Biggest CC0 dataset release – how Europeana got there

Submitted on 21 Sep 2012 by

Last week, Europeana opened up its huge cultural dataset for re-use under the Creative Commons Zero Universal Public Domain Dedication. We're making a special announcement at the Open Knowledge Festival this Thursday. To mark the occasion, Europeana's Jon Purday and David Haskiya tell us how they got there both strategically and technically…

The strategic perspective The release of Europeana's dataset of 20 million records under the CC0 waiver has its origins in 2003 when the European Union issued a Directive on the re-use of public sector information. At that time, data created by Europe's libraries, museums and archives was exempt. However, that exemption began to make less sense as the potential importance and value of cultural datasets started to become apparent. In 2008 Europeana was launched, beginning the aggregation of all the metadata for cultural materials digitised by Europe's memory organisations. The building of the vast dataset happened in parallel to the rapidly increasing uptake of smartphones and tablets, and the development of apps and web services. Taken together, a clear opportunity was emerging for digital innovation. Recognition of this was signalled in the publication of the Digital Agenda for Europe in 2010 with the action line to 'open up public data resources for re-use'. In January 2011, that action was reinforced by The New Renaissance report (PDF) which said that 'Metadata related to digitised objects produced by cultural institutions should be widely and freely available for re-use'. A proposal for a revision of the Directive was adopted on 12 December 2011. The main changes that the Commission proposed were the introduction of a general rule that 'all documents held by public sector bodies will be re-usable for both commercial and non-commercial purposes, unless covered by the exceptions provided for in the Directive', and that libraries, museums and archives would come under the Directive for the first time. Against that political backdrop, Europeana has been helping memory organisations recognise the value of open data and prepare for its implementation. We've run 35 workshops around Europe, and published two White Papers, setting out the theoretical case for releasing open data, Knowledge=Information in Context by Professor Stefan Gradmann, and the business case for doing so, The Problem of the Yellow Milkmaid: a Business Model Perspective on Open Data (PDF) by Harry Verwayen, Martijn Arnoldus and Peter Kauffman. At first, there was resistance in the sector to releasing data under CC0. For some libraries, for example, sale of their bibliographic data represented a significant revenue stream. The British Library was the first such institution to move to open data, making key files available as XML, meanwhile continuing to sell the MARC21 versions. Some museums felt that the considerable curatorial scholarship that went into very extensive descriptions of unique treasures had particular value that belonged to the institution. The solution to this was to provide to Europeana only the data fields that they were happy to see re-used. Europeana has a Data Exchange Agreement with every one of its data providers. This agreement had to be revised to incorporate the terms of CC0, and signed again by every provider. This has taken considerable time, so it's a major breakthrough to have secured the agreement of over 2,000 data providers to a dataset of over 20 million records. **The technical perspective** The first steps we have taken to make Europeana metadata available for re-use is to create data dumps that we have published at data.europeana.eu. Segmented by datasets, the dumps are available to download at http://data.europeana.eu/download/2.0/ An overview of the datasets is available in a spreadsheet. It is possible to preview a dataset in the Europeana portal by using the following pattern in the search box: europeana_collectionName: "dataset name", e.g. europeana_collectionName: "03486_L_DE_BSBMunchen" or europeana_collectionName: "08602_Ag_EU_EFG_InstituoLuce" The nt and rdf sub-folders contain the files corresponding to each individual dataset, expressed using

the N-Triples and RDF/XML syntaxes for RDF, respectively. In both datasets the data model used is the Europeana Data Model (EDM). The "links"-folder contains links to other Linked Data sources. These links are the results of the semantic enrichment done by Europeana. Co-reference links to Linked Data services maintained by Europeana partners (e.g. SOCH, the Swedish cultural heritage aggregator) are also provided in this folder. We're also planning to offer access to these data sources via a SPARQL-node. This work has just begun and we'll make sure to announce when it is ready. In parallel to working with making Europeana metadata available as Linked Open Data paradigm, we're currently developing a new REST-style API. This new API will allow querying of the Europeana repository with responses in JSON-format and the returned metadata modelled in EDM. Future versions of the Europeana portal will be developed using the same back-end as this new API so it will always keep pace with our own portal's search capabilities. We hope to make this new API available in late autumn. Please contact us if you want access to our current API and to receive news as soon as we're ready with the new API.