Licensing Library and Authority Data Under CC0: The DNB Experience

Lars G. Svensson

Effective from July 1st, 2012, the German National Library (Deutsche Nationalbibliothek, DNB) publishes most of its data under an open license. In doing so, the DNB followed the examples of Europeana and many national libraries, e. g. the Spanish National Library and the British Library, only to mention a few of the early movers. The road we took to arrive there was not free from issues and controversial discussions. This paper starts with a description of the current business model, outlines the road we took to arrive there and finally what are the plans for the future.

Current business model

The data the DNB curates is of two kinds: bibliographic information and authority data. While the bibliographic information is data describing the about 28 million media resources stored in the library's stacks, the authority data describes the entities surrounding the book, such as persons (as authors or as subjects of scientific discourse), places, subject, geographic entities etc. Currently, the DNB publishes the entities from both of those information sets as structured data in three formats:

- MARC (in the flavours MARC 21 and MARC-XML) – a data format specific to the library domain – features complete bibliographic and authority records including all internal cross-links and external references
- DNB Casual – title data available in CSV and XML – is a structured format with all content presented as literals, i. e. there is no cross-linking and also no external references to related data
- RDF – as the linked data format – contains *almost* the complete information contained in the authority and bibliographic records, except for

information specifically geared toward libraries (the RDF data is seen as a way to connect to non-library organisations).

All data – i. e. bibliographic and authority data – published in DNB Casual and RDF is available under a CC0 public domain dedication license and thus open data. The same applies to authority data in MARC. For bibliographic data in MARC, the DNB uses a moving wall: This moving wall is based on the so-called "volume number" of the Deutsche Nationalbibliografie (the former number of the weekly index) and is currently set at December 31st, 2011. Data indexed after this date is available against a fee; data from earlier bibliographic years is available under CC0. An exceptions to this rule is title data from the so-called Series O (online publications) which is also available under CC0. It is anticipated that this moving wall will remain until mid-2015, after that all data in all formats will be under CC0.

Orthogonal to the data formats are the methods to access the data. For machine-to-machine bulk access, we offer search over SRU and Z39.50, data harvesting per OAI-PMH, the possibility to assemble personal datasets through the catalogue, customised packaging according to specific customer needs – generally made available per HTTP or FTP download –, and finally the use in a linked data context per URI dereferencing and bulk download. Single metadata records in MARC are also available for download over the catalogue user interface.

For the use of all data access interfaces except for URI dereferencing and bulk download of RDF dumps, it is necessary to supply credentials when accessing the interfaces. This means, that even if the data is open data under CC0, we require data consumers to register with our Digital Services department – registration is free – who will create a customer account. From the DNB point of view, we do this in order to build customer relations. When we know who our customers are, we have better possibilities to interact with them and to adapt our services to their needs. Also we can contact them directly and provide information about maintenance leading to the systems not being available.

The road to CC0

Until autumn 2010, all data created and maintained by the DNB was licensed against a fee. Data production and packaging required human interaction, particularly since we lacked the user interfaces enabling end users to create and download individual datasets themselves, and the cost model was seen as a way to regain some of the costs. The first initiative to move towards a less restrictive license model was taken in the context of the publication of the authority data as linked data. The use of Creative Commons licenses for cultural heritage data was a relatively recent movement with little experience what were the actual consequences of freeing up your (meta-)data. Two areas received particular interest: the loss of revenue and the so-called unwanted spillover effects (the fear that someone else could take my data for free and earn money with it). For an institution that in 2010 made approximately a three-quarter million euro per year from the sales of bibliographic and authority data, the loss of this income stream would indeed have a noticeable impact on the balance sheet. The other aspect – the unwanted spillover effects – is more difficult to grasp. In our case it was less fear for loss of revenue, but rather the uncertainty what happens with our data once we give it away; in a time where "management by the numbers" is prevailing particularly the unability to directly measure use and impact of the data was an issue.

The first attempt to apply a more lenient licensing model was to create an own license based partly on the British Crown License and partly on CC BY-SA and to apply it to the authority data (in all formats) and the free-of-charge data formats for the bibliographic data. It soon became obvious that this was not a viable solution. On the one side it was difficult to explain what was the precise difference between the CC license and the custom DNB license (we did not forbid the commercial re-use per se, but wanted commercial entities to register with us). On the other side the open data community (e. g. the OKFN) made it explicit that data provided under this license could not be considered "open data" in its true sense which rendered the information unusable for re-use e. g. in the Wikipedia.

The result of the discussion was to re-consider the complete licensing framework deployed by the DNB. When doing this, we profited very much from the discussion around the new Europeana Data Echange Agreement which the Europeana Foundation adopted in 2011. The DNB participated in several workshops and together with other suppliers of cultural heritage data we could reach agreement that the unwanted spillover effects most likely were negligible. The other question regarding the loss of revenue was solved by introducing the moving wall which will disappear until 2015, thus enabling us to phase out that business model step by step (the fees become reduced annually) and find other revenue streams. On July 1st 2012, the DNB could announce the use of CC0 for all data except the two last bibliographic years in the MARC formats.

Lessons learned

Licensing is a difficult business. Particularly the use of a non-commercial clause causes problems for cultural heritage institutions since commercial use is hard to define and some of the key co-operation partners (e. g. Wikipedia) require that the data be free for commercial re-use. Institutions are also advised to use standardised licence (e. g. from the CC family) and not to create their own license models (and specifically not to extend CC licenses with custom clauses), since this will most likely cause confusion among licensees. The use of CC0 has had a very positive echo and has rendered us some very good publicity. The downside of the license is that it is very hard to track who uses your data. The requirement that institutions using our digital services need to register gives us some indication of that and the possibility to contact them and ask what they use the data for (this is particularly interesting for customers from outside of the library domain). When it comes to the use of our data as linked open data – however – we depend on anecdotal evidence or the customers contacting us when they have questions. We use any possibility we get to identify and interact with those customers, but we still have the feeling that we only know the tip of the iceberg.

Further, we need to explain to our board what we are doing , why we are prepared to accept the loss of revenue and how we intend to compensate that loss. In our case, we eventually were convinced that the market for library data would vanish anyway within the next few years. And if the open licensing makes people use our data and build new applications on top of it, but who would never have dreamed of paying for it, we definitely would say that it has paid off.