长尾数据共享研究进展

杨平 田野

[摘要]:长尾数据是一种重要的科研资源，由于缺乏关注度与技术支持，它的利用价值被长期忽视。文章简要概括其定义、属性以及重要性等。从壁垒与对策研究、基础架构研究、用户行为研究、图书馆与图书馆员责任能力研究、有机共享研究五个方面梳理长尾数据共享理论研究现状。此外，基于数据生命周期理论归纳促进长尾数据共享的 5 种常用管理工具，包括 DMP Tool、DataUp、EZID、Merritt repository、数据出版平台。最后，总结长尾数据共享所面临的社会和技术障碍以及相应的对策建议并提出未来的研究建议。

[关键词]:e-science 长尾数据 数据监护 共享因素 数据管理工具

[分类号]:G250

1. 背景

1.1 科研长尾理论

长尾(The Long Tail) 理论是 2004 年 C.Anderson 在给《连线》杂志的文章中首次使用的词汇, 专门用来描述网络环境下的某种经济模式如 Amazon.com 或 Netflix。长尾这一术语普遍使用于统计学中。长尾理论[1]的基本原理是:只要存储和流通的渠道足够大,需求不旺或销量不佳的产品所共同占据的市场份额可以和那些少数热销的产品所占据的市场份额相匹敌甚至更大，即众多小市场汇聚成可与主流大市场相匹敌的市场能量。

P.B.Heidorn[2]在 2008 年第一次将长尾理论应用到科学数据中，他指出，如果所有的科研项目排列在一条轴线上，而且沿着轴线由大至小的方式排列这些科研项目，那么由几十个或更多的研究人员所参与的非常大的科研项目在轴线的左侧，一些较小的项目将按照规模递减排序到右侧，整个曲线右侧的主要部分就是科研的长尾，如图 1 所示。与轴线左侧的数据不同，科研长尾上的数据通常是趋向于异质的、个人管理的或是未被管理的，未被保存的，一般是有条件获取的或者是受保护的科学数据。一项对美国国家科学基金会（National Science Foundation）2007 年的所有超过 500 美元的资助项目的研究发现[3]，80%的资金是用于资助 100 万美元以下的项目的，他同时还假设，如果每一美元的科研投入产生出恒定数量的科学数据，那么就说明大部分的科学数据是那些处于科研长尾上的项目所产生的。研究同时发现，这些处于科研长尾上的数据比较难获取并且很少被重用或者保存。

大多数研究人员所进行的项目都是相对较小的项目，主要是在一个人数相对固定的团队里进行的，团队通常是由一个研究员领导，组员的构成可能包括若干的助理研究员以及一些研究生等。研究人员的日常科研活动会产生大量的科学数据，在 P.B.Heidorn[2]的一项调研中发现近 80%的科学研究活动都是处在科研的长尾上面的，它们往往是规模较小、成本较低的科研项目。S.Carlson[4]也认为通常来说“小科学”产生的数据比大科学更多。与那些规模巨大、人数众多、投资庞大、并有相当大社会影响力的大科学相比，“小科学”中每个项目更倾向于单学科性的、人数较少、投入较小，但属于更前沿，更创新的研究。处于科研长尾上的项目数量众多，所以其研究人员的数量非常多，这些项目横跨多个学科领域。目前已有的研究主要集中在那些规模比较大的科研项目中数据的收集、保存以及重用问题，而对于多数研究人员产生的数据即长尾数据的关注度还远远不够，长尾数据的流失情况较为严重。研究表明[5]，处于科研长尾上的小项目通常是科学创新的源头，这种情况在文献计量学中得到很好的验证[6][7]，即那些具有很高影响因子的论文不一定出现在高影响因子的期刊中。笔者通过文献调研发现，目前针对长尾数据共享的研究稀少与分散，且只有很少一部分研究明确关注科研的长尾，而国内则几乎没有专门涉足此方面的探讨。因此，笔者从长尾数据的属性与共

享的理论研究现状出发，调研长尾数据管理工具，分析长尾数据共享面临的障碍与解决对策，也为后人对其进一步研究打下基础。

1.2 长尾数据、小科学数据与灰色数据定义辨析

小科学[8]通常是指由单个或者一个小组的研究人员在指定项目上进行工作的科学，“小科学”中每个项目的特点包括单学科性的、人数较少、投入较小等，小科学项目产生的科学数据就被称为小科学数据。

灰色数据[2]是指那些不容易被潜在使用者所发现的数据，这些数据很难被研究人员进行长期的保存。无论是“大” 科学还是“小”科学，灰色数据都是存在的。

综上所述，长尾数据、小科学数据、灰色数据都是科学数据的一部分。长尾数据与小科学数据的定义基本一致。灰色数据不仅仅存在于“小科学”项目中，同时也会存在于大科学项目中，只是小科学项目中的灰色数据量更普遍一些。

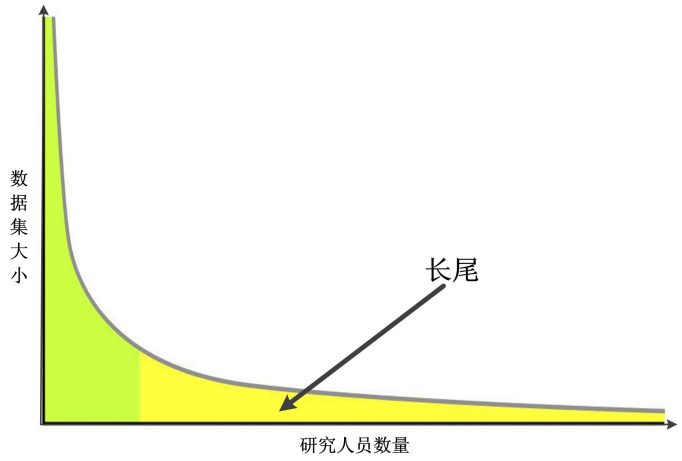

图 1 科研的长尾[9]

2.长尾数据的属性

正如 J.Porter[10]所言，那些在大项目中产生的数据趋向于更加同质，这类项目开始之前，研究人员一般都会事先设定产生的数据的格式、存储位置以及数据权限等诸多问题。大项目中的数据相对比较同质化的另一重要原因是，这些庞大的数据往往是使用一定设备进行机械化收集的，虽然各个项目之间使用的设备可能不太一样，但是同一个项目之间一般都是使用相同的设备，并且对于这些大项目的设备制造商来说，他们有能力来兼容不同的数据格式。另外，这些所产生的同质数据更容易被存储在结构化的数据库中，因此这类数据更容易访问、获取、维护、重用。此外，一般来说受政府资助或企业资助的大项目会更加注重该项目产出的数据，这样做的目的是为了确保这些项目有较大的影响力，相反，规模较小的项目没有客观环境的制约，因此所产生的数据往往都在研究人员个人的手中，很少被共享，很难被重用。关于长尾数据的属性的总结见表 2 所示。

表 1 长尾数据属性[10]

Head	**Tail**
同质的	异质的
集中监护	个人监护
有标准规范的仓储库	机构性仓储库或者没有
长期维护	很少或不维护
开放获取	有条件获取或受保护
随时可以重用	很少被重用

受到重视	目前很大程度上被忽视

3.长尾数据共享研究现状

长尾数据与小科学数据研究也受到了一些国际组织与科研院校的关注，2007 年，美国国家进化综合中心 (NESCent) 与北卡罗来纳大学元数据研究中心共同举办的 “数据保存、共享和发现—数字化时代的小科学挑战”研讨会[11]，共同探讨了小科学数据的属性以及当前保存、共享和发现小科学数据中面临的机遇与挑战；2013 年，有关小科学的国际会议在拉斯维加斯举行[12]，来自各个学科的研究人员探讨了小科学所取得的进展以及发展中的不足；同年，EarthCube[13]举办的研讨会探讨了大数据时代有关长尾数据共享的网络基础设施以及研究人员的数据共享的行为。

目前专门针对长尾数据的共享研究，国内尚未开展，国外的理论研究可以分为以下几部分：（1）壁垒与对策研究：代表学者是 P.BHeidorn[2]，他描述了长尾数据为什么是科学进步的关键、数据特性以及妥善管理这些数据的壁垒，最后从知识产权、投资机制、奖励机制、元数据标准和创建工具等 10 个方面提出了改善建议；他在另外一篇文章[14]中也论述了长尾的重要性并从技术层面、机构层面以及组织层面提出了长尾数据的解决方案；（2）针对特定学科的数据共享基础架构或网络架构研究：H.Onsrud 等[15]认为小科学数据是有着许多潜在价值，并且他们设计了一种基础结构，解决了重用许可、元数据生成、来源跟踪、归档以及同行评价等核心问题，让用户更方便的使用数据集；（3）数据共享用户行为研究：主要采用的是问卷调查法，进而基于调研报告得出结论，研究内容主要是影响某一领域研究人员数据共享的各种因素，研究对象主要是生物学、医学、生态学、地理学等领域的研究人员，如 H.A.Piwowar[16]等利用文献计量学分析了医学研究人员的数据共享行为与资助机构政策、期刊政策之间的关系；（4）图书馆、图书馆员与数据共享：P.B.Heidorn[17]在长尾数据的图书馆监护一文中探讨了图书馆与图书馆员在长尾数据共享中扮演的角色与承担的责任；（5）有机共享：探讨长尾数据共享的关联发布，Y.Gil[18]等建立了一个将科学文化习俗考虑在内的数据共享语义框架。

4.长尾数据管理工具

随着学术界越来越关注长尾数据的流失问题，一些科研资助机构、研究机构、期刊出版商等也发布了相关的政策，如美国国家科学基金会（NSF）[19]，其要求受到资助的个人必须提交一份数据管理计划（Data Management Plans），计划的内容包括产出的数据的类型、数量、存储方式、保存期限、访问权限等。为了使这一系列过程自动化处理，也为方便研究人员自身能够更好的对这些长尾数据进行监护，一些科研机构开发了一系列数据管理工具，这些工具可以全程对数据进行管理，也可以按照数据生命周期即数据收集—处理和分析——存储和共享——出版和重用等来进行分步骤分时段的管理，这其中目前最广泛使用的是由加利福尼亚数字图书馆[20]（California Digital Library）开发的 DMP Tool[21]、DataUp[22]、EZID[23]、Merritt repository[24]等，这些工具分别对应数据生命周期的几个不同的部分，如表 2 所示，

表 2 数据管理生命周期服务工具

工具名称	数据生命周期阶段	功能	受否开源
DMP Tool	资助申请	建立、编辑、共享和存储数据管理计划	是
DataUp	数据收集	利用 Excel 作为数据收集工具	是
EZID	数据管理和引用	创建和管理永久标识符	是
Merritt	存储和共享	作为一种数据仓储来储存、管理和共享数据	是

数据出版平台	出版	作为基础设施来发布数据以及确认学术优先权	

4.1 DMPTool

DMPTool 旨在帮助研究人员建立数据管理计划来符合一些资助机构的政策，注册该服务的机构可以得到定制化的服务，目前注册用户已经合计超过 450 个研究机构，总人数超过 2300 人次。DMPTool 为数据生命周期的其他阶段打下基础，其应用的数据类型可以是文本、表格、图像、音视频、3D 模型等。

4.2 DataUp

DataUp 对于科研长尾数据的收集具有很好的效果，作为数据收集的工具，其利用 EXCEL 来保存数据，进行监护，其功能包括标准化的标题、版本控制、自动归档并分配永久标示符等。DataUp 的应用流程是“上载表格数据→进行最佳实践审核→利用标准元数据描述数据→对该数据集生成带有永久标识符的引文→将文件存储在数据仓储库中”，DataUp 的最佳实践是指能够解析.xlsx 或者.csv 文件来检测是否存在潜在的文件，避免不必要的重复工作，其应用的数据类型主要是.xlsx 或.csv 格式的数据。

4.3 EZID

EZID 允许用户创建和管理永久标识符，其特点包括数据迁移能力强、文本或非文本资源都可以标识、不受时间地点限制等。对资源进行永久标识可以促进数据引用的规范，转而促进了数据的共享和重用，同时确保学术优先权。

4.4 Merritt

Merritt 作为数据仓储库，提供对科学数据的长期保存、共享和其他管理的功能，同时可以作为一个数据发现系统。其使用一种叫做“微服务（Micro-Services）”的设计范式，仓储的整体数据保存功能由一些小的独立的却具有高度互操作性的微服务组成。其特点包括永久储存、利用 URL 访问、共享方便以及应用程序接口友好等。Merritt 应用的数据类型包括文本、图像、音视频等各种数字对象类型。

5.长尾数据共享障碍和对策

造成长尾数据的共享困难的原因有很多，C.L.Borgman[25]认为，研究“小科学”中的科研人员的数据共享行为比大科学更为重要，“大科学”在数据管理上往往有着系统可靠的流程和相应的数据仓储。相比大科学，“小科学”往往没有任何统一的数据管理机制和相应的数据仓储来管理小项目中科研人员产出的不断增长的数据量，大部分的数据管理都由个人进行。关注长尾数据共享问题，有助于将数据共享的注意力转移到大多数研究人员正在进行的科研项目上，有利于确保科学研究的完整性和可重用性，有利于提高项目研究的影响力，减少研究人员的重复劳动并且加速科研创新。下面是笔者根据文献调研和一些调查问卷总结归纳出的目前科研长尾数据共享所面临的一些障碍。

5.1 缺乏学术报偿体系

目前的学术体系对于研究人员保存和共享科学数据没有任何激励机制，这点从很大程度上造成了长尾数据的流失。S.Lawrence[26]研究发现，没有共享的期刊论文的平均被引次数为 2.74 次，而共享后的论文的被引次数达到了 7.03 次。对于数据，很明显这也是可以适用的，但正是目前对于科学数据的计量研究开展的不够深入，也直接影响了研究人员对于数据共享后的学术报偿持一种观望的态度。

5.2 缺乏资金支持

长尾数据的监护（Data Curation）需要花费一定的成本，而目前的学术体系并没有给予研究人员足够的资金来维持这一科研活动。即便是一些学校或科研机构提供了这部分资金，但是这类资助一般只持续很短的时间。造成这种结果的原因可能是目前的学术体系并没有把

科学数据作为很有价值的知识资产（Knowledge Assets）。数据作为一种资产有着不可衡量的作用，特别是一些即时数据（Real-Time Data），比如气候数据等,他们的获取成本非常高，理应得到相应的保存和重用。

5.3 学科机构库建设不足

目前对于小科学产生的数据，专门性的学科仓储还远远不够。小科学项目种类千差万别，产出的数据非常的多，这些长尾数据因为一些客观原因不能够像大项目中的数据一样被很好的监护，因此需要建立更加细分的学科仓储来对此类数据进行学科性的集中监护。

5.4 缺乏培训

即便是一些研究人员有期望长期保存长尾数据的意向，也有相应的资金，但是不同于大项目中有专门的数据监护人员，研究人员自己由于专业知识的关系，他们缺乏专门的数据监护的能力，这些研究人员不了解机构仓储，也不熟悉相应的数据管理工具。如前所述，目前已经有一些比较流行的开源数据管理工具，比如 DMP Tool、DataUP、EZID service 等。

5.5 知识产权问题

一些研究人员认为共享他们的科学数据会带来一些知识产权（intellectual property rights）问题。随着开放运动的不断开展，研究人员应当逐步意识到知识产权和知识共享的区别，并且研究人员也可以采取逐步共享的做法，将数据共享问题分层次解决，首先实现从非正式共享到本团队、本项目组、本校之内的有权限的非正式共享，等到时机成熟或者制度规范明确的时候，再进一步推动非正式共享到正式共享的转变。

5.6 政策环境不明确

尽管目前有一些国家或地区性组织，甚至一些期刊都发布了论文的数据共享要求，比如向期刊投稿时必须向期刊编辑和同行评审专家提供相关的科学数据或者可以获得该涉及该研究的科学数据的第三方仓储库的存取号等方式，如 BMC Evolutionary Biology、PLOS One 等，但是执行力度的不统一，尤其是小科学项目中，由于没有细分的学科仓储，这些长尾数据的流失情况尤为严重。

5.7 没有统一的数据引用规范

一个较为规范的数据引用政策，可以促进数据集的合理引用、加快数据计量的发展，并且确保研究人员对该数据集的学术优先权，转而又促进了数据的共享，此过程是一个良性的循环。在大项目中，由于有专门的监护人员，所产出的数据可以按照一定的数据引用标准发布，比如 DataCite[27]、OECD[28]、DCC[29]等，而小科学中的长尾数据，由于较为分散，没有统一的引用标准，造成了在共享过程中诸多不便，渐渐这些数据很难被重用。

针对长尾数据在共享过程中面临的障碍，笔者进行总结后给出了一些建议的解决方案，如表 1 所示。

表 3 长尾数据共享障碍及解决方案

长尾数据共享障碍	建议解决方案
缺乏学术报偿体系	建立相应的数据计量体系
资金不足	对长尾数据长期保存提供专项资金
学科仓储建设不足	建立更加细化主题的学科仓储
缺乏培训	对机构仓储、数据管理工具等进行培训
知识产权	不共享到非正式共享到正式共享
政策环境不明确	加强政策层面的执行力度
数据引用规范不统一	规范引用标准、确认学术优先权

6.结语

长尾数据作为一种数量巨大，内容异质，呈现“小科学”特征，但受到的关注度却匮乏的

科学数据，在整个科研活动中发挥着重要的作用，它往往是科学创新的源头，与处于头部的数据不同，针对长尾数据共享的研究稀少、分散，共享现状也往往由于各种主客观因素而不容乐观，长尾数据管理工具的出现虽然可以很好的帮助研究人员进行数据管理与监护，但并不能从根本上解决共享问题，实现长尾数据的共享还需要政策、技术以及研究人员个人等各方面的支持，因此还有很长的路要走。对于今后的研究，笔者有以下几点建议：其一，将关注度从宽泛的数据共享浓缩到科研的长尾上，结合长尾数据的属性与国内外数据共享的研究成果，更加深入细致的探讨长尾数据共享的问题；其二，结合数据生命周期理论，将长尾数据共享研究从科学数据生命周期的下游（项目结束后，论文出版后）向上游（项目进行前和进行中，论文撰写中）拓展和延伸。不仅关注正式共享，更关注项目层面的非正式共享，为长尾数据共享研究提供新的思路；其三，由于各个学科产生的科学数据的类型、特性各不相同，可以将长尾数据共享的讨论纳入到学科分类体系中，对于各个学科的共享问题进行单独探讨。

参考文献

[1] Long tail[EB/OL].[2014-03-12].http://en.wikipedia.org/wiki/Long_tail

[2]Heidorn PB. Shedding light on the dark data in the long tail of science[J].Library Trends,2008,57(2):280-299.

[3]Peck SL.Science Suffers when Getting a Grant Becomes the Goal[J].Chronicle of Higher Education,2008, 55(7).

[4]Carlson S. Lost in a sea of science data[J]. The Chronicle of Higher Education, 2006, 52(42):A35.

[5]Edinburgh University.Data Library Research Data Management Handbook[EB/OL].[2014-03-16].
http://www.docs.is.ed.ac.uk/docs/data-library/EUDL_RDM_Handbook.pdf.

[6]Seglen PO.Why the impact factor of journals should not be used for evaluating research[J].BMJ:British Medical Journal,1997,314(7079):498.

[7]Sun B, Mitra P, Giles C L, et al.Topic segmentation with shared topic detection and alignment of multiple documents[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 199-206.

[8]Cragin M H, Palmer C L, Carlson J R, et al.Data sharing, small science and institutional repositories[J]. Philosophical Transactions of the Royal Society A:Mathematical, Physical and Engineering Sciences,2010, 368(1926): 4023-4038.

[9]Carly Strasser, John Kunze,Trisha Cruse.Big Data's Long Tail[EB/OL].[2014-04-06].
http://pt.slideshare.net/jakkbl/big-datas-long-tail

[10]PorterJ.Ecological data: Design, management and processing[M].John Wiley & Sons:Hoboken,2009.

[11]Data Preservation, Sharing, and Discovery: Challenges for Small Science in the Digital Era[EB/OL][2014-03-10].http://wiki.datadryad.org/Workshop_May_2007.

[12]The International Conference on Small Science (ICSS 2013) [EB/OL].[2014-03-15].http://www.icssci.org/

[13]EarthCube.Domain End-User Workshop: Engaging the Critical Zone community to bridge long tail science with big data[EB/OL][2014-03-10].
http://earthcube.ning.com/events/earthcube-domain-end-user-workshop-critical-zone

[14]Dark Data In the Long Tail of Science: Examples in Biology[EB/OL].[2014-3-10].
http://www.Slideshare.net/pbheidorn/dark-data-in-the-long-tail-of-science-examples-in-biology.

[15]Onsrud H J, Campbell J.Big opportunities in access to "Small Science" data[J].Data Science Journal,2007,6: 58-66.

[16]Piwowar H A, Day R S, Fridsma D B. Sharing detailed research data is associated with increased citation rate[J/OL]. PlOS One, 2007, 2(3)[2014-3-10].

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0000308#pone-0000308-g002
[17]Heidorn PB.Library curation of long-tail science data[EB/OL].[2014-4-6].
http://www.codata.org/10Conf/abstracts-presentations/Sessions%20J/J3/B.%20Heidorn.pdf
[18]Gil Y,Ratankar V,Hanson P.Organic data publishing: a novel approach to scientific data sharing[C]//Proceedings of the 2nd international workshop on linked science. Southern California:2012.
[19]National Science Foundation 2010."Scientists seeking NSF funding will soon be required to submit data management plans."
[20] CDL.[EB/OL].[2014-03-17].http://www.cdlib.org/
[21] DMP Tool.[EB/OL].[2014-03-18].https://dmp.cdlib.org/
[22] DataUp.[EB/OL].[2014-03-18].http://dataup.cdlib.org/
[23] EZID. [EB/OL].[2014-03-18].http://www.cdlib.org/uc3/ezid/
[24] merritt.[EB/OL].[2014-03-18].http://www.cdlib.org/uc3/merritt/
[25]Borgman CL.The digital future is now: A call to action for the humanities[J].Digital humanities quarterly,2009, 3(4):57-59
[26]Lawrence S.Free online availability substantially increases a paper's impact[J].Nature,2001,411(6837): 521-521.
[27]DataCite International Data Citation Metadata Working Group.DataCite metadata schema for the publication and citation of research data version3.0.[EB/OL].[2014-03-18].
http://schema.datacite.org/meta/kernel-2.1/doc/meta/kernel-3/meta/kernel-3/meta/kernel-2.2/index.html
[28]Green T.We need publishing standards for datasets and data tables[J].Learned Publishing,2009,22(4):325-327.
[29]Duke M,Ball A.How to Cite Datasets and Link to Publications:A Report of the Digital Curation Centre[C]. //In:23rd International CODATA Conference: Taipei,2012

Research of Long Tail Data Sharing

[Abstract]:Data in the long tail is an important research resource. Due to the lack of social attention and technical support , its use value was long neglected.This article briefly summarizes the definition, properties and importance of data in the long tail. Generalize the theoretical research of data sharing in the long tail from the research of barriers and countermeasures, web infrastructure,users'behavior,libraries and librarians'responsibilities,organic data sharing.Based on data lifecycle,summarize commom management tools of data sharing in the long tail， including DMP Tool,DataUp,EZID,Merritt repository,data publishing platform.Finally,the article discusses the social and technical barriers of data sharing in the long tail and then propose recommendations.

[Keywords]:e-science,long tail data ,curation,data ,sharing factor,data management tools