http://dx.doi.org/10.18243/eon/2016.9.11.7

ARTICLE

开放数据：资料库和政策

Meredith Morovati (ORCID 0000-0002-5383-7643)
Dryad 执行总监
director@datadryad.org

我很开心可以为你们提供一些有关发表数据的背景资料，还有解释这个议题值得你留意的原因。我自 2014 年的秋天开始担任 Dryad 的执行总监。我的背景是学会和董事会管理，但我也具有出版领域的工作经验，主要是在学术出版业的期刊部门。当我还在出版界工作的时候，开放获取（Open Access, OA）还未形成一个商业模式。在我离开这行一段时间后，OA 期刊才开始出现，很快的到了 2016 年，出版商组织迎来了它的第一百位会员，而 OA 期刊的潮流则继续快速增长。其中包括大型期刊例如 PLOS、Scientific Reports 和 PeerJ。十年的时间可以带来很多转变。

不知不觉间，OA 由「新的潮流」变成真正的商业模式。我不知道这个转变是由于某个特定的时间点，还是一个由新颖和值得注意的东西到逐步逐渐为大部分人接受，缓慢转变的结果。从这方面来看，开放数据的发展有一点不一样，除了在某些例子和领域中开放数据经历跟 OA 一样的转变。

开放研究数据是将科学研究背后的数据让任何希望取得数据的人获得，数据是科学的核心，如果我们可以将研究再现，那么一个显而易见的先决条件就是我们要评估某研究。另一方面，开放研究数据的运动是希望解决一个我们深知存在已久的问题：就是私下和同行分享数据不是一个好方法。如果你读了 David Crotty 最近在 Scholarly Kitchen 发表的文章，你可能对动画发出会心微笑，因为动画反映了一个研究员在希望获取数据时经常面对的悲哀经历。

数据政策

另一个分析这个问题的角度是以更严谨的目光去审视和比较开放数据和 OA 出版之间的明显分别。如果一篇学术文章不能以开放方式获取，你也许还可以透过图书馆和付费获得。可是如果数据并没有正式收藏在档案库内，那么我们可能永远失去数据。试想象一下，这是一个对科学来说很严重的问题。如果数据没有被归档或跟文章连结，那么每年我们会损失 17% 科学成果基于的数据。正是这个原因促使联合数据收藏政策（Joint Data Archiving Policy, JDAP）的始创者创立组织，以下一段简单，清晰的段落说明：

> 作为出版的条件之一，[期刊] 需要要求支持文章结论的数据应该被储存在一个合适的公共资料库中，例如 [认可资料库的清单]。数据是科研的重要产品，它们应该被保留，而且在以后几十年可以被使用。作者可以选择在发表文章时让公众取得数据，或者，如果资料库的技术容许，作者可以选择数据在文章出版后一年才让公众取得数据。编辑可以批准例外的例子，尤其是敏感的资料，例如实验对象的资料和濒危动物的地点。

JDAP 是由一群编辑在 2011 年以联合和协调的方式在演化和生命科学领域形成

的。结果是没有任何人需要承受不必要的惩罚 — 因为期刊有一致的政策，所以鼓励了作者们去遵守原则。在 2011 年的时候，有些专门和以社群形式运作的资料库切合了当时的部分需要。Dryad 的成立是为了照顾那些在传统资料库范围以外的数据，并继续支持社群。自此以后，Dryad 专注在连结和支持学术文献的数据，而 Dryad 奉行筛选数据的做法并以开放方式提供数据。时至今日，在这个政策实施不过五年的时间，开放数据成为很多演化和生命科学领域甚至以外的学科感到骄傲的一点。五年时间可以成就的变化真大。

今天资助机构开始有开放数据的要求，数据处理政策在很多期刊和大型出版商也变得常见。通过大型出版商例如 Springer Nature 和 Wiley 的数据共享服务，这个政策得以传播。 Elsevier 甚至买下一个初创资料库，Mendeley 数据，作为其首选的资料库，从而保障 Elsevier 发表的文章数据。而结果是？如果你没有类似的数据政策，你就落后了。

那么，一个数据政策包含什么？而你该如何创立数据政策呢？这个问题有很多个答案，有些人也开始指出不同的期刊和刊物之间没有政策标准。我的推测是来自例如 Springer Nature 的标准政策最终会成为常态。可是，有些领域总会有特别的需要。例如，考古学和博物馆研究有很多在研究过程中被破坏的样本，因此，研究方法和笔记可能是最需要保留的重要物件。

一个政策包含一些很基本的要素，例如，很多期刊要求研究员在提交稿件时将数据归档。在实际操作上，作者需要在提交稿件时表明数据的存储位置。通常需要保留的数据是支持文章理据的资料，而不是所有收集的数据。而且通常开放数据比非公开数据较容易取得。Dryad 几乎将我们所有所有的数据以知识共享 CC0 形式发表，其中有很多原因。可是主要的原因是 CC0 形式不会影响数据的法律地位，因为在大部分国家，数据本身不在版权保护的范围内。而且，以 CC0 形式发表数据不代表你不需要引用数据出处。

引用数据

越来越多编辑在关注于引用数据的做法，这是出版商和编辑很容易就可以参与的面向。引用数据原则联合声明（Joint Declaration of Data Citation Principles）受到很多编辑拥护，而工作仍在持续进行中，目标是为作者提供更多有关如何遵守这些原则的指示。尽管如此，这些原则的主要宗旨包括：

- 重要性；
- 嘉许和归因；
- 证据；
- 特别辨识；
- 取得性；
- 坚持；
- 特别性和确认性

根据这些原则，数据本身被看作是重要的学术成果，值得被嘉许和支持文章的理据。数据需要持续和独特的辨认码，因为这些东西的寿命比单一作者和文章长，而且其中应该包含人们可以理解和机器可以读取的元数据。

事实上，Crossref 建议论文本身的参考资料中应该引用数据，这做法通常称作自我引用。此外，其他地方的引用可以参考这些范例，它们跟编辑每天已在处理的文章引用和引用来源很相似。不过你仍然需要小心处理，确保你不会剔除看似跟其他一般引用不同的资料。

采纳开放数据的做法跟围绕这 OA 发生的其他改变很相似。虽然做法的部分细节可能看似很技术性又很复杂，成立一个基本政策和严谨引用审查可以作为编辑和出版社的起始点。