中国科学院文献情报中心科技信息政策中心

系列编译资料

促进数据统计：

数据计量先导计划

Making Data Count:

A Data Metrics Pilot Project

原编著者：Lin, Jennifer, Cruse, Patricia,

Fenner, Martin,Strasser, Carly,

2014 年 12 月

使用须知

中国科学院文献情报中心为促进学术交流、促进文献情报服务创新发展，特组织编译《促进数据统计：数据计量先导计划》，供个人学习和研究使用。文献情报机构可以在保证《促进数据统计：数据计量先导计划》完整性和所有编译者信息完整准确性的条件下在网站上整期上载和传播《促进数据统计：数据计量先导计划》的 **PDF** 版本，并明确说明来源。

任何机构或个人在引用《促进数据统计：数据计量先导计划》内的具体编译内容时，请按照学术规范注明来源，包括原始文献的著者、题名和来源网址等，也包括编译者姓名和编译内容来源。如果任何机构或个人要直接整条采用具体编译内容（包括仅对文字进行非实质调整后的采用）、或者要对较长编译内容直接采用其较大篇幅内容（例如超过五百字以上），应事先征得编译者的同意。任何机构和个人，未经中国科学院文献情报中心许可，不能直接把《促进数据统计：数据计量先导计划》的内容大规模直接编撰为新的作品或作品的一部分。

编译者：王璐

审校者：王微一校，顾立平二校

审核者：顾立平

促进数据统计：数据计量先导计划

项目概述

概述：

项目概述：促进数据统计：开发数据计量的试点

加州数字图书馆（the California Digital Library，简称 CDL），美国科学公共图书馆(the Public Library of Science,简称 PLOS)和地球数据观测网络(Data Observation Network for Earth，Dataone）将重点关注什么样的计量指标可以以一种可靠有效的方式捕获活动周边的科研数据。这些组织将展开设计开发并进行指标模型的数据实践，测试自动跟踪的机制；探索一些方法，可以使原始未经加工的动态的数据集级别计量（DLM)数据通过数据类型和科研问题转换成数据挖掘；形成一个灵活的报告的给资助者、机构和人员分享 DLM 的结果。

背景：

长期以来，科研人员一直在努力解决学术上投入产出的评估和跟踪问题。随着一些新的工具出现，例如论文级别计量（article-level metrics，ALMs）给学术交流的生命周期带来了新的观点，在评估的实践过程中，已经开始不断结合此类数据以及所制定的方法体系随之而来的变化。这类指标使用率的增加有助于识别科研成果中需要借助不同的方法进行评估的单个对象，以获悉其工作成果的功能效用。

这项建议将形成一份解决科研数据溯源的公约。DLM 将提供关于科研数据的传播和影响范围的周边的、直接的、第一手观点的一个清晰的不断增长的前景。这些指标反映了数据自进入知识库之时起到随时间动态演变的整个路径。这些结果可用来通过启用新的过滤与推荐机制实现数据发现，从而使科研人员能够更好的找到与其特定需求最相关的数据。DLM 将提供一个多维度的视角以满足广泛的投资者和机构报告需求。DLM 通过自动跟踪来实现，因此可以减轻报告的负担并且不断提升潜在一致性。

学术价值：

学术价值：在最高层次上来讲，这项计划工作是为了理解数据的使用和共享的大型科研生态系统以及跟踪数据生命周期不断演化目标的贡献。该项目还为如何利用早期最佳共享科研成果以推动科学进展目标提供了新的评估工具。该项目将增进科研人员对自动效果跟踪的数据价值发现的理解程度，科研人员可以再次利用它来促进自己的工作。

广泛影响：

广泛影响：该项目的主要影响是增加现有的学术基础设施，现阶段的重点是期刊文章并将数据作为一个有价值的学术成果引入。一旦科研的影响展开，数据计量将建立激励机制以支持数据共享，提高跨学科的信息传播的速度。该项目还将激励其他一系列广泛的转变。这个项目中的团体参与将促进从业人员、资助者和机构之间对调研数据实践的使用和再利用进行更深入的探讨。软件将通过开放源码许可证实现公开分享，使团体能持续地提高和重用这种技术来收集指标和让指标更有用。所有收集的数据都是人机均可使用。由于科研领域中数据使用和活动收集的系统化规范化是首创的，它将会引起在这方面有很多科研实践以及领域内相关科研中学科带头人的高度关注。该项目将可以识别数据计量与期刊计量的差异，并将形成 DLM 更进一步的开发。这个建议为经济模型奠定了基础，作为一个深入洞察科研影响和总生产率的数据来源，可以用来确定科研产品的范围。提供数据计量将有助于对科研人员或科研工作进行评估，有助于在论文发表之外，拓展对科学生态系统中成就和名声的界定。在一个越来越由数据驱动的世界，该项目将提供更多了解如何建立公共机构和基础设施的基础来促进可靠的数据实践，进一步促进实现整体公共利益。

1.背景

数据是科学的基础，对科学科研工作起着根基性作用。这些科研形成的成果往往通向新的假设，启动新的科学发现和创新驱动。然而，在学术生态系统的利益相关者，包括从业人员、机构和科研资助者，越来越关注科研数据的共享和再利用 [1-3]。一旦错失优先投资的机遇，会阻碍再现性，影响资源配置的效率和公平 [4]。根据两份针对科研人员在数据共享方面的观点和行为的重要科研显示，这些问题并非毫无根据。PARSE 科研发现，25%的科研人员将数据公开，而 Carol Tenopir 等人的著作也显示出类似的结果（46%的科研人员开放部分数据在线获取，6%的科研人员共享全部数据）[5]。与此同时，总的来说，可供参考的

数据共享实例还未得到现存的学术基础设施完好的支持，包括数据产品的获取、可重复成果的共享，以及对科研中数据生产者、数据中心和资助方等角色的识别[7-11]。

最近这些利益相关者群体之间的发展，显示正在有着广泛变化。近期白宫科技政策办公室的政策备忘录[12]要求政府资助机构确保支持的所有科研成果对外公开。在英国，科研理事会政策[13]要求数据可获取并保存十年，而且科研出版物需要包含一条如何获取基础材料（如数据，实例或模型）的声明。更广泛的，欧盟地平线 2020 项目（ European Union Horizon 2020 ）包括一个开放科研数据试点[14]，这将需要授权人的数据共享。投资者乐于支持所资助项目中的数据再利用。新的激励措施现在都开始出现了，如由美国国家科学基金会（NSF）的数据管理计划的需求和收益，如广泛引用的元分析[15]，所有这一切都表明改变的时机成熟了[16]。

解决数据访问的问题，只能部分地解决科研人员共享和重用数据时面临的挑战。一旦数据是可获取的，科研人员将越来越多地接触到多样化的数据和来源不明的相关内容。这些数据都位于不同的期刊和数据档案中。接下来的问题是，“科研人员如何发现最能满足需求和支持学术研究的科研数据”？在一个数据密集型科研驱动的世界，质性研究受阻就是由于缺乏可靠的工具来公开相关科研数据。许多工具已经被提出或者具有原型，但除了谷歌学术中简单的搜索之外，很少取得进展。对相关内容识别的常见解决方案通常要求有随手可得的更好的自动搜索程序或改进的内容语义理解，但目前尚未能满足这一需要。如果用户社群有一个基本的工具集和操作框架以及灵活性来进行实验和开发自己的解决方案，那么这些和其他创新的方法将获取更大的成功[20]。

然后，这一进程可以动态反馈到系统之中，并在社群团体中全面共享。对于行动和关键挑战的号召是明确的：我们必须提供可靠的技术来评估科研，进行灵活的组织，使科研人员能够找到并获取与他们最相关的高质量信息[20]。

科学界也缺乏一个系统机制，使产生和共享的科研数据可以得到评价和影响信誉。学术成果评价一直专注于学术论文的出版，这从根本上来说是不同的工作成果。学术论文中可能附有也可能不附有数据，而且一定是用不同的方式。此外，科研数据是高度结构化的，因此往往是由更小的单位（如单一数据表或空间层）组成，并结合多个来源数据集的数据聚合成派生数据集。当这些子集或衍生产品的得到使用时，他们经常被分配唯一标识符并得到原始数据集的独立引用。此外，随着时间的推移，独立数据对象的版本因为新的数据添加或错误发现而得到不断修订。引用实例仍然是不断变化的，而且没有既定解决方案来处理特定数据的无限复杂性。

新兴科研领域的实践表明，理解数据的使用需要对数据源有深刻的理解。数据衍生也远超出在文章描述中所传达的最小工作范围。从一个对科研人员的更全面的贡献角度来看，数据计量是必不可少的。基于正式发表文献的参考文献的通用数据跟踪工具逐渐产生（例如：汤森路透的 Web 科学数据索引）。但他们不允许对数据集的广泛使用和重用，因此受到视野范围的限制。由于对文章出版的依赖性，他们也经历着指标积累的重大时间滞后。

到目前为止，创建一个全面的自动化数据影响跟踪系统的一贯努力还无法提供证明跨越广泛的科研领域的影响力。科研新的服务，如科学引文索引的 Web 数据是专有的，基于订阅的方法不太适合一个充满活力的科研信息环境。一个开放的框架有利于数据验证和数据对比，并促进程序和工具的发展，产生越来越多的增添原始数据附加值的新方法。指标数据必须向文献和数据集公开，且允许最大程度的重复使用。科研信息管理系统有一个完整的框架，将能更有效地捕获科研资源和产品的管理，其中包括科研数据输出的跟踪，可以自由访问、使用和共享。

2.项目目的和目标

从现状来看，科学界一直需要数据发现的解决方案和更广泛的采集工作价值，即科研者学术追求的核心。我们打算对数据跟踪和测量使用的指标进行设计和开发，即“数据级别计量”（data-level metrics，DLM）。

我们将通过收集指标和调查如何最好的利用收集的数据来调查指标的可行性和有效性，促进这些学术产出的发现和学术报告。该科研项目将从收集及使用 DLM 数据的角度探讨以下问题：

数据计量（DLM）
像单篇文章评价指标一样，DLM 是一套多维的指标，测量周边广泛范围内的活动和使用数据使用作为科研成果。

DLM 生成
什么样的指标是可获取的，可以通过自动跟踪收集并对机构团体产生作用？ 在我们试点中确定的指标的局限性和风险是什么？他们会创造或加深现有的偏差吗？ 机构团体对数据有效性和可靠性的要求是什么？博弈是一个议题么，如果是这样，它将如何解决指标的实施问题？

要建立一个长期 DLM 框架，什么样的数据通道和相关的资源（网络服务）是可持续的？ 什么样的数据收集方法对提高 DLM 数据标准化和确保获取数据交叉比较而言是必需的？
DLM 应用（数据发现、评价和计量分析）
谁将使用指标和用什么方式使用（使用案例和限制）？ 指标能告诉我们什么，不能告诉我们什么？ 科研团体对数据的使用不同吗？人口统计学的差异有哪些（学科、地理、机构联系等）？ 怎样让指标得到最有效的沟通和共享（聚合和推荐报告）？ 哪些指标试点突出了开放数据和封闭数据之间的差异？数据是按照开放存取政策和试点指标中的区别共享的么？ 什么样的网络基础设施缺口（元数据漏洞，系统桥接等）是为 DLM 能够集成到更大的信息科研生态系统的而存在的（知识库，资金，人员，机构信息）？

该项目提出的科研和模型代表了国家艺术的基本进展，但也建立在现有的能力上，因此结果是既费劲又易于处理的工作。为实施这项科研，我们将：

- 设计、开发和原型化一个数据计量参考模型
- 自动跟踪测试机制
- 通过数据类型和科研问题，探索用传送的原始 DLM 数据来驱动数据发现的方法
- 形成一份灵活的报告，与资助者、机构和科研人员分享 DLM 的结果。

我们会对现有的调查进行深入的科研和分析，包括数据共享的态度、看法和认知，确定数据使用和重用的关键价值以描述周围的现有规范使用和数据共享。我们将让社群团体参与进一个开放的需求收集过程，然后完善并将结果转换入一个行业数据指标平台的计量框架。我们将开发可以通过广阔的科研领域和机构实践进行概括的指标，包括生命科学，物理科学和社会科学。由此产生的框架和原型会反映数据之间的联系和各种数据产生联结的渠道。

我们将用真实数据实地测试 DLMs 的有效性，并探讨在何种程度上，实现自动跟踪是一个可行的方法。以 2009 年起就被学术界吸收利用的论文级别计量论文级别计量应用为基础[21]，我们将扩大技术平台来跟踪和收集数据指标。我们也将通过利用 Dataone 现有的网络基础设施及其构件库，在试点语料库获得数据集的“使用”活动（访问和下载次数）。我们也将通过连接作为渠道的各种 Web 服务应用程序，在正式和非正式的学术交流中跟踪数据引用。试点实施将覆盖 Dataone 总库约五十万个数据集。我们将基于多样性在以下几个方面选择语料库：科研领域覆盖，数据公共可访问性以及学术论文关联。我们将建立一个基于单篇

文章评价指标报告的引文工具[22]，建立一个 Web 界面，允许用户发现需要的相关数据以及用一个方便的和令人信服的可视化方式来报告收集一系列数据集的活动。这将帮助我们证明 DLMs 的价值，增加高质量数据的价值，并测试其效用。

3.战略伙伴关系和人员

这个项目的合作伙伴包括项目的成功所需的所有专业领域国际公认的专家。同时，我们在现有的基础设施、专业知识、出版物的获取、不同领域的数据、科学团体的关系与全球的数据基础设施项目中，积累了丰富的经验和能力。加利福尼亚数字图书馆（California Digital Library，简称 CDL）的加利福尼亚大学管理中心与 PLOS 和 Dataone 合作，他们在各自的领域中众所周知，并且通过战略联盟获得更广泛的联合（例如 ESIP、 Research Data Alliance、Global Earth Observing System of Systems，International Geosphere Biosphere Program 等）。在科学家、出版商和图书馆之间的这种创新型的合作伙伴关系中，正确的时机和工具的发展将是渐进的，如果不是革命性的。由于开发数据管理工具和对数据的管理实践的科研[24]CDL 与列入试点的科研社群之间结成了一个深入的联系。PLOS 由于因其出版革新的历史而在学术交流领域具有悠久传统的主导地位。他们是论文级别计量论文级别计量的先驱，用于跟踪使用和学术论文在线的技术平台在整个行业越来越多地得到采用。他们通过大量的 Web 服务获得数字活动的经验与在学术交流应用上的深层知识相结合。作为一个数据网络程序 DataOne 促使其工作提供核心服务和网络基础设施以改善长期的数据可访问性和国家科学基金会资助的科研数据的再利用。

4.项目结构

为期一年的项目由五部分组成：

第一单元：我们对现有的关于数据共享的态度和看法（忧虑）的调查进行深入的科研和分析，识别数据的使用和重用的核心价值观以及阐明周围使用和共享数据的现有规范。

第二单元：我们延伸Dataone使用跟踪能力，用于服务符合行业使用标准的数据统计。

第三单元：我们将第一部分的科研形成了一套指标进行测试，扩展现有的技术，使我们

能够开始通过指标评价收集数据。

第四单元：我们为团体机构开发工具，以使用在数据发现和评估报告中的指标以及在可能的情况下整合指标纳入科研论文。

第五单元：我们对在多个科研领域收集的DLM数据的可行性和适应性进行分析评估。

4.1第一单元：数据计量领域的研究

指标的设计是保证指标满足社群科研群体需求的一个重要部分，因此了解这样需求的领域科研是极其重要的。由于试点的结果将对所有的领域都是有用的，因此DLM试点的目标受众将是环境学家、海洋学家、生态科学家。需求收集将在本质上是迭代的，为所需数据指标特征集提供更丰富的输入。考虑到最终用户整体理解的需要，我们将采用改良的人类学科研方法。具体的活动包括：从UC和Dataone社群识别候选人；访谈与观察行动和工作流程，获得初始要求，在国家会议召开两次焦点小组；并使用结果制定初步的要求。我们将评估什么样的价值和利益能够使利益相关者愿意接收来自数据指标以及任何他们认为关键的重要特点或特征。我们将为不同类型的科学家创建用例，使之更好地处理与不同类型的数据和科研人员相关的挑战。

科学社区的需求评估将依靠两个方法：

1．定量评估包括简短的调查（网络或者纸质），用5-10分钟完成，以及一个快速调查（Web或纸质，或口头执行），包括1-3问题，要求科学家找出最理想的数据指标特性。

2．定性评价，通过深入访谈（网络或会面）观察科学家对数据度量的使用，询问他们关于当前的对指标的看法和使用情况，并确定数据指标特征兴趣点的主要特征。

在招募评估的科学家时，我们将从所有领域（地球、环境、海洋、地质、生态）和各类子领域对个人进行查询。我们将采访各种级别的专业科学家，从科研生到学术领头人，还包括一系列机构（学术界、博物馆、非营利组织）。为了招纳科学家，我们会使用多种工具，包括社交媒体（Twitter、博客、电子邮件）和邮件列表，大会、会议和其他场地的面对面交流，以及加州大学研讨会和会议。

4.2 第二单元：数据使用跟踪

我们将扩大现有Dataone技术平台来处理、隐匿数据作者姓名并传播存入Dataone库的数据，使服务可以集成到DLM系统。Dataone提供机器级别的访问入口，通向一个联合网络成员的数据存储库，包含成千上万的数据集，从而提供方便的集中接入科研数据存储库的高度分布式网络。这些数据集的使用贯穿整个联盟，而这些使用统计聚集在Dataone的协调节点（CN）。新的DLM平台将与Dataone网络以两种方式相互作用：

1.在Dataone网络获取和更新的数据集的时候，数据的目录元数据和永久标识符（PIDs）将从Dataone协调节点（CN）获取数据并在DLM注册。

2.随后，DLM会定期向协调节点请求关于所有DLM注册数据集的当前使用统计数据，以更新项目建立的所有关键影响指标。

这两种相互作用不仅将充分利用现有Dataone APIs作为有形的用例，也推动持续改进这一核心功能的设计和部署。

单个资源存储库已收集这些信息，但要求客户（如DLM）发布大量并行查询单个存储库的来源是不可行的。在可以轻易地被回收的CN上，自动化处理集中管理的统计集合会更好一点。请求和提供相关的元数据API的方法已经存在，包括PIDs。作为这项活动的一部分，Dataone团队将定制自己的新兴使用统计API来满足DLM平台的需求。此外，API将启用以时间和空间聚集的数据使用的检索。

4.3第三单元：数据活动聚合

伴随数据使用跟踪的发展，我们建立、试验、并促进 Component Three的开源数据指标平台聚合基于测试指标的数据集的影响以及测试指标的有效性和可行性。DLM的支撑技术将基于现有的不断发展的、开源的、机构论文级别计量论文级别计量的项目进行构建。

PLOS开创了论文级别计量（ALM），在2009年3月成为第一个收集和显示关于文章本身的使用和发表信息的出版机构，因此整个学术界可以借此评估自己的价值。PLOS使用ALM应用来聚集学术论文相关的数据和统计，包括在线使用、引文、网络书签、社会媒体提及、博客订阅、评论或建议。ALM应用是一个为公众实施和提升的免费开源社群项目，并且借此建立的第三方应用程序。它还包括一个API，使这些数据可以被任何人利用和混搭。多年来，ALM应用不断成熟，扩张成一套指标以及一系列辅助工具，允许用户搜索、排序、评价和发现在这些指标基础上的文章。这个平台的基础设施也已发展的更强大，规模扩张以容纳更多的文献记录和可扩展为其他标识符。

因此，开放源码的ALM应用作为最优化技术，服务于快速构建的数据指标平台和测试所提交的数据指标。我们将（1）扩展对数据集的多标识符方案支持的能力，（2）通过建立正式Web服务连接，进行跟踪下载、对话反向依赖，并可以直接获取（包括对Dataone数据加工和包装的使用活动），以及（3）使他们可以通过Web应用程序接口和API进行访问。除了正式的Web服务捕获之外，我们也调查了技术解决方案以解决标准化渠道的缺乏。为了解决这个问题，我们将开发新兴的可追溯性工具，设计现场数据使用工具以跟踪替代机制。通过建立新的标记（数据文件的数字碎片），在发生的活动中留下踪迹，我们可以建立跟踪电

话来接收匿名信号的活动，不依赖于这些整理数据的目的的能力。我们将遵循这一方法的应用，在其他领域实施合适可行的技术产业。

在一个知识库接着一个知识库里翻箱倒柜地进行数据的收集是一项繁重的工作，可能不会有效满足上述目标。但随着Dataone网络集成，将允许DLM平台在一个真实动态的数据环境访问多样化知识库，在其所持有的数据积累更多使用和再利用活动的同时，持续扩大规模。

4.4第四单元：数据计量集成和展现

DLM的原始数据将形成他们支持的项目和用例的基础。数据能成为信息和洞察力，但是，需要通过使用和共享来实现。我们将为科研者、资助者和机构开发一个基于Web发现和报表工具的参考实现。此工具将展示组织DLM数据的大量方法，从而支持学者找到科研需求的相关数据，以及支持投资者和机构来跟踪这些科研成果。参考实现也将证明捆绑数据的价值，并将这些指标和相应的科研论文及相关的论文级别计量相关联。

产生的数据将很容易被人类和机器使用。为此，我们将开发一个直观的视觉化语言系统的Web界面，在功能上结合DLM内容，阐释科研在更为广泛的世界范围被使用和整理的方法途径。消费者将能够生成报告，以清晰地阐述公开的科研数据的力量以及不同的扩散方式及其影响。此外，该工具还将推进的对DLM数据价值的发现过程，提高科研经验。用户将能够利用DLM更好的过滤和对大量数据及进行排序，从而解决“过滤失效”在信息过载中的根本问题。让DLM数据可适用于数据集，将使科研人员能够在其他科研人员留下的更广泛的“碎片路径”的知识范围基础上，更有效地、准确地揭示相关的数据集。

在设备损耗方面，我们将投资于一组广泛的DLM APIs。大多数的API将包括自然语言查询，每一个都旨在提供一个特定的数据视图，而我们也将为原始数据的复杂查询的提供一个更强大的，开放的API。我们将展示元数据如何将数据档案和出版商联系在一起，以及作者和投资人的可用标识符的标准化。试点中一小部分的数据集将被选择作为PLOS出版的关联。同时，API和Web接口的工具将提供一个受资助的科研活动的广阔图景。

4.5第五单元：文献计量学分析

DLM将提供关于科研数据的传播和影响范围的周边的、直接的、第一手观点的一个清晰的不断增长的图景。我们将引导对数据使用和重用的动态分析，更好地了解一个根本性问题，即影响对于作为学术产出的科研数据而言意味着什么。为了实现更广泛的文献计量学分析，测试集将包括一组不同的数据类型，可以划分为跨学科领域、知识库、访问许可以及文章科研协会（开放或基于订阅）。一旦收集了足够的数据，我们将全面检查的试验指标的有效性和可靠性，允许更充分的调查，包括指标能够告知什么和不能够告知什么。我们也将开

始探索代表性科研机构数据的使用行为，钻研人口学统计差异表达（学科、地理、机构联系等），突出在开放数据和封闭数据之间的DLM显现的差异。最后，我们将探讨科研论文的周边活动及其相关的数据集的语料库适用的子集之间的关系。

我们认为，将科研人员、数据知识库、机构和资助科研成果的资助机构紧密结合进行全面分析，所面临的困难可能是基于元数据可用性的限制。我们将揭露这些缺口，提供一个网络基础设施以开发可持续的和强大的科研信息系统，可以为科研和利益相关者间的组织需求服务。

5.人员角色和资格

Patricia Cruse，加利福尼亚大学管理中心（UC3）主任，CDL，将作为PI负责项目的所有方面，设置项目的优先事项，建立项目重点的利益相关者的关系（PLOS，加州大学的科研人员和Dataone团体），指导和评价结果，并确保项目成功完成。Cruse是全球DataCite联盟的创始成员，DMPTool项目的主要协调员，负责根据每个国家科学基金会和其他机构的要求生成数据管理计划，并属于Dataone的领导团队。

Patricia Cruse，UC3科研数据专家，CDL，将管理社群互动、收集需求、用户测试，并将带领团队与科学机构及Dataone进行相互合作。Strasser 也将与团队合作以传播成果。Strasser 曾担任DataUp 项目的这个角色[25]，现在担任UC-wide Dash项目的项目经理。

Martin Fenner，高级产品经理，PLOS他将作为项目经理服务于项目的所有方面，包括驱动项目的优先事项，建立合作关系等。此外，她将监督工作发展的五个阶段，从指标的科研和需求收集，到参考工具及应用。她目前管理的是单篇文章水平指标项目和出版商的数据项目。

Martin Fenner，ALM技术总监，PLOS，他将作为数据指标应用技术的数字总监。他目前是作为DLM的基础的论文级别计量应用的技术总监。他已经是欧盟资助的ORCID DataCite Integration（ODIN）项目成员，他是Dryad Digital Repository数字资源库的董事会成员，他也是Dataone的成员之一。

应用程序开发商，国家生态分析和综合中心 (NCEAS) （TBN）NCEAS 雇用应用程序开发商。该开发商将会开发数据使用统计 API，并且为了实现 DLM 在 Dataone 环境中的应用，持续扩展统计 API。应用程序开发人员将与 Matthew Jones, Dataone and NCEAS, University of California Santa Barbara 一起工作，确保 DLM 应用的统计 API 的拓展工作可应用于 Dataone。

6.适用性期望

试验性和高回报：

我们提出的项目是一具有盈利潜力的高度实验性项目。虽然分享自己的数据的做法被广泛认为是“有益的”，并不是所有的人都能做到[6]。出于现有数据的缺乏，对他人科研数据的应用不是一个普遍的实践。非正式的实例可能是有限的，而且可能会影响如何广泛建立正式的渠道，与现有的Web服务聚集这样的行为。目前还不清楚现阶段数据共享和重用中的缺乏标准化将对建立一系列更全面效果的证据产生什么样的影响。

该项目对现有的NSF程序可能风险很大，因为自动获取的数据源尚未建立和认可学术价值。机构团体尤其是基金会，为了能够实现这些措施的有效性，需要对数据集和其他活动的使用做初步准备。作为试验场，该试点旨在通过一系列的试验观测启动交流对话。此外，该项目对基础设施建设投入是在一般的美国国家科学基金会资助的领域之外。尽管该项目是一个试点，开放源码的DLM应用也要做好投入生产使用的准备。在科研信息系统中的科研实体可以通过系统将数据聚合链接到科研单位（科研员、基金、机构、数据仓库、学术指标等）。数据发现和报表工具的参考实现，将通过这些联系展示一个科研活动不断拓展的发展图景，包括从授予奖项或者科研职位以及其他新的见解的等方面的影响科研。

广泛利益：

该项目的主要影响是增加现有的学术基础设施，特别着重学术论文，将数据作为一个有价值的学术成果引入框架。DLM服务将允许任何人获得一个完整意义上的数据理解，包括数据如何被使用，以及在整个网络的集中讨论中如何展示数据使用指标。例如，它将支持所有科研者提出有意义的影响证据作为评估其工作的依据。另外，科研赞助商、数据生产者和使用权推广委员会可以使用指标工具来跟踪项目和科研人员的生产力，或通过书签、讨论量、下载量和数据集使用等来探讨观众的多样性。在可识别的综合生态系和可追踪的科研数据中，这些数据的工具可能成为数据丰富的科学的重要基础。

在这一转变服务中，该项目还将激励其他一系列广泛的转变。数据级别计量将建立激励机制，支持数据共享和使用增加在广泛的学科信息的传播速度。这个项目中的团体参与将促进从业人员、资助者和机构之间对调研数据实践的使用和再利用进行更深入的探讨。软件将通过开放源码许可证实现公开分享，使团体能持续地提高和再利用这种技术来收集指标和让指标更有用。所有收集的数据都可以提供给人类和机器使用。由于科研领域中数据使用和活动收集的系统化格式是首创的，它将会引起在这方面有很多科研实践以及领域内相关科研中主导的学者很大的兴趣。最后，这个建议为经济模型奠定了基础，作为一个深入洞察科研影响和总生产率的数据来源，可以用来确定科研产品的范围。对于确定基础设施、支持、时间以及机构、资助者、科学家自身的科研的花费问题，跟踪数据的引用是理想化的。在一个越来越由数据驱动的世界，该项目将提供更多了解如何在一个道德的框架，建立公共机构和基础设施，提供一个使用可靠的数据指标，进一步促进实现整体的社会公共利益。

7.时间表

Mo 1-3	开始进行关于数据使用活动的科研社群会话，讨论试点的目的 进行关于数据使用和重用的规范性价值和行为的领域研究 建立 Dataone 的数据集使用公共层、API 与 DLM 应用的连接 识别和筹备 Dataone 的存档数据集 规模基础应用（ALM） 提供领域报告和数据指标需求文档
Mo 4-6	定义反映社群科研输入的初始指标 测试 Dataone 参与成员节点的使用渠道 DLM 架构技术设计
Mo 7-9	扩展现有的应用程序，收集数据指标 扩大与其他外部 Web 服务之间的数据指标的相互作用 收集 Dataone 的科研数据集
Mo 10-12	继续完善和扩大 DLM 开发 DLM 数据的交互发现和报表工具 DLM 数据的文献计量分析 加强 DLM 的文档编制、安装和配置程序，提高系统重用性 下一步建议的最终报告，内容关于开放、自动跟踪试点数据使用的成就、挑战和机会

8.现有结果

Patricia Cruse, Carly Strasser：NSF Grant No. OCI 0830944 (W. Michener PI)“DataNetONE（地球观测网络）”。预算总计：15257190美元。时间：2009年8月1日至2014年7月 31日。Dataone项目已经达到了它创造一个借助接口和软件的数据档案开放网络的里程碑目标

（Dataone.org）以及一个科学家、图书馆员和学生组成的社群，这些都是将数据引用变成常规做法的关键。这个安全、可扩展的网络有（a）scheme-agnostic的永久标识符，（b）数据和元数据的复制，（c）搜索和发现，和（d）身份认证和访问控制。除了一个日益增多的用户组（DUG），Dataone也具有一个公开保护政策和一系列培训活动、实习以及涵盖一些领域内容如教育、社会参与、元数据、语义和可持续性的工作组。

9.参考文献

[1] Nelson, B. 2009. Data sharing: Empty archives. Nature 461: 160-163.

[2] Tenopir, C, S Allard, K Douglass, AU Aydinoglu, L Wu, E Read, M Manoff, and M Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS One 6: e21101

[3] LeClere, F. 2010. Commentary: Too Many Researchers are Reluctant to Share Their Data. The Chronicle of Higher Education, 3 August.

[4] Nature Editorial, 2009. Data's shameful neglect. Nature 461:145.

[5] PARSE Insight (2009) Available: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf. Accessed 2010 Oct 12.

[6] Tenopir, C, S Allard, K Douglass, AU Aydinoglu, L Wu, E Read, M Manoff, and M Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS One 6: e21101

[7] Cook, R. 2008. Editorial: Citations to Published Data Sets. FLUXNET Newsletter 4:1-2.

[8] Costello, MJ. 2009. Motivating Online Publication of Data. BioScience 59: 418-427.

[9] Hey, T, S Tansley, and K Tolle (Eds.). 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.

[10] Parsons, MA, R Duerr, and JB Minster. 2010. Data citation and peer-review. Eos, Transactions of the American Geophysical Union 91(34): 297-98. DOI: 10.1029/2010EO340001.

[11] Whitlock, MC. 2011. Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution 26: 61-65.

[12] White House Office Open Data Policy. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf. Accessed June 1, 2014.

[13] Research Councils UK Policy on Access to Research Outputs. http://roarmap.eprints.org/671/1/RCUK%20_Policy_on_Access_to_Research_Outputs.pdf. Accessed June 1, 2014.

[14] Horizon 2020 Open Research Data Pilot. http://europa.eu/rapid/press-release_IP-13-1257_en.htm.Accessed June 1, 2014.

[15] Hampton, SE and JN Parker. 2011. Collaboration and Productivity in Scientific Synthesis. BioScience 61: 900-910. DOI:10.1525/bio.2011.61.11.9

[16] Alsheikh-Ali AA, W Qureshi, MH Al-Mallah, and JPA Ioannidis. 2011. Public Availability of Published Research Data in High-Impact Journals. PLoS One 6: e24357. DOI:10.1371/journal.pone.0024357

[17] Parr, CS and MP Cummings. 2005. Data sharing in ecology and evolution. Trends in Ecology & Evolution 20: 362-363.

[18] Data Replication and Reproducibility: Science special issue, 2 December 2011.

[19] Smith, VS. 2009. Data publication: towards a database of everything. BMC Research Notes 2:113.

[20] Enriquez, V, SW Judson, NM Weber, S Allard, RB Cook, HA Piwowar, RJ Sandusky, TJ Vision, and B Wilson. 2010. Data citation in the wild. 6th International Digital Curation Conference.

[21] Article-Level Metrics website. http://articlemetrics.github.io/. Accessed June 1, 2014.

[22] ALM Reports website. http://almreports.plos.org/. Accessed June 1, 2014.

[23] Kratz, J and C Strasser. 2014. Data publication consensus and controversies. F1000Research, 3:94.DOI: 10.12688/f1000research.4518

[24] Kratz, J and C Strasser. Data publication practices and perceptions. In preparation for submission to PLOS ONE.

[25] Strasser C, J Kunze, S Abrams, and P Cruse. 2014. DataUp: A tool to help researchers describe and share tabular data. F1000Research 3:6. DOI: 10.12688/f1000research.3-6.v1

数据管理

1.产生的数据类型

该项目的课程将生产或开发五种类型的数据。

第一种是社会科学数据，用定量和定性调查结果形式收集，作为第一单元的一部分。鉴于这些调查将涉及人类受试者，我们将确保我们的科研是符合伦理审查委员会的政策，并且在进行调查之前事先得到批准。这些数据将通过纸质和电子手段收集。本文的调查结果将通过手工进行数字化，在项目持续期间，原件将得到保存。

第二种是数据集使用指标，通过第二单元建立的机制，从Dataone网络收集。个人Dataone成员节点采集原始Web日志的使用统计。因为该数据可能被用来识别用户的个人数据，在被dataione协调节点聚集之前，先进行匿名，才能由DLM工具收集。

第三种包括由软件生成或修改的数据，是项目的一部分。在项目结束后，由PLOS 创建的新DLM软件将继续保存在一个GitHub库中。Dataone的代码扩展将保存在公众Dataone Subversion版本库。Fenner将负责与DLM工具相关的代码。

第四种是数据集引用数据，由DLM工具收集，包括传统和替代的（例如，博客、推特等）学术交流途径，作为第三单元的一部分。

第五种包括所有其他科研的产品（例如，第四单元开发的额外的最终用户工具，第五单元的分析结果，社群的反馈和项目的状态交流），这将通过同行评审期刊文章或该项目的网站进行公开，这部分数据由California by Strasser进行维护。

2.数据和元数据标准

项目开发的DLM软件将符合社会的最佳标准，包括带有主要版本标签的版本控制（使用Git），允许开放源代码许可证（Apache 2），机构库的公开获取（GitHub库），注释、参考文献和下载和安装说明的教程文档，这些可以从该软件以及一个社群网站上获得，并可以有广泛的测试覆盖率。

作为这个项目的一部分，Dataone开发团队将以与COUNTER标准一致的形式，评估效用以及为此付出必要的数据统计水平的数据。这将允许对数据使用指标和其他类型的在线资源指标之间的有意义的比较，包括传统系列和专题出版物。

3.访问和共享政策

第一部分列举的五个主要数据类型，作为项目完成后产生的成果，将面向公众开放访问、

评估和使用。有关软件和数据的可用性的声明将通过各种渠道（如博客、Twitter、电子邮件）发送给所有感兴趣的利益相关者。

4.政策的重用，再分配

同样的，在项目中或者项目之后，所有的软件产品将被重新使用和重新分配。软件的唯一限制就是软件被重新编排时，著作权和许可声明必须保持完整的Apache开源许可证。该软件将赢得出版商，数据中心和存储库，科研人员和机构管理人员的兴趣。

5.归档和保存计划

除了基于GitHub库的机构管理，所有DLM软件的主要版本将被归档在CDL’s Merritt库，并由ezid服务提供持久标识符。只要对科研人员和项目合作者有持续价值，科研数据和记录将会得到有效保存。

加利福尼亚大学管理中心（UC3）的Merritt Repository Service有能力管理、归档，并分享数字内容，并提供持久的访问网址，搜索接口和长期的数据管理工具。Merritt relies依赖于一个带有显著的计算和存储组件冗余的高容错微服务体系结构。在其五年的生产经营中，虽然目前已经管理过超过16TB的数字资源，Merritt并没有丢失过任何数据。

由美国国家科学基金会资助的Dataone项目，包括一系列意义重大的活动，调查大量能够确保组织、金融和技术上的延续性及可持续发展的选择。基本完整的以Dataone基础设施为基础的软件是由一个公共的Subversion 代码库进行管理，并由大量不同的开发者社群的支持。Dataone的基础设施依赖于一个高度容错的分布式体系结构。虽然Dataone网络上的所有单个节点都经历了许多周期系统软件升级和更新周期，在任何情况下全球建设一直都能正常工作，自动将故障转移到其他服务器上的冗余系统的容量。自从第一次投入生产运营，最终用户都没有中断与Dataone网络的连通。