

英国《开放科研数据协议》的解读与启示*

王丹丹

(河南科技大学管理学院, 洛阳 471023)

摘要: 在开放科研数据方面, 英国是世界领先者之一, 本文通过全面解读《开放科研数据协议》的目的和意义、构建方法和过程、主要原则与要求, 阐释其中重要新概念和术语表达, 从英国《开放科研数据协议》中汲取对我国开放科研数据有价值的营养。研究发现, 《开放科研数据协议》的许多理念和观点对我国推进开放科研数据、开展科研数据管理理论研究与实践探索有重要启发。

关键词: 科研数据; 科学数据; 开放科研数据协议

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2018.08.005

随着数据密集型科研范式的到来, 科研数据作为新的学术交流的基本要素, 在世界各国均得到广泛探讨和深入研究。实现科研数据的有效管理, 开放科研数据已成为学术界、教育界乃至社会各界广泛关注的重要课题。开放科研数据是实现开放科学抱负的重要步骤, 以英国为代表的发达国家陆续提出相应的科研数据开放要求, 其中, 英国一些知名机构, 包括面向英格兰的高等教育资助委员会 (Higher Education Funding Council for England)、英国研究理事会 (Research Councils UK, RCUK)、英国的大学和惠康基金会 (Wellcome Trust) 于2016年7月28日联合推出的《开放科研数据协议》(Concordat on Open Research Data, 以下简称《协议》), 引发各界的广泛关注。

《协议》的目的是建立尊重各方需求的合理原则, 确保英国科研人员收集和生成的科研数据, 以符合相关法律、道德、学科及监管框架和规范的方式, 尽可能开放以供他人使用^[1]。该《协议》必将成为开放科学环境下, 利益相关方制定科研数据管理方案和设计科研数据管理服务的指导性文件。深入解读《协议》的制定理念与丰富内涵, 了解在开放科研数据方面处于世界级领先地位的英国开放科研数据理念与实践的最新进展, 对推进我国科研数据的开放, 实施科研数据的管理大有裨益。

1 《协议》的主要内容

1.1 相关概念的界定

《协议》的主体内容由定义、引言和10个核心原则组成。定义部分给出了科研数据的定义, 指出科研数据包含的类型及其主要目的, 同时解释开放科研数据的概念及其范围。《协议》对科研数据的界定: 是支持研究问题答案的证据, 无论是何种形式 (印刷、数字或物理), 均可用于验证研究结果。科研数据是科研人员在工作过程中通过实验、观察、建模、访谈等数据收集方法或其他方法从现有证据中获得的定量信息或定性陈述。其可以是原始数据 (测量或观察访谈), 也可以是二次数据 (从原始数据中导出的数据, 或者从其他人或机构拥有的数据中提取的数据)。具体包括统计数据、数字图像集合、录音、访谈记录、调查数据、具有适当注释的实地观察、解释、艺术品、档案、出版的文本或手稿等。《协议》指出, 考虑到数据机密性、防止不合理的成本、保护个人隐私、尊重同意条款及管理安全性或其他风险等, 并非所有的科研数据都适合开放。所谓开放科研数据指在明确科研数据所有权的前提下, 可以自由获取、使用、修改和共享的数据。

*本研究得到国家社会科学基金项目“社会科学科研数据的管理服务研究” (编号: 18BTQ070) 资助。

1.2 《协议》的目的与作用

《协议》指出开放不是简单地发布数据,而是所有参与研究的人员都有责任确保收集和生成的数据得到妥善管理,并使其他人可访问、可理解、可评价和可使用。开放科研数据可以为科学研究乃至整个社会带来实际利益,如有助于揭示新的研究途径、改善并创新应用于用户生活的技术和服 务。然而,实现科研数据的开放并不容易,甚至具有重大挑战,需要参与研究与学术交流过程的每个人员的承诺。《协议》是确保承诺的重要保证,其目的是建立尊重各方需求的合理原则,提出一套良好做法的期望,旨在增加获取科研数据的机会。《协议》有助于确保英国科研人员收集和生成的科研数据尽可能以符合相关法律、道德和监管框架及学科规范的方式开放,以供他人使用,同时适当考虑涉及的成本。

对《协议》原则的承诺将有助于向政府、企业、国际合作伙伴、其他科研人员和更广泛的公众证明,签署协议的机构:①以适当的方式对科研数据采取行动;②遵守与其工作相关的道德、法律和专业义务;③形成一种科研环境,使科研数据在任何可行和负担得起的地方开放;④使用透明、健全和公平的流程来制定有关数据开放性的决策;⑤有合适的机制来保证其科研数据的完整性;⑥认识到数据引用和承认贡献的重要性。

《协议》明确指出科研人员、科研人员所属机构和科研资助者的不同责任。同时指出,专业/法定监管机构、期刊、出版商、学术团体在这方面也发挥重要作用。

《协议》通过概述利益相关方的责任,帮助其清楚地了解各自在提升科研数据社会效益和经济效益方面应发挥的作用。

1.3 《协议》的10个核心原则

开放科研数据不断发展主要是由自下而上的需求及自上而下的政策共同推进。随着开放获取的发展,科研人员认识到真正的开放获取不仅是对全文的获取,还应包括对数据的获取。2010年2月,科研人员发布了关于科研数据共享的“潘顿原则”,提出开放科研数据应该允许任何人不论何种目的均能通过互联网免费获取,对于下载、复制、分析、重新处理和导入软件没有资金、法律、技术的障碍^[2]。为增进科研数据的可重用性,2014年学术界、行业、资助机构和学术出版商等不同利益相关者设计并共同认可了一套简明易懂的原则,

称为“公平数据原则”(FAIR原则),作为希望提高其所持有数据的可重用性的个人或机构的指南。此外,FAIR原则还特别强调了提高机器自动查找和使用数据的能力^[3]。

2016年,由多个知名机构联合推出的《协议》也围绕开放科研数据提出了10个核心原则,来指导整个科学研究界确保以加速发现和社会利益最大化的方式获取和使用科研数据。

原则1: 开放获取科研数据是高质量研究的推动者、创新的促进者和良好科研实践的保障。

原则2: 有充分理由说明科研数据的开放性可能需要限制,但任何限制都必须合乎规定。

原则3: 开放获取科研数据会带来巨大成本,应得到各方的尊重。

原则4: 确认科研数据创建者首次合理使用的权利。

原则5: 使用他人的数据应始终符合法律、道德要求和监管框架,包括适当的承认。

原则6: 良好的数据管理是科研过程所有阶段的基础,应在开始就建立起来。

原则7: 数据监管是确保数据有用性及长期保存的重要保障。

原则8: 出版物的数据应在出版日期之前提供,并且应以可引用的形式提供。

原则9: 帮助科研人员提升数据管理技能是所有利益相关方的责任。

原则10: 应定期审查开放科研数据的推进情况。

2 《协议》的整体解读

2.1 《协议》引发的回应

实际上,《协议》草案是在2015年8月19日由RCUK代表英国开放科研数据论坛(Open Research Data Forum)发布的。草案由跨部门工作组制定,其成员来自惠康基金会、英国的大学、罗素集团、联合信息系统委员会、大英图书馆和自然出版集团等机构。草案发布后,英国皇家统计学会^[4]、政治研究协会(Political Studies Association, PSA)^[5]、布里斯托尔大学、剑桥大学、曼彻斯特大学、诺丁汉大学和牛津大学等^[6]都对草案内容进行了评论。Elsevier也针对10个原则提出了一些具体意见^[7]。在广泛听取各界回应的基础上,RCUK对草案进行了反复修订,形成最终《协议》。

2.2 《协议》的显著特征

2.2.1 认识到数据类型的复杂性, 强调科研人员的自主权

《协议》中概述的原则适用于收集和分析科研数据的所有学科。在结晶学、遗传学、考古学和语言学等领域, 科研数据已被广泛共享。然而在其他一些学科中, 支持科研成果的科研数据是公共和商业机构数据与科研人员产生数据的结合, 这使得开放获取科研数据并不总是合适的。此外, 有多种合理的理由限制开放获取数据, 如确保机密性、防止不合理的费用、保护个人隐私、尊重同意条款及管理安全或其他风险等。

《协议》对于可能有合理理由限制开放获取数据的一些情况进行了详细说明, 如对于个人相关的数据,

《协议》指出应基于已建立的模型和管理数据获取的良好实践, 考虑与持有的具体数据相关的风险水平, 在保护隐私和确保机密性的前提下, 确定是否及如何提供与个人相关或来自个人的数据。推荐采用分级方法, 使较不敏感的数据更容易获得, 并且更严格地控制对更敏感数据的访问。要求相关机构的数据监管方案要充分考虑法律和道德要求, 具体包括数据保护法和科研伦理及科研诚信的相关法规。对于商业敏感数据, 《协议》指出应考虑为合作科研项目提供资助的数量和性质, 在开放和商业激励之间适当平衡, 进一步制定关于是否开放、何时开放及如何开放商业敏感数据的协议。对于第三方数据, 《协议》承认第三方数据提供者在更广泛的研究环境中的作用, 但同时指出第三方数据提供者为使数据更广泛可用而施加的合法限制(数据许可协议), 导致科研数据的开放变得更加复杂。

《协议》承认与科研数据相关的政策和程序没有“一刀切”的方法, 认为应在合适的机构层面上, 在可核查和透明的监督过程的支持下, 由科研人员决定保留和开放的数据。《协议》提供了一个灵活的框架, 以确保科研人员能够完全履行职责, 并确保其所在机构拥有符合最高标准的机制。

2.2.2 认识到协作的重要性, 强调合作与责任并正视开放数据的成本问题

《协议》意味着不同利益相关方之间需要合作, 并在支持开放科研数据方面发挥不同作用。《协议》致力

于确保在英国开放科研数据的最佳方式成为现实, 即所有参与者承认并履行其具体职责, 共同努力建立可持续发展的开放科研数据环境。为此, 《协议》针对科研人员、所属机构和资助机构分别提出了期望。①科研人员在尽可能短的时间内开放并供他人使用该科研数据, 具体因学科领域和研究领域而异; 支持出版物的数据应在出版日期之前提供, 并应以可引用的形式提供; 如果部分科研数据无法正常开放, 科研人员应说明理由。②科研人员所属机构应创建一种科研环境, 该环境需承认开放科研数据的价值, 并提供基础设施和相应服务, 以使科研人员在“物有所值”的前提下, 考虑科研数据的开放和可用; 该环境下良好的科研数据管理实践应被视为科研人员职责的一个重要方面。③科研资助机构应通过适当认可, 以及解决开放科研数据的成本问题来支持开放科研数据, 通过适当的政策和资助活动支持更广泛地开放科研数据的议程。

《协议》指出开放科研数据的优势虽然真实且可实现, 但IT基础设施和服务、行政管理和专业支持人员及科研人员培训等成本问题同样不容小觑。在科研数据开放的早期阶段, 对于大学或研究机构而言, 这些成本很可能是主要影响因素, 特别是在尚未建立成本回收机制的情况下。IT基础设施和服务成本、科研人员和相关专家(如数据管理专家)的持续培训成本会随时间的推移而显著增加。此外, 定期审查开放科研数据的推进情况也会产生成本, 并且这一成本可能超出科研项目研究周期的研究周期。就可持续发展而言, 所有这些费用应与实际利益成比例, 即开放科研数据的利益和成本必须与整个研究组合的利益和成本紧密相关, 所有这些成本必须与整个研究组合的收益相平衡。

2.2.3 认识到学科的差异性, 强调必须保障数据创建者权益

《协议》指出创建原始科研数据可能需要科研人员多年的专业知识及辛勤工作。任何破坏进行此类工作的动机都会对全球科研和知识进步产生严重的负面影响。因此, 强调向开放数据过渡而不降低科研人员收集和生产原始数据的意愿至关重要。

在天文学和基因学等学科, 科研人员会立即共享科研数据并产生明显的益处, 但并非所有学科。如果要求所有学科的科研人员都立即提供新生成的数据或对该数据进行分析, 可能会有很多人认为主动收集、测量

和分析数据有弊无利。相反,等待他人开展工作然后利用其数据会更容易。这种情况显然是不可取的。为防止这种负面现象,《协议》指出必须让生产原始数据的科研人员在合适且明确的时期内,拥有首次使用权,并且时间期限因研究主题和学科领域而异,应该由学科社群协商确定,并作为学科规范来执行。与此同时,《协议》也指出,即使在某些立即共享不是常态的学科中,某些情况下,科研数据也应为了公共利益而立即开放共享,如当科研数据对处理突发公共卫生事件具有重要意义和价值时。

此外,学科的差异性还体现在数据监管方面。《协议》指出数据监管对确保数据有用性及长期保存至关重要。数据监管可以通过多种方式实现,如同行评审、遵守科研社群的特定数据格式和标准、恰当描述存储在具体知识库中的数据或在期刊出版物中发表专门的数据文章。但是,选择何种方法应考虑学科领域、研究主题甚至数据类型的差异,不应强制规定方法。

2.2.4 认识到数据管理的重要性,强调从关注结果向关注过程转变

《协议》指出,如果科研项目产生的科研数据要可发现、可访问、可理解、可评价和可使用,在整个科研过程中仔细管理科研数据至关重要。因此,必须从科研过程开始就考虑科研数据的管理问题,重点考虑如何管理科研数据。为此,期望科研机构提供必要的基础设施,以支持科研人员有效地管理其数据,同时为科研人员提供有关该研究领域正确和相关数据管理和存储方法的指导。但是,《协议》同时指出基础设施的建设是整个研究界的共同责任,而不应仅由某一科研机构来承担。

科研人员个人可以从相关专家那里获得有关其领域最佳实践的建议,考虑如何管理他们收集的数据(或在研究早期阶段产生的数据)。《协议》指出在许多学科中科研人员还需要更具体的指导,因此应该让科研社群在制定相关学科具体指南方面发挥重要作用。建议科研人员在具体科研项目开始之前,制定深思熟虑的、合适的科研数据管理计划,以避免数据丢失或存储不当等问题。在科研数据管理计划中,应尽可能指明是否、何时,以及如何开放数据。尽管并非所有数据都值得保留,但具体保留哪些数据,科研人员需要在适当的指导下进行判断。因此,进行科研数据管理培训,并让

所有科研人员都在其职业生涯的早期阶段接受此类培训,并进行动态跟进至关重要。

监管、归档和分析数据需要一系列技术,这些技能与用于收集、生成或测量数据的技能不同。在某些情况下,个别科研人员很可能通过自主学习获得必要的技能,但对大多数人来说,专业学费是必不可少的。因此,帮助科研人员提升数据管理技能是所有利益相关方的责任。对于科研机构而言,这应包括以有组织和专业的方式提供给科研人员培训的机会。资助组织必须通过适当的筹资途径支持此类培训;科研人员必须确保其数据技能处于足以履行自身义务的水平。总之,数据科学家的专业技能对于支持科研人员和机构的数据管理需求至关重要。科研机构和资助机构应共同努力,协作发展这一能力,并通过为数据专家设计良好且可持续的职业道路来吸引和留住他们。

2.2.5 建议考虑学科差异性,设计更灵活的数据管理方案

科研数据的定义与学科密切相关。《协议》草案在征求反馈意见时,PSA成员指出普适性的数据管理要求存在以下问题。

(1) 将科研数据定义为“定量信息或定性陈述”存在问题,因为在政治学领域定性数据可能包含陈述或数字,因此,建议将科研数据定义为“定量或定性结果”。

(2) 在政治学领域中,与经验丰富的学者相比,早期职业研究人员分析数据或撰写论文都不老练,如果让他们在能够充分有效利用数据之前就开放数据,会对其职业发展造成严重影响。因此,指出“科研人员在短期内开放并使其科研数据可使用”这一期望,应该考虑学科、研究领域和科研人员所处的职业阶段的差异。

(3) 在人文和社会科学领域,许多科研人员并没有得到科研机构的资助。一些研究人员,特别是早期职业研究人员,可能不会被正式雇用。由于缺乏机构支持和基础设施,这些人可能很难开放他们的数据。因此,建议《协议》增加“科研人员所属机构应提供合适的基础设施系统和服务,以使科研人员能够开放其科研数据”。

(4) 如果科研人员的研究主题是政治冲突问题,那么保护受访者的个人信息安全就至关重要。《协议》草案虽然指出“在遵守机构层面适当的、可核实的、透明的程序的前提下,由科研人员自行决定是否开放数

据”，但对研究参与者构成的潜在威胁程度强调不够，建议增加“如果科研人员或其科研参与者的安全受到威胁，在决定数据是否开放及何时开放时应该谨慎”。与此同时，对于此类型的研究所产生的科研数据，资助机构和期刊应考虑给予其豁免权。

(5) 在定性研究中，数据被理解为由数据被监管的个人“拥有”，而不是获得数据的科研人员所有。如果数据的背景发生了显著变化（如政治研究中当权者的变化），这意味着研究对象之前的陈述会成为批评压迫政权，此时应考虑数据撤回权，以确保数据提供者的安全。因此，建议《协议》包含如下段落或类似内容：“如果个人希望收回他们提供的数据，特别是在威胁到个人安全的环境中，必须更新数据以反映此撤销”。因此，《协议》鼓励存储开放数据的机构制定流程，来确保在数据发生变化时自动联系开放数据的获取者。

(6) 人文科学和社会科学出版物的半衰期比科学、技术、工程和数学学科要长得多。因此，建议将10年期限改为“因学科和学科范围而有所不同的有限时间段”。

3 《协议》对我国开放科研数据的启示

《协议》制定的背景是数据密集型科研范式的转变及开放科学的新特征。它针对数据密集型科研环境和开放科学的需要，将注意力聚焦到面向科研数据的明确和实用的原则，涵盖支持科研过程所需的多种角色，让利益相关群体更加清楚地认识到开放科研数据的时代紧迫性与面临的挑战。它是RCUK发布的一个指导性文件，不是规则手册，而是研究界本身对最佳实践的一系列期望，因此，非官方、理论指导性是其基本特征。《协议》是“以实现英国开放科学抱负为目标，改善合作并加强英国作为全球科学领导者的地位”，适用对象是所有学科参与科研过程的所有利益相关群体。它提出的基本理念和一些核心概念，对利益相关群体探索科研数据管理方案、实施科研数据管理服务、推进开放科研数据具有启发意义。

首先，我国应尽快出台类似《协议》的宏观架构性和理论指导性文件，以此为载体，建立利益相关方之间的对话和协作平台。《协议》的起草邀请了不同利益相关群体的参与，在意见征集阶段又广泛吸纳了不同学科科研人员的参与，融合多学科科学研究的独特性，将多个开放科研数据的相关概念和理念融汇成一个连贯

的整体，因此，它是实现科研数据开放和深入开展科研数据管理的工具性文件。我们需要学习其思想精髓，研究探索并形成适合我国国情的开放科研数据协议或者相关文件，以指导国家科研数据管理实践的开展。需要注意的是，开放科研数据指导性文件的制定应将宏观架构性和理论指导性作为起草出发点，因为这种文件形式具有很大的灵活性和可拓展空间。

其次，应借鉴国外科研数据基础设施的部署经验，全面规划面向我国科研数据的基础设施建设方案。无论是开放科研数据还是实施科研数据管理均需要完善的基础设施作支撑，基础设施也是实现与相关利益群体真正意义上合作的前提。根据科学数据联盟（Research Data Alliance）的划分标准，基础设施可分为技术基础设施和社会基础设施，前者是指实现数据获取、存储、处理、发布与共享所需要的一系列相互联系的平台、硬件、软件、工具组成的基本架构及相关技术标准（如数字标识符、共享元数据框架等）；后者主要指促进不同机构间协作的措施和手段，如公共政策和组织实践、统一的标准、通用的数据获取和保存方法等^[8]。《协议》的顺利出台与发展，离不开英国相对完善的数据基础设施建设^[9]，因此仅出台宏观架构性文件还不够，更应该重视面向数据的基础设施建设，这是解决开放数据成本问题的关键。

最后，深入具体学科领域，研究科研数据管理服务，助力开放科研数据目标的顺利实现。科研数据管理服务是协助科研人员创建和实施数据管理计划及应用管理科研数据的最佳实践，在科研过程的任何阶段接受合适的管理服务，以确保其科研数据达到可发现、可获取、可互操作、可重用标准。它是由利益相关群体在不同层面以协作型方式组织开展的一个服务体系，是创造科研数据价值、提升科研数据能力、保持科研数据优势的必要前提和关键所在，也是一系列科研数据管理政策、制度得以顺利实施的重要保证。然而，不同学科领域科研数据的规模、结构和形式均存在较大差异。实践表明，科研数据管理的最大挑战在于学科不同的数据实践。科研数据的性质和使用与学科密切相关，学科特征会影响管理数据的方法，即学科的研究问题、方法和过程，对创建数据的性质、数据的操作和存储方式、数据共享和重用的可能性，以及科研数据管理实践的构成都会产生深远影响^[10]。因此，应该面向具体学科领域研究其科研数据的特征，以及科研人员在研究生命周期中所展现出的服务需求与期望。以此为基

础, 揭示面向学科的科研数据管理服务的具体过程, 包括服务体系的框架构建、服务流程设计、服务模式及其实现等关键问题。

参考文献

- [1] RUCK. Concordat on Open Research Data [EB/OL]. [2018-06-28]. <https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/>.
- [2] 黄永文, 张建勇, 黄金霞, 等. 国外开放科学数据研究综述 [J]. 现代图书情报技术, 2013 (5): 21-27.
- [3] WILKINSON M D, DUMONTIER M, JAN AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship [EB/OL]. [2018-07-01]. <https://www.nature.com/articles/sdata201618>.
- [4] Royal Statistical Society. The Royal Statistical Society's response to the UK Open Research Data Forum's Draft Concordat on Open Research Data [EB/OL]. [2018-06-18]. <http://www.rss.org.uk/Images/PDF/influencing-change/2015/RSS-response-UK-open-research-data-Sept-2015.pdf>.
- [5] PAUL T, HELENA D, Will J. Political Studies Association Response to the Concordat on Open Research Data [EB/OL]. [2018-06-18]. <https://www.psa.ac.uk/sites/default/files/PSA%20Response%20on%20ORD28.9.15.pdf>.
- [6] Joint response on the draft UK Concordat on Open Research Data [EB/OL]. [2018-06-18]. <https://unlockingresearch-blog.lib.cam.ac.uk/?p=285>.
- [7] Elsevier. Elsevier response to Concordat on Open Research Data [EB/OL]. [2018-06-18]. https://www.elsevier.com/___data/assets/pdf_file/0015/116313/Elsevier-response-to-Concordat-on-Open-Research-Data.pdf.
- [8] BERMAN F, WILKINSON R, Wood J. Building global infrastructure for data sharing and exchange through the research data alliance [J/OL]. [2018-07-18]. http://www.dlib.org/dlib/january14/01guest_editorial.html.
- [9] Open Research Data Taskforce with Michael Jubb. Research Data Infrastructures in the UK [R/OL]. [2018-06-30]. <https://www.universitiesuk.ac.uk/policy-and-analysis/research-policy/open-science/Documents/ORDTF%20report%20nr%201%20final%2030%2006%202017.pdf>.
- [10] LYON L, RUSBRIDGE C, NEILSON C, et al. DCC SCARP: disciplinary approaches to sharing, curation, reuse and preservation - final report [R/OL]. [2018-07-18]. <http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf>.

作者简介

王丹丹, 女, 1980年生, 博士, 副教授, 研究方向: 网络化信息服务, E-mail: wangdan430806@163.com。

Interpretation and Enlightenment of UK' Concordat on Open Research Data

WANG DanDan

(College of Management, Henan University of Science and Technology, Luoyang 471023, China)

Abstract: In terms of open research data, the UK is one of the world leaders. We will obtain valuable nutrition for open research data in China from the *Concordat on Open Research Data* through a comprehensive interpretation of the purpose and significance, methods and processes of building, main principles and requirements, explaining important new concepts and terminology expressions of the *Concordat on Open Research Data*. The study found that many of the ideas and viewpoints of the *Concordat on Open Research Data* have important implications for China to promote open research data, theoretical and practical research of research data management.

Keywords: Research data; Scientific Data; Concordat on Open Research Data

(收稿日期: 2018-07-20)