地球表层科学数据共享程度现状调研

一、FAIR 原则概述

2016 年，“科学数据管理的 FAIR 指导原则”在《科学数据》上发布。F、A、I、R 4 个维度是评估 FAIR 应用状况的主要维度，旨在为改进数据资源的 F 可发现性，A 可访问性，I 可互操作性和 R 可重用性提供指导。这些原则强调机器可操作性（即计算系统在没有或最少人为干预的情况下查找、访问、互操作和重用数据的能力），分别注重为数据分配持久标识符和丰富元数据；遵循标准数据访问协议；使用 RDF 等机器可读格式增强关联数据语义；明确数据重用协议与溯源信息。目前 FAIR 原则已经有用于多个领域的开放共享评价，例如生物、医疗等，此文的目的是通过 FAIR 原则对地球表层科学数据共享程度做出评价。

FAIR 原则的一级指标	FAIR 原则的二级指标
可发现性（Findable）	是否使用持久标识符进行标识
	元数据是否包含数据来源、格式和质量
	数据是否可以被搜索和发现
可访问性（Accessible）	数据是否易于访问，同时考虑隐私和安全方面的限制
	数据是否具有适当的访问控制和权限
	数据提供了可靠的质量保证
可互操作性（Interoperable）	是否使用公认的词汇表和语义标准
	是否使用公认的数据格式和协议
	是否提供了与其他相关数据的链接和关联
可重用性（Reusable）	是否允许使用的明确许可
	是否提供了明确的引用方式
	是否可用于在不同的环境中重复使用

1、可发现性（Findable）

使用数据的第一步是找到它们，元数据和数据应该很容易被人类和计算机找到。机器可读的元数据对于自动发现数据集和服务至关重要，因此这是 FAIR 原则的重要组成部分。

1.1 F1：元数据是否被分配一个全局唯一且持久的标识符

F1 可以说是最重要的，因为如果没有全局唯一和持久的标识符，就很难实现 FAIR 的其他方面。全局唯一且持久的标识符通过为元数据的每个元素和数据集中的每个概念、度量分配唯一标识符来消除已发布数据含义的歧义。F1 为标识符规定了两个条件：

标识符是否是全局唯一的（即其他人不能在不引用了的数据的情况下重用、重新分配相同的标识符）。可以从注册表服务获取全局唯一标识符，该服务使用算法来保证新铸造标识符的唯一性。

标识符是否是持久的。保持网络链接处于活动状态需要时间和金钱，因此随着时间的推移，链接往往会变得无效。注册管理机构服务至少在某种程度上保证了该链接在未来的可解析性。

1.2 F2：是否使用丰富的元数据描述数据

在评估 FAIR 数据资源时，元数据的内容是否是广泛的，包括有关数据背景，质量和状况或特征的描述性信息。

1.3 F3：元数据是否清晰明确地包含它们所描述的数据的标识符

元数据和它们描述的数据集通常是单独的文件。元数据文件与数据集之间的关联应通过在元数据中提及数据集的全局唯一且持久的标识符来明确。连接应该以正式的方式进行注释，例如在 RDF 元数据的情况下使用 foaf：primaryTopic 谓词。

1.4 F4：数据是否可在搜索资源中注册或索引

仅靠标识符和丰富的元数据描述并不能确保互联网上的“可查找性”。完全好的数据资源可能仅仅因为没有人知道它们的存在而未被使用。如果不知道数据集、服务或存储库等数字资源的可用性，那么没有人（也没有机器）可以发现它。可以通过多种方式发现数字资源，包括索引。例如，谷歌发出“读取”网页并自动索引它们的内容，这样它们就可以在谷歌搜索框中找到。

2、可访问性（Accessible）

可访问性是指一旦用户找到所需的数据，用户需要知道如何访问它们，可能包括身份验证和授权。

2.1 A1：数据是否可通过其标识符使用标准化通信协议进行检索

A1.1：该协议是否是开放、自由且普遍可实施的

为了最大限度地提高数据重用率，该协议应该是免费（免费）和开放的（来源的），因此可以在全球范围内实施，以促进数据检索。任何拥有计算机和互联网连接的人都可以访问至少元数据。例如是否是 HTTP，FTP， SMTP， ...协议，这样的协议就是开放的，一个反例是 Skype，它不是普遍实现的，因为它是专有的。

A1.2： 该协议是否允许在必要时进行身份验证和授权程序

这是 FAIR 的一个关键但经常被误解的元素。FAIR 中的“A”并不一定意味着“开放”或“免费”。相反，它意味着应该提供可访问数据的确切条件。因此，即使是受到严格保护和私人数据也可以是公平的。理想情况下，可访问性的指定方式使机器可以自动理解需求，然后自动执行需求或提醒用户注意要求。请求用户为存储库创建用户帐户通常是有意义的。这允许对每个数据集的所有者（或贡献者）进行身份验证，并可能设置特定于用户的权限。

2.2 A2：即使数据不再可用，元数据是否可访问

数据集往往会随着时间的推移而降级或消失，因为维护数据资源的在线状态会产生成本。发生这种情况时，链接将变得无效，用户会浪费时间搜索可能不再存在的数据。存储元数据通常更容易且更便宜。因此，原则 A2 指出，即使数据不再持久，元数据也应保留。

3、可互操作性（Interoperable）

数据通常需要与其他数据集成。此外，数据需要与应用程序或工作流互操作，以便进行分析、存储和处理。

3.1 I1：数据是否使用正式的、可访问的、共享的和广泛适用的语言进行知识表示

人类应该能够交换和解释彼此的数据，所以不要使用死语言。但这也适

用于计算机，这意味着机器应该可以读取的数据，而无需专门的或临时的算法、转换器或映射。该原则的主要目标是通过用于表示这些对象的知识表示语言来提供对数字对象的“共同理解”。原则 I1 定义了这些语言应具有的一些属性。所选择的语言应该有一个正式的规范，即语言的语法和语法以精确的方式定义。另一个要求是知识表示语言规范应该是共享和可访问的，以便其他人可以阅读规范并学习语言。

3.2 I2：数据是否使用遵循 FAIR 原则的词汇表

在数据或元数据中使用词汇表，我们应该确保它们本身也是属于 FAIR 的，以便人类或机器，可以找到、访问、互操作和重用它们。用于描述数据集的受控词汇需要记录下来，并使用全局唯一且持久的标识符进行解析。使用数据集的任何人都可以轻松找到和访问此文档。

3.3 I3：数据是否有对其他（元数据）数据的限定引用

限定引用是解释其意图的交叉引用。目标是在（元）数据资源之间创建尽可能多的有意义的链接，以丰富有关数据的上下文知识，此外，所有数据集都需要正确引用（即，包括其全局唯一和持久的标识符）。

4、可重用性（Reusable）

FAIR 的最终目标是优化数据的重用。为此，应详细描述元数据和数据，以便可以在不同的环境中复制或组合它们。

4.1 R1:数据描述是否丰富，具有多个准确和相关的属性

这个指标用来判断数据的描述是否丰富，如果数据附加了许多标签，则查找和重用数据会容易得多。

R1.1：数据是否以清晰且可访问的数据使用许可证发布

在“I”下，涵盖了技术互操作性的要素。R1.1 是关于法律互操作的。数据附加了哪些使用权限，这应该清楚地描述。歧义可能会严重限制对数据的重复使用。随着涉及更多许可考虑因素的自动搜索，许可状态的明确性将变得更加重要。

R1.2：数据是否有详细来源

为了让其他人重用数据，用户应该知道数据来自哪里即关于起源、历史的清晰故事，引用谁和、或你希望如何被承认。包括对导致数据的

工作流的描述：谁生成或收集了数据,它是如何处理的,以前发表过吗,它是否包含可能已转换或完成的其他人的数据,此工作流是否以机器可读的格式描述。

R1.3：数据是否符合与领域相关的社区标准

该指标判断此数据是否符合领域相关的社区标准，如果是符合标准的，它们的数据集相似，有相同类型的数据、以标准化方式组织的数据、完善且可持续的文件格式、遵循通用模板和使用通用词汇的文档（元数据），则更容易重用数据集。

二、国外研究现状

A. Dunning[1]等分析了 40 多个数据存储库遵循 FAIR 原则的情况。不过这些分析并未使用某个正式的评估模型或依据某个评估标准，而是由作者基于数据存储库的帮助页面、元数据记录等相关资料去评估其是否符合 FAIR 原则，评估过程较为主观。因此，建立明确的、有识别力的、可测量的并且通用性强的评估指标评估数据实施 FAIR 的程度成为迫切需要，为此，FAIR 指标小组、荷兰数据存档与网络服务、澳大利亚研究数据共享组织、澳大利亚联邦科学与工业研究组织、研究数据联盟、FAIRsFAIR 、欧洲开放科学云等组织提出了各自的评估 FAIR 实施程度的模型或工具。

国外研究机构和科研团体等已开发 FAIR 评估方法和指标体系，具有代表性的主要是 Go FAIR Metric Group (GFMG)的 FAIR 通用指标框架。该评价体系围绕 FAIR 准则中技术细节定义了各类别评估指标,此外还有荷兰数据存档与网络服务(Dutch Data Archiving and Networked Services, DANS)、欧盟 Horizon 2020 及澳大利亚研究数据共享组织(Australian Research Data Commons, ARDC)的 FAIR 原则评估指标体系、澳大利亚联邦科学与工业研究组织(Commonwealth Scientific and Industrial Research Organisation, CSIRO)的 5 星级数据评估工具。

名称	提出方	评估方式	设置维度
GFMG	M. D. Wilkinson 等	通用指标框架	F、A、I、R

DANS	荷兰	打分评估	F、A、I、R
Horizon 2020	欧盟	自评估模板	F、A、I、R
ARDC	澳大利亚	问卷评估	F、A、I、R
CSIRO 5 Star Data Rating Tool	澳大利亚	星级评定	自设维度，可与F、A、I、R映射

基于这些评估框架开展FAIR的应用评估有助于监测科学数据管理情况、FAIR原则的应用情况和数据FAIR化的程度。在可发现、可访问、可互操作、可重用4个一级指标下共设12个二级指标和17个三级指标，用于评估我国地球表层科学数据开放平台的FAIR应用情况。

一级指标	二级指标	三级指标	指标解释	指标借鉴情况
可发现 Findable	标识符	标识符类型	数据集是否使用唯正式、持久标识符	ARDC、CSIRO
	元数据	元数据格式	元数据是如何描述数据的	ARDC
		元数据丰富度	元数据详细描述数据内容情况	CSIRO
	资源标识符	元数据包含标识符	数据集标识符是否含在描述数据的所有元数据记录/文件中	ARDC
	在可搜索资源中索引	搜索引擎可发现	在平台中的登记注册情况	CSIRO
可访问 Accessible	访问协议	访问注册	有无访问数据用户条款	ARDC
		访问协议	访问数据遵循的协议	ARDC、CSIRO

	访问授权	用户审核机制	用户注册情况与对应的数据访问权限	欧盟、ARDC
	元数据寿命	(元)数据寿命承诺	数据存储方式	CSIRO
可互操作 Interoperable	知识表示语言	数据文件格式	是否使用通用机器可读标准格式表示	ARDC\CSIRO、DANS
	FAIR 化词表	(元)数据元素描述	元数据元素与标准词汇表关联情况	ARDC、CSIRO
	合格引用	(元)数据关联情况	元数据关联其他(元)数据以明确、增强上下文指示关系的方式	ARDC
可重用 Reusable	许可声明	许可声明明确性	是否明确表达可重用的条件	ARDC、CSIRO
		声明格式标准性	许可声明遵循的许可标准情况	ARDC、CSIRO
		限制声明明确性	是否明确表达不可重用的条件	欧盟
	溯源规范	溯源信息明确性	是否描述数据的溯源信息	ARDC
		溯源格式标准性	溯源信息遵循的格式情况	CSIRO

三、国内研究调研

杨啸林的“FAIR 准则与生物医学数据标准应用服务”[4]中将 FAIR 原则的共享理念应用到生物学领域，并做出了综述性概论。李春秋的“医学科学数据开放平台 FAIR 原则的应用评估与调查分析”[3]制定了面向我国医学科学数据开放平台的 FAIR 原则应用评估框架。叶兰[5]在“FAIR 数据评估模型与工具研究”对比分析了 FAIR 数据评估模型与工具,为数据建设和数据管理过程中利益相关者评

估 FAIR 数据的遵循度提供参考。从评估指标和评估方法两方面介绍国际上 7 个评估 FAIR 数据遵循度的指标模型与工具,采用比较分析法从评估方法的类型、评估方法的自动化程度、评估方法的可操作性、指标数量与分布、元数据指标设置、指标清晰度等 6 个方面对比分析各模型与工具。

四、总结展望

目前 FAIR 评估的研究涉及面向评估内容的指标设计与测量、面向评估技术方法的自动化、面向评估结果的可视化等，从目前调研得出，基于地球表层科学数据的共享程度以 FAIR 原则为评价还是要以该学科来补充构建评价指标，结合全球地球表层科学数据开放平台特征,充分考虑指标设计规范性、适用性等要求。再通过对全球地球表层科学数据开放平台的预调研,持续调整和完善指标及其可选项。最后得出针对地球表层开放科学的共享程度评价模型。

[1] DUNNING A,DE SMAELE M,BÖHMER J.Are the FAIR data principles fair? [J].International journal of digital curation,2017,12(2):177-195.

[2] Wilkinson M D,Sansone S A,Schultes E,et al.A design framework and exemplar metrics for FAIRness[J].Scientific Data,2018,5:180118.

[3] 李春秋,杜博雅,耿骞,宋宁远,李子璇,原顺梅.医学科学数据开放平台 FAIR 原则的应用评估与调查分析[J].图书情报工作,2022,66(03):72-82.DOI:10.13266/j.issn.0252-3116.2022.03.009.

[4] 杨啸林,杨晟,潘虹洁,王哲,王志刚,何勇群.FAIR 准则与生物医学数据标准应用服务[J].中国医学伦理学,2020,33(02):153-159.

[5] 叶兰.FAIR 数据评估模型与工具研究[J].图书情报工作,2021,65(16):138-147.DOI:10.13266/j.issn.0252-3116.2021.16.015.