Date submitted: 04/08/2010

Step one: blow up the silo! - Open bibliographic data, the first step towards Linked Open Data

Patrick Danowski
Senior Expert Information Services
IST Austria
3400 Klosterneuburg, Austria
E-mail: patrick.danowski@ist.ac.at

Meeting: **149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

More and more libraries starting semantic web projects. The question about the license of the data is not discussed or the discussion is deferred to the end of project. in this paper is discussed why the question of the license is so important in context of the semantic web that is should be one of the first aspects in a semantic web project. Also it will be shown why a public domain weaver is the only solution that fulfill the the special requirements of the semantic web and that guaranties the reuseablitly of semantic library data for a sustainability of the projects.

In this paper I will describe why the discussion about licenses should be a first part of any semantic data project. Also this can help to do early an step in a semantic project, e.g. to publish the non-semantic raw data to get an early feedback and to get in discussion with others.

Motivation

There are three steps, in my opinion, if libraries like to be successful in the semantic web:

1. Library data must be presented in a broad (not only by libraries) supported semantic format
2. Library data must be reusable
3. Library data is reused by other projects

The first aspect of formats won‘t be cover in this text. The other two aspects will be in the main focus. First I like to discuss the question why the questions of the license should be step one. In my personal view we can directly stop the whole discussion about semantic data and all the technical aspects if our is goal is not that library data should be reused by

others. For the communication in the „library only world“ we have perfectly working exchange formats. The big improvement of the semantic web is tat this data can be understood by others. If reuse by other is the big goal of going semantic, one of our first question should be how we approach this goal. This can be still the question about formats but even more important is the question what others are allowed to do with this data.

The context of the Open Access movement

Today library data is hidden away in databases where most times even deep links to a dataset are not possible. The data is hidden in the deep web, closed in a silo. Even if libraries have already an tradition of sharing data most times it is combined with a fee, that is most times described as „small“. In Europe the database as whole (or an important part of it) is protected by copyright law. A dataset is a fact and as this not protected by copyright law. One of the problems of the semantic web is that the boarders of a dataset and a database are fading away. This makes a copyright discussion about data more difficult.

At the same time libraries are doing lobbying for Open Access since year. In the Berlin Declaration is written:

> „We define open access as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community.“[1]

Also libraries have started now to argue for Open Access on research data. Library data is part of cultural heritage and can be also seen as research data. This means open access to library data should be a natural step. Tim Berners Lee asked in is talk at the TED conference for „Raw data now!“[2]. This means that we can start without a format discussion to publish our data, but for that we have first to decide about a license.

The Options for a License

If we like to share data the classical copyright is certainly not an option. A little bit more difficulty are the different Creative Commons licenses. Very popular in the library world is the non-commercial license, but the impact of this license can be very tricky. A commercial search-engine would be not allowed to use this data (or to mix it up with other data). Which application commercial is and which is not is not that easy to decide. Imagine today the situation we wouldn`t allow google to index your webpage. I don‘t think that this is the way we like to go.
Also the other creative commons license can be tricky in the context of data. Just because they force all that you have to give attribute to the creator. Certainly this sounds nice and in context of texts and pictures the author should be always cited, but data is based on facts and the creator plays quite a less important role[3]. Also a data set from one resource can be mixed up with quite a lot dataset from different resources. In this case you have to quote all used datasets and the lists of used datasets can be longer than the result.

---

[1] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 22.10.2003 http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
[2] Berners-Lee, Tim: Tim Berners-Lee on the next Web, TED Talk, Februar 2009 http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
[3] exept if the issue of trust enters the discussion, but this discussion can fill an own paper

An other reason why you should think before a project about the right license is that it will affect on the whole design process of the project. Maybe later one if you decide to change the license is can be quite a big project like learned by wikipedia where it took years to change to the more modern Creative Commons license from the GFDL.[4]

This means for data the best solution is that there would be no copyright at all, or that data is published in the public domain. Public domain as a license is not possible in Europe because it is no possible to abandon all rights. There is already a solution for that. Instead of a license a public domain weaver like CC0[5] can be used to declare that all depending rights are cleared and there will be no rights claimed by the owner.

The Movement of Open Bibliographic Data

In January of 2010 CERN Library published there bibliographic data[6] in MARCXML under the Public Domain Dedication License[7] and CCO, with a clear statement for open access. Other libraries like the Ghent University Library[8] and a number of libraries from cologne, members of North Rhine-Westphalian Library Service followed[9]. More libraries are planing to join this movement during this year. As part of this movement libraries are able to show how serious they take Open Access for there own products. The idea behind this projects is just to put out data will put others in the situation to transfer the data are already to create something with this data. After CERN published the data Dan Brickley published a visualization of the used UDC classifications at CERN. This visualization showed that classifications the 400 area is already in use, but up to now the UDC haven‘t used the 400 classification area. This visualization helped easily to find potential errors in the data.

[4]Walsh, Josh: Wikimedia community approves license migration 21.05.2009 http://blog.wikimedia.org/2009/wikimedia-community-approves-license-migration/
[5] CC0 Waiver Text: http://creativecommons.org/choose/zero
[6] http://www.cern.ch/bookdata
[7] ODC Public Domain Dedication and Licence (PDDL) Text http://www.opendatacommons.org/licenses/pddl/1-0/
[8] Ghent University Library Exports http://lib.ugent.be/info/en/exports.shtml
[9] University and Public Library of Cologne (USB), the Library of the Academy of Media Arts Cologne, the University Library of the University of Applied Science of Cologne Library Centre of Rhineland-Palatinate, the Public Library of Cologne and the Central Library and Library of the German Sport University

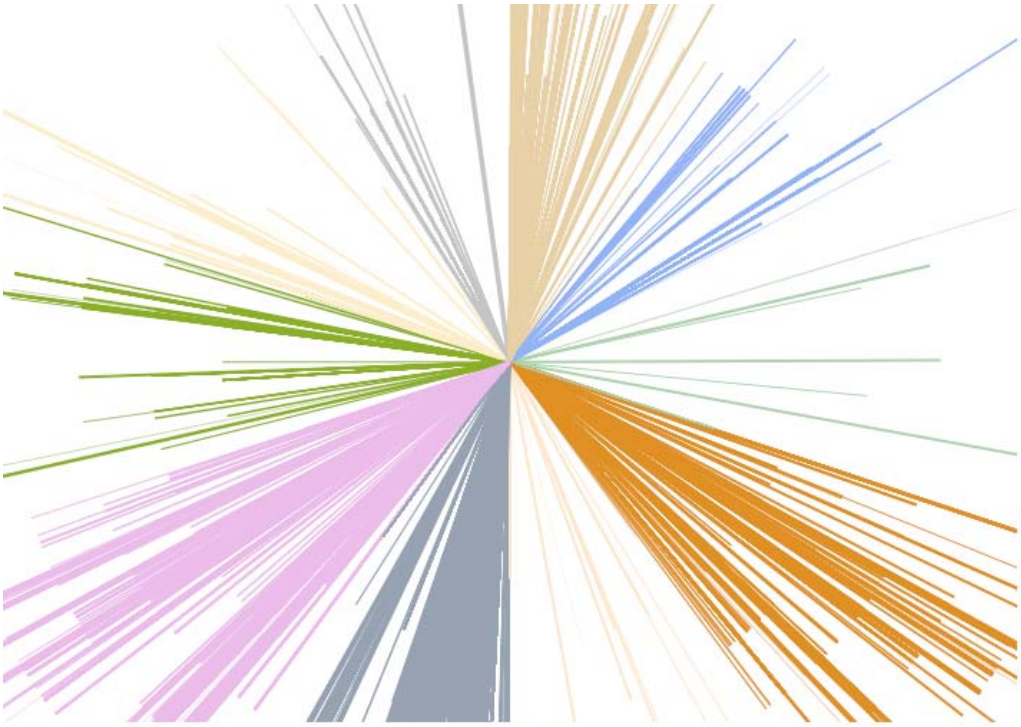

Figure 1: UDC Star by Dan Brickley http://www.flickr.com/photos/danbri/4326955233/sizes/o/ „1st cut viz of cern.ch/bookdata where each segment is from the core UDC decimal divisions.“

Openness of library data (not only bibliographic data, also classification systems and authority data) is the base that libraries can play an important role in the semantic web. In this area we can say „Open Access to library data is driving access to knowledge“[10]. Every limitation will stop different project to use our data which makes this resource less worthy. We should allow all projects and specially commercial project to extend and use our data as much as possible to create a rich public domain. Or how the WSIS say:

> „A rich public domain is an essential element for the growth of the Information Society, creating multiple benefits such as an educated public, new jobs, innovation, business opportunities, and the advancement of sciences.“[11]
> (WSIS Declaration of Principles, Section 26 12.12.2003)

[10] Presidential Thema of the IFLA President Ellen Tise (2008-2010): „Libraries Driving Access to Knowledge“
[11] http://www.itu.int/wsis/docs/geneva/official/dop.html