国家科学图书馆科技信息政策研究服务中心

英国皇家学会：科学是开放事业（节录）

Science as an open enterprise

2013 年 4 月

编撰说明

英国皇家学会于 2012 年 6 月 20 日发表了《科学是开放事业》（Science as an open enterprise）研究报告。这些对我们思考开放科学、开放数据、科技数据管理等方面有重要的借鉴意义。我们节录重点，翻译成中文，供人们在教育、科研和个人学习中合理使用。

原文：The Royal Society. Science as an open enterprise[EB/OL]. http://www.docin.com/p-586474923.html

翻译：顾立平

科学是一项开放事业：用于开放科学的开放数据

目录

总结

Summary

科学的实践：

公开质询是科学事业的核心。发布科学理论及其实证和观察数据是其他人判断、同意、拒绝、理解该项工作的基础。

各种变革正促使可理解的开放成为标准：

快速普及的技术挑战了现存科学行为的模式：数字技术对纸本交流模式的冲击、大规模数据集成与分析对个人研究者的挑战、专家和业余科学家在互联网上的网络交流等。与此同时，科学结论的可信度影响人们生活、日常判断和政府治理；而向公众释放数据，可以增加公众信任和商业活动实践。

本文深思科学活动和传播如何在新的信息技术时代进行调整，建议科学治理如何完善、科学家如何回应公众期待和政治文化的转变，以及如何增进公众从科研中所获得的利益。这些转变直指科学事业的核心，需要更多数据公开共享。

数据必须容易接触而且便于找寻，让想检查数据的人能够清楚地理解数据，数据必须可以被评价，好让人们检查数据的可靠性和研究者的能力；数据也必须让其他人能够使用。这些要求有赖解释性元数据的支持。朝向开放性的第一步是把期刊论文所依据的数据通过可以访问的数据库与期刊论文同时开放。我们正处在取得以下目标的突破口：让所有科学文献上网、让所有数据上网，并且保持两者的互操作性。

开展研究的新方法：计算和传播技术：

数据驱动的科学研究是充满潜力的新的知识来源。例如已经有人通过药性化合物数据库发现了新药，商业服务可以通过销售数据发现客户行为、从而改变其服务模式，关联数据可以跨多个数据集深度集成数据和加强自动分析能力来发现新信息，传播技术可望产生新的社会动力学等。例如，菲尔德数学奖得主 Tim Gowers 在 2009 年把一个未解题放到博客上邀请人们解题，结果一个月内有 28 个人给了 800 多条意见，解决了该问题，截至目前为止，还有 10 多个数学项目透过这种方式解决问题。

开放科研不仅能有效促进科学发现，还有助于发现、遏制和清除“坏的科学”。开放性促进系统化的科学诚信，有利于早期发现各种错误、玩忽职守、欺诈，从而遏制这些行为。但是要达到这种透明性，开放性必须与可理解性和可评价性的标准相结合，做到可理解的开放（Intelligent openness）。

促使变革：

要有效利用上述新方法，需要六个方面的变化：（1）转变那种把数据作为个人独占品的研究文化；（2）扩展科研评价的标准，支持和奖励数据传播和新的协作方式；（3）开发共同标准，用于传播数据；（4）要求与科学论文相关的数据实行了理解的开放；（5）强化数据科学家队伍，管理和支持数字数据的使用（包括私营的数据分析和政府的开放数据）；（6）开发和使用新的软件工具，促使数据集创建和探索的自动化和简化。实现这些变化的方法已经具备，但要实现它们还需要科学家、研究机构和科研资助者对使用这些方法的有效承诺。

收集数据、扩展数据库、开发工具涉及财务成本和机会成本。但是，本文提到的不同领域的许多案例都表明，按照规范标准管理数据的成本小于收集新的数据的成本。例如，国际蛋白质数据库的年度运行管理成本小于生成这些数据的成本的百分之一。

与公众沟通：

数十年来，公众、市民团体、非政府组织不断加大对科学研究结论的证据的检查评价。有些领域，已经有公众参与到一些研究项目中，成为公众科学家，进一步模糊了专业和业余科学家的界限。矛盾的是，科学交流的一个原则是“别轻易相信任何人的话”，但要理解许多领域的知识所需的技巧和知识却远远不是多数人、包括不是研究该领域的科学家等所能掌握的；例如，免疫学者对宇宙学的知识有限，反之亦然。所以，多数公众只好依赖那些它们可以判断的科研实践和标准，而非个人对具体内容的熟悉度与评价。如果需要对困难费解的科学问题开展公共政策的民主协商，那么公众的信赖度在很大程度上取决于与专业科学社群之间以及它们与公众之间的开放而有效的交流上。

数据开放的一种现实的做法是要确保公众所需要的数据被开放，而且这些数据能被获取、被理解、被评价和被使用。要做到这点，需要花费比简单地把数据向科学界同行开放更多的努力。开放数据是公众参与科学、促使科学成为健壮的公众事业的一个必要条件。

国际层面

研究者们往往在某处测试、驳斥、增进、扩展来自其他地方研究者们的结果和结论，这种国际交流往往演化为复杂的合作网络，驱动人们竞相发展出新的认识。所以，一国的知识和技术不仅来自该国的纳税人的支持，而且也来自国际上广泛的研究结果。试图控制这种开放交流，为了自身利益的短视而耗尽公用资源，会冒“共同体悲剧（tragedy of the commons）”的风险，而互联网的存在也会使得这种控制徒劳无功。

有条件的开放（Qualified openness）

开放性也存在一些法律的界限，需要保护商业价值、隐私、安全和保密要求。保护知识产权的法律在许多行业中依然很重要，这时保持数据封闭的理由应该得到尊重。如果商业数据具有潜在公共影响，例如临床数据，则更适合加大开放。

在促进个人创造知识以获得利益的各种激励措施和开放知识以便社会通过各种办法来创造性地利用知识的宏观经济利益之间，有一个平衡点。重要的是，大学从知识产权中的短期获利不应抵制国家经济的长期利益。

共享包含个人信息的数据集对医疗和社会科学研究都很重要，但面临信息治理和隐私保护的挑战。如果开放数据能被置于合适的治理框架下，这将符合公共利益。

仔细评估开放的范围很重要，以便防止信息被滥用于威胁保密、公共安全和健康；在此情况下，本文建议一个平衡和适度的开放方式，而非全面禁止这类数据的开放。

建议

Recommendations

本文分析新技术对科研活动和交流的冲击，相关建议涉及完善科学过程、回应变化中的公众期待和政治文化，以及促使研究者们最大化他们的科研影响力。这些建议致力于在海量数据时代中确保科学研究的可重复性和科学知识的自我修正，致力于促进交流与合作来做大化数据密集型导向的科学研究，采取措施在商业和公共政策中最大化地利用科学。不过，不是所有数据同等重要和具有相同价值。有些数据的确需要因为商业、隐私、安全和保密等原因而不开放，而且完整呈现数据和元数据也带来机会和财务双重成本。这里的建议是一组原则，在本文内文探讨如何判断怎么应用这些原则以及谁该负责。

建议 1：科学家应该传播他们所收集和建模的数据，使其自由和开放获取。

建议 2：大学和科研机构应该在支持开放数据的文化中扮演重要角色。

建议 3：评估大学研究时应像奖励出版物那样奖励开放数据建设工作，应包括激励协同工作的措施。

建议 4：学会、学院和专业团体应该在成员中推进开放科学的重要性，寻求对开放获取期刊的稳定财务支持。

建议 5：研究理事会和慈善资助机构应该改善他们所资助项目的科研数据的传播。

建议 6：科学期刊应该加强要求支持文章论点的数据的可访问、可评估、可使用和可追溯。

建议 7：工业部门和监管部门应该共同努力来确定符合公众利益的数据、信息、知识的共享机制。

建议 8：政府应该认识到开放数据和开放科学提升卓越科学基础的潜力。

建议 9：数据集应纳入一种适度治理的体系内；只有在有潜在的高度公共价值的情况下，个人数据才能被共享；可共享的数据类型和规模取决于研究项目的特殊需要，并采取同意、授权和安全港等措施。

建议 10：在安全与保密方面，更广泛地采用基于现有商业标准的信息共享协议和良好实践。

数据术语

Data terms

数据关系	**定义**
Data 数据	设计为某个现象的某个变项的数字、文字或图像。
Information 信息	当数据按照有可能揭示这个现象的模式而被组合起来时。
Knowledge 知识	当信息支持完整的、声称具有真实性的现象时。

数据类型	**定义**
Big data 大数据	需要大规模计算能力去运行的数据。
Broad data 告示数据	机构化的大数据，在网络上提供任何人自由访问。
Data 数据	质化或者量化的文字或数字；数据是裸数据或者第一手数据、也可能是第一手数据的衍生数据，但是不是经过计算分析后的某类产品。
Data-gap 数据鸿沟	当数据从出版物的结论中拆解出来。
Data-intentsive science 数据密集型科学	涉及大量数据集的科学。
Data-led approach 数据导向	研究假设在定义数据集的关系后所组成。
Data-led science 数据导向科	使用大量数据去发现模式的基础研究。

学	
Dataset 数据集	真实数据以电子形式的集合，不是经过分析后的产品，也不是统计数字，是不可改变以及不可调整的记录。
Linked data 关联数据	为了网络访问而统一识别命名和部署，用以表示相关数据，并且允许数据之间的链接。
Metadata 元数据	数据的数据，有关数据集的信息。
Open data 开放数据	用以评估可解读的开放性的数据，
Semantic data 语义数据	与能够揭示数据之间关系的元数据一同被标注的数据。

Intelligent Openness 术语	**定义**
accessible 可访问性	数据必须配置在能够被发现的实物以及能够被使用的形式上。
accessable 可评估性	能够判断数据或者信息的可靠性。数据必需因为不同受众而有差异化。数据必须提供科学工作结果的账目，用以解读和仔细检查它们。
intelligible 可解读性	仔细检查某事。受众需要能够对交流内容产生某些判断和评价，他们需要判断论点的本质，需要判断产生这些论点的完整性和可靠性。
useable 可使用性	数据或者信息能够被使用的形式。

第一章 科学的目的和实践

Chapter 1 – The purpose and practice of science

数字革命改变了科学与社会，本文关注科学结果和智识的有效传播，涉及如何改变过程以符合新技术带来变化中的公众期待和政治文化。

1.1开放性在科学中的作用

同行评价对理论和证据的严格分析，是一项重要的自我审查机制。科学期刊在17和18世纪对科学知识的探索具有贡献。

1.2数据、信息和有效沟通

有时人们会混肴数据、信息和知识，在此做个说明...。裸数据和衍生数据在科学中各自扮演不同角色；裸数据是被测量的数据；为了解释数据，通常需要其他背景信息或者元数据，通常包括谁创建数据、采集数据方式，使用数据的技术细节、如何选择数据，以及为了什么科学目的如何分析这些数据等。

元数据是为了数据能够被数学建模而准备，而且是重复验证所不可或缺。实现开放数据的好处，需要可解读开放性（intelligible Openness），而这需要四个条件：可访问性、可评估性、可解读性、可使用性。

1.3可解读的开放数据的力量

在2011年5月德国汉堡爆发的肠病毒，传染欧美4000人，导致50多人死亡。在中科院北京基因组研究所的深圳团队初步分析了病毒，并且与汉堡合作，三天之后在开放数据许可下公布基因组，又在一周内公布了24份报告，提供关于抵抗该病毒的相关信息；在2011年7月科学家发表了此事件的相关论文。

从临床实验中提供匿名的病历数据给医学科学家们，能够合理保障病患的隐私。它允许人们利用统计技术发现科学造假行为，它协助排除同行评议的科学论文中的不完整数据的瑕疵，并且促使更多建立在裸数据而非摘要上的后设分析研究。

开放数据能够帮助仔细审查科学结果。例如OPERA团队从欧洲粒子研究中心发射一束中微子到730公里远的Gran Sasso 国家实验室，而在2011年9月科学家们惊奇发现，中粒子看起来比光速还快，可能改变物力定律；欧洲粒子研究中心于是将其结果和细节发表在arXiv.org上，超过200篇的论文开始对此展开讨论，在2012年2月23日OPERA宣布两个可能的时间延迟缘由，最后经过四个独立文件的确认，疑似实验误差导致原始结果的错误。

1.4开放科学：理想与现实

许多科学不具有可解读性，因为它们的专家训练以及科学传播的评鉴数据并不坚持可评价性。甚至许多科学虽然日复一日从事数据分析，但是它们像是活在纸张

时代而非数字时代。

开放科学在此定义为科学出版的开放获取和有效传播内容的开放数据（可访问性、可评估性、可解读性、可使用性）组合（Open science is defined here as open data (available, intelligible, assessable and useable data) combined with open access to scientific publications and effective communication of their contents）。

最近十年开展的期刊文章的线上公共典藏，例如PubMed Central和and arXiv.org等，自2003年Public Library of Science (PLoS)开始线上开放共享论文，以及Wellcome Trust允许资助的研究论文储存在PubMed Central里。

瑞典有个著名的OpenAcess.se计划；冰岛允许所有公众在国家的ISP位置上，自由访问数量广泛的电子期刊；最近美国政府一项剥夺开放获取的研究工作法案Research Works Act (House Resolution 3699)受到科学社群的抨击而中止。

开放数据也可称作一种开放第一手科学文献的理想，包括所有已出版论文的全文立即访问等。去除缴费订阅的限制，可以增强新的文本挖掘技术和多元学科研究。这类全球政策势不可挡而且代表了科学界的渴望。不过，出版者的文献加值工作，包括为科学的准确性和完整性而遴选和编辑、增加元数据，以及为多数用户发现有价值或者精华的数据处理等。这些都是需要付出成本的劳动。

为了替换资助订阅出版的模式，出版成本将由研究者的资助方或者雇主推动到作者上，可参考英国政策的Finch工作小组报告。

1.5开放科学的重要性：科学界以外的价值

在什么情况下，英国或者其它国家有意朝向开放数据呢？

1.5.1全球科学、全球效益

单纯依赖其它人的科学不是一个选项（Simply relying on the science of others is not an option）。本地科学基础越强，越能够吸收和从别的地方的科学中获利（The greater the strength of the home science base, the greater its capacity to absorb and benefit from science done elsewhere）。有能力和天份的科学家受益于国家计划而实现走向国际网络，在那里他们能够轻易取得网络中已经成形的科学知识。这类国际协作的开放性刺激创造、散播影响力，以及容易产生对创新的关注，无论他源自何处。国家基金带来从国际交流所得的国家和全球利益。

在1997年美国研究理事会声明“对科学数据的完整开放获取，应该作为一种国际上交流公共资金资助科研成果的规范”。2007年OECD发表《Principles and Guidelines for Access to Research Data from Public Funding》报告。在2009年美国科学院的报告指出“所有研究者应该把科研数据、方法，以及资助项目研究结果的相关信息，实现公共共享。以及时方式，允许相关发现和结果被其他研究者核实，除非在特殊情况

下提出令人信服的理由，研究人员应该解释在公共访问方式下为什么某些数据被扣留而不释放”。在2010年欧盟议会的高级专家小组也建议为科学数据建立一个共通数据基础设施。

在增长中的多极世界（multi-polar world ）里，中国、印度、巴西这些科学发展快速发展的国家，以及中东、东南亚、北非这些受益科学影响的地区，许多作为国际科学理事会（International Council of Science, ICSU）的成员，已经签署了开放数据原则。OECD全球科技论坛的专家小组将在2012年秋季发布数据与科研基础设施的报告。

即便是气候变迁的DVD数据和数据集，在不同国家也有不同的访问获取形式。

许多国家无力负担国际期刊的高额订阅费。困扰许多开发中国家还有，这着威胁到建立在知识更新基础上的相关研究，以及培养下一代科学家等问题。

能够理解开发中国家获取数据的困难性。许多发展中的开放期刊，如非洲健康科学（Afirica Health Sciences）很难看到前景，它们担心更多科学资源惠及海外，而损害本地研究者。例如，印尼在2007年停止提供供流感样本的访问，因为担心已开发国家会利用这些数据制造病毒，而不利于印尼。直到世界卫生组织对此制定为疫苗和医疗的公正访问协议后，政策才有所逆转。

此外，还有一些限制国际开放的障碍。例如美国国家安全试图限制具有加密功能的软件出口到其他OECD国家，这创造了一个需要出口许可证的复杂体系。美国国家科学院2009年声明这些手续程序太过抑制，为了研究应该给予例外。然而，涉及国家安全的法律将会继续限制国家之间的开放数据。

1.5.2经济效益

科学在知识经济中扮演重要角色。基于科学和新技术的产品和服务，能够创造新的工作岗位。英国作为世界领先的科学基地和优秀的大学体系，扮演转移技术到制造的重要角色。皇家协会2010年的报告《The Scientific Century: Securing Our Future Prosperity 》提出两个重点：科学和创新是英国长期经济发展的战略核心，以及英国面临来自其他国家大规模和高速的挑战。

与此同时，数据左右我们英国的未来经济。根据英国数据资产的分析，在2011年商用数据占据251亿英镑，预估在2012到2017年增长2.3%而达到2160亿英镑。而其中过半数（1490亿英镑）是数据使用。数据驱动研发的开支将有240亿英镑。

在2004年英国政府利用语义数据整合公共部门的公共信息，英国政府数据项目在2009年创建data.gov.uk 网站提供政府的公共信息； 作为UK Strategy for Life Sciences的一部分，在2011年年中和同年12月英国首相宣布，为研究目的而允许访问

病历数据，包括开发新的医疗保健产品和服务。开放数据的目的是为了提高英国在医疗研究和数字技术。

政府主动发布资助项目的数据，使之自由而且开放重用，会刺激经济活动。美国国家气象服务将气象数据放入公共领域，估计能够驱动超过15亿美元市场的私营企业活动。英国气象局和土地登记处，在开放许可协议下公开数据，并且保持与IBM、帝国大学商学院、Grantham气候变迁研究所等合作。

在NASA开放数据，与Google取得良好成果后，大英地理调查（BGS）的三维模型也实现开放数据，开发了iGeoglogy移动APP程序，自2010年以来，已经被来自56个国家的用户下载60,000次。

在英国领导下，欧盟议会最近通过一项开放数据协议，预计每年投入1400亿欧元（**Review of recent PSI studies.**）。

从宏观经济评估而言，困难在于哪些研究数据驱动经济发展。对此，澳大利亚公共部门对于开放获取的影响分析，仔细评估了开放科学信息在经济上的价值：一次性开放公共部门的研发信息（前提是公共部门产生有用的知识，而且工厂会使用它）能够为国家经济回馈90亿澳币超过20年。

1.5.3公共和公众的效益

过去20年科学社群努力与公众共同发挥影响，特别是在公共利益科学的领域(如医疗、经济、气候、生命等科学)。英国研究理事会制作一份可以指导科学家与公众对话的纲要。

更高的透明度能够打击腐败，并且提高公众对政府的信任感（greater transparency combats corruption and improves citizens' trust in government）。科学的治理，提供可解读的开放性，是指容易理解和评价的沟通，而不仅仅是信息披露。美国2000年的自由信息法案，开创了公众从大学和研究机构等公共共部门获取信息的公共权利。英国政府在2010年对自己"踹开公共机构的大门，让人民监督政治家和公共机构"发布每一位员工和高级官员的工资信息。英国首相2011年表示"政府透明度的革命"受到创造经济价值的公众问责（public accountability）所驱动；他认为政府应该使得公民能够容易进行知情选择（ informed choices ）以及对政府实施公共服务的效能监督；科研数据属于这个范围之内；英国内阁办公室发布《正确对待数据（Right to Data ）》白皮书，给与相当压力。

支出和服务的数据往往来自各个部门或者机构，是大量、非结构化、不统一的数据集，这些是政府自己的"大数据（Big data）"，通过data.gov.uk 等协议，可以机构化数据，提供任何人在网上贴上告示数据（Broad data）。科研数据从小型定制到

复杂模型输出各有不同，它们的使用和管理方式也大不相同。科研数据不同于大数据和告示数据，应该分为不同等级。全球政府都在调整对数据的看法，包括印度的Data.gov.in在2012年2月印度发布《国家数据共享和获取政策〉公布国家计划和发展的数据，目的是最大化利用数据、避免重复劳动、最大程度地整合、信息拥有权、增加更好的决策以及公平获取。该网站没有登记或者权限控制的用户友好界面，而且元数据经过标准化处理。

第二章 为什么需要改变：挑战与机遇

Chapter 2 – Why change is needed: challenges and opportunities

本章讨论科学论文出版在海量数据时代为何与如何支持开放数据原则、开放数据与协作如何意味着新的科学和技术探索，以及有效的开放数据政策应该作为与公众进行科学交流的一个部分。大部分讨论公共和慈善基金资助的科研成果开放共享，但是也考虑私营资助的科学的交流。

2.1在充满数据的世界中的开放科学数据

2.1.1消除数据鸿沟：维护科学的自我纠正原则

如果数据不能访问和评论，那么如何挑战和修正一个理论呢？在30、40年前发表一篇论文需要包括完整的数据以提供重复操作，然而新的科技研究所需的大量数据远非期刊能够刊载。这使得科学成就的两个主要部分：思路和证据，被拆开来，形成了一道数据鸿沟（data-gap）,不利于科学的自我审查机制。

超过50%的组学（基因组、转录组、代谢组、蛋白质组）和生物医学类期刊，被接受出版的论文的数据，储存在特殊的数据中心，这个比例仍然太低。

英国皇家学会最近改版了期刊数据政策，颁布数据和元数据的标准以提供应用、研究资助者需要具体明确科研过程中的数据和元数据汇编成本，提供参考的数据集储存在Dryad (http://datadryad.org/)里。

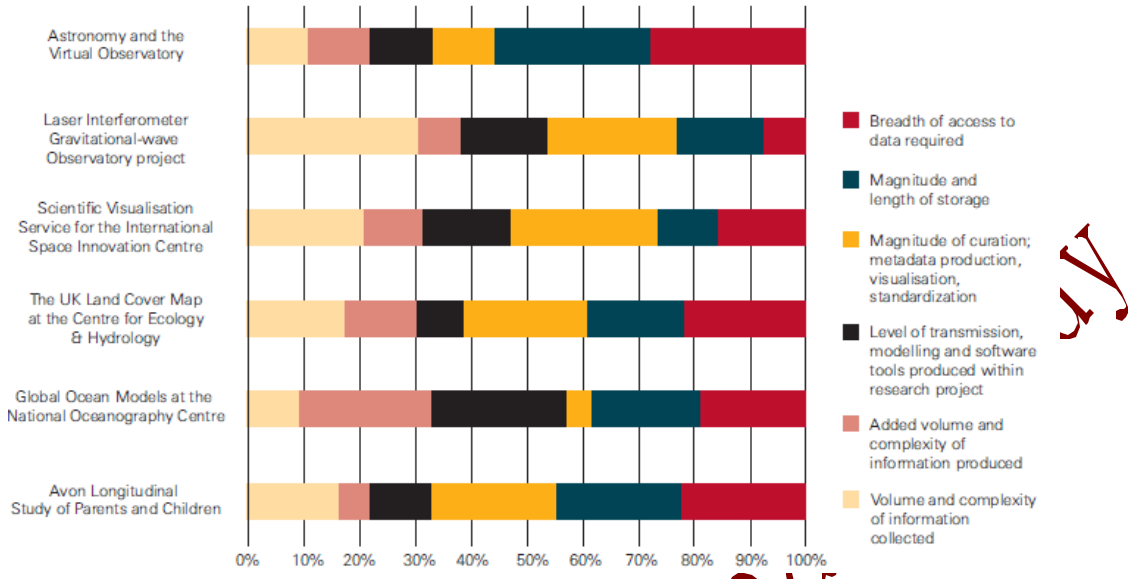
2.1.2建造信息访问：多样化数据和多样化需求

数据是构成科学论文的基础，集结数据到机构化数据集中能够支持探索和发现可能有的潜在缺失。在公开可访问的数据库中进行合作数据管理，成为一种趋势。但这并不意味所有科学领域内的数据，可以因此受益。

理解和弄明白生物问题越来越依赖在主要数据库中进行的数据分析。自从1996年的Bermuda Principle 后，基因排序数据必须立即释放到公共领域里。保持这种数据的稳定成长是一项艰难挑战。在2012年来自1000多个项目的200T的数据将被上传到Amazon网络服务云，可能导致溢出。开发储存、获取和分析的工具成为生物计量学社群的另一项工作，这些分析基因序列的工具将被用于了解人类疾病，以及用于识别新的分子以开发新药。

复杂建模的工具建立在开放数据资源的基础上。国际合作In Silico Oncology利用医学模型和病历数据开发一种描述癌症扩散的数学模型。根据新的病患的临床条件，这个oncosimulato 模型会改变参数，协助医生和病患决定采用哪种威胁较少但是效果较好的治疗方式。

目前许多科学领域收集和整合数据，用来测试理论和检验新提出的假设。根据数据类型以及获取和重用的需求，可以分为几个层次：



许多科学领域内的数据是限制在封闭的、相互信任的小群组里面。然而，有人认为，小科学的长期研究也能够创造出与大的、多的、数据充沛的科学一样有价值的数据。在这些领域的科学家们，逐渐转向大学图书馆和机构知识库去支撑他们的数据保存。以史丹佛大学为基础的国际共同协议LOCKSS里，提供一套工具来支持机构收集和呈现它们所拥有的电子内容。在机构层面上的支持，颠覆传统意义上的小科学的成功模式，使之进入到中型、大型科学的新的发展机遇上。数据保存和数据密集型科学，不容忽视。

2.1.3科学的第四范式？

信息计量不仅仅是支持传统意义上的查询，而且将会改变整个科学发展。（informatics need not merely support traditional ways of conducting inquiry in a particular discipline, but can fundamentally change the development of a discipline.）

不仅仅用所收集的数据检验假设，而且还从数据集的关系识别来建立假设。

（Rather than hypotheses being tested and developed from data collected for that purpose, hypotheses are constructed after identifying relationships in the dataset. ）

2.1.4从数据链接到出版以及关联数据技术的保障

不仅拉近出版的思想和所凭借的数据之间的鸿沟，而且也让出版物和数据之间的关系更加活跃（对于namic）。例如，用以分析化学结构的PubChem 数据库、聚集

生物医学文献和生命科学期刊与电子图书的PubMed 数据库等。

由学者、图书馆员、出版商和基金资助者所组成的国际团体对Force 11的评价是，期刊作为一种知识交换的形式，主要有赖能够转移为数字对象的研究对象，例如数据集、工作流、软件包等，动态链接文章和数据能够让读者阅读文章的同时，也调用数据；而Force 11让期刊论文成为“可操作的（executable ）”文章。

在出版论文的主体上，衍生交互性元素，读者能够操作第一手数据而且重新计算论文结果，在基础架构上，这种服务能够被整合到出版商的门户。与文章相关的数据库链接，也可链接到其他相关的数据库上。

关联语义数据的技术，可以更深层次地链接，因为语义数据是揭示元数据和元数据之间关系的数据。机械可读的信息可用来描述数据之间的关系类型（types of relationships），而不只是数据之间有关（related）的事实。识别符和描述符的相关全球标准有URIs、RDFs等。而BBC、Thomson Reuters和美国国会图书馆目前共有240个数据集，以及W3C自2007年开始的SWEO Linking Open Data Community Project等。

如果搜索引擎是第一代排序网络型知识的工具，那么语义分析则呈现第二代不仅是文件列表，而且是揭示它们彼此关系的工具。

然而，关联数据集的问题在于深层次和更优良的整合理解，在异质性关联数据集中的元数据非常不同，而由此产生的词汇含义则差别甚远。解决之道是提升搜索元数据的系统，自现存复杂搜索引擎的自由文本索引中取经。

2.1.5复杂计算模拟的出现

数学模型是科学理论正式的量化阐述。它使科学家们能够对一个问题进行近似精确的定量分析、理解和预测。然而，爆发性增长的计算能力将数学模型提升到一种更为复杂的层次。它让科学家们能够模拟复杂系统中的行为，例如气候、基因结构或者城市变化。

当前的模拟技术渗透各种科学实践，从理论到经验研究。计算机模拟可以比喻为一场物理实验，然而实验的是数学模型而非物理实体。模拟出来的是一个高度理论化的模型。模拟的预测能力和再现真实的能力，取决理论表述所达到的最佳准确度；不少建立在不确定性之上的假设，带来模拟的预测失误；因为不确定性是对参数和关系缺乏理解，而它们却是运行计算模拟的重点。最常见到的是非线性方程，可能由于数据小变，而导致输出结果的大变。

模拟是不断发展和纳入新见解的主要工具。相关工具有BLAST、SPiM等。科学社群开始发展共同标准，如SBML、CellML和EBI等，接触模型共享和模型交流的议题；这些标准的最小要求是能够比较和测试相反的数据。而且最好是由研究资助单

位推动软件代码、对科学社群和同行评议的开放和获取。

科学实践中的模拟应该要能够重复，British Atmospheric Data Centre 对保存信息和数据以供模拟，提供了指引。

2.1. 技术促使网络化和协作化

自1940年代起，意识到科学对国家经济起到重要作用。John Ziman 声称科学本身所环绕的优先议题将不再集中在个人或者小团体的研究工作，而是集体协作。发表在期刊上的论文有35%是属于国际合作，在15年前是25%。

免费网络资源和搜索引擎，取代了信息、搜索和目录等来源的图书馆。许多工具，如myExperiment提供分享和实现科研工作流的功能。在Wiki和博客上，能够及时而且开放地进行动态的学术讨论。2009年菲尔德讲得主，数学家Tim Gowers 在博客上，提出问题，得到来自27位800多条留言，在一个多月内解决了核心问题，他感慨地说道“这就像开车和推车那样不同（“It felt like the difference between driving a car and pushing it）”。

2.2开放科学和公众

2.2.1透明度、沟通和信任

好的传播是可评价的传播，不仅允许人们理解和发言，而且也允许人们评价支持发言内容的理由和证据。如果科学界不能对科学理论评价，那么就不能判断何时以及为何这个论点具有真实性。真正需要的是良好交流，而非简单的透明化举措。

错误地传达与公共利益相关的科学信息，会损害公众对科学研究的信任。东安吉拉大学对气候变迁的研究，缺乏支持论点的数据；并不支持自由信息法案所倡导的数据共享作法。更大的数据集的可获取性，可以减轻自由信息法案下的数据获取声明，并且产生更多可解读和可重用的数据。然而，自由信息法案只提供一种数据共享的方式，而且常常不能令要求数据的人和被要求释放数据的人满意。它所要求的数据往往没有上下文脉络，而且不利于数据分析和重用。

在2012年自由保障法案Protection of Freedoms Act 中，回应了这种义务，要求所提供的信息必须是在合理可行范围内（reasonably practicable ）可以被重复使用的电子格式。但是，对“合理可行（reasonably practicable）”或者经济实惠（affordable）的需求被严重低估了。有些科学数据很容易被不熟悉该研究项目的人拿来重用，这对许多数据而言并不合理。解决这个问题的方法是科学家们为公共利益的数据准备好一套经济实惠的数据集和元数据。如果考虑成本，不得不同意这种公共利益数据将会进一步强占自由信息法案所要求的相同的数据集（从而使得相关经费豁免）。未来国际和国家数据中心的一部分工作，如英国研究理事会的研究协议等，应该是为这些数据集建立更为广泛获取的机制。

2.2.2公众对科学的参与

如果查询者（inquirer）的知识源头来自所查询的数据，那么还需要一个科学背景的描述。鉴于数据产生者（data originators）和数据科学家们提供数据给 从领域高手到非专家们的用户，有必要选择哪些数据优先提供公共使用。

公共开放性必须采用受众感知的方式（Openness to the public must be audience-sensitive）。它必需正视公众需求的多样性。未经正式训练但是对某些议题怀抱兴趣的“数字声音（digital voice）”正在形成，例如纳米、HIV流感、AIDS病毒、气候变迁等。

几个项目包括：Fold.it让公众解谜和设计蛋白质氨基酸的结构、Galaxy Zoo让公众为太空望远镜的照片分类和分析、BOINC让自愿者的家用计算机参与大科学的计算。

2.3体系健全：揭露差劲和欺诈行为

获得科研资源和作出重要发现的奖励是相当可观的。无论在公共或者私营部门，这些奖励对科学家有极大诱惑，不少人为了突出某个特定假设，沉沦在造假数据和选择性呈现研究结果的贫贱手段（poor practices）里。

好的科学，简言之，是“直视所有证据（而不是选择性呈现）、掌握变量以便我们确认真正运行的部分、盲法观察（blind observations）以减少偏见的影响，并且内部逻辑要有一致性”。忽视这种严谨性就属贫贱手段（Poor Practice）。

科学造假在医学和物理学领域的严重案例有Jan Hendrick Schön和Woo Suk Hwang两个事件。Schon在两年内造假了17篇文章，如果他的研究正确，将对固体物理学产生革命；而Hwang对干细胞的研究产生破坏性影响，许多不治之症的患者期待干细胞带来新的希望。

在2000年到2010年间，估计有8万多个病患手术后取消了与研究相关的实验。近几年，撤回临床实验论文的速度比论文还快。然而，仍然不清楚这样是否提升了论文品质。

新的协议，如CrossMark等，能够告诉读者论文的最新版本以及是否修改变动过。这种新的服务是确保修订后的结果准确传达到研究人员的重要举措。

只是无效而非造假的科学论文应该不被撤销，然而，撤销了（Scientific papers that are merely wrong rather than fraudulent should not, however, be retracted. ）。了解错误的增加是科学自我审查的一部分，也是探索的一部分。诺贝尔奖得主Richard Feynman认为“如果您决心测试一种理论，或者解释一些想法，您应该发布它。如果我们只发布一种结果，会让论点看起来很好。我们应该发表两种结果”。

误差来自不良的经验设计、数据收集、数据分析或者数据呈现，以及早期错误

或者统计错误等。对误差的某种扭曲理解的形式，是强调在医学领域不发布负面的观察结果。而British Medical Journal 曾经有过声明“有意隐瞒数据是严重违反道德的行为，未能披露临床数据的研究人员应该交由专业团体纪律处分”。

第三章 开放性的范围边界

Chapter 3 – The boundaries of openness

开放并非无条件式，有四个领域存在必要限制：商业利益、个人信息、安全性和国家机密。有效管理这四个领域可以优化开放效益。

3.1商业利益和经济效益

如果商业力量占据主导优势，数据可能只有关系人同意后才能共享，而且理论上会引起另一波国际淘金热。在1997年联合国通过《人类基因与人权宣言》对此作出贡献。然而效果不明显，国家专利局批准上千件包括人类DNA在内的专利。基因发现是否符合专利性的法律条件，或者是否属于全人类共同资产，对此需要寻求平衡。

数据越来越被看作一种商业资产，囤积而非共享的力道增加。“数据是最大的商务原料，如同资本和劳力（data is the greatest raw material of business, on a par with capital and labour）”渐渐成为共识。

Google的数据比欧洲生物信息研究所和大型粒子对撞机的数据加在一起还多，而且形成一个数据分析行业。专门从事数据管理和分析的公司资产，估计超过1000亿美元，而且每年以10%的速度增长，比整个软件商务快了两倍。他们的服务能让公司了解潜在客户的习惯和优先顺序，并且确认采用那种最有效的销售方式。但是，勾引客户的数据集的价值，与科学数据集的价值不同。估计在2012到2017年英国经济从数据相关行业中获益2016亿英镑，而来自客户情报、供应链管理和其他商务比数据驱动科技研发的收益多了六倍。

看似公共部门和私营部门对数据的开放性是分歧的，但是他们的经济效益实际上是交叠的。

3.1.1数据所有权和知识产权的实行

政府应该重视公共资金资助的研究数据重用的潜在商业价值。政府政策不利其他人重用数据，而且不必顾虑数据所有权，实际上他们所资助的研究结果反倒激励了商业剥削（Government policies are not inimical to data reuse by others and does not necessarily require the funders of research to assert ownership of resultant data, indeed most encourage commercial exploitation of research results by those they have funded.）。与此同时，扩大政府的数据集并且提供重用，如气候或者社会行为等时间序列的数据，显得极为重要。

英国经济与社会研究理事会，有一项相关政策，最大化重用它们所资助研究项目的数据，同时保有知识产权以便谋求国家效益。

知识产权的界定所有权。它是一种控制方法，但是不必因此而限制访问。数字材料容易复制的特性，破坏了传统的知识产权的成效并且创造出新的拥有权类别。GNU General Public Licence 和and Creative Commons ShareALike licences 维护着信息自由流通的权利。

专利经常被视为开放性的障碍，尽管它的核心目的是共同使用信息，而非用于商业机密。专利所有人控制专利，透过许可证允许自由流通信息，提供进一步的研究和创新。

许多国家在知识产权法中写入研究例外，然而在欧洲极不协调，而且美国法院如此狭隘地定义的专利例外，使得几乎没有例外。

最近UK Hargreaves Review of Intellectual Property’s 建议豁免非商业性的文本和数据挖掘，这将授予英国科研人员新的利器：在期刊里搜索某个化合物的论文，或者二次分析统计数据。英国政府已经原则上接受这项建议，该报告期待实现在英国法律中的合理例外的改变。

排斥或者取消知识产权不太可能使得更多科学数据可用。英国和其他 21 个国家签署了在 2012 年 6 月由欧洲联盟进行辩论 防伪贸易协议（ACTA）。普遍担忧和抗议对网上内容的不当限制。此外，还有争议的美国Stop Online Piracy Act (SOPA)和 Protect Intellectual Property Act (PIPA)等。

科学界可以集体行动，推动知识产权与开放性的共存。数据库权持有者，可以发布授予非独家许可（non-exclusive licences）和使用条件的意愿。专利池可以设置专利拥有者同意协调许可权的举动，并且协助避免错综复杂的专利问题，当某个科学领域内充斥着知识产权就不可能导航。专利交流中心可以控制特定领域的专利，当其他人便于获取的时候，为专利拥有者索费。在Hargreaves Review 里，并没有改变与科研社群相关的知识产权（除文本挖掘外）的建议。与知识产权相关的问题主要不是它的形式、如何部署，而是谁在使用它。

Human Genetics Commission (HGC) 关于知识产权和DNA诊断报告反应出典型的紧张局势因为强制执行专利的结果导致，一些北美的临床实验室停止测试。HGC建议生物医学研究的资助者应该检讨它们对许可证的指南。

3.1.2知识产权在高校研究中的实行

Engineering and Physical Sciences Research Council (EPSRC)调查公司报告，研究公司与大学合作的障碍，因为知识产权导致的潜在冲突从2004年的32.4%涨到2008年的55.6%。

高校严格控制知识产权的经济理由是可疑的。从2003年第4季度到2009年第4季度这7年间，英国大学的收入上升35%其中有2.6%来自知识产权，包括销售，但是整

个期间没有明显增加。正规的技术转让也是明显的低回报，这表明研究价值不在它的所有权，也不保证在技术转让办公室里严格控制。一些知识产权的商业价值被高估，而且过早在知识生成阶段行使知识产权权利。短期内有利于大学财政，但是长期不利于国家经济。知识产权局2011年5月更新的知识产权战略指南，建议大学采取更为灵活的管理方式。

3.1.3公共-私人的互惠关系

许多公司所增加的开放创新过程，采用外部思想从事产品和服务开发。学术研究建立了研究者和工厂之间的新关系，但也产生一些特殊问题。在各种合作关系中，必须明确定义包括数据在内的开放和受限制的素材，以及知识产权的范围边界。

根据2012年的自由保护法案，大学必须向任何申请者提供数据共享，在可重复使用的格式中，允许申请者重新发布信息。这些要求可能会破坏企业和英国大学合作的信心。应该为此界定谁拥有这些数据，以及谁拥有它们所储存的位置（Arrangements will need to specify who owns the data, and perhaps who owns the locations in which they are stored.）。

四个案例：

InnoCentive：公司把问题或者科学难题放在网上，并且提供悬赏金额。

The Structural Genomics Consortium (SGC) ：促成公共和私营部门合作开发药物的三维蛋白质结构基础科学的非营利机构。

Rolls-Royce University Technology Centres：始于1980年代末，受企业资助工程类别的研究计划，知识产权归大学所有，但是许可证授予资助者使用。

Imanova：医学研究理事会和三所大学的联盟，发展和加强利用影像数据从事生物医学研究的新方法论，它训练科学家和物理学家，并且希望成为国际药厂的合作伙伴。

Syngenta：与技术战略委员会（The Technology Strategy Board ）共同建立一套帮助科学家们模拟与可视化ChEMBL分子的开放源码Ondex 软件以及用于药品开发的数据库。

数据往往不是这些合作关系的通用货币，例如Syngenta对开放源码的可视化软件具有贡献，尽管软件是开放的，但是往往难以开启数据。在英国，鼓励企业和大学合作生产数据的政策还在起步阶段，而且根据先前立法遗留的不确定性，可能还会停留在起步阶段。在Tim Wilson爵士的评述中，集中论述了人类网络所需的协同工作没有仔细查看网络共享和控制数据的方式。

英国的Catapult Centres 提供大学和企业合作的物理基础设施，还需要强化数字基础设施以及基于数据的知识经理人（data-based knowledge brokers ）。不过这些都

将在英国信息自由法案所设立的框架下执行，

法律实际上影响着数据密集型科研的合作类型。

3.1.4揭露公共利益中的商业信息

Medicines and Healthcare Product Regulatory Agency 决议：公众代表产品可用性（以及可持续性）；在保护合法商业利益、个人隐私与安全、国家机密的情况下，公众能够获取信息。

公开披露从私营资助的研究所产生的数据和信息的合法商业利益管理，因为符合公众利益，应该认真考虑（Managing legitimate commercial interests in the public disclosure of data and information from privately funded research that is of public interest warrants careful consideration.）。

例如，在获得知识产权以及成为一种特定商品或者服务后，应该公开信息并且使得数据可用（Managing legitimate commercial interests in the public disclosure of data and information from privately funded research that is of public interest warrants careful consideration. ）。然而，应该承认，商业秘密是知识产权的重要组成部分，某些类型的研究（例如，制造工艺的研究）是有限的公共利益。涉及安全和紧急安全的研究，所需的信息和数据应该优先于商业上的考虑。

临床实验登记制度，或可平衡商业利益积累，以及在研究数据出于安全评估或者公共决策审议的目的而访问研究数据的公共利益。在ClinicalTrials.gov中，要求工业赞助厂商的临床实验的结果摘要在它们的数据库上向公众披露，而公开披露的数据不会影响赞助商取得专利；从而使得公众利益不会因为商业掌控研究数据的优势而受到压迫。

3.2隐私权

包括个人信息的数据集，对医学和社会科学研究而言十分重要，但是也有危及个人隐私的信息治理的挑战。公众有合法保护个人隐私的权利，避免个人信息被利用、羞辱、歧视、侵犯人身自主权。“尊重私人和家庭生活的权利”的法律框架建立在欧洲各国遵守的欧洲人权公约第8条上。隐私权编纂自欧盟数据保护条例（EU Data Protection Directive (95/46/EC)）和英国实施的数据保护法案（Data Protection Act 1998 (DPA)）。

在医学研究中使用个人数据：

Huntington’s disease：该疾病通常在中年发生神经快速退化的症状，源于染色体基因突变，取自HD家庭的数据包括许多详细的个人信息，如家庭成员的性别和年龄等。

UK National Cancer Registration：英国癌症患者通过医院、 癌症中心、 安养院

的筛选计划和 GPs 被收集数据，在特定癌症情况下，会透露病人姓名给研究者调查；这类研究若要发布信息，必需通过研究伦理委员会及其伦理和保密委员会。

UK surveys on access to patient records：在2006年医学研究理事会的 Ipsos MORI survey 报告说69%（包括14%肯定）的英国民众同意他们的个人健康信息用于医疗研究，有四分之一的人不同意，包括7%的人表示决不同意。有89%的人不信任公共部门研究人员处理他们的医学研究记录，而且 96%不相信私营机构的研究人员。要向研究人员开放医疗记录，还需要更多如何保持机密性的公共信息内容。

DPA定义个人数据为可识别或潜在可识别个人的数据，规定收集、 存储和处理个人数据的规范。个人数据的处理必须符合欧洲普通法和ECHR 第 8 条的保密原则与责任，在已获得知情同意，并且数据主体已清楚了解事实、所涉问题和潜在后果后，才能处理个人数据。

这一方面并不免除遵守安全标准和告知义务的数据控制；另一方面在不明确同意的情况下因为具有重大的公众利益而处理数据。

它假定数据当事人的隐私可能经过匿名、 除名等处理过程。然而，大量计算机科学的工作证明不能通过匿名程序保证个人记录在数据库中的安全。所有包含有关个人信息的数据集，即使匿名化，也能够用与主题相关的其他信息进行推断。

公共记录匿名的破解（ad hoc） 方法有个著名案例：研究人员设法通过链接两个或多个独立的、看似无害的数据库取得个人信息。Latanya Sweeney提供了麻省GIC的案例作为有力证明。GIC健康保险大约拥有135000 名员工和家庭。上世纪 90 年代中期GIC向研究人员和商界提供去除详细内容的匿名（如姓名、 地址和社会安全号码）病患数据。所公布的数据，包含每个人的邮政编码、 出生日期和性别，以及诊断和处方等详细信息。Sweeney 花了20美元向麻州购买选民登记表，其中记载每个选民的姓名、 地址、 邮编、 出生日期和性别。对它们的公共字段 （邮编、 出生日期和性别）关联两个数据集，匹配出特定的个人诊断、过程和药物。例如，从州长的出生日期找到六个人，其中三个男人和州长的五位数邮编匹配，有可能找出州州长的医疗纪录。

很难作出平衡个人私隐和潜在利益的判断，例如，释放给公众更多公共卫生福利的数据。Joseph Rowntree Trust 的数据库状态报告指出：在扩大政府持有我们生活的个人数据，以及扩大GPs和医学研究人员访问病患记录的国立卫生研究院条例的变化上，公众既不被服务也不被保护。在RUCK对开放数据的公共对话上，与会者普遍看待数据保密性只要在适当之处予以治理，但是关注这个议题的人极少。这些涉及公共价值的议题，不是技术能够轻易解决，应该得到更为广泛的公共辩论。

在隐私利益的背景下，将不利科学承诺开放会直接进入公共领域的个人数据。

本文主张一种为研究目的而共享、编译和链接个人数据的数据集方法。公共利益和保密风险需要评估，在个别情况下，重视没有无法完全平衡，开发各种治理机制的目的是为了促进研究和其他目的数据访问，同时减少隐私风险。例如，知情同意以及安全地使用限制访问的数据，如果研究者们违反合法的保密权利应该受到惩罚。尽管法律、 伦理和实际的考虑因素所拼凑的结果往往让研究人员很难顺心如意（navigation；导航），应该让一个独立的监督机构提供可使用特定形式数据的途径的建议。

同意使用个人数据通常被认为是信息治理中的黄金标准（Consent to use a person’s data is often thought to be a gold standard in information governance.）。然而，既不必要也做不到保护所有利益方，无论从道德或者法律上来看。

许多研究所涉及的研究问题，没有预料到同意采集原始数据的问题，对所有学科的原始数据重新订约（re-contact）不仅困难而且昂贵。

英国生物库（UK Biobank）以及它的道德治理委员会，定义数据的适当用途并且判断是否需要进行 重新许可（re-consent）的作法，或可为其他机构授权经营公共利益提供借鉴。作为2008年数据共享评论（Data Sharing Review）的一项建议，获得授权的研究人员只能在数据库的安全站点访问包含了敏感的个人信息的数据。核可的研究人员必需遵守严格规则，如果违反保密，有最多两年的监禁刑罚与刑事制裁。保密协定防范滥用信任，它制定研究人员访问敏感数据集的行为；访问包含个人数据的数据集的所有人员都要签署保密协议。 苏格兰纵向研究（The Scottish Longitudinal Study (SLS) ）是共享复杂和敏感数据的较好范例。SLS整理来自日常行政和统计调查的数据，包括人口普查 (1991 年和 2001年） 的数据、 重大事件数据 （出生、 婚姻、 死亡）、 NHS 中央登记 （迁出入苏格兰） 和 NHS 数据 （癌症登记和出院）等，用来审视迁移模式、不平等现象、健康、 家庭重组和其他人口、流行病和社会经济的问题。这类数据是一个宝贵的社会决策信息来源。为了保护人们的隐私，具有一系列措施：第一，补充劳工计划所预定获得的个人数据，只有一小群研究人员知道这些日期；第二，数据集是匿名，调查中涉及到的个人没有姓名或者地址保留在数据库上；第三，实际的数据存储在一个独立网络，受密码保护，只能在受保护的位置访问数据；第四，督导委员会，负责维护和采纳研究理事会审查的每一个研究建议，将不授权任何可能会确认个人的研究。第五，数据不是公开。此外，严格控制访问数据，如果研究人员想要远程分析数据，可以由补充劳工计划中心代表他们运行统计程序。

未来的数据治理，需要反映数据分析技术正在改变的速度。因为重组数据的技术提高，保护隐私只会更难。治理过程需要权衡潜在的公共利益以及抵御最新技术

带来的风险。

3.3保密性和安全性

开放科学信息对系统与硬件工程师提出挑战，要求开发共享机密的、敏感的、特级数据的安全方式（防止意外事故和避免蓄意攻击）。数据密集型未来可能会增加信息安全的担忧。敏感数据的泄露是一场不可避免的赌注。个人数据的持有者面临不仅是管理，而是即使采用密件格式，去匿名化技术（de-anonymisation）也可添加更为详细的数据。已有迹象显示数字安全的落后，只有不到1/3的数字信息受到最低限度的安全保护，而且只有一半应该受到保护的数据被保护。

保有源代码和系统架构的机密，不是确保信息安全的可靠作法。现代密码学来自保证系统安全的必要条件是潜在攻击者知道内部运作方式。开放源码和系统架构虽然能够让攻击者分析漏洞，但是也允许对系统进行更为彻底的测试。这种“开放性最终形成更好的安全性（openness ultimately breeds better security ）”作法，可以同样运用在科学数据。

科学探索通常具有两种潜在用途-效益或者破坏（Scientific discoveries often have potential dual uses - for benefit or for harm.）。基于国家安全拒绝公布数据集的案例很少见，自然出版集团在2005年到2008年间收到74,000份意见书中，只有28件标示具有双重用途。虽然本文没有具体出现危害的案例，但是不至于傻到认为它没发生过。目前，英国通过Export Control Organisation限制敏感信息的出口。

英国皇家学会、跨学科小组和国际科学理事会在 2006 年的一份联合报告的结论：限制新的科技进展信息的自由流通，极不可能防止潜在滥用，甚至可能鼓励滥用（restricting the free flow of information about new scientific and technical advances is highly unlikely to prevent potential misuse and might even encourage misuse）。

脊髓灰质炎病毒的序列在 1980 年出版。在 1981 年从克隆 DNA 重新创建活脊髓灰质炎病毒。在1981 和 2001 年间，其他病症从克隆 DNA 中受益，重新创建脊髓灰质炎病毒的能力增强对病毒的认识，并允许派生更稳定的疫苗。

资助者目前屏蔽有潜在双重用途的研究。但不寻常的一系列环绕禽流感的事件，引发对科学素材的安全和保障的普遍关注。

H5n1 型禽流感病毒很少感染人类，而且不容易在人与人之间传播。病毒可能演化而传染人类，从而严重威胁全球公共卫生。研究影响 H5N1 病毒的人际传播能力，对了解这种可能的威胁至关重要，但是帮助预防病毒的信息也可以出于有害目的而被滥用。两份提交出版的手稿，描述了实验室的结果， 对h5n1 型病毒有更大破坏潜力的结论。美国国家科学咨询委员会的生物安全措施 （NSABB） 建议，作者和编辑与全球流感监测和研究社群应该在期刊上发布一般结论，但是不能开启复

制实验，以避免那些可能伤害人类的细节。

《Sciences》与《Nature》期刊最初支持 NSABB 裁决，但同时强调，研究人员有必要获取这些工作内容，以便维持适当的科学审查。研究人员在 2012 年 2 月在世界卫生组织会面后达成共识，认为减少文件内容并不是一个实际选项，最好的解决办法是发布的完整文本。《Nature》期刊目前已经版他们的论文。

早在 2005 年的美国科学家小组就解开1918 年 virus流感的序列，另一个小组接着发表了一篇论文描述了如何使用此序列重建完整的病毒。

虽然帮助理解病毒，文章内容也让恐怖主义团体更容易使用。因此，出现了督促谨慎小心的声音。NSABB 检查之前出版的两篇文章，并且得出结论是发布的好处大于潜在危害。

这个案例表明，国家科学咨询委员会对物种安全而言，并没有权力对在公共领域内的论文版本进行存储，例如，大学服务器上的文件版本。这打开了对“（cyberhygiene）”研究数据的安全议题。

JISC开发的美制Shibboleth单一登入系统，省去内容提供商的用户帐密，允许机构控制用户访问信息的权限。从历史上看，数据的机密性已几乎是安全的代名词。保护个人数据的安全，是创建安全系统的主要动机。最近更多的事态发展表明，确保数据的完整性和来源，以及保持数据可重用，也是创建安全系统的重要动机。有大量接受标准的商业团体采用这种做法，这些应被视为最起码的科学数据安全管理标准协议。

专业科学家行为守则有一部分鼓励发挥个人责任（Codes of conduct for professional scientists also have a part to play in encouraging individual responsibility ）。尽管如UK’s Universal Ethical Code for Scientists报告指出，应该个别看待，所有科学家应该与雇主签署约定，遵守所有国家和地区的安全法律。

第四章 实现开放数据文化：管理、责任、工具和成本

Chapter 4 – Realising an open data culture: management, responsibilities, tools and costs

第1章认为数据应该是可解读开放性，而不是封闭性。第2章探讨新兴的通信技术创造开放机会，而第3章讨论如何看这个开放边界。本章讨论开放的实际准则。共享研究数据，可能复杂而昂贵，需要估计这些数据的实际需求现实。 4.1以分层分类方法对研究数据进行不同层次的保存。 4.2数据持有者应当如何在这个分层方式中操作。此外，考虑数据管理的工具（4.3）和运营成本（4.4）的限制。

4.1数据管理的层级

了解目前的科学数据管理模式，在一定程度上，有助于思索四个层次的活动，它们反映了规模、成本和管理数据的国际影响力，每一层都需要不同的财务和基础设施的支持。

Box 4.1 The Data Pyramid – a hierarchy of rising value and permanence

Details of examples given in appendix 3.

第1层包括主要的国际案例，例如在大型粒子对撞机生成自己的数据，又如全球蛋白质数据银行聚集大量来。前者依靠在一个复杂网络上释放数据，而后者则依赖工具提交、保存和释放数据。

第2层包括数据中心和托管资源的国家机构，如英国研究理事会或WelcomeTrust基金。

第3层是由个别机构、大学和研究机构保存研究计划所产生的数据。他们彼此之间的差异很大。

第4层是个别研究人员或者研究小组。研究小组整理和存储自己的数据，往往只提供值得信赖的合作者，但也可能让公众通过自己的或机构的网站获取所储存的数据。研究人员通常在传统的、现成工具的、小范围内里提交如Excel或MATLAB等数

据，但是缺乏有效保管、使用和可持续性的功能。

以上层次不包括私营公司基于商业用途所创建的数据库，例如业务流程、客户和供应商的数据等。这些公司在上述每层共享数据的活动里，其实也有利可图。最有前途的私营企业关心数据的访问和重用，而不是囤积数据的工具。例如，搜索引擎作为公共数据源和数据用户之间的中介。

Scratchpad Virtual Research Environment 是一个在线系统研究网络，支持自然历史科学的人，成立于2007年3月。目前已有340个网站被创建，原来的目标是支持植物和动物物种的分类，但是这已经扩大到包括当地活动和兴趣小组的网站。Scratchpad Virtual Research Environment 正在试行一种机制，从一个网页内容生成一个XML文件，提交期刊出版。在出版商那里呈现XML的PDF文件草稿，并且自动转发给评审，如果收到正面评审意见就印刷出版。

4.2责任

一个有效的数据生态环境，必须适应不断变化的研究需求、行为和技术。动态系统需要一套原则：

开放标准：开放的和负责任的共享立场，提供访问以及合理限制。

明确的政策：数据质量和访问应该有透明的政策。

清除路由访问：建立明确路径的数据寄存器。

预规范数据：事先指定如何重用新近获得的数据。

尊重价值观：结合保护个人隐私和商业利益的治理机制。

共享的规则：进行数据共享的明确条款及条件。

4.2.1机构战略

数据层次的第3层有两个当前问题：他们应该有什么样的责任，以支持他们的研究人员所需数据的保存；他们应该有什么样的责任，以保存它们产生的数据？

研究创建的数据经常被丢弃，因为它几乎没有长期价值，于是，失去了许多具有重用潜力的重要数据，尤其是当研究人员退休或者转换到别的机构。本文建议机构应该建立并实施旨在避免这种损失的策略。

大学的一个特别困境是：确定他们的科学图书馆在数字化时代的作用。图书馆作为一个帮助学者访问数据、信息和知识的传统角色仍然存在，但在数字时代，完成相同的功能的过程和所必需的技能是根本不同的，他们应该为学者和研究人员提供那些网上科学文献和所有线上数据两者的互操作。

4.2.2引发数据释放

公布数据的时间是开放数据文化的一个重要问题。在一些领域，如基因组学，科学家已经适应了立即释放数据的规则。但是，研究人员或多或少不愿意公布他们

已经收集的数据集，担心被“挖出”来出版，所以有短期和明确期限的独占访问好让研究人员分析和公布结果是适当的。

本文表明，机构应该预先制定数据释放的时机和做法，作为研究资助拨款申请时必要的数据管理计划的一部分。海量数据的增加无疑会增加数据共享的进一步实际障碍，但是这不是断然拒绝提供数据的借口。重要的是要建立一个广泛的共识：囤积数据不利于科技进步。不过，促使所有数据即时释放既不现实，也不是一个理想目标。

4.2.3熟悉数据的科学家的需求

科学数据有个快速增长的规律。数据科学家的技能是支持研究人员和机构的数据管理关键。他们擅长数学，而且还是对信息工具和数据管理流程训练有素的某类科学家。美国国家科学基金会（NSF）的报告描述结合专业和信息分析师（informaticians）技能。

4.3数据管理的工具

现代数字数据库的数量和复杂性，导致了归档、整理和提取数据已成为苛刻的技术任务，而需要能够有效运行的复杂的软件工具。此外，数据归档的现有水平的可持续性是个严重问题。

附录2总结了一些软件工具。需要有确保可随时更新、链接数据库、提早因应“过时”的方法。索引数据需要有更好的工具，才能在需要精确数据的时候，可以很容易地从更大的数据资源中提取。

跟踪数据源，对其评估以及将数据归属发起人，是至关重要的。需要为可解读开放标准（方便、清晰、可评估和可用性）设置共同的标准和结构，还需要允许数据重用的人不只操纵数据，还能与其他数据集整合。

4.4成本

除了大规模的数据集，如大型粒子对撞机或欧洲生物信息研究所的数据，数据存储和备份的成本相对总成本而言是很小的。例如，全球蛋白质数据银行（wwPDB）是世界范围内的大型生物分子的三维结构信息库，拥有约80,000结构，但它拥有的数据，占地不超过150GB（一个比笔记本电脑的硬盘空间）。

同样在第2层，大型数字储存通常约占研究费用的1-10%之间。在第3层，在2011年75所英国大学的调查表明，大学图书馆平均雇用了1.36全职员工进行管理、行政和技术结合的相关工作。

澳大利亚国家数据服务成立于2008年，创建和发展的澳大利亚研究数据共享（ARDC）的基础设施。他们的主要活动之一是数据采集、建立机构内的基础设施、收集和管理数据，以及提高元数据的管理方式。这向英国的双重系统抛出问题：如

何划分研究数据、研究项目和机构之间的管理责任呢？

第五章 结论和建议

Chapter 5 – Conclusions and recommendations

研究数据的可解读开放性的机遇体现在众多科学领域，有了这些经验作为指导，应该加速变革和协调科学家、研究机构、基金、公众的意见 。这份报告建议：为提高研究工作，一个新的科学时代将改变研究人员的行为以及传播成果的方式。这些变化将提高科学行为和应对不断变化的公众期望。但并非所有的数据都应一视同仁。有些具有机密、商业、隐私、安全或保密的理由，应用数据和元数据进行有效的沟通必须谨慎判断。

当务之急是要确保在科学界的科学家、研究机构、投资者、出版商和政府，认识到六大变化：（1）远离私人保存数据的研究文化、（2）扩大评估的研究，给与数据通信和协作的信任（credit）标准、（3）发展数据传输的共同标准、（4）授权与出版科学论文相关的可解读开放数据、（5）强化支持和管理数字科学数据使用（私营部门的数据分析以及政府的开放数据战略），以及（6）发展和应用新的软件工具，以自动化和简化数据集的创建和抽取。

5.1 国家科学院的角色

英国皇家学会将与其他国家科学院和国际科学联合会，鼓励国际科学界实现可解读开放数据的政策。它还支持建立全球科学数据和元数据标准的举措。在2012年4月包括英国皇家学会的会员院校的ALLEA签署了一项促进出版物、研究数据和软件开放的科学原则的承诺。

英国皇家学会支持全球科学界的努力，确保研究能力相对有限的国家能够公平受益，努力扩大全球研究数据的访问。在低收入和中等收入国家的研究经费应包括进行数据管理和分析的可持续方式，以提高国家能力的努力。

至关重要的是：共享方式要平衡生成和使用数据的权利和责任，并承认参与研究的个人和社群贡献。本文认为，英国可以在分享研究数据的同时也创造价值。虽然开放存取出版模式将有助于确保世界各地的研究人员可以自由访问研究出版物，但是关键在于确保作者自付费用的过渡，不限制发展中国家的科学家公布其工作的能力。

5.2 科学家及其机构单位

5.2.1科学家

这份报告认为囤积数据，严重阻碍了科学进展。它抑制个人数据集、复制实验、理论测试和重用别人新颖方式的数据验证。除了已经成为立即释放的领域，研究人员应该有一个明确的独占访问期间，研究资助应该预先指定数据得发布时机和条件。

5.2.2机构（高校和科研院所）

如果开放科学的好处是扩散到新的研究领域，在高校和研究机构的奖励和晋升系统，需要认识到这些发展。机构需要使信息和知识管理成为他们的组织战略的一部分，包括数据和文章出版时间表。这也是一个机会，更新知识产权战略，帮助机构采取更加多样化的方式行使知识产权。

5.3 评价高校科研

对国家卓越大学研究的影响评估，有增长趋势。公共资金达到大学和研究所的资金使用的问责制，直接通过政府部门或通过中介机构的渠道，进行评估。本文认为，成功地获得数据所需的技能和创造力，代表了科学的卓越层次，应该作为奖励。

生成大型数据集的设备和设施，需要团队运作。同理，为大型数据集的处理、存储、保存、再现和底层元数据信息的要求，需要团队合作。

5.4 学会、学术和专业团体

科学家往往双重效忠他们的学科和雇用他们的机构。他们研究习惯带来的纪律是最强的，具体反映在学术团体、院校和专业团体所代表的学科标准、价值观和优先事项上。学术团体在促进开放数据的文化上，处于有利地位。英国皇家学会主办一系列有关数据共享的几个讨论会，将在2012年9月由英国皇家学会和国际科学理事会在科学行为自由与责任委员会将举行联合讨论，在数字化时代的科学产出的价值。

5.5 研究资助者：研究理事会和慈善机构

自2006年以来，英国研究理事会开展研究成果的开放存取政策。在2011年美国国家科学基金会（NSF）更进了一步要求，建议包括数据管理计划的如何符合NSF的政策："在合理时间内与其他研究人员共享，不低于增量成本的国家科学基金会资助的工作过程中所创建或聚集的主要数据、样品、实体收藏或其他配套素材。（to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants）"。

国家资助机构 – 英国研究理事会 – 发挥领导作用，逐步拒绝资助那些不共享数据的申请。这应该在在不同学科内进行敏感数据的管理规范。

5.6 科学期刊的出版

理想的情况下，在一篇文章中所提出的论点，都应该呈现所有的数据，考虑空间因素，应该可以通过电子文章的链接做到。此外，刊物应说明何时以及如何将数据供他人访问。在特例上，研究人员应该解释为什么扣留释放数据的开放访问。

5.7 企业研究资助者

这份报告介绍了如何更大程度的开放，加强和提供商业价值。更大程度的开放，

也可以提供开发利用数据的机会，信息和知识是免费提供的商业产品和服务。开放研究和政府数据提供了创新和开展新业务的机会。封闭过程有时是必要的 - 无论是暂时为了吸引更多的投资或者永久保护商业秘密。

5.8 政府

许多国家的政府都把本国的科学基础视为未来民族发展的一个关键因素。但是实际上，具有有效地利用数据密集型科学的机会吗？需要支持企业的科研能力？如何平衡免费发布的政府数据以及先进的信息产品的收费机制呢？ 这些问题，需要与世界各地的政府以及国家优先事项互相达成一致的方式加以解决。

5.9 隐私、安全和保密的规则

所有的监管和治理机构，以及数据托管，应采取以风险为基础的方法，以促进开放数据政策和保护隐私利益。部署最适当的治理机制，以实现更大程度的开放，同时保护隐私和保密。

虽然关注安全性尤其必要，但是不应该被当作逃避开放数据的借口。一直有新的科学发现很少在比较开放研究记载了公共利益的双重用途。与开启研究的公共利益相比较，具有双重用途的科学发现显得微不足道（There has been very little dual use of new scientific findings in comparison with the documented public benefit of opening up research. ）。

术语表

Glossary

附录 1 – 各种数据库

Discipline-wide openness – major international bioinformatics databases 学科-广泛的开放性 – 主要的国际生物信息学数据库

Processing huge data volumes for networked particle physics 处理网络粒子物理的海量数据

Epidemiology and the problems of data heterogeneity 流行病学和数据异质性的问题

Improving standards and supporting regulation 提升纳米技术标准化和相应规范

In nanotechnology

The avon longitudinal study of parents and children (alspac) 雅芳的长期亲子研究

Global ocean models at the uk national oceanography centre 英国国家海洋中心的全球海洋模型

The UK land cover map at the centre for ecology & hydrology 生态与水文中心的英国土地覆盖图

Scientific visualisation service for the international space innovation centre 为国际太空创新中心的科学可视化服务

Laser interferometer gravitational-wave observatory project 激光干涉引力波观测台项目

Astronomy and the virtual observatory 天文学和虚拟天文台

附录 2 – 为开放数据的技术考虑

Dynamic data 动态数据

Indexing and searching for data 索引和搜索数据

Servicing and managing the data lifecycle 维修和管理数据

Provenance 出处

Citation 引用

Standards and interoperability 标准化和互操作

Sustainable data 可持续的数据

附录 3 – 数字储存库成本的案例

International and large national repositories (Tier 1 and 2)　国际和大型国家储存库（等级 1 和 2）

1. Worldwide protein data bank (wwpdb) 全球蛋白质数据银行
2. UK data archive　英国数据档案
3. Arxiv.Org　Arxiv.Org 储存库
4. Dryad　Dryad 储存库

Institutional repositories (tier 3)　机构典藏库（等级 3）

5. Eprints soton　Eprints soton 储存库
6. Dspace@mit 麻省理工 Dspace 储存库
7. Oxford university research archive and databank　牛津大学的研究档案和数据库

附录 4 – 致谢、事例、工作会和咨询

Evidence submissions 提交事例

Evidence gathering meetings 事例采集会议

Further consultation 进一步咨询