Open Access to Public Sector Information: Making Data Effectively Available

Melanie Dulong de Rosnay

Open Access to Public Sector Information : Making Data Effectively Available

Mélanie Dulong de Rosnay
ISCC/CNRS, 20 rue Berbier-du-Mets, 75013 Paris, France
melanieddr@gmail.com

Keywords: information science, copyright law, licenses, terms of use, metadata, e-infrastructure, interoperability, Creative Commons, governance, technical regulation, public domain, public policy, open access

1. Introduction

Implementing an effective open data policy implies making sure datasets will be available without economical, technical, social or legal restrictions for any purpose, including mash up and carry persistent semantic annotation so that others may continue to build upon augmented datasets. Technical and legal accessibility requires the use of open formats, the inclusion of information for reusability and licensing conditions which avoid legal interoperability bottleneck.
While numerous declarations[1] provide definitions and requirements for open data, technical accessibility has to be accompanied by institutional platforms. Public policies should be enacted by governments in order to make sure these platforms get populated by data, for instance by mandating data produced and digitalized with public funding to be uploaded in an open format.
This contribution proposes a technical, policy and legal framework which could be implemented in an integrated platform encompassing the complete workflow of public sector information from upload by the contribution to transformative successive reuses, allowing a persistent availability of data for citizens, researchers and innovators. Recommendations are proposed, building on incremental criterias which are being analyzed in each section of this article.

2. Technical accessibility public deposit mandate

2.1 Open format

According to common open licenses and users expectations, it is expected that the original data can technically be modified, not only to make derivatives, but also to transfer it to other medias or formats, or to reformat it to include it in broader collections, which requires delivery in a format that effectively enables reuse. In the same way fees and rights clearance burdensome process, technology can create barriers to access, redistribution and reuse of data. But technical choices can also help remove barriers. Technical accessibility should ensure that materials can be actually and effectively reused, mined, processed, aggregated, integrated, and searched by both humans and machines.
Technical barriers can include the following: protection measures that prevent copying, compulsory registration before download, design features that add hidden costs to search and automatic processing, complexity of all sorts prior to full accessibility of the content in a data format allowing any sort of processing. For example, it can be more or less easy to interact with a text document because of the publication format. HTML pages are more convenient to browse a large amount of articles compared to PDF files which require download. HTML and wiki allow comments and editing; two-column articles are difficult to read quickly on most screens but are the norm for scientific articles.
These technical restrictions, which go beyond making data available in an open format, have to be avoided, so that researchers and the public can not only access, but also redistribute and reuse data in any way, including ways that initial creators had not considered. Technical restrictions can be embedded in the design of the database. Technical restrictions affect databases that cannot be searched or processed in any possible way. Technical openness includes the possibility of downloading the whole dataset and reusing and integrating data, allowing machine-readable representation and interpretation of data, for instance through

1 For instance the open definition http://opendefinition.org/okd/, Science Commons Protocol for Implementing Open Access Data http://sciencecommons.org/projects/publishing/open-access-data-protocol/, or the GNU Free Documentation License article 1 providing a definition of open format crafted for textual software documentation towards a "Transparent" copy http://www.gnu.org/copyleft/fdl.html

Basic Local Alignment Search Tool (BLAST) to find similarities between sequences. Semantic web processing applied to data will improve the way science and evidence-based policy decisions are performed and will allow network effects by connecting knowledge from various datasets. Databases that require registration before access, or offer only a batch processing or a query-based mechanism to retrieve data after a specific search, do not comply with the technical requirements necessary to make data open. The website provides a file transfer protocol or a link to download the whole dataset without registration. The ability to download the whole dataset without registration constitutes the double requirement to be considered as technically accessible.

2.2 Informational context

Poor indexing and lack of metadata also prevent some modes of use, because the data will not be cognitively accessible. The presence of standardized fields for annotation, contextualisation or comments will allow users to contribute and better understand data collected by others across disciplines.
Building on the state of the art of open science, "community standards for sharing publication-related data and materials should flow from the general principle that the publication of scientific information is intended to move science forward. An author's obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also to provide them in a form on which other scientists can build with further research."[2]
Data must be provided in an intelligible language, which is not as straightforward as making sure to use accessible drafting in the case of other media types than text and numeric data fields. Taking the example of audio recording and music, having an audio file in a free and open format may not be sufficient to remix it. Instructions such as information enclosed in a MIDI file and other data such as music notation or explanation for performers could, when they exist, also be released. Some of these issues can be solved through project-oriented online communities which would encourage or require uploading complete project-files in addition to the media.[3] Community, disciplinary and cross-disciplinary guidelines could advise the contributor to release the data in a format suitable for manipulation and with the information necessary for its manipulation in a reasonable manner appropriate to the media.

Public sector could require that all data which is produced and digitalized with public funding should be uploaded in an open format, without technical barriers and with appropriate informational context to allow full reuse.

3. Legal accessibility

Many open licensing framework exist to distribute open data: Creative Commons licenses, CC0 protocol, OKFN databases licenses and finally terms of use and licenses which have been written by governments, for instance in the UK and in France. They all aim at making data as broadly available as possible, but nevertheless may contain hidden legal restrictions.

3.1 Contractual requirements on the reuser

Legal restrictions to redistribute and modify of data can be diverse, for instance among the Creative Commons provisions: Attribution, Non-Derivative Use, Non-Commercial Use, disclaimer of warranties.
The Attribution requirement may constitute a restriction on the reuse of data. Instead of strong contractually binding requirements on how data should be attributed, attribution could become optional and a request of acknowledgment according to best practices, at most, should be sufficient.
The Non-Commercial and Non-Derivative requirements prevent many types of data use. They are defined as restrictions based on the commercial nature of the user or of the usage, and as restrictions on the distribution of modified versions of the database.

The following checklist may assist data curators in opening their data, and to make sure that the database's design and terms of use will allow others to access, reuse and build upon their data. All

2 Board on Life Sciences (BLS), Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences (2003). http://books.nap.edu/books/0309088593/html/R1.html

3 Cheliotis Giorgos, "From open source to open content: Organization, licensing and decision processes in open cultural production", *Decision Support Systems*, Volume 47, Issue 3, June 2009, p. 229-244.

answers should be positive.

A. Check your database technical accessibility
A.1. Do you provide a link to download the whole database?
A.2. Is the dataset available in at least one standard format?
A.3. Do you provide comments and annotations fields allowing users to understand the data?

B. Check your database legal accessibility
B.1. Do you provide a policy expressing terms of use of your database?
B.2. Is the policy clearly indicated on your website?
B.3. Are the terms short and easy to understand by non-lawyers?
B.4. Does the policy authorize redistribution, reuse and modification without restrictions or contractual requirements on the user or the usage?
B.5. Is the attribution requirement at most as strong as the acknowledgment norms of your scientific community?

3.2 Open licenses interoperability issues

Open licenses have been designed to facilitate the use and reuse of data. However, the multiplication of licenses in order to answer to specific problems of correct bugs of previous licenses versions may lead to legal incompatibility, contrary to the purpose of all these licenses because of the differences among open licenses that have the same purpose, but use a different language. Attempts are being led to reach compatibility by accepting that derivative works and data may be re-licensed not only under the same license but also under licenses that will have been recognized compatible.
This risk coming from differences between licenses which are declared compatible is enhanced by copyleft or Share Alike clause allowing the relicensing of transformed data under another license declared compatible, such as Open Data Commons Open Database License (OdbL). A contractual issue could arise, as a contributor will be supposed to consent to the licensing of derivatives under conditions unknown as yet. Because of the Share Alike option transmission, a contributor is expected to consent to the Adaptation of his or her Work to be licensed under different, future, unidentified terms along the chain of derivatives and Licensees, and inconsistencies may grow exponentially after several generations. The problem is still theoretical, but a Licensor could sue a downstream Licensee who would still have respected the terms of the license received – only it was different from the license initially used by the Licensor.

Therefore, not only Attribution but also Share Alike provision contain risks and limit the effective openness of data. The easiest solution recommended is to place data in a legal status which will be as close as possible to the public domain. This will avoid requirements for reusers, and legal insecurity for the data providers, curators and users.

Developing a single European open data license drafted in plain language, and based on the public domain, will avoid transaction costs and interoperability issues between datasets licensed under different open licenses.

4. An integrated platform

4.1 Terms of use to secure contributions

Databases can be offered free of technical, informational and legal restrictions by the database curator, following above recommendations, but nevertheless be provided without warranties on the legal status of the data and successive transformations and annotations submitted by public authority and contributors. This situation creates another source of uncertainty and may jeopardize easy reuse of data for any purpose. Indeed, data may contain elements protected by copyright or any applicable right.
The database curator may not want to be endorse liability for copyright infringement. But a copyright warranties disclaimer can be seen as a hurdle to the usage of data. Both uncertainty for the reuser and absence of responsibility for the curator might be avoided by offering contributors a seamlessly integrated data sharing agreement prior to submission, requested from the contributors to assert that it is their contribution and to license it under the aforementioned conditions (public domain without technical or

informational restrictions). Although the procedure might disincentive some upstream contributors, the burden of checking the legal status of data and avoiding possible claims by downstream third parties should not rely on the data user, forcing her to hire a lawyer.

It is recommended to include in the deposit process a rights clearance declaration, placing the burden on the upstream contributor rather on the downstream reuser.

4.2 Metadata following derivatives reuse

Finally, I propose to accompany open content licensing with a technical framework facilitating the transmission of annotation and accompanying information along the lifecycle of data which is being processed, transformed, visualized and repurposed. An initial technical solution could be to better assist contributor by providing fields to host appropriate information as described in section 2.2 of this article. This task can be facilitated by applications that would automate the process for both (1) contributors, who when uploading data, should enter correct and complete data in the deposit interface which already contains fields for optional additional information, and (2) reusers when editing and redistributing modified data.
Based on Creative Commons licensing infrastructure, it is already possible for the licensor to include metadata about 1) the format and the title of the work, 2) the name users of the work should give attribution to, 3) the URL users of the work should link to, the source work URL and 4) an URL for additional permission. As an example, a simple specification of attribution elements[4], would help contributors to be attributed the way they are entitled to request, and to help licensees to respect these requirements. Then, attribution elements would follow the work along its life-cycle.
If this option was more widely used and further developed beyond creative works and repurposed for public sector information and scientific data, any contributor would be able to include proper information in the databases metadata, and further processing applications could help maintain the persistence and the update of metadata when they redistribute the work or reuse it otherwise.

In the context of data which do not require attribution as suggested in section 3.1, but which must be accompanied by information to help reuse as proposed in section 2.2, the deposit infrastructure for data should contain fields for metadata and comments which may follow the data transformation after transformation.

References

This contribution builds upon prior work of the author

Melanie Dulong de Rosnay, From free culture to open data: technical requirements for open access, in Danièle Bourcier, Pompeu Casanovas, Melanie Dulong de Rosnay, Catharina Maracke (eds.), *Intelligent Multimedia. Sharing Creative Works in a Digital World,* European Press Academic Publishing, Florence, June 2010, pp. 47-66.

Mélanie Dulong de Rosnay, Creative Commons Licenses Legal Pitfalls: Incompatibilities and Solutions, *study of the Institute for Information Law of the University of Amsterdam*, September 2010.

Alison Macdonald, Philip Lord, Damian Counsell, Melanie Dulong de Rosnay, Isabel Galina, Neil Beagrie Daphne Charles, Robert Beagrie, Pawel Plaszczak, Krzysztof Wilk, Pawel Jarosz, Paul Pillar, Richard Sinnott, European Study Towards a European eInfrastructure for eScience digital repositories, *study for the European Commission, DG Information Society and Media Unit F – GÉANT and e-Infrastructure, e-SciDR, Project reference no: 2006 S88-092641*, 2008.

4 Which can take the format of trackbacks and of RDF tags supported by the CC Rights Expression Language (ccREL) described at http://wiki.creativecommons.org/CcREL and in Hal Abelson, Ben Adida, Mike Linksvayer and Nathan Yergler, "CC REL: The Creative Commons Rights Expression Language" in Melanie Dulong de Rosnay, Juan Carlos De Martin, (eds.), *The Digital Public Domain. Foundations for an Open Culture*, Open Book Publishers, Cambridge, UK, March 2012.