基于数据生命周期的图书馆科学数据服务研究

师荣华 [a,b]　刘细文 [a]

（a 中国科学院研究生院，北京 100190）

（b 中国科学院国家科学图书馆，北京 100190）

摘　要：数据生命周期是依据科研过程发展而来，从数据产生、加工到数据发布、再利用的一个循环过程。本文首先归纳了数据生命周期理论，在此基础上推演出 e-Science 环境下图书馆可以尝试开展的科学数据服务方式，划分为数据初次加工、数据再加工和知识抽取三类服务。并且，提供相应案例支持，考察了服务的实践开展情况。

关键字：科学数据　数据生命周期　服务方式

分类号：G250

Research on Scientific Data Services of Libraries Based on Data Life Cycle Theory

Shi Ronghua[ab]　Liu Xiwen[b]

(a Graduate School of CAS, Beijing, 100190)

(b National Science Library of CAS, Beijing, 100190)

Abstract: Data life cycle is developed along with the research life cycle, a full cycle from data creation, data processing, data dissemination to data repurposing. This paper concluded the data life cycle theory firstly, on the basis of which deduced the scientific data service modes the libraries can try to start under the e-Science environment, including three types of services: data first-processing, data reprocessing and knowledge extraction. Besides, the article provided relevant cases to survey the situation of data service activities.

Key words: scientific data　data life cycle　service modes

进入 21 世纪，e-Science 的产生改变了科研方式，科技创新越来越依赖于对海量数据的再利用。因此，如何融入 e-Science 环境满足科研人员的数据需求是图书馆界亟需探索的一个问题。国外已经有学者探讨 e-Research 中图书馆参与数据领域的角色定位问题，本文在总结前人观点基础上，利用数据生命周期模型推演了 e-Science 环境下图书馆可以开展的科学数据服务方式，并考察了各项服务的实际开展情况，辅以案例分析。

数据生命周期是指从数据产生，经数据加工和发布，最终实现数据再利用的一个循环过程，其实质是依据科研过程来管理数据。本文从来源、类型、基本流程、特色、实质等方面分析了各种数据生命周期理论，从中归纳出数据生命周期的一般基本流程，以此为指导思路探索了 e-Science 环境下图书馆可以开展的科学数据服务方式。

1 数据生命周期理论归纳

生命周期（life cycle）的概念源于生物领域，科学家描述了寄生物扁虱从一个宿主转换到另一宿主的生命周期过程，宿主为扁虱的整个生命周期提供支撑生存的环境。作为一种比喻，数字对象也可以看作扁虱，从一个数字加工环境到另一环境，最终生成数字产品供用户使用。数字生命周期（digital life cycle）的提法在 IASSIST（International Association for Social Science Information Services & Technology）2006 中多次出现，Ann Green 总结了各类数字生命周期理论，并讨论了数字化生命周期的内涵：首先，“生命周期”（life cycle）不同于“生命期”（life span），即从产生到消亡，一个生命周期意味着一种数据加工环境，经过数据管理和长期保存，实现资源发现和再利用。具体来说就是对数字化资源进行保存及长期保存、提供获取，最终用于支持研究、政策制定等再利用活动[1]。本文归纳了各派数据生命周期理论，从来源、基本流程、类型等方面进行列表对比，具体见表 1：

表 1：数字生命周期各派理论总结

模型	基本流程	特色	实质	来源	类型
DDI 3.0 的元数据创建模型[3]	研究课题→数据收集→数据处理→数据存档→数据发布→数据发现→数据分析→数据再利用→数据处理……	最简单、最基本	数据加工流程： ①数据加工 ②知识抽取	DDI [2]根据社会科学数据的处理加工流程来创建元数据	**数据管理流程**
e-Research 下的数据生命周期模型[4]	数据生产→数据管理→学术交流体系→增加附加值→知识抽取	涉及数据和知识层级，由低到高	e-Science 下数据利用过程： ①数据加工 ②知识抽取	Liz Lyon, UKOLN（UK Office for library Networking）	
DCC 数据保存生命周期模型[5]	核心：数据 ；第一层：数据描述；第二层：数据长期保存计划；第三层：团体活动参与；第四层：数据管理和长期保存； 最外层活动包括：创建或接收数据→评估和选择→数据传递到数据中心等→长期保存行动→存储→获取\再利用→转换	分层次； 全面、具体；	数据保存模型： ①数据加工	2004 年英国 JISC 和 e-Science 核心计划发起的数据管理项目	
知识创造的生命周期模型[6]	①研究课题→数据收集→数据处理→数据发布→数据发现→数据分析→知识传递循环 ②e-Science：数据再利用和数据发现	①系统描述知识传递循环；②增添 e-Science 环境下的数据处理活动	e-Science环境下的科研经历： ①数据加工 ②知识抽取 ③知识传递	Charles Humphrey，加拿大阿尔伯塔大学数据图书馆（The University of Alberta Data Library）	**科学研究流程**

从上述各派理论可以看出，科研生命周期是数据生命周期的来源，同时 e-Science 环境下一个完整的数据生命周期涉及数据加工和知识抽取两个层次，数据加工是知识抽取的基础。其中，数据加工的过程基本达成一致，各家基本都涉及数据收集、数据处理、数据发布、

数据发现等；另外，Liz Lyon 的 e-Research 下的数据生命周期模型中有增加附加值环节，实质上属于数据加工的高级阶段，即再加工；而各家基本都涉及的数据分析则是最高级的数据加工，即知识抽取阶段。

综上所述，从内容层面看，一个完整的数据生命周期包括数据加工和知识抽取两个层次，数据加工是知识抽取的基础；数据加工又包括数据初次加工和数据再加工，前者包括数据存储系列环节；后者则是在一次加工基础上增加附加值。在知识抽取方面，主要在数据获取基础上进行的一系列高级活动，包括数据挖掘等知识发现活动。

2 数据生命周期流程下的图书馆数据服务拓展

由前文总结可以看出，数据生命周期由科研周期发展而来，通过数据生命周期图书馆可以宏观把握科研人员的科学数据需求，结合自身实际开展服务。笔者也试图通过数据生命周期理论推演 e-Science 环境下图书馆可以开展的科学数据服务方式。在此之前，国外已经有很多学者、机构对 e-Science 环境下图书馆参与数据领域的角色定位进行了探索，这与科学数据服务方式的探索是异曲同工的，因此本文首先总结了已有研究，在此基础上提出自己的服务推演类型。

2.1 图书馆在数据服务领域的角色探索

Anna Gold提到科学研究生命周期理论和学术交流系统结合起来，很容易推导出数据和文献的生命周期流程，具体见图1[7]。

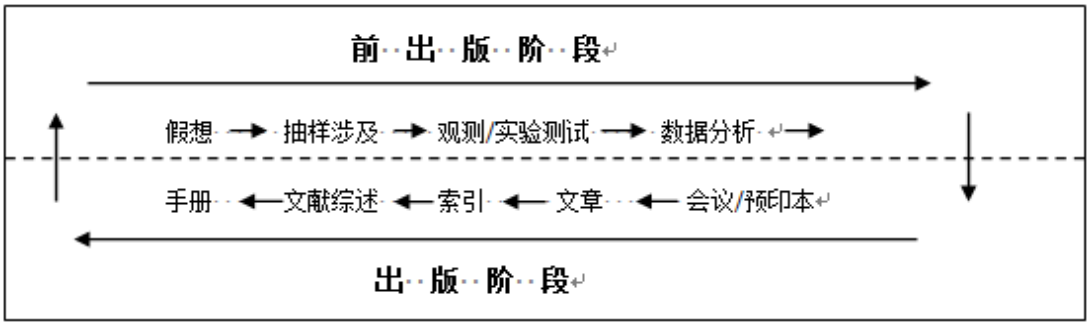

图1: 数据和文献生命周期

作者以此为线索探索了图书馆员在科学数据服务中可以担任的角色。在前出版阶段，主要角色是：①选择数据集并发放许可②制作元数据（或标准）描述数据集③数据保存服务④评审、挑选长期保存资源⑤协助用户数据发现⑥发展数据出版标准和系统⑦呼吁出台知识产权保护文件⑧建立学术成果储存库，如数据仓储。在后出版阶段，作者主张图书馆员要争取成为研究者的合作者，如参与创建数据管理原型等。

另外，Liz Lyon 也提到科学数据服务中涉及到的各主体的职责，其中，数据馆员可涉及的工作包括数据评估、数据长期保存、协调机构合作、宣传数据服务、发展标准等[8]。一家专门从事学术交流领域咨询的公司也谈到图书馆介入科学数据领域的几种方式：培训研究者的“数据意识”；数据存档和保存；培训和提供数据馆员[9]。Rick Luce 也提到图书馆在 e-Science 环境下参与数据领域可尝试的新角色包括：改变传统的文献和学术交流视野；数据存档、机构库；发现相关资源、数据保存、教育和培训等[10]。

由以上研究可以看出，国外图书馆界已经意识到在数据领域要参与新的分工，并开始将服务边界拓展到传统学术交流的上游即数据阶段。以上各家讨论的e-Science环境下图书馆在

数据领域可以尝试的角色存在一定重叠，例如数据保存、数据获取等，这一定程度上是由图书馆存储信息资源的传统和优势决定的；其他如发展标准、知识产权等则是个别学者提出，主要是因为目前数据服务还没有产生成熟模式，讨论这些问题缺乏一定的实践基础。

2.2 基于数据生命周期的数据服务拓展

上述学者在图书馆数据服务探索方面积累了一定成果的同时，也存在很多不足。例如，大部分学者仅发散地列出图书馆员在数据服务领域可以尝试的角色，缺乏理论支持和系统性，只有个别学者按照学术交流系统的流程展开讨论；另外，上述研究没有归纳为服务方式。因此，本文利用数据生命周期模型归纳了图书馆可以开展的科学数据服务方式。如图 **2** 所示：

由前文可知，一个完整的数据生命周期应该历经数据初次加工、数据再加工、知识抽取阶段。其中数据初次加工包含数据收集、数据描述、数据存储、数据获取等环节，实际就是数据存储服务，这本质上和图书馆的文献保存性质相似。数据再加工则是对已经存储的数据进行二次加工，以增加附加值；而知识抽取则是在数据加工基础上的服务升华。由此可见，从数据初次加工到数据加工再到知识抽取是一个由低到高的循环过程。笔者在各个服务模块下又划分为具体的服务方式，下文主要讨论了每项服务的内涵、开展情况，并辅以案例支持。

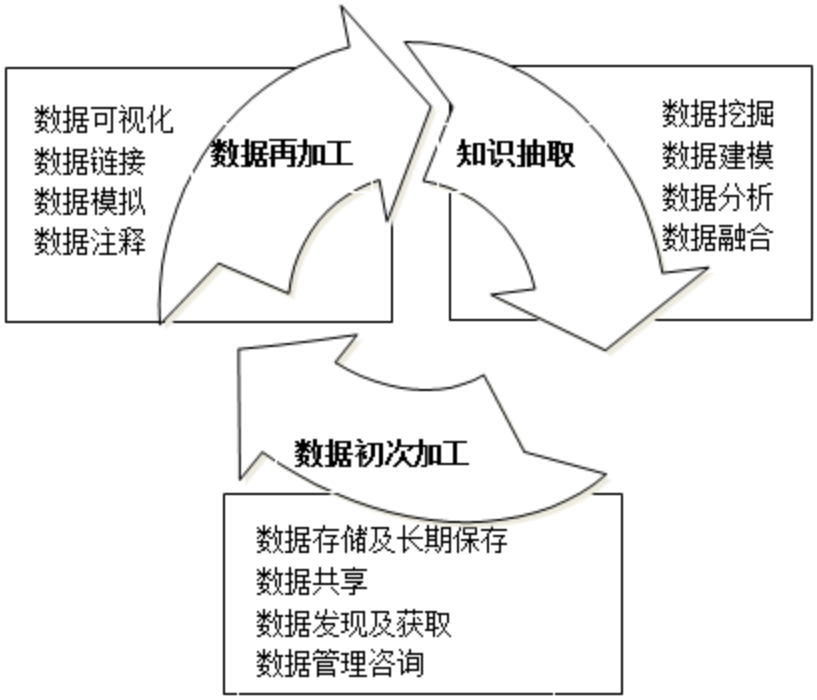

图 2：e-Science 环境下图书馆开展的科学数据服务方式

2.2.1 数据初次加工服务

数据初次加工的核心服务方式是数据存储服务，数据存储服务其实是数据管理（Curation）的一个重要环节。Curation 是指从数据的被生产出来起就开始的管理和促进其被利用的行为，目标是使得数据能够符合现实的需要，或能被用于发现和重用数据。[11] [12]。数据存储服务中包括的具体服务形式如数据存储及长期保存、数据发现及获取和数据管理咨询等。具体来讲，图书馆可以探索的服务项目可以包括以下几类：

①数据存储服务

一类服务是暂时性的数据存储，例如建立机构数据仓储，方便机构内部的数据共享，同

时有部分数据可以转移到更高一层的机构库中。例如康奈尔大学图书馆建立的 DataStaR [13] 就是一个临时的存储库，用户可以上传数据、选定特定同事进行数据共享、选择一个长久保存的机构库、数据出版等，支持小型研究团体的数据共享。普渡大学图书馆的 D2C2 [14]是一个分布式数据保存中心项目，其中 e-Data 作为数据管理服务的试验平台。e-Data 实现了对远程机构库的以及网格上数据集的分布式存取。其本地存储容量大概是 30Tb，图书馆员已经和各个领域的研究者进行合作收集数据。

②数据长期保存服务

另一类是永久性的数据存储。长期保存是一种基于存档的活动，数字保存需要解决的问题是即使随着时间的流逝、在技术已经变化了的情况下，还能够对文档的数据进行存取[12]。例如由美国航空航天局（NASA）1990 年开始着手建设的国家级分布式数据存档中心（DAAC's）[15]包含海量卫星观测数据，由于这些数据具有不可重复性，因此必须保证可以永久获取，这对数据长期保存提出了很高的要求。在处理技术方面，涉及在原有数据存储基础上增加一些长期保存活动，例如制定长期保存计划；数据评估来决定哪些数据需要长期保存；数据清洗、分配保存元数据、文件格式等等。

③数据发现及获取服务

帮助用户在海量信息中发现关联信息一直是图书馆的优势所在，类似于检索文献，e-Science 环境下图书馆员也可以开展数据发现服务，形式如数据检索、数据导航、集成融汇服务等。例如加拿大科技信息研究所（CISTI）就提供对加拿大科学、技术和医学数据（STM）的数据导航服务。它整理了加拿大范围的科学数据，进行分类整理、元数据描述，有些还提供科学数据库的链接等[16]。集成融汇方面，2006 年以来，中国科学院国家科学图书馆提出并开展了科学数据与科技文献跨界集成服务、数据融合技术的研究和开发[17]，利用数据 SRU 技术实现了科技文献、科学数据、字典等的服务融合[18]。

④ 数据管理咨询服务

除此之外，图书馆还可以提供数据管理咨询服务，数据馆员可以全程跟随科研项目，进行数据管理，从规划、收集到存档、发布的系列活动。另外，图书馆员也可以在存储格式、存储流程、标准等方面协助科研用户进行科学数据管理。国外的一些科学数据中心会跟随科研团队进行数据管理协助服务，例如美国 NASA 的地球观测实验室（EOL）[19]。麻省理工大学图书馆就提供社会科学数据、地理 GIS 数据以及生命科学数据的咨询服务[20]。

2.2.2 数据再加工服务

为了在更大范围内发挥数据的作用，图书馆可以对数据进行再加工，提供数据增值服务。例如数据可视化、文献和数据的链接等。在数据再加工服务方面，已经有图书馆开始探索数据增值的新形式，例如图书馆尝试给科学数据添加注释以及来源出处，实现了科学文献和科学数据的交叉链接。在实践方面，德国国家科技图书馆（TIB）就利用 DOI 系统，通过分配数据集数字对象唯一标识符，实现文献和科学数据的链接[21]。

数据加工的最高等级即知识抽取活动，包括例如数据挖掘、数据分析、数据融合等。知识抽取服务将是科学数据服务的未来发展方向。在初期，图书馆可以协助科研用户进行数据挖掘、数据融合等服务，并可以提供相应的数据分析软件等；在后期，图书馆员应该尝试和研究者进行合作，参与到科研的前出版过程，提供数据分析等服务。

3 结论

由前文研究可以得出，数据管理的生命周期源于科学研究的生命周期。数据生命周期实质是将传统学术交流的链条拓展到前出版时期的数据阶段，从数据产生、整理到数据发布和获取，拓展了图书馆的服务范畴，可以有效地指导图书馆开展科学数据服务。图书馆现在的服务主要围绕已经出版的文献资源，e-Science 环境下图书馆应该将服务链条拓展到上游数据处理阶段，并将知识服务作为未来服务发展的方向。

根据数据生命周期模型，本文将 e-Science 环境下图书馆可以开展的科学数据服务可以划分为两个层次：数据加工和知识抽取，其中前者可以划分为初次和二次数据加工；后者的发现方向是知识服务。数据初次加工主要依照数据管理生命周期思路展开，可以分为数据存储及长期保存、数据共享、数据发现及获取以及数据管理咨询等服务；数据再加工主要是前一阶段基础上的升级，具体分为数据再加工，例如数据可视化、增添数据链接、数据注释等；知识抽取形式如数据挖掘、数据建模、数据分析和数据融合等。

从实践开展情况来看，目前科学数据服务主要停留在数据初级利用阶段，部分图书馆开始探索数据再加工服务，通过增加附加值、充分发挥图书馆自身优势；而在高层的知识服务方面，服务实践比较少，还处于探索阶段。

参考文献：

[1] Anna Gold. Conceptualizing the Digital Life Cycle. [2010-7-13]. http://www.iassistdata.org/blog/conceptualizing-digital-life-cycle

[2] Data Documentation Initiative.[2010-7-20].http://www.ddialliance.org/what

[3] Overview of the DDI Version 3.0 Conceptual Model. [2010-7-10]. http://opendata foundation n.org/ddi/srg/Papers/DDIModel_v_4.pdf

[4] Liz Lyon. Dealing with Data: Roles, Rights, Responsibilities and Relationships. [2010-7-8].http://www.jisc.ac.uk/media/documents/events/2007/06/liz_lyon.pdf

[5] Sarah Higgins .The DCC Curation Lifecycle Model, the International Journal of Digital Curation Issue 1, 2008(3):136-138

[6] Charles Humphrey. E-Science and the Life Cycle model of Research. [2010-7-8]. http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

[7] Anna Gold. Cyberinfrastructure, Data, and Libraries, Part1 A Cyberinfrastructure Primer for Librarians, D-lib Magazine, 2007(13):5-6

[8] Liz Lyon. Open science at web-scale optimising participation and predictive potential. [2010-7-9].http://www.jisc.ac.uk/media/documents/publications/research/2009/open-science-report-6nov09-final-sentojisc.pdf

[9] Alma Swan, Sheridan Brown. Skills, role and career structure of data scientists and curators. [2010-8-1] .http://eprints.ecs.soton.ac.uk/16675/

[10] Rick Luce. the role of academic research libraries in the digital data universe. [2010-7-9]. http://www-s.cic.net/programs/centerforlibraryinitiatives/Archive/ConferencePresentation/Conference2008/e-ScienceSpeakerPresentations/RickLucePPTpres13May08.pdf

[11] Lord, P. Macdonald, Lyon & Giaretta. From data deluge to data curation. [2010-8-9]. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf

[12] 孙坦.数字化科研——e-Science 研究.第 1 版.北京:电子工业出版社,2009.8:187
[13] DataStaR.[2010-7-8].http://datastar.mannlib.cornell.edu/
[14] D2C2.[2010-7-9].http://d2c2.lib.purdue.edu/
[15] DAAC's. [2010-9-8]. http://nasadaacs.eos.nasa.gov/about.html
[16] CISTI. [2010-8-10].http://data-donnees.cisti-icist.nrc-cnrc.gc.ca/gsi/ctrl?action=catba&lang=en
[17] 李春旺,图书馆集成融汇服务研究.现代图书情报技术,2009(12):1-6
[18] 李春旺等,基于 SRU 的集成服务平台设计与实现,现代图书情报技术,2007(2),12-15
[19] EOL data management. [2010-8-8].http://www.eol.ucar.edu/about/our-organization/cds/dmg
[20] Geographic information systems services@ MIT. [2010-8-8].http://libraries.mit.edu/gis/
[21] Michael Lautenschlager, Heinke Höck, Jan Brase. Publication and Citation of Scientific Primary Data at WDC Climate.[2010-9-8].http://colab.mpdl.mpg.de/mediawiki/images/3/30/ESci08_Sem_1_Primary_data_registration_Lautenschlager.pdf

作者简介： 师荣华，女，1985 年生，硕士，发表论文 2 篇；
刘细文，男， 1965 年生，研究员，发表论文 20 余篇。