《专题：数字资源长期保存》

经受住时间的考验[1]

——科学与工程领域数字资源的长期管理

美国研究图书馆协会 著　刘菊红 编译

信息技术和泛在网络的迅速采用已经改变了研究和教育事业，科学和工程领域数字资源馆藏在这场变革中处于中心地位，与其相关的生命周期管理也是需要解决的问题。

本次会议讨论了研究性和大学图书馆及其它相关机构在科学和工程领域数字资源管理中的角色问题。与会者提出了新的议题如：为更好的管理数字资源，在领域科学家、图书馆员和数据专家中间建立新的伙伴关系和合作关系；支持数字资源必要的基础设施的发展；为国家网络基础设施，建立可持续发展的经济模型，支持科学和工程领域的数字资源管理。

这次学术讨论会建立在由 NSF（National Science Foundation）支持的一些前期研究基础上，有众多研究团体的参与。它反映了一种认识：数字资源管理是科学、工程研究以及教育事业的基础，因此也是创新和竞争的基础。总的来说，强调面向个人和机构，设计共享基础设施的组织生态，满足异质数据和多样化的学术和专业文化的需要。

1 研讨会的成果

- 数字资源的生态反映了参与机构, 制度安排和机构库的分散性，政策和实践的多样性。
- 数字资源管理所面临的挑战，要求包含组织、学科领域和跨学科领域在内的多样化的实体和参与者都要承担起责任。
- 一直以来，大学通过其图书馆在知识进步方面起着领袖作用，在知识的长期保存方面承担着重要责任。就一些研究性和学术性图书馆、大学及其它机构所承担的数字资源管理这一新的的使命，展开了重要的讨论并给予了肯定。
- 各分散的数据中心和机构库应该共同承担数字资源长期保存的责任，在认识到数据的异质的同时保证联合和互操作的可能性。
- 数据资源管理中的各参与者对数据都有自己不同的经验，观点，设想和目的。紧密的合作关系将要求超越和调和文化的差异。建立合作模型共享经验和资源将是极端重要的。
- 数字资源的管理涉及保存和管理两方面。保存要求有基于标准的、主动管理的实践。这些实践能够在研究的生命周期中引导数据,使数字资源的长期保存成为可能。管理涉及到组织、显示和以不同目的重新使用已被保存的数据的方式。
- 为数字资源服务的基础设施是共享的公共物品，由联邦资助的研究所生产的数字资源也是公共物品
- 研究和教育团体生产的数字资源的管理和共享，要求有技术和经济支持的可持续发展的模型。
- 数字资源的存档，学术出版和相关学术信息交流之间需要建立紧密的联系，在数字

---

[1]本文为 2006 年 9 月 26 日—27 日，ARL 学术研讨会上提交给 NSF 的报告，主题为“新的合作关系：学术性图书馆在数字资源中的角色”。译文根据该报告的总结性摘要和结论进行编译。

资源管理领域主动承担责任为研究性图书馆提供建立这些重要联系的机会。

- 如果支持数字资源保存和管理的项目和目标走向成功，则和数字资源管理相关的联邦政府资助机构和研究事业的文化都有必要改变。
- NSF 和其它资助机构提高认识，满足研究团体进行数字资源管理和共享的需要将是极端重要的。

2 概要、结论和建议

数字资源的管理对研究事业来说起着基础性的作用。数据拥有广阔的使用范围，它提供了复制一个实验的能力；提高了数据重复利用的效率；在新的集成和管理数据的条件下，为提出新的问题提供了可能。因此问题就产生了，我们作为社会有选择性的收集者，在尊重大量法律法规和面对有时会出现的利益冲突的情况下，该如何来存储、管理、并使数据可获得呢？正像高等教育系统自己建议的那样，要求建立起组织生态学。但是建立在几个世纪的传统和实践之上的，赋予目前高等教育特色（包括它们所依赖的继承和文化制度）的组织生态已经在过去的几十年中进化了。因为有新的机构和学科进入系统，研究事业也是活跃的，这反映了人们研究新的领域、继续学习和新的安排的需要。

数据信息是易损坏的，我们不能只听其自然发展。在这个报告中已经出现的自组织存档显示了研究者将建立和使用这种类型的数据集合.但是这种补充式的存档和数据中心，大多是完全靠志愿者的努力来支持的，但对于很多学科来说，长期支撑数据并使之拥有较高的置信度以激发更广泛的使用是不够的。对任何想获得成功的系统来说，一定要获得用户的信任。NSF 的威信为建立一个成功的系统提供了良好的基础，但是就其本身来说还不充分。潜在的用户必须看到系统的价值，系统也必须能运行；必须能提供可靠的服务。在 NSF 呼吁建立一个全国性的数据中心系统的框架下，提出了一个很简单的问题：怎么样才能实现这个目标呢？

3 全体会议讨论概要：沟通文化和创造激励

尽管各个分组有自己独立的观点，但他们提出的讨论意见和建议比较集中。几乎所有的小组都要求一个反映新的伙伴关系的数据管理设施。这首先要通过实验的方式达到，通过原型和短期安排来学习成功的模型，汲取已经获得的经验和最好的实践。针对以下问题达成了共识：

- 跨学科研究和支持数据长期保存的技术能力和工具。
- 适当的存取监测和控制，保护机密和个人隐私，保护系统安全和降低风险。
- 自动生成元数据。
- 异构数据和系统之间的互操作。

此外，开展即期和长期的研究，促进相关学科专门知识的创造的重要性。例如：数据设施建立引发的相关的商业和经济问题应该成为重点之一。因此，无论是操作性的还是原型化的设施，都应该足够灵活以应对技术和组织的变化，包括在系统更新换代时的交接。

与会者也认为对科学家、图书馆员和公众进行教育，使他们了解数据资源管理是非常重要的。这将包含：在图书馆员和档案工作者工作的传统上，培训新型信息专家的课程；使科学家认识到数字资源管理的价值的策略；为还不经常重复使用数据的领域的研究创造可能。

最后，所有的小组都承认在跨学科和跨组织文化的交流和工作中存在挑战，在为个人提供恰当的激励和为组织和数据收集建立恰当的管理政策方面存在压力。这个议题在专门的论文中、全会和分组讨论会上都曾被提出。解决的方法无可避免地要通过建立混合性的跨机构、跨学科的组织来实现。它可以采取多样化的形式，能够覆盖国际性和国家性的组织如蛋白质数据库（Protein Data Bank）及与之相似规模的数据库中心，也能覆盖规模小一些的地方中心，这些地方中心可能和现存的位于大学校园内的提供特别馆藏的中心同为一体；也可能包含在现存的正在为特定社区服务的博物馆和图书馆里面。如以人类学著称的费尔德博物馆

（Field Museum）就是一个例子。号召数字资源管理机构积极主动的参与是关键的和不可或缺的，因为这些机构的存在为科学家存贮自己的数据提供了可能，也为科学家利用机构提供的数据进行实验提供了机会。同时，这些参与使数据资源管理机构日益合理，同时为管理服务提出了新的要求。

关于跨文化存在的挑战和激励的讨论主要从两个维度展开：一是关于 NSF 的内部文化；二是研究事业内部的文化，因为它影响了单个研究者及推动他们参与数字资源管理的机制。总的来说，NSF 作为一个基本的研究管理机构，数字资源管理对它的文化提出了挑战。特别地，反映基础性研究价值的标准评审过程，应用性研究的项目很难通过并得到资助，它们可能和原型数据管理设施有紧密的联系。对 NSF 来说，讨论这种担心是非常重要的。

从学术的角度来说，以宣传推广为目的、以任期考评的方式建立的声望系统可能会对数字资源的长期管理造成消极的影响，他们暗暗鼓励研究者保留对自己数据的控制权，但同时又不支持研究者对自己的数据做哪怕做极其微小的处理。研究者时间的有限和元数据的广泛存在共同呼唤帮助研究者的工具和自动生成元数据的出现，NSF 应该以某种方式提出识别数据集生成的办法并给予授权来促进这种改变。例如，在一个获得授权的机构里提交论文，可能被当作高质量的出版物，NSF 的编辑的数据保存可能当作科技指标之一。因此，对个人来说，数据资源管理适合了现存的声望系统。

人们提供了一些能够改进 NSF 资助和报告过程以激励研究者的想法，既有奖励也有惩罚。一些草案呼吁将数据管理计划和预算论证联系起来，同时和结论中要求的报告联系起来。后者更有优势，因为它提供信息术语集，NSF 可用来理解数据生产、贮存、再使用的模式。

数据管理计划对 NSF 来说并不是新概念。国家科学委员会（National Science Board）在一篇名为 “长期存在的资料收集：保证 21 世纪的研究和教育的能力”中说：“任何想创造和管理数据的个人的或小组的研究者，从数据的产生到消亡，需要一个策略来处理数据”。NSB 的建议比目前 NSF 在《拨款建议指南》中提出的要求有更深入的细节和指导。同时，与会者大多数同意要求数据管理计划的策略，只有少数人反对。一些人建议说要求申请人提供即使是简短的数据管理计划，对申请人已经紧张的资源也是一种额外的负担；一种反对观点则认为，为珍贵的实验数据制定并阐明计划（将预算论证和前期工作包括进来），是对学术负责的一个方面，但在提交论文的过程中，并不要求包含数据管理计划，数据资源管理和保存的文化将不太可能发生改变。

4、建议：

经过各小组的激烈讨论形成了各小组建议。在全体会议讨论过程中，这些小组建议被完善并被去除重复，并得到了一条总的建议。此外，三条由小组讨论产生的普遍性建议对总的建议进行了更清晰的阐述。最后，在更普遍性建议的基础上，形成了六条目标性建议。

4.1 总的建议：

NSF 应该为数字资源长期管理的可持续发展框架的建立提供方便。这个框架应该通过以下方式包含广泛的成员:

- 支持对数字资源管理进行的研究和开发，这些研究和发展要求理解、塑造技术和组织的能力，并建立模型；包括研究大规模地、长期的可持续发展的策略。
- 支持培训和教育计划,为 NSF 和其它合作机构培养数据科学领域的新一代工作人员。
- 考虑到管理所有科学和工程领域生产的数字资源的重要性，需要发展、支持和促进教育，使研究事业发生有效的变化。

总的建议认识到了潜能（设施和资源）和动力、供给和需求的自发的相互依赖。它同时认识到能力还没有完全开发（尽管在许多学科领域内有很多方法的例证），要求在相关的技术、组织、行为和经济商业问题方面，有更多原型和研究。需要有实质性的努力，来创造环

境和文化氛围来促进数据的管理和保存。这些努力必须被 NSF 和其它管理机构、不同文化领域内独立的学科和组织、包括专业学会、图书馆、档案馆和其它传承机构、及出版商和大学共同承担。最后，因为数据资源管理要求跨学科的合作，同样地，仅仅靠 NSF 的强烈要求无法来完成数字资源管理的使命。因此，这些建议提出了一个与数字图书馆计划不一样的跨机构要素。Digital Library Initiative 将大量机构的信息集合起来使之成为一个追求共同目标的资源整体。

4.2 三条普遍性的建议：

1. 资助项目应致力于解决数字资源被大量团体提取、存档和再使用的问题的项目。促进在大量的数字资源管理参与方之间的合作和分工，包括研究性和学术性图书馆，学术界（scholarly societies），商业伙伴，科学、工程和研究的相关领域，正在发展的信息技术和组织。

2. 鼓励对数据科学的新一代工作者的培训和发展。这包含为了新的目标，支持培训信息科学家、图书馆学专家、科学家和工程师，使其能在数字资源管理工程中，更有智慧地工作。

3. 支持可用的和有用的工具的建设开发，包括：

- 使数据的理解和管理更方便自动化服务。
- 数字数据登记。
- 参考工具，记录常用的术语和概念。
- 自动生成元数据。
- 权利管理和存取控制方面的其它方面。

4.3 六条目标性建议：

这些普遍性的建议被下述目标性建议进一步细化了：

1．NSF 应该制定计划来资助科学和工程领域数字资源的管理和保存。资助对象应该包括研究性学术性图书馆、科研领域、现有的技术支持者和其它合作者。以不同模型进行实验的多样化的项目应该得到资助。

NSF 应该建立一个可持续发展的框架，为科学和工程领域由联邦投资研究的长期管理提供方便。为了实现这个目标，得到投资的工程应该包含多样化的成员，包括最关键的成员，如大学、学术性研究性图书馆、领域专家、计算机科学家、专家协会、标准制定主体、出版商、营利和非营利的销售商、投资机构。下面的话表明数据资源管理包含成员的广泛性：“保存研究数据需要一个研究性的团体”。考虑到面临挑战的范围，项目应该反映数据资源管理的复杂性、使用的分散性、职责的多样性、成员变化的利益和需求等。

跨学科合作遇到的挑战在很多地方都是明显的，在数据集的异质性、收集和记录数据所遵循的文化的、组织的、技术的框架方面更加显著。这为数字资源管理机构和员工带来了一个非常实际的问题。与会者对采取分布式的方法达成了共识。但是有效的分布式组织和技术结构要求有共享的工具、标准、协议和程序，使有效的和互操作的系统和合作成为可能。当考虑到要使相关数据的提交、存档和管理、重用对未来的研究者有用时，这些需要特别明显。在很多地方比如元数据和本体方面已经有了相当的研究，但对组织内部和组织之间的信息流程理解得还不是很好，特别是在需要跨学科合作的地方。因此，为这些流程建立原型，使研究者知道什么有用什么没有用，交接在哪里发生，怎样改进支持特定步骤的界面和工具。

2．NSF 和其它联邦机构如博物馆和图书馆、图书馆学院和信息科学学院应该支持培训，这些培训的目标是使信息专家、图书馆专家和科学家，作为团队的成员，在数据管理、保存方面更可靠更高效地工作。

普遍认为，我们需要一种新型的专家，他们的专业知识对数字资源的成功管理是关键的。数字环境要求新的工具和技术。例如，科学家典型的不是被培训来管理数字资源的，

因此，对研究数据收集和数据分析的领域来说，现在研究数据资源管理具有同样重要的意义。数据管理不只是对实验结束后数据的存贮产生影响；它同时要求科学家理解存档的数据是如何存贮和管理的。它和科学家在做实验的过程中应该完全理解和掌握实验用的工具那样同等重要。社会科学家已经开始检验对数据的偏见是否合理，下一步应该是在消除偏见的前提下，理解数据资源管理的含义。因此，对数据资源进行更有效的管理，对研究者和数字资源管理者来说都是有好处的。因此，无论是作为将来的数据资源管理者，还是作为将来的数据资源使用者，数据资源管理都应该被广泛重视。

3．NSF 应该支持可用的和有用的工具和自动化服务的发展（如：元数据的生成、收割和确认），这些工具和服务使数字资源的理解和管理变得更容易，应该制定促进社会使用的激励措施。

第一条建议提出了对科技和工程领域信息完整流程的原型和模型的需求。它呼吁应该对已经在前面描述过的几个包含在组织流程中的特别的研究领域给予关注。所有这些问题在数字资源管理领域都是知名的问题，尽管也有了很多研究项目，但这些问题还远没有被解决。而且，在包含分布式组织和多样化知识团体的集成解决方案方面，这个问题也没有被解决。因此，根据这条建议被资助的项目将继续完成由早期建议所产生的工作，并且可能直接解决某个特别的问题。这可能包括：数据注册、元数据自动生成、参考工具以适应不断出现的常用术语和概念、权利管理。

4．在建立可持续发展的数字资源管理的经济模型的时候，经济学专家和社会学专家应该被包括进来。在这些领域的研究应该最终产生模型，这些模型在合理的时间范围内，在多个不同的科学领域内，在多样化的项目中，应该经得起时间的检验。

技术和经济的可持续性被认为是关键性的。它们是建立信任的基础，是可行的数字资源管理机构和更大规模的包含单个机构的数据资源管理组织的必要前提。综合考虑本地的、地区的、国家的和国际的范围、学科内及跨学科的内容，很多方法都被讨论了。但是，一般认为，建立一个好的模型要求研究相关的组织的和行为的问题；也要求使在信息经济学、公共物品和基础设施投资方面已经做过的研究发挥作用。一些可能被发表的论题包括：无形评估；模型化合作；在多样化的假设、动机、激励和声望系统下，考虑价值组成；以及信任的建立。

5．在提交申请的过程中，NSF 应该要求有数据管理计划；并在申请评议的过程中，对这些计划的适用性赋予更重要的意义。一个数据管理计划应该识别数据是否有更宽泛的价值；在潜在的分布上是否存在限制，如果有，限制的本质特征是什么；如果数据是相关的，分布的机制是什么，生命周期支持和保存是什么样的。数据管理的报告应该包含在关于 NSF 奖励的中期报告和最终报告中。提供恰当的培训手段和工具，保证研究团体能有效地发展和实施数据管理计划。

6．NSF 应该鼓励包含社区数据的项目的数据共享政策的发展。对发展这些项目机制的讨论应该作为数据管理计划的部分之一。此外，NSF 应该努力保证所有数据共享政策对公众是可得的和可获取的。

这些建议的目标是：使 NSF 提高研究领域对数字资源管理的认识；承认对数字资源管理机构提供的服务有需求；鼓励研究人员通过申请资助和提交报告的方式参与。这包括：

- 数据将被怎样管理，被谁管理，以怎样的机制管理；
- 数据是否要和社会共享（如果不共享，原因是什么）；
- 数据是否要保存下来供将来使用，如果是，怎样实现；
- 数据管理在项目被资助的过程中及在项目被资助之后得到怎样的支持是恰当的。

从报告中得到的信息因为包含了数据管理方面的政策对基金会来说是有用的。重要的是，NSF 应该支持培训目标，在提供给研究界足够的激励来获得他们的认同的同时，来确保研究界能实现这个要求。同时，也应该认识到，这样的要求可能使获取资源不是那么充分

的大学的申请人处于不利地位。因此，将和一些现存的已经恰当处理好这一问题相似的项目一起实施，如 EPSCoR 寻求纠正这种不平衡。

NSF 有机会为其它机构可能采用的内部商业过程建立模型。NSF 长期以来一直是基础研究机构中的领袖。与会成员催促 NSF 在数据管理领域再次担任领袖角色。正如基金会已经认识到的那样，在数字化时代，数字资源管理是科学和工程研究以及教育事业的基础。

编译自：ARL. To Stand the Test of Time. Long-term Stewardship of Digital Data Sets in Science and Engineering. http://www.arl.org/bm~doc/digdatarpt.pdf.[2007-4-19]