通过机器学习增强数据空间语义互操作性：愿景展望

Zeyd Boukhers 弗劳恩霍夫应用信息技术研究所 FIT 和科隆大学医院生物医学信息学研究所，德国
z
eyd.boukhers
@fit.fraunhofer.de

Christoph Lange 弗劳恩霍夫应用信息技术研究所
FIT 和亚琛工业大学，德国
christoph.lange-
bever@fit.fraunhofer.de

Oya Beyan 科隆大学医学院和科隆大学医院生物医学信息学研究所，德国
oya.beyan@uni-koeln.de

arXiv:2303.08932v1 [cs.DB] 15 Mar 2023

摘要

我们的愿景论文提出了一项计划，旨在通过应用机器学习，改善数据空间语义互操作性的未来。数据空间让成员能够在自我监管的环境中交换数据，其应用正日益普及。然而，目前在这些空间中管理元数据和词汇表的人工操作既耗时，又容易出错，且可能无法满足所有利益相关者的需求。我们相信，借助机器学习的力量，可以显著提升数据空间中的语义互操作性。这包括自动生成和更新元数据，从而形成更灵活的词汇表，以适应不同子社群所使用的多样化术语。我们对数据空间未来的愿景，旨在突破传统数据交换方式的局限，让数据更易于访问，并为社群所有成员创造更大价值。

CCS 概念

• **信息系统** → **数据交换**；**数据访问方法**；**语义网描述语言**。

关键词

数据空间、语义互操作性、机器学习

ACM 参考格式：

Zeyd Boukhers、Christoph Lange 和 Oya Beyan。2023年。通过机器学习增强数据空间语义互操作性：愿景展望。载于*ACM* 会议论文集（*Conference'17*）。ACM，纽约，纽约州，美国，6 页。
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

数据空间是应对数据交换与集成难题的前沿解决方案。它们充当联邦平台，供各类实体共享和交换数据，并提供必要的工具与安全措施，确保数据能够安全地共享和使用。数据空间的最终目标是提升广大利益相关者对数据的访问能力和使用水平，从而推动数字经济中的创新与研究。这些创新型基础设施让所有用户受益，使数据集成更加简单、灵活且高效。与传统的数据交换系统相比，数据空间的优势在于能够轻松接纳新成员。

数据空间的理论重要性在于，它能够无缝集成具有不同模式和结构的数据源，而无需采用共享模式，也不需要投入大量前期工作来规范数据。这得益于语义技术的广泛应用 [3], ，这些技术有助于对齐数据，并使数据能够以一致的方式使用。因此，使用数据空间可以减少建立数据集成系统所需的时间和精力，同时提高灵活性和可扩展性。然而，要在数据空间中实现令人满意的语义互操作性，仍是一个复杂且持续推进的过程。这需要不断开发用于数据集成和处理的标准、工具和技术。

国际数据空间[1] 和 Gaia-X[2] 是欧洲乃至全球两项重要倡议，旨在汇聚各方利益相关者，共同打造安全、可互操作且去中心化的数据基础设施。IDS 和 Gaia-X 的最终目标是建立可信、安全、高效的数据基础设施，在保护隐私和安全的同时支持数据驱动型创新。这些倡议着重提供基础设施服务，以增进实体间的信任、推动数据主权并保护数据隐私。因此，预计它们将简化企业和组织安全、透明地访问和利用数据的流程。

随着参与各种数据空间的组织和个人数量不断增加，对有效的数据管理和交换解决方案的需求也将随之上升。成功进行数据交换的关键在于各方对所共享数据达成相互理解，这称为语义互操作性。因此，随着加入不同数据空间的成员不断增多，寻找提升语义互操作性的方法至关重要。

尽管机器学习在解决上述一般数据交换互操作性问题方面展现出潜力，但其在数据空间中的应用尚未得到充分探索。本文旨在阐明利用机器学习的潜力

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字版或纸质副本，前提是副本不得为营利或商业利益而制作或传播，且副本须包含本声明及首页的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的内容，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新出版、发布到服务器或分发至列表，则须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。*Conference'17*，*2017*年7月，美国华盛顿特区 © 2023年计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://internationaldataspaces.org/ [2] https://gaia-x.eu/

来提升数据空间中的语义互操作性，使其成为数据交换的首选方案。

本文以 IDS 为具体示例，阐述我们对于如何利用机器学习解决方案提升数据空间语义互操作性的观点。尽管语义互操作性涉及诸多方面，我们着重讨论数据交换中普遍存在的六项关键挑战。值得注意的是，本文并未提出任何定论，而是全面概述了将机器学习解决方案直接整合到数据空间中的框架。

2 相关工作

数据空间的概念已存在数十年，但近年来受到广泛关注，人们也投入了大量精力，致力于促进当今数据驱动生态系统中的数据交换，例如国际数据空间（IDS）[3] 和欧洲共同数据空间 [23]。迄今为止，实际数据空间的重点主要在于法律、技术和元数据互操作性，而对数据语义方面关注甚少，因为该领域仅有少数研究 [3]。这意味着，在语义互操作性方面，目前的重点仅限于元数据，并假定其已经存在。然而，数据语义互操作性已经经过数十年的研究。例如，Ouksel 等人 [17] 探讨了如何在复杂的异构信息系统（如互联网和万维网）中找到准确信息这一问题。他们提出了一种互操作性框架，通过将信息与现实世界实体关联起来，并承认语义会不断变化。十多年后，Kiljander 等人 [12] 讨论了采用通用方法实现异构物联网设备之间高级互操作性的必要性，以实现普适计算和物联网愿景。他们将互操作性挑战划分为两个层面：连接性和语义。连接性层面涵盖开放系统互连（OSI）模型中从物理层到传输层的传统层级。语义层面则涵盖实现通信各方之间共享含义所需的技术。作者指出，语义层面的互操作性已被确定为语义网的一项主要目标，并且语义

物联网相关项目可以利用语义网技术来表示有关物理世界的知识。

数据交换中的语义互操作性也已在特定领域得到探讨。Lin 等人 [14] 评估了逻辑观测标识符名称与代码（LOINC）的使用情况，以及其对使用 LOINC代码的不同机构之间实验室数据互操作性的影响。他们发现，不同机构使用 LOINC代码进行相同实验室检验时，数据格式存在差异。在调查汇总此类数据时出现的常见问题后，他们指出，要实现更高的互操作性，需要为实验室结果编码的最佳实践提供更多指导。

3 机器学习增强型数据空间

数据空间中的语义互操作性是一个涉及多个方面的复杂问题，如图 1 所示。尽管机器

[3]https://www.fraunhofer.de/en/research/lighthouse-projects-fraunhofer-initiatives/international-data-spaces.html

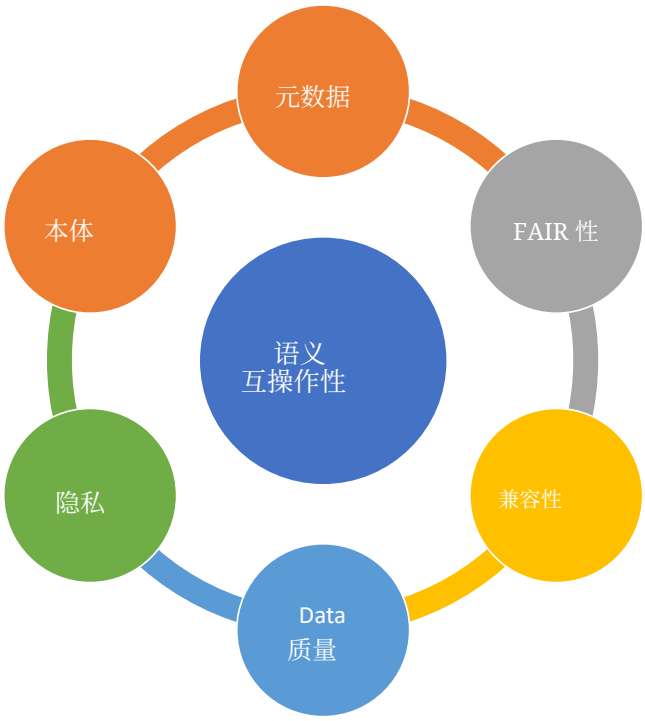


图 1：机器学习可以增强的数据空间语义互操作性方面

学习有望改善这些方面，但传统方法主要是孤立地运用机器学习技术，而不是将其置于数据空间的更广泛背景中加以应用。至关重要的是，要全面考虑语义互操作性的各个方面，并在数据空间环境中以综合、整体的方式整合机器学习。

图 2 概述了国际数据空间环境中的机器学习增强型数据空间，展示了数据管理的六个关键方面，涉及数据提供者、消费者和服务提供者这三类利益相关者。这些方面包括：

- **自动元数据提取（1）**：如果尚无元数据，机器学习模型可以自动从数据中提取关键属性值，帮助数据提供方为数据交换和使用做好准备，而无需手动准备元数据。

- **本体与词汇对齐（2）**：数据空间的词汇与数据提供方的词汇相对齐，使数据消费者能够理解所交换的数据。这样一来，数据空间的成员就不必采用相同的内部词汇，因为这往往是一项颇具挑战性的工作。

- **FAIR 性评估（3）**：根据提供或提取的元数据评估数据的 FAIR 性水平，使数据提供者能够提升其数据的 FAIR 性，并让数据消费者了解数据的易用程度。

• **数据质量评估与增强（4）**：根据数据的格式评估数据质量，并在可能的情况下加以改进。机器学习可用于评估和增强结构化数据及表格数据；不过，必须注意，质量指标可能因数据格式而异。例如，评估非结构化数据（如文档语料库）的质量可能颇具挑战。

• **隐私保护（5）**：基于机器学习的匿名化和掩蔽技术可应用于包含私人、敏感或个人信息的数据，使其能够共享。敏感数据可以自动检测，也可以由数据提供方提供，从而使数据提供者能够在无需担忧隐私问题的情况下共享数据。

• **兼容性改进（6）**：将数据转换为数据消费者可读的格式。如果数据需要与消费者的数据合并，消费者会告知数据的结构和格式，以便据此转换数据。这样，消费者便可以使用收到的数据，而无需额外花费精力去读取和理解数据。

下面，我们将讨论这些方面：

3.1 自动元数据提取

元数据在数据交换中发挥着至关重要的作用，因为它能帮助数据消费者了解数据，并判断数据是否满足其需求。然而，许多数据提供者可能因缺乏能力或知识，无法为其资源准备必要的元数据，因此不愿提供这些元数据。这可能成为数据交换的一大障碍，因为它限制了消费者访问和使用所需数据的能力。

为克服这一挑战，可以利用机器学习从资源中（半）自动提取元数据。机器学习算法可以使用包含资源及其相应元数据的数据集进行训练，从而学习数据与元数据之间的模式和关系。随后，可将这些算法应用于新资源，以提取相关元数据。这种方法的优势在于能够处理数据与元数据之间复杂且细微的关系。随着数据及其需求的变化，这种方法也可以轻松更新和调整。不过，需要注意的是，数据空间中的一个典型挑战是资源采用了不同的异构格式。

在数据空间中交换的不同资源可能具有不同的元数据属性，因此可能需要针对不同资源和元数据属性使用不同的机器学习（ML）模型。例如，对于文档语料库，可以采用自然语言处理（NLP）技术提取标题和描述。具体而言，可以使用 [5, 28] 中的自动元数据提取技术，从每份文档中提取元数据，例如出版日期、作者、语言等。随后，可以利用这些元数据推导出整个集合的元数据，例如出版时间范围、作者、语言等。

3.2 本体与词汇对齐

国际数据空间参考架构[4] 强调了通用词汇对于在数据空间内实现有效数据交换的重要性。然而在实践中，数据提供者可能各自拥有独特的词汇，因此难以将其与数据空间使用的词汇对齐。这可能是因为将现有词汇映射到数据空间词汇需要投入成本和精力，也可能是因为数据提供方会参与多个使用不同词汇的数据空间。

为应对这些挑战，可以利用机器学习算法，自动建立数据提供方的本地词汇与数据空间所用词汇之间的映射。这有助于实现无缝且可互操作的数据交换，而无需数据提供者采用新的词汇。

基于机器学习的本体对齐 [16]和本体匹配 [7] 方法可用于自动将一个本体或词汇中的概念和术语映射到另一个本体或词汇。这些算法运用语义相似度度量 [25], 、基于图的方法 [24], 和深度学习模型 [4, 10, 11] 等技术，识别不同本体或词汇中的概念之间的对应关系。其目标是生成映射，使使用不同本体或词汇的系统能够进行数据交换，同时保留数据的含义。

3.3 FAIR 性评估

FAIR（即**可查找、可访问、可互操作**和**可重用**）原则在数据交换与共享中日益重要。这些原则旨在确保数据资源易于发现和访问，能够轻松整合其他数据源，并可用于多种用途。遵循这些原则更有可能促使数据得到使用和重复使用，因为这能提升资源的整体质量和可用性。

评估资源的 FAIR 性是确定其是否适合使用的重要一步，因为这有助于识别可能阻碍再利用的因素。这些因素可能包括许可限制、数据访问条件和数据互操作性问题。提前开展这项评估可以节省宝贵的时间和资源，因为这样有助于避免为获取可能不适合预期用途的数据而进行耗时的谈判或长时间等待。

如第 3.2 节所述，使用本体等共享词汇对于提高资源的可发现性和互操作性十分重要。然而，仅使用映射技术（参见第 3.2 节）可能还不够，因为描述元数据的内部本体可能并未采用通用类表示。为了解决这一问题，可以使用 BERTMap [9], 等机器学习技术评估提供方本体与数据空间本体之间的兼容性。此外，还可以使用基于规则的方法和语义网技术来评估元数据的结构，从而进一步提高资源整体的 FAIR 性。

[4]https://internationaldataspaces .org/use/reference - architecture/

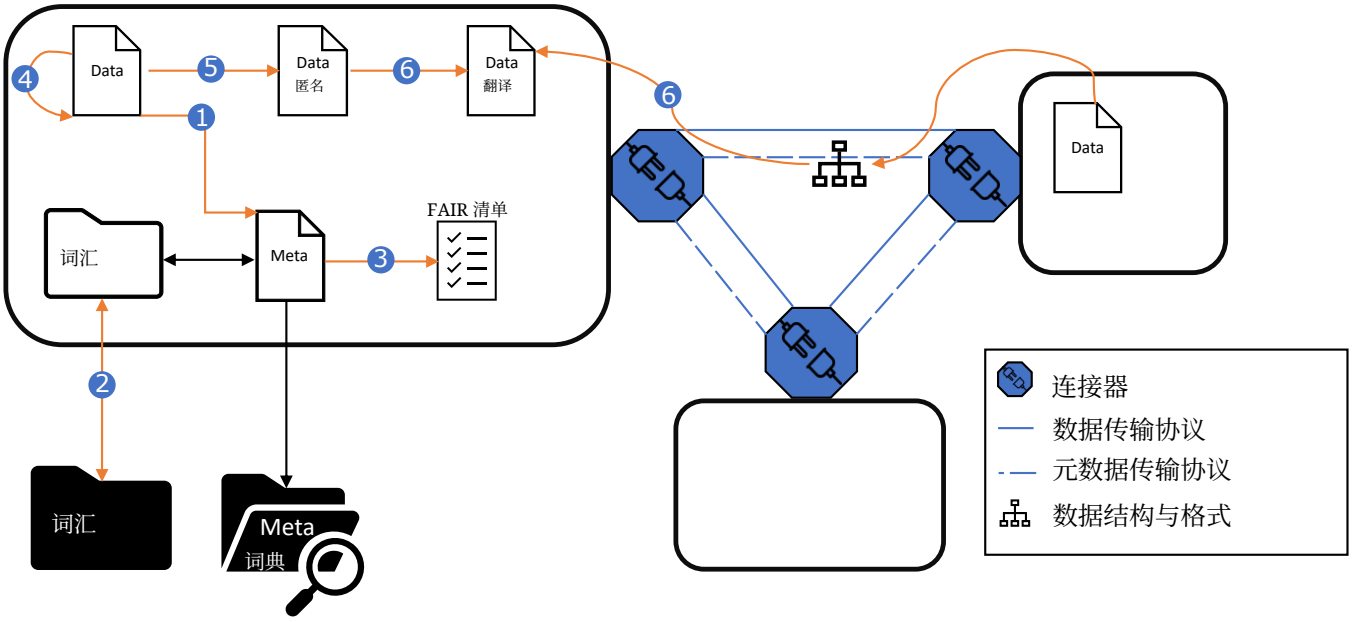


图 2：包含三个成员的机器学习增强型数据空间概览。（1）自动元数据提取，（2）本体与词汇对齐，（3）FAIR 性评估，（4）数据质量评估与增强，（5）隐私保护，（6）兼容性改进。

3.4 数据质量评估与增强

数据质量是数据消费者尤为关注的问题，因为它会影响数据的可信度和实用性。遗憾的是，仅凭元数据无法判断数据质量。为确保数据质量，必须考虑多个维度，包括准确性、完整性、正确性、有效性、完整性和唯一性。各个维度的重要性可能因数据的预期用途和数据消费者的需求而异。

准确性是指数据对其所代表的现实世界现象的反映程度。完整性是指必要数据的齐备程度。正确性是指数据遵循既定规则的程度，例如语法、语义或数据约束方面的规则。有效性是指数据遵循预定义格式、结构和领域规范的程度。完整性是指数据免受未经授权更改的保护程度。最后，唯一性是指每个数据项彼此不同且可识别的程度。

为确保数据质量，数据提供者必须采取措施评估并提升其数据质量。这包括实施数据验证和质量检查，采用数据剖析和数据清洗等技术，以及落实数据治理政策和流程。数据消费者也应采取措施评估所接收数据的质量，例如评估数据的来源与溯源信息、执行数据质量检查，并监测数据是否存在异常。

机器学习算法在确保数据的质量和准确性方面发挥着至关重要的作用。实现这一目标的一种方式是将数据与其他来源进行比较，以验证其准确性。此外，机器学习算法还可以经过训练，以识别数据中的模式和异常 [1, 18], , 从而标记出任何潜在的不准确之处或错误。

使用机器学习算法的另一项优势是能够补全缺失数据。通过分析数据中的模式和关系，机器学习模型可以预测缺失值并将其填补 [8, 20, 27]。在手动填补缺失数据耗时较长或具有挑战性的情况下，这尤其有用。

此外，机器学习技术还可用于识别并删除数据中的重复项，从而提高数据整体的唯一性和一致性 [6, 19, 26]。

3.5 隐私保护

个人信息、医疗记录和金融数据等私有且敏感的数据，通常受到严格的保护和访问法规及准则的约束。为了让不同系统能够交换和使用私有数据，它们必须能够准确解读并理解数据的含义和上下文，并确保数据的使用符合适用的法律法规。确保私有数据的语义互操作性，需要结合技术解决方案（例如安全数据交换协议和数据匿名化技术）以及严格的治理与合规机制。

为实现这一目标，数据提供者可以使用机器学习技术，自动检测其系统中的私人和敏感数据 [2, 21]，并采取适当措施对数据进行屏蔽 [29] 或匿名化 [15]。这有助于在促进数据共享和互操作性的同时保护个人隐私。例如，数据去标识化、数据屏蔽和差分隐私等技术可以

用于移除数据中的身份识别信息，同时保留其分析价值。

3.6 兼容性改进

此外，即使数据提供方和消费者使用相同的词汇和本体，如果资源与需要集成这些资源的消费者系统不兼容，它们在语义上仍不可互操作。为了解决数据交换中的资源不兼容问题，可采用数据映射和数据转换等方案。机器学习技术在这些任务中表现出色。

当需要访问资源的系统无法理解或使用这些资源时，资源在语义上便不可互操作。如果资源采用不同的数据格式或结构，就可能出现这种情况，导致系统难以集成并利用其中的信息。

为了解决数据交换中的资源不兼容问题，可采用数据映射 [13] 和数据转换 [22] 等方案。数据映射是指将一种资源中的数据元素与另一种资源中对应元素进行对齐的过程。数据转换是指将数据从一种格式或结构转换为另一种格式或结构的过程。这两种方案都有助于提高资源的兼容性并实现数据交换。机器学习也可用于将数据从一种格式转换为另一种格式，例如将自然语言文本转换为结构化数据 [30]。

讨论

提升数据空间的语义互操作性是一项复杂的任务，涉及不同的层面和方法。本文聚焦于可在国际数据空间背景下通过机器学习加以改进的特定方面。为此，我们提出开发由机器学习驱动的软件，使其能够作为智能数据应用轻松集成到数据空间连接器中。这将使软件更加易用且可访问，从而实现与现有系统的无缝集成。

此外，随着Gaia-X在欧洲及其他地区日益普及，这类软件也可以作为服务在Gaia-X框架内提供，为成员提升语义互操作性提供宝贵资源。通过将机器学习融入数据空间，组织能够确保其数据得到妥善组织，并确保其系统能够与其他系统有效通信和交换信息，从而实现更高效、更有效的数据管理与交换。

4 结论

本文提出了我们关于通过机器学习提升数据空间语义互操作性的创新观点。我们重点探讨了国际数据空间架构中互操作性的六个关键方面，阐明了每个方面的重要性，以及机器学习如何增强这些方面对成功实现数据交换的作用。

作为这项工作的后续计划，我们将把本文提出的一些概念和解决方案整合到国际数据空间和 Gaia-X架构中的真实世界数据交换场景中进行测试。这将为我们提供宝贵的见解，帮助我们了解所提方法在实际实施中的情况和有效性，并进一步推动数据互操作性与交换领域的发展。

参考文献

[1] Shikha Agrawal 和 Jitendra Agrawal。2015。使用数据挖掘技术进行异常检测的综述。Procedia Computer Science 60（2015），708–713。[2] Hadeer Ahmed、Issa Traore、Sherif Saad 和 Mohammad Mamun。2021。使用深度学习自动检测非结构化上下文相关敏感信息。物联网 16（2021），100444。[3] Vladimir Alexiev。2022。数据空间与知识图谱。在数据空间与语义互操作性研讨会。TRUSTS可信安全数据共享空间。[4] Alexandre Bento、Amal Zouaq 和 Michel Gagnon。2020。使用卷积神经网络进行本体匹配。在第十二届语言资源与评估会议论文集。5648–5653。[5] Zeyd Boukhers 和 Azeddine Bouabdallah。2022。利用视觉和自然语言从科学PDF文档中提取元数据：一种多模态方法。在第22届ACM/IEEE数字图书馆联合会议论文集。1–5。[6] Peter Christen 和 Karl Goiser。2019。迈向自动化数据关联与去重。Computer 16，17（2019），22–24。[7] AnHai Doan、Jayant Madhavan、Pedro Domingos 和 Alon Halevy。2004。本体匹配：一种机器学习方法。本体手册（2004），385–403。[8] Md Kamrul Hasan、Md Ashraful Alam、Shidhartho Roy、Aishwariya Dutta、Md Tasnim Jawad 和 Sunanda Das。2021。缺失值插补会影响机器学习的性能：文献综述与分析（2010–2021）。Informatics in Medicine Unlocked 27（2021），100799。[9] Yuan He、Jiaoyan Chen、Denvar Antonyrajah 和 Ian Horrocks。2022。BERTMap：一种基于BERT的本体对齐系统。在AAAI人工智能会议论文集，第36卷。5684–5691。[10] Vivek Iyer、Arvind Agarwal 和 Harshit Kumar。2020。VeeAlign：一种用于本体对齐的监督式深度学习方法。在OM@ ISWC。216–224。[11] Meriem Ali Khoudja、Messaouda Fareh 和 Hafida Bouarfa。2018。使用神经网络进行本体匹配：综述与分析。在2018年应用智能系统国际会议（ICASS）。IEEE，1–6。[12] Jussi Kiljander、Alfredo D'elia、Francesco Morandi、Pasi Hyttinen、Janne Takalo-Mattila、Arto Ylisaukko-Oja、Juha-Pekka Soininen 和 Tullio Salmon Cinotti。2014。普适计算与物联网的语义互操作架构。IEEE Access 2（2014），856–873。[13] Dingcheng Li、Ming Huang、Xiaodi Li、Yaoping Ruan 和 Lixia Yao。2018。MfeCNN：用于数据映射的混合特征嵌入卷积神经网络。IEEE纳米生物科学汇刊 17，3（2018），165–171。[14] Ming-Chin Lin、Daniel J Vreeman 和 Stanley M Huff。2011。研究使用LOINC代码在三家大型机构间交换的实验室数据的语义互操作性。在AMIA年度研讨会论文集，第2011卷。美国医学信息学协会，805。[15] Abdul Majeed 和 Sungchang Lee。2020。用于隐私保护数据发布的匿名化技术：全面综述。IEEE Access 9（2020），8512–8545。[16] Azadeh Haratian Nezhadi、Bita Shadgar 和 Alireza Osareh。2011。使用机器学习技术进行本体对齐。International Journal of Computer Science & Information Technology 3，2（2011），139。[17] Aris M. Ouksel 和 Amit Sheth。1999。全球信息系统中的语义互操作性。ACM SIGMOD Record 28，1（1999），5–12。[18] Guansong Pang、Chunhua Shen、Longbing Cao 和 Anton Van Den Hengel。2021。用于异常检测的深度学习：综述。ACM计算调查（CSUR）54，2（2021），1–38。[19] Jisung Park、Jeonggyun Kim、Yeseong Kim、Sungjin Lee 和 Onur Mutlu。2022。{DeepSketch}：一种用于{去重后}增量压缩的新型{基于机器学习的}参考搜索技术。在第20届USENIX文件与存储技术会议（FAST 22）。247–264。[20] PS Raja 和 KJSC Thangavel。2020。使用无监督机器学习技术进行缺失值插补。Soft Computing 24，6（2020），4361–4392。[21] Soumya Ray、Kamta Nath Mishra 和 Sandip Dutta。2021。云与物联网网络中的敏感数据识别与安全保障。Int J Comput Netw Inf Secur 13（2021），1–25。[22] Saliha Sajid、Bjørn Marius von Zernichow、Ahmet Soylu 和 Dumitru Roman。2019。使用机器学习在grafterizer中提供预测性数据转换建议。在《元数据与语义研究：第13届国际会议，MTSR 2019，意大利罗马，2019年10月28日至31日，修订版论文集》。Springer，137–149。[23] Simon Scerri、Tuomo Tuikka、Irene Lopez de Vallejo 和 Edward Curry。2022。欧洲共同数据空间：挑战与机遇。《数据空间：设计、部署与未来方向》（2022），337–357。[24] K Manjula Shenoy、KC Shet 和 U Dinesh Acharya。2013。使用图匹配进行安全本体匹配。载于《计算与信息技术进展：第二届计算与信息技术进展国际会议论文集》。

计算与信息技术（*ACITY*）*2012*年*7*月*13*日至*15*日，印度金奈——第*2*卷。Springer，11–18。[25] Rita T Sousa、Sara Silva 和 Catia Pesquita。2022。《监督式语义相似度工具包》。载于《语义网：*ESWC 2022*卫星活动：希腊克里特岛赫索尼索斯，*2022*年*5*月*29*日至*6*月*2*日，会议论文集》。Springer，42–46。[26] Sashi Tarun、Ranbir Singh Batth 和 Sukhpreet Kaur。2021。分布式系统中使用先进机器学习架构进行数据去重的方案。载于*2021*年国际计算科学会议（*ICCS*）。IEEE，53–60。[27] Tressy Thomas 和 Enayat Rajabi。2021。基于机器学习的缺失值插补技术系统综述。数据技术与应用 55, 4（2021），558–585。[28] Dominika Tkaczyk。2017年。科学文献元数据提取的新方法。*arXiv* 预印本 *arXiv:1710.10201*（2017）。[29] Vicenç Torra。2022年。数据隐私：掩蔽方法。载于数据隐私指南：模型、技术与解决方案。Springer，159–209。[30] Saurav Verma、Khushboo Jain 和 Chetana Prakash。2020年。物联网（IoT）中使用机器学习算法将非结构化数据转换为结构化数据。载于国际创新计算与通信会议（*ICICC*）论文集。

Rec收稿日期：2007年2月20日；修订日期：2009年3月12日；接受日期：5 2009年6月