Open Access to Scientific Data

Melanie Dulong de Rosnay
ISCC – CNRS
CERSA – CC France
COMMUNIA
melanieddr@gmail.com
@melanieddr

Centre for Research and Interdisciplinarity

CRI Fabelier Seminar

11/04/2012

The different forms of Open Access

- Economic OA
- Legal OA
- Technical OA

Budapest OA Initiative

- Literature
- Free and unrestricted online availability
- Economic, legal and technical barriers
- http://www.soros.org/openaccess/read.shtml

Open Knowledge Definition

Open Knowledge is material that others are free to access, reuse and redistribute

Open Knowledge Foundation

http://opendefinition.org/

Tools facilitating sharing and reuse

- CC licenses: The rights granted “may be exercised in all media and formats (and) include the right to make such modifications as are technically necessary to exercise the rights in other media and formats”
- Science Commons Biological Material Transfer Agreements (MTAs)
- Science Commons Protocol for Implementing Open Access Data: “allow massive-scale machine integration of data” with “the lowest possible transaction costs on users”

Science Commons Open Data Protocol

- Not a license
- A methodology for databases interoperability and marking data in the public domain
  - Simplicity
  - Absence of control
  - Attribution

- A project to assess and define openness for life science databases
- A taxonomy for legal and technical OA
- Methodology and concepts applicable to other domains?

Open Data: Freedom to Integrate

- Which freedoms and restrictions for open data?
- Life science data and the public domain
  - Access interface
  - Terms of use

Technical barriers

- Protection measures
- Registration
- Design
- Complexity of all sorts prior to full accessibility of the content in a data format allowing any sort of processing
  - Format
  - Metadata
  - Searchability

Results

- 20% of the databases are open
  - Molecular Biology Database hosted by the Nucleic Acids Research Journal
  - Life Science Resource Name (LSRN) Schema registry
- Submit your database
- User interface by Shirley Fung http://labs.creativecommons.org/demos/mbdb/

Molecular Biology Databases

Is the data really open?

Home Search Browse Submit Contact

Choose any of the options below to start:

Find Open Data

Browse a list of databases compliant with the Science Commons Open Access Data Protocol

Browse Policies

View databases categorized by their technical and legal accessibility regimes

Classified Databases

Find all the databases classified by the project. You also may want to use the "Browse Policies" section for a specific kinds of databases.

Submit Policy

Use the questionnaire to submit a database policy to our system

Why does this Web site exist?
This work is being developed under the auspices of the Science Commons Data project and builds upon the Science Commons Open Access Data Protocol proposing requirements for interoperability of scientific data. Legal simplicity and predictability can be achieved by waiving copyright and other

Results

Databases that are Science Commons Open Data Access Protocol Compliant

7 databases found.

The following list of databases meets the requirements listed in the Science Commons Open Access Data Protocol. The databases listed below are:

- In the public domain with a published terms of use policy, and
- Downloadable in whole without registration.

Click here to see databases in a detailed list

Acronym	Technical Accessibility				Legal Accessibility	
	Downloadable	Offers Batch Processing	Offers a Query Interface	No Registration Required	Policy is Available	Public Domain
AceView_WormGenes	✓			✓	✓	✓
AGI_LocusCode	✓			✓	✓	✓
CDD	✓	✓	✓	✓	✓	✓
COG_Cluster, COG_Function, COG_Pathway	✓			✓	✓	✓
dbEST	✓	✓		✓	✓	✓
dbSNP	✓	✓		✓	✓	✓
dbSTS	✓	✓		✓	✓	✓

AceView_WormGenes - AceView Worm Genome

Links	Direct Download Terms of Use
Terms of Use Summary	National Center for Biotechnology Information (NCBI) License
Technical Accessibility	Downloadable in Whole? YES Offers Batch Processing? NO Offers a Query Interface? NO No Registration Required? YES
Legal Accessibility	Policy available for use and redistribution of the database Public domain

AGI_LocusCode - Arabidopsis Genome Initiative (TAIR, TIGR, MIPS)

Links	Direct Download Terms of Use
Technical Accessibility	Downloadable in Whole? YES Offers Batch Processing? NO Offers a Query Interface? NO No Registration Required? YES
Legal Accessibility	Policy available for use and redistribution of the database Public domain
Terms of Use Summary	Most information at this site is in the public domain. Unless stated otherwise, documents and files on TAIR Web servers can be freely downloaded and reproduced. However, you may encounter documents or portions of documents that were contributed by private companies and other organizations, who may retain all rights to publish or reproduce these documents or to allow others to do so. Some documents available from this server may be protected under the U. S. and foreign copyright laws. Permission to reproduce these documents may be required.

1. DOWNLOADABILITY
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. TECHNICAL RESTRICTION: the database can be accessed only through registration, batch or query-based system.
Technical accessibility is not achieved.

3. PUBLIC DOMAIN POLICY: the website provides simple and clear terms of use informing users that the data are in the public domain.
Data are thus free to integrate. Legal accessibility is achieved.

4. NO POLICY: the website does not provide terms of use.
Legal accessibility is not achieved.

5. LEGAL RESTRICTIONS: the terms of use impose contractual restrictions, such as heavy contractual requirements for attribution, limitation to non-commercial usages, prohibition to modify data, or other constraints on their redistribution or modification.
Legal accessibility is not achieved. The data are not free to integrate.

Figure 2. Databases qualification

Checklist to assess databases openness

- *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?

Can these requirements be transposed outside life science databases?

- Software
  - GNU-GFDL: release of the source code & documentation in an open format
- Music: what is open media?
  - ogg? mp3? MIDI? separate tracks? music notation or explanation? Media + project-files?

Which regulation(s)?

- A clause in the licenses? Is a format requirement too specific?
- Guidelines?
  - Games, comics, information literacy
- Community usages or social norms?
- Platforms technical design?
  - Format
  - But also attribution
  - And maybe other issues raised by reuse

The Polar Information Commons, a collective resource

- Appropriate behavior
  - Set of guidelines: ethics and norms for contributors and users
  - Acknowledge authorship and recognition
  - Incentive to contribute high-quality material: documentation, errors, PD & rights clearance

http://www.polarcommons.org/ethics-and-norms-of-data-sh

Biodiversity and Agricultural Research data: institutional policy

- What data? What technical requirements? What pitfalls? What incentive?
- Plant Genetic, but also policy and data on development, climate change, “aesthetic, cultural, ecological, economic, educational, environmental, genetic, medical, recreational”
- The CGIAR Collections: “the CGIAR Centres are responsible for conserving the samples and making them available without restriction on the understanding that no intellectual property protection is to be applied to the material.”

A geographic dataset...

- Raw data
- Data sheet
- Model, taxonomy, ontology, structure
- Database
  - not all of them are protected
  - copyright for compilation
  - EU Directive on sui generis rights
  - Access-control mecanisms
  - Licence or terms of use

Licenses for geodata

- Geoscience Australia: CC BY 2.5 Australia
- data.gov.uk: Open Government licence v1.0
- Open Street Map: migration from CC BY SA 2.0 to the Open Database licence
- IGN Belgium: ARR

IGN

- Les cartes ne peuvent être reproduites ou adaptées par quelque procédé que ce soit (scannage, réimpression, emploi sur Internet) pour être publiées et/ou diffusées, sans accord préalable de l'IGN.
- Aucune partie substantielle du contenu de la base de données géographiques (numériques ou analogiques) ne peut être réutilisée ou saisie, sans accord préalable de l'IGN.
- La saisie répétée et systématique et/ou la réutilisation de parties non substantielles du contenu de la base de données géographiques sont interdites, si elles sont contraires à une exploitation normale de la base de données géographiques ou si elles portent atteinte aux intérêts légitimes de l'IGN.

NC

- belief that the monetization of reserved usages will bring additional funding to an expensive digitization process
- uncertain
- transaction costs may deter initiatives
- and be higher than royalties
- deserves further research
- the public domain generates value through its reuse for further creation, research, education, private consultation or commercial use

More requirements in the BY attribution clause than the FAQs and the law

•Section 4(c)If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Ssection 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

•If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

UK Open Government License

You must, where you do any of the above:

acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence;

If the Information Provider does not provide a specific attribution statement, or if you are using Information from several Information Providers and multiple attributions are not practical in your product or application, you may consider using the following:

UK OGL

Contains public sector information licensed under the Open Government Licence v1.0.

* ensure that you do not use the Information in a way that suggests any official status or that the Information Provider endorses you or your use of the Information;

* ensure that you do not mislead others or misrepresent the Information or its source;

* ensure that your use of the Information does not breach the Data Protection Act 1998 or the Privacy and Electronic Communications (EC Directive) Regulations 2003.

Open Street Map

- migration
- interoperability
- consent problem
- OdbL

Open licence Database

- A Share Alike licence for data and databases
- Collective database
- Derivative database under same licence
  - Means a database based upon the Database, and includes any translation, adaptation, arrangement, modification, or any other alteration of the Database or of a Substantial part of the Contents. This includes, but is not limited to, Extracting or Re-utilising the whole or a Substantial part of the Contents in a new Database.

OdbL

- Produced work
  - a work (such as an image, audiovisual material, text, or sounds) resulting from using the whole or a Substantial part of the Contents (via a search or other query) from this Database, a Derivative Database, or this Database as part of a Collective Database.

OdbL

- Share Alike
  - Any Derivative Database that You Publicly Use must be only under the terms of:
  - i. This License;
  - ii. A later version of this License similar in spirit to this License; or
  - iii. A compatible license.

Share Alike

- A legal bottleneck
- What is a compatible licence
- Small differences
- Consent problem
- Expectations

Collaborate with scientists and OA actors

- Identify barriers, needs and incentives to sharing
- A comparative approach
  - Adaptation to domains, cultures, disciplines
- Define appropriate regulations
  - Diversity of communities norms and usages

- Develop models for legal and technical interoperability
  - Infrastructure
  - Public policies, ethics and institutional policies
  - Support by technology and pedagogy

What are data? What is reuse?

- Mash up
- Visualisation
- Derivative
- Translation
- Integration of databases
- ?

Thanks a lot!