开放数据政策：框架与实践

Open Data Policy : Framework and Practice

报告人：顾立平 （Alan Ku）

时间：20-09-2017

地点：兰州市

前言Foreword

- 在2016年科学大数据国际培训课程中，我们以【开放数据政策与实践】为名，为CODATA发展中国家青年人才计划进行两小时的培训课程。培训课件在http://ir.las.ac.cn/handle/12502/8778 实施开放获取。**In the 2016 International Training Workshop on Big Data for Science, we have a two hours training course with title "Open Data Policy and Best Practice" for CODATA developing country young talent program. The course materials is open access at** http://ir.las.ac.cn/handle/12502/8778

前言Foreword

- 在两周之前，我们以【开放科研数据研究与实践】为名，为中国科学院研究所图书馆和信息中心进行半小时的课程。课件在 http://ir.las.ac.cn/handle/12502/9420 实施开放获取。**Before two weeks, we have a half of hour lecture with title “Research and Practice on Open Research Data” for library and/or information center of institutes where in Chinese Academy of Sciences. The lecture materials is open access at** http://ir.las.ac.cn/handle/12502/9420

前言Foreword

- 在一个月之后（10月），我们将以【开放科研数据的权益管理与政策建议】为名，在第六届中国开放获取推介周上进行演讲和专题讨论。After one month(October), we will present and discuss on the topic of “Rights Management and Policy Recommondation of Open Research Data” in the 6th China Open Access Week.

前言Foreword

- 在两个月之后（11月），我们将在国际开放获取知识库培训班上，进行系统性的开放数据管理服务课程。After two months (November), we will make a systematic training course of open data management service in the international open access repository course..

前言Foreword

- 这四次报告的内容都不一样，但是核心主旨相同，就是如何更好地促进我国的开放数据实践。The contents of these four presentations are different, but the core theme is on same, which is how to better promote China's open data practices.

- 感谢这次和大家交流的机会！Thanks for this opportunity to communicate with you!

前言Foreword

- 首先，请让我们一起回顾一下两个核心定义和一个概念。To Beginning, please Let's review the two core definitions and one concept.
- 稍后，它会带出今天报告的三个问题。Later, it will brings three issues in this presentation today.

前言Foreword

- 开放的定义：如果任何人都可以自由地访问、使用、修改和共享它---主题，甚至，保存出处和开放性的措施--- 那么知识就是开放的了。
- **The definition of open: Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.**

前言Foreword

- 开放数据的定义：开放数据是任何人都可以自由使用、重复使用以及重新散播的数据 --- 最多只有署名以及以相同方式共享的要求。
- **The definition of open data: Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike.**

前言Foreword

- 开放数据基于数据共享的实践经验但不完全等于数据共享，开放数据主要关注： **Open Data is based on data sharing practice, but it is not equal to data sharing. Its concern are:**
  - **发现数据的能力The ability to find the data.**
  - **使用数据的能力The ability to use the data.**
  - **重新调整数据的能力The ability to repurpose the data..**

前言Foreword

- 因此，开放数据实践上有三个关键议题：

  Therefore，there are three key issues of open data on practice as

  - 数据难以发现

    Data is hard to find

  - 数据难以使用

    Data cannot readily be used

  - 许可授权缺乏标准

    Licensing is lacking standards

《The State of Open Government Data in 2017》
《Open Data Management》

前言Foreword

- 如果我们想使公开数据成为现实，关键是如何让人们觉得开放的数据存放的是高质量的数据对他们有好处。If we want to make open data be real, the crux is how let people feel that the open data what people deposited the high quality data has benefit to them.

大纲Outline

- 数据难以发现

  Data is hard to find
- 数据难以使用

  Data cannot readily be used
- 许可授权缺乏标准

  Licensing is lacking standards

问题Problem

- 人们仍然需要从网络上许多不同的地方来拼接数据。**People still need to check many different places on the web to stitch data together.**

方案Solution

- 了解哪些数据可以相互关联（基于数据标准、咨询或者类似过程）。Understand what data can be related to one another (either based on data standards, consultations, or similar processes).
- 对于特殊数据集的附加文件的数据予以标识。Highlight the data where to find any additional document that is related to a specific dataset.

方案Solution

- 创建全国性发布开放数据的门户网站；数据可以托管在一个可以为每个人访问数据的平台上。**Create a data portal to publish open data nationwide. The data can be hosted on a platform where the data is accessible for everyone.**

问题Problem

- 数据可能隐藏在网站的深处，链接的名称既没有意义也没有自我解释。**Data may be hidden deep in websites and names of links are neither meaningful nor self-explaining.**

方案Solution

- 使用良好标注的链接，使得数据浏览更加直观。如果开放数据不能放在主页上，需要确保用户可以直观地点击网站。Use well-labelled links to make browsing for data more intuitive. If open data cannot be placed on a homepage, it is necessary to make sure that users can intuitively click through a website.

问题Problem

- 糟糕的命名或网站索引迫使用户不断查询尝试。**Bad naming or website indexing forces users to experiment with queries.**

方案Solution

- 给出可以让搜索引擎和人们理解的数据文件命名。Give data files comprehensible names that the names matter for search engines and humans.
- 标记数据文件，以便用户在使用搜索引擎时不需要知道文件的确切名称。Tag data files so that users do not need to know the exact name of a file when using a search engine.

问题Problem

- URL是不是永久性的，导致没有内容或者链接失效。URLs are not permanent and lead to empty or broken websites.

方案Solution

- 促使数据永久可访问。数据应在稳定的网络站点上，无限期地提供访问，并且尽可能保持稳定一致的数据格式。**Make data permanently accessible. Data should be made available at a stable Internet location indefinitely and in a stable data format for as long as possible.**

大纲Outline

- 数据难以发现

  Open Data is hard to find
- 数据难以使用

  Data cannot readily be used
- 许可授权缺乏标准

  Licensing is lacking standards

问题Problem

- 因为在机构内部产生数据的程序和优先级不同，数据变得不可用或者不被出版商视为关键数据。**Data is not available or hasn't been considered by publishers to be crucial data because there are different routines and priorities to produce data inside the institution.**

方案Solution

- 理解和思考数据的生产链。检查是否有机构常规或部门指南，以说明数据是如何产生和传递的。**Understand and rethink data production chains. Check if there are institutional routines or sectoral guidelines that inform how data is produced and communicated.**

问题Problem

- 数据可以根据研究团体的相关标准发布，但这些数据可能与主要用户的需求不一致。**Data may be published according to relevance criteria of research group, but these might not align with the needs of primary user groups.**

方案Solution

- 理解用户需求。这可以通过焦点团体或者发展【型人】来实现；型人可以理解数据需求以及开放数据的风险。**Understand user needs. This may be achieved by running focus group sessions or developing user personas; User personas enable to understand data needs as well as the risks of opening data.**

方案Solution

- 如果您不知道什么是【型人】的话，您可以参考某人的学术著作或者他的博士论文，在 http://ir.las.ac.cn/handle/12502/1514 开放获取。

  If you don't know what means “Personas”, you may read someone's book or his doctoral dissertation that open access at http://ir.las.ac.cn/handle/12502/1514 .

问题Problem

- 已发布的数据有多种形式，从地图到预算互动图表。其中还可以帮助非专家更容易理解数据，但是它们没有显示原始数据。**The published data in many forms, from cadastral maps, to interactive bubble charts of budgets. Some of these help non-experts to make sense of data more easily, but they do not show raw data.**

方案Solution

- 发布的原始数据要精确和准确。在可能的范围内，所释放的数据是初始的、未修改的形式的，链接到相关指南、文件、档案、可视化或者分析的数据。Publish raw data that is accurate and precise. To the extent possible, release data in its original, unmodified form, and link data to any relevant guidance, documentation, visualisations, or analyses.

问题Problem

- 数据以PDF或HTML格式发布，而不是以机器可读的格式发布，人们难以进行大数据分析方面的应用。**The data is published in PDF or HTML, and not in machine readable formats, people is hard to use it in the big data analysis application aspect.**

方案Solution

- 确保数据能够处理。原始数据必须以机器可读格式发布，这些格式需要具有一致的键值，使得可以通过检查丢失的数据值和相似性来验证数据的一致性。Ensure that data is processable. Raw data must be published in machine-readable formats, which need to have consistent values. This also to verify consistency of data by checking for missing data values and alike.

问题Problem

- 数据集不容易理解，缺少有用的上下文。文件以难以置信的方式命名，或者具有不可理解的结构。**Datasets aren’t easy to understand and lack more context to be useful. Files are named in implausible ways, or have an incomprehensible structure.**

方案Solution

- 任何特殊术语都需要使用元数据（关于数据的数据）进行解释。电子表格不该自我解释，尤其是如果它们包含日常语言中不使用的表达式。**Any special terminology needs explanation by using metadata (data about data). Spreadsheets are not self-explaining especially if they contain expressions that are not used in daily language.**

方案Solution

- 添加元数据以确保数据可以被人们理解。在任何情况下，元数据应该是机器可读和容易找到的。元数据必须在数据源附近发布，并清楚地引用一段数据。**Add metadata to ensure that data can be understood by people. In any case, metadata should be machine-readable and easily findable. Metadata must be published close to a data source and clearly refer to a piece of data.**

大纲Outline

- 数据难以发现

  Open Data is hard to find
- 数据难以使用

  Data cannot readily be used
- 许可授权缺乏标准

  Licensing is lacking standards

问题Problem

- 不清楚著作权保护是否适用于数据。**It is unclear whether copyright protection applies to data or not.**

方案Solution

- 有必要申报数据和/或数据集是否属于知识产权保护范围，使用指南或咨询科技信息政策小组。**It is necessary to declare the data and/or dataset whether fall under the scope of intellectual property (IP) protection by using guideline or consult the scientific and technological information policy teams.**

问题Problem

- 项目负责人和科研人员选择不属于【开放定义】的许可条款或者选择还未正式认可具有开放性的许可条款。**Principle Investigator and researchers choose license terms that do not fall under the 【Open Definition】 or are not officially acknowledged as being open.**

方案Solution

- 推荐使用标准化的开放许可证。开放数据许可协议和知识共享许可协议容易理解，应该是第一选择。这类许可协议提供彼此之间能够达到互操作的协调一致的条款。**Recommondation to use standardised open licenses. Open data licenses or Creative Common are easily understandable and should be the first choice. Those licenses provides conformant terms that are interoperable with one another.**

问题Problem

- 许可协议不完全梳理那些数据的适用性。**The license does not entirely clarify what data it applies to.**

方案Solution

- 准确地指出许可授权下的数据，并在提供数据时，提供一个时间戳。 **Exactly pinpoint within the license what data it refers to and provide a timestamp when the data has been provided.**

问题Problem

- 许可条款可能不会立即发现，或是发表在不同的网页与数据没有关联。**License terms may not be immediately findable, or are published on a different webpage that is not linked to the dataset.**

方案Solution

- 明确开放许可细节下的数据。标明许可授权的版本，以及提供数据可以怎么使用的上下文背景信息。**Clearly publish open licensing details next to the data. Highlight the license version and provide context how data can be used.**

方案Solution

- 维持许可协议的链接，以便用户可以随时访问许可条款。它也有助于“在”数据中具备一项许可通知。Maintain the links to licenses so that users can access license terms at all times. It also helps to have a license notice 'in' the data.

问题Problem

- 混杂一起的信息，包括在网址上的著作权信息和数据存储的平台上的信息。**There are mixed messages about copyright in the sites and platforms where data is stored.**

方案Solution

- 重新评估页面设计，以及避免在网站注脚的著作权声明、免责声明和使用条款的混乱和矛盾。**Re-evaluate the web design and avoid confusing and contradictory copyright notices in website footers, as well as disclaimers and terms of use..**

问题Problem

- 部分附加条款的编写方式会让用户感到困惑。**The way some additional clauses are written can be confusing for the user.**

方案Solution

- 只要有可能，就尽力避免那些没有在标准许可协议中出现过的限制性条款。**Whenever possible, avoid restrictive clauses that are not included in standard licenses.**

结语Foreword

- 开放数据政策、框架和实践，不可避免的议题：

  Unavoidable issueas of open data policy, frame and practice as

  - 数据难以发现

    Data is hard to find

  - 数据难以使用

    Data cannot readily be used

  - 许可授权缺乏标准

    Licensing is lacking standards

结语Foreword

- 人们不会把数据资产交给一个没有数据政策的知识库。People don't give data assets to a repository without data policies.
- 人们不会理会那种毫不考虑人性的数据政策。People ignore that kind of data policy without regard to human nature.
- 然而，人们需要开放数据。However, people need open data.
- 这就是奋斗的理由。That's why we fight.

结语Foreword

- 方法、工具、技能，甚至资金，都很重要；

  但是如果您失去人们的信任，就失去一切。

- **Methods, Tools, skills, and even money are important,**

  However, if you lost people trust, you lose everything.

附录Appendix

- 数据治理政策：分布式大数据的知识资源整合与知识加值服务的产权问题与对策（PPT+Video）

  http://ir.las.ac.cn/handle/12502/8645
- 全球开放数据政策与知识共享的模组成形与运转观察（PPT+Video）

  http://ir.las.ac.cn/handle/12502/8779
- 数据权益：跨越开放获取的边界（PPT+Video）

  http://ir.las.ac.cn/handle/12502/8551
- 科学数据开放共享的权益政策问题与基础设施需求（PPT+Video）

  http://ir.las.ac.cn/handle/12502/7163

附录Appendix

- 科研模式变革与科研数据管理（PPT+Sound）
  http://ir.las.ac.cn/handle/12502/9331
- 数据资产管理的法律与政策问题（PPT+Sound）
  http://ir.las.ac.cn/handle/12502/9126
- 国际出版商的文本和数据挖掘政策（PPT+Sound）
  http://ir.las.ac.cn/handle/12502/9093
- 科学数据管理与数据资源建设中的权益问题与解决方案（PPT+Sound）http://ir.las.ac.cn/handle/12502/7060

附录Appendix

- 国家科技信息政策研究与服务的介绍（PPT+Video）
  http://ir.las.ac.cn/handle/12502/9406
- 分布式大数据知识服务的数据资产管理问题与对策（PPT+Video）
  http://ir.las.ac.cn/handle/12502/9389
- 科研教育机构及其图书馆的合理使用政策（PPT+Sound）
  http://ir.las.ac.cn/handle/12502/9095
- 科技信息政策研究咨询与服务（PPT+Sound）
  http://ir.las.ac.cn/handle/12502/9109

附录Appendix

- The open search.org in open science era- A communication platform for everyone building their repositories and using others （ PPT+Sound/English ）
  http://ir.las.ac.cn/handle/12502/9385
- Exploring data librarian development in open science（PPT+Video/English）
  http://ir.las.ac.cn/handle/12502/8741
- Repository Network in China（PPT+Sound/English）
  http://ir.las.ac.cn/handle/12502/9099
- Country update: Institutional Repository in China（PPT+Sound/English）
  http://ir.las.ac.cn/handle/12502/8918

- 希望大家能够加入这里的两个社群：

- 中国数据馆员交流群（CDLG）
  - 数据科学
  - 数据政策

- 中国机构知识库学术交流群（CIRC）
  - 开放获取
  - 合理使用

诚邀大家一起努力：

Welcome to

第六届中国开放获取推介周

6th Chinese Institutional Repository Conference

25-26 October 2017, National Science Library CAS

第五届中国机构知识库学术研讨会

5th Chinese Institutional Repository Conference

21-22 November 2017, National Agriculture Library.

谢谢聆听
欢迎联系

联系邮箱sipc@mail.las.ac.cn

办公电话10-62537995