and programme, but also started using open (meta)data from MusicBrainz.

Using music metadata and artist identifiers from MusicBrainz allowed us not only to enrich the experience on our music site - it enabled us to link up some of our web silos and create journeys between /programmes and /music. Another example of our engagement with the LOD cloud is the usage of GeoNames identifiers to support our Weather site and most recently the Olympic Venues for 2012.

Following the Domain Driven Design approach has been very much about the possibility of dynamic aggregations around content, which let us maximize exposure to our content by placing it in different contexts (wildlife, music, food, football, etc.).

Because bbc.co.uk has content in so many domains, it tends to be like a microcosm of the web. One of our goals with this work was to move from a set of silo'ed sites to a coherent service which we can only do if our content is well described and interlinked.

Finally, by using domain-native URL keys we could generate more inbound link density and make our content more findable on search engines.

Current Approach

In the past few years, our interest in open data has evolved from the initial goal of breaking silos to a more complex array of objectives. We still consider open data to be a powerful tool to link and reach across sites and boundaries, either on the world wide web or in the microcosm of a large organisation.

A significant amount of effort is also spent extracting and creating data from some of the large media archives which the BBC has accumulated through the years. Ultimately, we believe that beyond data, meaning is where the value creation lies.

How do we detect a story from the always moving, always fluctuating stream of data from the web? What is our role in the growing LOD community? How can we encourage story-driven and data-driven journalism? How can we make large, noisy, ill-documented archives into usable and enjoyable sources of discovery?

Some of our recent work attemps to answer some of the questions above. Projects in the area include:

Linked Data Platform

The Linked Data Platform is now extending the linked data services in the BBC beyond sport coverage to more domains of interest e.g., News, Learning, Programmes, Music, e.t.c. Offering the technology (quadstore, tagging, aggregation and data management services) and policies on linked data in the BBC.

The LDP has the unique opportunity of serving as a centralised repository of metadata about all of the BBC's content that currently resides in various (and some times legacy) applications across different parts of the organisation. To this end we feel the responsibility of opening up this content to the LOD community and the public in useful and meaningful ways.

News and Open Data

The BBC News website publishes hundreds of articles every day, produced using a flat page-publishing system with no underlying data model. We are working on replacing this system with a linked-data driven model.

Initially we are extracting the concepts (people, organisations, places and themes) mentioned in the article at publishing time, and creating triples to state the relationship between the article and the concepts. We will then publish aggregation pages showing the most recent News articles about that concept, and link to them from the article page. Articles and aggregation pages will carry embedded RDFa (schema.org/rNews) to express these statements as linked open data.

We are also collaborating with the Press Association, The Guardian and Google on developing an open model for describing stories and events in News. Starting with the Event Ontology we are exploring how News organisations move from basic factual reporting to narrative development and story telling.

As well as facilitating the sort of concept-based aggregations mentioned above we hope this model can facilitate story/data-driven journalism, unlocking the data in our back-catalogue of millions of BBC news articles to aid future content development.

Generating and opening data from large media archives

We have also been working on opening a large audio archive (BBC World Service) on the web. The project started by processing a large audio archive to automatically assign topics to each radio programme, in an efficient manner.

The second phase of the project sees us generating open data about the programmes by combining the noisy automated transcripts of the audio with linked data identifiers in order to generate key topics, then combining this with a crowdsourcing experiment to validate the quality of the automatically generated data.

The result is published as an open archive.

Issues to consider

Both our long experience in this area and our recent efforts have raised a number of yet unanswered questions, challenges and issues. Among them:

- How do we build a compelling business case for publishing open data so that we can secure the appropriate resources?
- How do we decide on abstract yet extensible enough models for key concepts (e.g., people, places, events, etc.) which we can use to link content together and open it up to the cloud?
- How can we make the most of existing LOD datasets (freebase, wikidata) to help us improve our internal production processes? How to select the highest quality datasets that we can reuse for this?
- How do we track the provenance of datasets originating from internal systems and how does our approach translate outside the organisation where policies cannot be enforced in the same manner?
- How do we move from a model of publishing open data (linked data about our content) to being able to publish open content (where some rights reside with us but some with external content providers)?
- Who really benefits when we publish large volumes of data? There is a nagging concern among open data publishers that we may just be giving away our livelihood to search engines which could ultimately make the original websites irrelevant, unless attribution is done properly.
- What is the best way to define shared models and identifiers on a per-industry basis? Our efforts in News show this is a long and painful process. Can we fix that?
- How do we measure the health of an open data community? A large portion of open data is community generated. What happens if the community gets bored or goes away? How can we make sure that does not happen?

We are looking forward to participating in the upcoming "Open Data on the Web" workshop. We would be delighted to share our experience and present our recent projects, and are certain that many more will be interested in several of the challenges we have identified as worthy of discussion in the workshop.