arXiv:2602.01822v1 [cs.DB] 2 Feb 2026

ChemDCAT-AP：通过 DCAT-AP 的情境扩展实现语义互操作性

Philip Str¨omert[1], Hendrik Borgelt[2], David Linke[3], Mark Doerr[4], Bhavin Katabathuni[1], OliverKoepler[1], 和 Norbert Kockmann[2]

[1] TIB——莱布尼茨科学与技术信息中心，德国 philip.stroemert@tib.eu[2] 多特蒙德工业大学，德国[3] 莱布尼茨催化研究所[4] 格赖夫斯瓦尔德大学

摘要。跨领域数据集成推动跨学科研究数据再利用和知识转移。然而，各学科均维护各自的元数据模式和领域本体，并采用不同的概念模型和应用配置文件，因此语义互操作性难以实现。W3C 数据目录词汇表（DCAT）提供了一套广泛采用的 RDF 词汇表，用于描述数据集及其数据分发，但其核心模型有意保持轻量。为增强 DCAT 的表达能力，人们开发了许多领域特定应用配置文件，其中最知名的是面向公共数据的 DCAT-AP。为促进研究数据的跨领域互操作性，我们提出 DCAT-AP Plus，即关联用例特定情境的应用配置文件（DCAT-AP+）。这一通用应用配置文件能够全面呈现研究数据生成的来源信息及其情境。DACT-AP+ 引入了一个上层，可由各个领域加以专门化，同时不牺牲兼容性。我们展示了 DCAT-AP+ 及特定配置文件 ChemDCAT-AP 的应用，以说明化学与催化这两个相邻学科开展数据集成的潜力。我们采用基于 YAML 的建模框架 LinkML，以支持模式继承、生成领域特定子模式，并提供数据类型协调、验证和格式转换机制，从而确保 DCAT-AP+ 与 ChemDCAT-AP 能够顺利集成到现有数据基础设施中。

关键词：DCAT-AP、ChemDCAT-AP、DCAT-AP+、LinkML。

1 引言

过去十年间，研究格局发生了深刻转变，开始以前所未有的规模共享开放数据。欧洲开放科学云（EOSC） [1] 和德国国家研究数据

基础设施（NFDI） [2] 通过公开资助研究数据，推动可重复、高效且跨学科的研究。随着 FAIR原则（可发现、可访问、可互操作、可再利用） [3], 日益普及，研究人员如今期望研究数据能够被发现，并依据最低信息标准实现再利用；这些标准应记录完整的实验背景，例如化学或催化领域中的仪器设备、反应条件、来源信息和分析方法，从而使研究结果能够被理解、复现并跨学科整合 [4]。对于化学和催化领域的科学界而言，物理科学数据基础设施（PSDI） [5] 计划、NFDI4Chem[6] 和 NFDI4Cat [7] 联盟致力于制定最低信息标准。

数据目录应用规范（DCAT-AP）是一个成功范例，它为元数据提供了坚实基础，使欧洲数据门户之间能够实现数据互操作性 [8]。然而，DCAT-AP 规范并未涵盖详细的科学语境，即数据集的实际内容以及数据的生成方式。对于化学或催化数据集而言，需要针对化学结构表示法（例如通过 InChI [9] 或 SMILES [10]）、反应条件、样品溯源信息以及特定分析方法（例如核磁共振波谱或气相色谱法）进行领域特定的扩展。

为弥补这一缺口，NFDI4Chem 和 NFDI4Cat 合作开发了化学 DCAT 应用规范（ChemDCAT-AP）。该规范为两个联盟提供共同基础，并构建于 DCAT-AP 的领域无关核心扩展之上，该扩展名为 DCAT-AP 用例特定上下文扩展（DCAT-AP+）。这一核心扩展将 DCAT-AP 中已有但定义不够充分、基于溯源本体（PROV-O）[11] 核心类的溯源模式规范化。它还增加了一种轻量级分类机制，以及一种通用属性模式，可为新引入的类附加任意定量或定性属性。ChemDCAT-AP 引入 DCAT-AP+，并添加与既有化学本体（即 ChEBI、CHEMINF 和 CHMO） [12–14] 对齐的化学与催化专用概念（化学实体、反应、分析方法），从而为 NFDI4Chem 和 NFDI4Cat 提供共享的、基于语义的元数据基础。

我们的框架采用关联数据建模语言（LinkML）[15], 这一灵活的建模框架，使我们能够创建模块化架构，其中基础的 ChemDCAT-AP 可轻松导入，并根据化学和催化领域的具体需求进行扩展。这种方法通过将架构编译为多种实用格式（包括 Python/Pydantic 数据类、SHACL 形状和 JSON Schema），让语义标准更易于采用，从而弥合语义严谨性与存储库开发者实际实施之间的差距。本文的主要贡献如下：

– 将 DCAT-AP 3.0规范 转换为 LinkML 模式。– 设计领域无关扩展 DCAT-AP+ ，以便更好地基于来源信息描述数据集的生成方式及其内容。

– ChemDCAT-AP 作为共享基础配置文件，导入并针对化学和催化领域专门扩展 DCAT-AP+ 。– 为核磁共振波谱实现化学研究最小信息（MIChI）标准的原型，以展示 ChemDCAT-AP 对实验情境的适应能力，以及其在 NFDI4Chem 搜索服务中的实现。

本文结构如下：第 2 节介绍相关元数据标准和 LinkML 框架的背景。第 3 节概述用于设计和构建扩展的方法。第 4 节介绍研究结果，详细说明 ChemDCAT-AP 的架构及其在 NFDI4Chem 搜索服务中的实现。最后，第 5 节讨论本研究的影响和局限，并总结结论及未来方向。

2 背景

2.1 领域特定元数据的必要性

随着关联开放数据的发展，领域特定元数据的需求日益凸显。DataCite [16], 都柏林核心 [17],和 METS [18]等标准支持以 XML、JSON 或 RDF 描述数据集，而 DCAT [19], PROV-O [11], 和语义传感器网络本体（SOSA [20] 等 W3C 词汇表则增添了语义层。这些标准和词汇表共同支持跨领域的机器可读元数据。

然而，这类通用框架仍过于宽泛，无法捕捉科学语境中的详细信息。CDIF（跨领域互操作框架） [21] 指出了支持可发现性（例如 Schema.org[22], DCAT）的标准与面向内容或变量的标准（例如 SOSA、I-ADOPT[23]）之间的差距。在化学、催化和材料研究领域，弥合这一差距尤为重要，因为数据集通过反应、催化剂和实验条件紧密相连。

2.2 DCAT-AP：与本体对齐的元数据模式

DCAT-AP 是 DCAT [24], 的欧洲应用配置文件，为跨门户描述公共部门数据集提供了通用规范，同时允许各目录保留自己的存储和文档。它通过与受控词汇表、分类体系和本体对齐，确保语义互操作性。

DCAT-AP 的扩展必须遵循严格的实施指南 [25]:

– 使用说明只能收窄，不能扩大。– 可以添加新的类和属性，但不得与现有内容重复。– 属性的基数可以在一定范围内修改，但必须保留强制要求。

作为一种应用配置文件，DCAT-AP 对通用 DCAT 词汇作出限制，但不重新定义概念。它以 SHACL 表达，规范本体术语的使用，同时确保不同配置文件之间的互操作性 [26]。

2.3 现有的 DCAT-AP 扩展

多个领域特定的配置文件展示了 DCAT-AP 的灵活性：

GeoDCAT-AP [27] 扩展了 DCAT-AP，用于描述地理空间数据集、数据集系列和服务。它满足地理信息特有的元数据需求，改善了欧洲数据门户中的发现与再利用。其主要特性包括：提供有关参与数据集创建、处理和使用的代理的更丰富元数据，以及一些领域特定属性，例如以文本表示的空间分辨率和主题类别。作为官方扩展，它遵循与 DCAT-AP 相同的方法，以确保互操作性。

StatDCAT-AP [28] 扩展了 DCAT-AP，适用于统计数据集，使统计领域内部以及统计领域与开放数据门户之间实现互操作性。1.0.1 版本支持统计数据与元数据交换规范（SDMX）、RDF数据立方和 CSV 等格式的元数据。其显著特征是能够描述多维结构，捕获跨地理、时间和领域特定维度的变量。它通过与 DCAT-AP 相同的工作组和审查流程开发而成，确保广泛的利益相关者参与和兼容性。

MLDCAT-AP [29] 扩展了 DCAT-AP，以支持机器学习模型、其数据集、质量度量及相关出版物。它整合了专门的词汇表，包括 mls 和 mlso（机器学习）、biro（引文）、dqv（数据质量）以及 lpwc/lpwc-p（带代码论文）。MLDCAT-AP 仍处于候选建议阶段，是标准化人工智能资产元数据的早期尝试，重点关注模型与数据集之间的关系，而非数据集内容。

HealthDCAT-AP [30]扩展了 DCAT-AP，以支持健康数据集，满足特定领域的要求，并确保与《欧洲健康数据空间》（EHDS）法规保持一致。其目的是提升欧洲各地健康相关数据的互操作性和可发现性。

这些扩展主要为数据集类增加属性，但并未采用统一的设计模式。此处提出的 DCAT-AP+ 方法旨在通过基于来源信息的模型协调此类工作，并与 DCAT-AP 开发者 [31]提出的建议保持一致。

2.4 DCAT-AP的替代方案

在可发现性语境中描述数据内容时，DCAT和DCAT-AP的两个主要替代方案是Schema.org和DataCite。

由主要搜索引擎提供商开发的 Schema.org 将元数据直接嵌入网站，以支持搜索功能。虽然它可以描述

与 DCAT 类似的内容，但它针对网页索引进行了优化，对研究数据的表达能力有限，且扩展灵活性较低。生命科学数据领域一个值得关注的领域专用扩展是 Bioschemas。

DataCite [16] 提供与数字对象标识符（DOI）关联的元数据，确保研究成果能够稳定、持久地被引用和发现。然而，其元数据与这种持久标识符（PID）基础设施紧密绑定，因此相比 DCAT-AP 这类开放模型，更难进行调整。

两者都发挥着重要作用，但都不足以支持详细的领域特定元数据。

2.5 LinkML 框架：基础性选择

选择关联数据建模语言（LinkML）作为实现框架，是因为它简化了符合FAIR原则的标准的创建，同时让语义网技术更易于使用。其基于 YAML 的语法支持通过单一事实来源进行模式设计，并据此生成多种机器可解读的表示形式（Python/Pydantic类、带有 JSON-LD 的 JSON Schema、SHACL 形状、Markdown）。

模式元素（类、属性、枚举和数据类型）可以通过其术语IRI映射到OWL本体或 SKOS叙词表等外部资源。这使 LinkML 模式扎根于共享语义，并促进互操作性。由于这些映射是可选的，用户可以受益于语义网工具栈，而不必受其限制；对于许多用例，其他自动生成的表示形式或许已能提供足够的标准化。不过，使用此功能时，LinkML 的数据转换库可以成为快速生成 RDF知识图谱的易用工具。

通过将已选定的术语 IRI 映射替换为等效映射，单个 LinkML 模式即可整合基于不同本体承诺构建的 RDF 数据，前提是所用本体遵循相同的建模模式。例如，具有 given_name 和 family_name 属性的 Person类可以与 FOAF本体 [32] 或 Schema.org 对齐，因为二者定义了等效结构。

这种灵活性，加上实用的工具和活跃的社区，使 LinkML 非常适合扩展 DCAT-AP，也适合开发人员和数据管理员等不同用户群体采用。

3 方法

本节概述 ChemDCAT-AP 开发过程中采用的设计原理、技术实现和验证流程。其协作式迭代开发遵循了若干核心原则：

1. 符合性：所有扩展均严格遵守官方 DCAT-AP 3.0规范的强制性约束。2. 以可发现性为中心：新增属性旨在直接提升特定化学和催化应用场景下的数据集可查找性。

3. 语义接地：每个新增类或属性都映射到既有本体（例如 PROV-O、QUDT、BFO、IAO、OBI、CHMO） [33–37]。4. 简化非专业人士的使用：数据管理员和软件开发人员无需具备深厚的本体知识，也能使用该模式，同时仍可获得丰富的机器可读表示。

将这些原则转化为最终 LinkML 模型的具体工作流程分为三个紧密关联的阶段：(i) 将 DCAT-AP 忠实地自动移植到 LinkML；(ii) 系统地扩展该 DCAT-AP LinkML 模式，加入溯源模式；(iii) 由持续集成驱动 ChemDCAT-AP 模式的开发。源代码、文档和所有模型制品均托管在 GitHub[5] 上，向公众开放。

3.1 建模框架与实现

我们选择 LinkML 作为实现框架的主要原因已在 2.5 节中详细阐述。使用官方 LinkML 项目复制器模板[6] 为项目搭建初始框架，为我们提供了稳健的起点，并带来诸多优势，例如：

– 标准化的代码仓库布局，其中包含示例模式和测试文件，– 预配置的 GitHub Actions，用于持续集成 (CI) 和质量控制，– 或一个 just 命令运行器，其中包含用于验证模式和数据示例、测试和部署 HTML 文档[7], ，以及更新 Python 虚拟环境和模板代码的标准化命令

完整的工具链基于 Python 3，并使用 linkml-runtime 进行数据处理、使用 pydantic 进行数据验证，以及使用 rdflib 操作 RDF。

3.2 将 DCAT-AP 移植到 LinkML

为确保符合我们的首要指导原则，我们编写了一个 Python 脚本，将 DCAT-AP 3.0.0 的 SHACL 形状转换为 LinkML。我们将 DCAT-AP 规范的 JSON-LD 序列化形式作为事实标准，因为这样可以使用 Python 内置的 JSON 库更简单地进行解析。我们特意采用了 DCAT-AP GitHub 代码仓库主分支中的版本，该版本比标记为 3.0.0 的发行版更新，而且也是官方规范网站链接到的版本。[8]

[5] https://github.com/nfdi-de/chem-dcat-ap [6] https://github.com/linkml/linkml-project-copier [7] https://nfdi-de.github.io/chem-dcat-ap/ [8] 另请参阅 https://github.com/SEMICeu/DCAT-AP/issues/428

该脚本读取 JSON-LD 文件，并将每个 SHACL 节点形状转换为 LinkML 构造。targetClass 指向本体类的节点形状会转换为 LinkML 类，而指向 XSD 数据类型的节点形状则会转换为 LinkML 数据类型。这些节点中的属性形状会被解释为派生 LinkML 类的槽。SHACL targetClass 和 path 属性中的原始术语 IRI 均按原样保留，分别作为类_uri 和槽_uri，从而确保生成的模型在语义上与原始 SHACL 表示完全一致。

原始 SHACL 形状中存在联合值域，这带来了一个特殊挑战。对于 dcat:primaryTopic 这类属性，我们使用 LinkML 的 any_of 关键字来支持对象类的联合。不过，由于当前 LinkML 实现对数据类型联合的支持尚不稳定，我们采取了保守做法，将日期相关槽的值域限制为 XSD date 数据类型。这是一种有意采用的、更严格的解释；对于我们当前的使用场景而言，这种做法是可接受的，同时我们也在等待未来的 LinkML 版本提供完整的数据类型联合支持。

3.3 模块化双层设计

我们采用了模块化双层设计：首先，使用同一个 Python 脚本，以程序化方式将 DCAT-AP 忠实转换为 LinkML 的成果扩展为 DCAT-AP+，并以 wasGeneratedBy 属性及其值域类 Activity 作为切入点。第二步，将 DCAT-AP+ 导入手动创建的 ChemDCAT-AP 中。

溯源核心层 DCAT-AP+ 层是一个最小化、与领域无关的扩展，用于规范化数据溯源的描述。它引入了一种通用模式，用于描述数据生成活动，并与 W3C PROV 本体（PROV-O）及其他成熟的词汇表保持一致。Prudhomme 等人 [38] , 提出了 PROV 与基本形式本体（BFO）之间的形式化映射，进一步验证了这一选择；这与我们在 NFDI4Chem 和 NFDI4Cat 中使用 OBO Foundry 本体的做法相契合。

我们首先通过扩展 DCAT-AP 中定义不足的 Activity 类，实现了核心 PROV-O 活动模式，并为其添加 PROV-O 属性（为清晰起见，将其重新命名为：has_input_entity、has_output_entity、carried_out_by 和 has_input_activity），以及各自对应的、尚未纳入 DCAT-AP 的值域类 Entity 和 AgentEntity。后者之所以更名，是因为 DCAT-AP 中的 foaf:Agent 类已用于其他目的。随后，我们通过添加 DataGeneratingActivity 类（Activity 的子类）对这一核心模式进行了细化，用以表示测量或模拟等过程。此类活动的主要输入为 EvaluatedEntity（例如化学样品）或 EvaluatedActivtiy（例如催化反应），并使用专门的子属性 evaluated_entity 和 evaluated_activity 对其进行描述。此类活动的 carried_out_by（例如反应器）还可以通过 AgentEntity 的 Software 或 Device 子类进一步细化。

为了详细描述实体、活动或代理，我们添加了用于表示定性和定量属性的类。QuantitativeAttribute 类（与 qudt:Quantity 对齐）包含用于表示数值、数量类型和单位的属性。QualititativeAttribute 类（与 prov:Entity 对齐）允许为其字面表示提供字符串值。这些类通过 has_quantitative_attribute 和 has_qualitative_attribute 属性使用，而这两个属性均映射到 dcterms:relation。

DCAT-AP+ 的一个关键特性是分类器混入类，它为所有 DCAT-AP+ 核心类提供了两个用于语义注释的属性：

– rdf_type：映射到 RDF 属性 rdf:type，使实例能够额外标注为外部本体中的一个类（例如，将一个数据生成活动实例归类为 CHMO:0000595，即 13C 核磁共振测量）。这是 DCAT-AP 开发者认可的一种方法 [31]– type：映射到 dcterms:type，使其能够使用 SKOS 词汇表或其他受控术语集中的术语进行分类，而不像 rdf_type 那样作出强本体承诺。

对于可选元数据，DCAT-AP+ 重用标准 DCAT-AP 和都柏林核心属性，以便提供标识符、标签、描述和部分—整体关系。它还以较为宽泛的方式重用 PROV-O 中的计划和位置类，详情可见 DCAT-AP+图 1 所示的 UML 图。

领域专用层 ChemDCAT-AP 层作为领域专用扩展，将 DCAT-AP+ 的通用溯源模型调整为满足化学和催化研究领域的明确需求。它通过导入 DCAT-AP+ 中的所有实体，并引入直接映射到既有化学本体的专用类和属性来实现这一目标。这种分层方法确保元数据完全符合 DCAT-AP+ ，因而也符合 DCAT-AP 标准，同时获得支持高级领域专用发现用例所需的语义精度。ChemDCAT-AP 的首要设计目标是从通用键值描述转向丰富的自描述数据模型，以反映化学研究的核心实体，例如物质、样品和反应。

ChemDCAT-AP 的一项关键策略是对 DCAT-AP+ 中的通用类进行专门化。通用的经评估实体被设为子类，以创建具有化学意义的概念，例如 SubstanceSample（映射到 SIO:001378）和化学实体（映射到 CHEBI:23367）。这使得元数据记录能够直接表示化学样品或分子实体，消除了通用模型带来的歧义。例如，SubstanceSample 通过 ChemicalSubstance 混入类进一步细化，该类提供描述其组成（composed_of）以及浓度（has_concentration）或 pH 值（has_ph_value）等关键化学属性的属性。

此外，ChemDCAT-AP 引入了一组丰富的领域特定属性，这些属性被定义为相应通用 DCAT-AP+ 属性的子属性。这符合创建应用配置文件的建议，

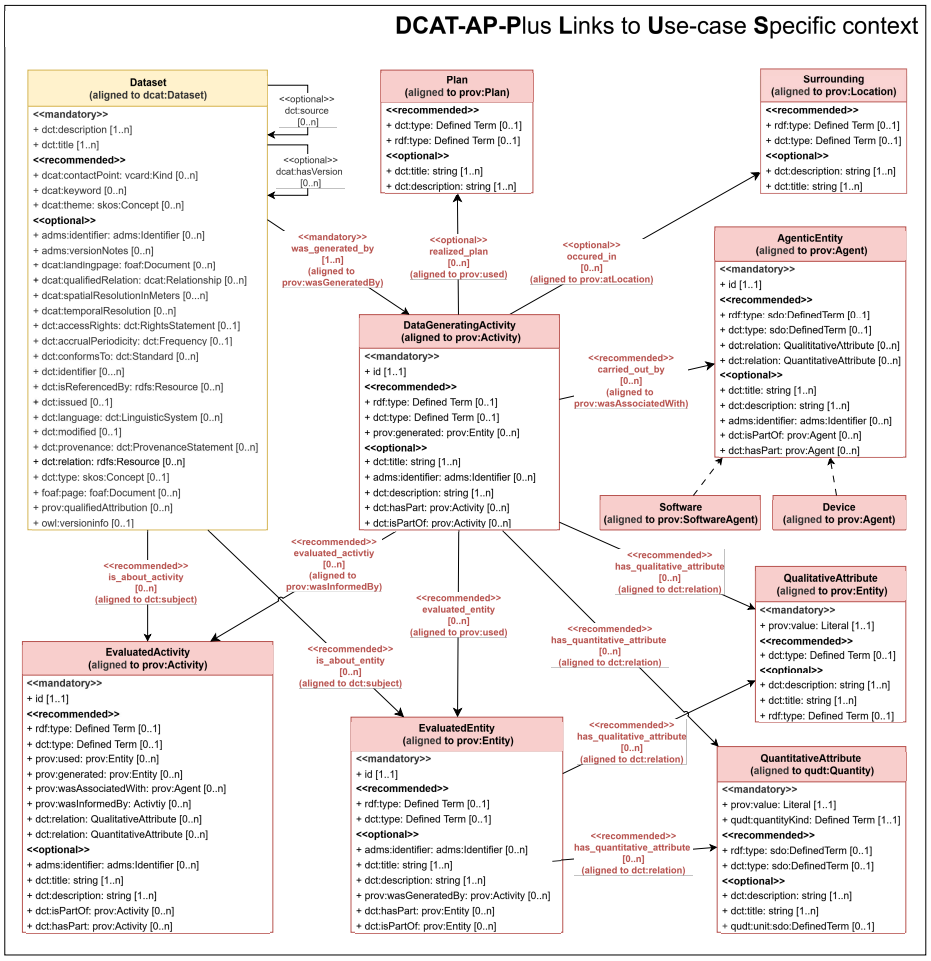


图 1. 此可视化图展示了我们如何通过 *prov:Activity* 类扩展 DCAT-AP，从而能够为作为 *dcat:Dataset* 主要主题的实体或活动提供额外的领域特定元数据，同时也为创建数据集时所用的工具和计划，以及该活动发生的环境提供元数据。图中还展示了我们采用的映射，这些映射将此 LinkML 模式纳入 PROV-O、DCTerms 和 QUDT 的本体承诺中。[34]

确保继承模型清晰 [25, 31]。例如，为了描述化学反应的组成部分，ChemDCAT-AP 提供了高度具体的属性。used_catalyst 属性被定义为通用属性 carried_out_by 的子属性，并被更精确地映射到命名反应本体（RXNO）中的谓词 RXNO:0000425（“具有催化剂”） [39]。同样，generated_products 属性专门化了 had_output_entity，并映射到 RO:0004008。这种分层属性设计在保持与更通用的 DCAT-AP+ 模型完全向后兼容的同时，提供了更丰富的领域特定语义。任何能够理解 DCAT-AP+ 的客户端仍可将 used_catalyst 解读为参与该活动的某种代理，而具备化学知识的客户端则能理解其作为催化剂的确切作用。

比较两种模型下化学物质样品的元数据记录时，这种专门化的实际优势显而易见。[9] 如示例所示，符合 DCAT-AP+ 的记录必须依赖通用的 has_qualitative_attribute 属性，并使用 rdf_type 槽来指定某个字符串表示 InChIKey（CHEMINF:000059）或 SMILES 字符串（CHEMINF:000018）。相比之下，符合 ChemDCAT-AP 的记录使用专用的 inchikey 和 smiles 属性。这些属性在 LinkML 模式中将 CHEMINF 本体类定义为其值域，因此无需在实例数据中反复添加 rdf_type 注释。这不仅让实例数据更加简洁易读，也大大简化了从现有化学软件和数据库映射数据的流程，因为它们通常具有名称相似的字段。ChemDCAT-AP 通过提供更明确且受约束的模型，降低了歧义的可能性，并提升了自动化数据验证以及将数据整合到更大型化学知识图谱中的潜力。

3.4 验证与应用

ChemDCAT-AP 模式的语法正确性通过自动化测试流水线持续确保，该流水线会在每次拉取请求时通过 GitHub Actions 执行。该流水线使用 linkml-validate 工具检查 LinkML 模式本身的完整性，并据此验证示例数据实例。

为评估其适用性，我们将该模型应用于表示 Chemotion存储库中的真实世界元数据。Chemotion存储库是一个电子实验记录本和存储库，在 NFDI4Chem 联盟中得到广泛使用。我们成功地将 Chemotion 记录中的所有现有元数据字段映射到 ChemDCAT-AP 架构。这项评估证实，ChemDCAT-AP 不仅能够容纳所有当前元数据，还提供了捕获更详细语义信息所需的结构，从而实现了我们的首要设计目标。

[9] https://github.com/nfdi-de/chem-dcat-ap/blob/main/tests/data/valid/NMRAnalysisDataset-001.yaml

4 结果

我们的主要成果是一个灵活的双层元数据框架，成功弥合了通用数据编目与化学研究特定描述需求之间的差距。我们通过两项关键成果展示了其有效性：为核磁共振波谱制定了化学研究最小信息（MIChI）标准，并成功将该框架部署于 NFDI4Chem 搜索服务中。

4.1 使用 ChemDCAT-AP 形式化 NMR MIChI

我们工作的一个重要成果，是成功地对用于 NMR 光谱分析的 NFDI4Chem MIChI（化学研究最小信息集）MARGARITAS [41] 进行了形式化。ChemDCAT-AP 框架证明了其卓越的效用，使协作式迭代开发成为可能。在这一过程中，主要挑战在于正确解读标准的语义，而非技术建模本身。

这一过程促成了 MARGARITAS 标准中所有必需属性及大多数推荐属性的形式化。值得注意的是，我们的模型能够捕获比 Chemotion 存储库目前通过应用程序接口公开的更丰富的元数据。例如校准化合物或探测核等细节，虽然存在于原始数据文件或着陆页中，但此前首次得以使用我们的模型进行结构化。

除了模式本身，LinkML 工具链生成的一系列自动生成成果也是一项重要成果。供人阅读的文档便于领域专家审阅和贡献，而多种机器可读表示（例如 Python/Pydantic、JSON Schema）则降低了仓储库管理员和软件开发人员采用该标准的门槛，帮助他们与 MARGARITAS 的这一形式化版本保持一致。如今，整个工作流程已成为系统化编码其他 MIChI 标准的经过验证的模板，有望加速 FAIR 化学研究数据的创建。

4.2 在 NFDI4Chem 搜索服务中的实现

ChemDCAT-AP 的实用价值已通过其成功集成到 NFDI4Chem 搜索服务[10], 中得到证实，该服务基于 CKAN [40]构建。通过为现有的 ckanext-dcat 插件[11], 开发自定义配置文件，我们使该服务能够解析并导出丰富且符合 ChemDCAT-AP 标准的元数据。在这一过程中，根据我们的 LinkML 模型自动生成的 Python 数据类发挥了关键作用，为验证搜索服务采集的元数据并将其转换为标准化 RDF 图提供了直接简便的机制。

[10] https://search.nfdi4chem.de/ [11] https://github.com/ckan/ckanext-dcat

这一实现并非单纯的技术实践；它直接促成了 NFDI化学知识图谱（ ChemKG）的创建。我们的模型所捕获的语义精确的元数据支持高级分面检索查询，而这类查询无法通过通用目录元数据实现，例如查找所有使用了特定溶剂的数据集，或分析具有指定分子骨架的化合物的数据集。能够直接从搜索服务导出这些结构化元数据，并通过其 SPARQL端点提供访问，是推动研究基础设施之间实现更高互操作性的重要一步。

5 讨论

我们证明，通过 DCAT-AP+ 和 ChemDCAT-AP 实现的 DCAT-AP 分层扩展，为描述特定领域的研究数据提供了稳健且可扩展的解决方案。作为 DCAT-AP+ 核心的基于溯源的模式，提供了一项其他领域也可采用的统一设计原则；而 ChemDCAT-AP 则证实了该模式适用于化学和催化领域的复杂需求。

尽管本研究奠定了坚实的基础，未来仍有若干技术和战略方向值得进一步探索。在技术方面，我们当前的首要任务是通过 w3id.org 将这些模式发布为持久URL（PURL），并将模型作为软件包发布到 PyPI，以促进更广泛的复用。我们还计划将 DCAT-AP+ 拆分为独立的代码库，以便 NFDI联盟及其他相关方采用。此外，我们的研究发现，LinkML 当前的 SHACL 生成器在处理节点和属性形状的 IRI 命名方面存在局限。我们计划为 LinkML 社区贡献力量，以解决这一问题，确保生成的形状完全符合现行 SHACL 规范。

从战略层面来看，长期愿景是将 DCAT-AP+ 纳入未来版本的官方 DCAT-AP 标准中，因为它对 prov:wasGeneratedBy 属性进行了合乎逻辑且必要的形式化。不过，这需要经过全面验证，并得到更广泛社区的采用。我们已通过在 NFDI网络 内收集反馈启动这一进程，并诚邀更广泛的研究数据社区测试和批评 DCAT-AP+。

最终，ChemDCAT-AP 框架是 NFDI4Chem 第二阶段资助计划中拟建元数据模式服务（MSS）的基石。该服务不仅将托管化学领域专用及其他相关模式，还将提供这些模式之间的映射和自动转换；ChemDCAT-AP 的多种代码表示形式非常适合完成这项工作。通过将其与目前正在开发的 EOSC 元数据模式与交叉对照注册表（MSCR） [42], 集成，这项工作有助于构建一个互联性更强、语义互操作性更好的研究数据生态系统。

致谢。本研究由德国研究基金会（DFG）资助，资助编号为 441926934（NFDI4Cat）和 441958208（NFDI4Chem）。

参考文献

1. 欧盟委员会。欧洲开放科学云（EOSC）战略实施计划。（2019），https://www.go-fair.org/wp-content/uploads/2019/08/European-Open-Science-Cloud-EOSC-strategic-implementation-plan.pdf，访问日期：2025-09-26 2. Hartl, N., W¨ossner, E. & Sure-Vetter, Y. 国家研究数据基础设施（NFDI）。 *Informatik Spektrum*. 44, 370-373 (2021,10) 3. Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Du mon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J., 't Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons, A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., Van der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. 科学数据管理与管理维护的 FAIR 指导原则。 *Sci. Data*. 3, 160018 (2016) 4. Herres-Pawlis, S., Bach, F., Bruno, I., Chalk, S., Jung, N., Liermann, J., McEwen, L., Neumann, S., Steinbeck, C., Razum, M. & Koepler, O. 化学最低信息标准：改善研究数据管理实践的倡议。*Angew. Chem. Int. Ed Engl.*. 61, e202203038 (2022,12) 5. Bicarregui, J., Coles, S., Matthews, B., Frey, J., Montanari, B., Bunakov, V. & Knight, N. 连接基础设施：英国物理科学数据基础设施（PSDI）。 *Proceedings Of The Conference On Research Data Infrastructure* . 1 339 6 2023( ), https://www.tib-op.org/ojs/index.php/CoRDI/article/view/ 。6. Steinbeck, C., Koepler, O., Bach, F., Herres-Pawlis, S., Jung, N., Liermann, J., Neumann, S., Razum, M., Baldauf, C., Biedermann, F., Bocklitz, T., Boehm, F., Broda, F., Czodrowski, P., Engel, T., Hicks, M., Kast, S., Kettner, C., Koch, W., Lanza, G., Link, A., Mata, R., Nagel, W., Porzel, A., Schl¨orer, N., Schulze, T., Weinig, H., Wenzel, W., Wessjohann, L. & Wulle, S. NFDI4Chem——迈向德国化学国家研究数据基础设施。 *Research Ideas And Outcomes* **6** 55852 2020,6 . pp. e ( ) 7. Horsch, M., Petrenko, T., Kushnarenko, V., Schembera, B., Wentzel, B., Behr, A., Kockmann, N., Schimmler, S. & B¨onisch, T. 催化研究数据基础设施的互操作性与架构需求分析及元数据标准化。 *Data Analytics And Management In Data Intensive Domains*. pp. 166-177 (2022) 8. 欧盟委员会 / 互操作欧洲。 *DCAT Application Profile for data portals in Europe* 26 2015 。发布于：10月。https: // interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe 访问日期：2025年9月26日。9. Heller, S., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI，即 IUPAC 国际化学标识符。 *J. Cheminf.*. 7 pp. 23 (2015)，https://doi.org/10.1186/s13321-015-0068-4

10. Weininger, D. SMILES，一种化学语言与信息系统。1. 方法学与编码规则简介。 *J. Chem. Inf. Comput. Sci.*. 28, 31-36 (1988 005 11 10 1021 00057), https://doi.org/ 。/ci a . 11. Lebo, T., Sahoo, S. & McGuinness, D. PROV-O：PROV本体。（W3C,2013,4），https://www.w3.org/TR/2013/REC-prov-o-20130430/ 12. EMBL 欧洲生物信息学研究所：CHEBI:35701—2-羟基苯甲酸。在线获取：https://www.ebi.ac.uk/chebi/searchId.do?chebiId=35701 访问日期：2025年9月26日。13. Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C. & Dumontier, M. 化学信息本体：生物语义网上化学数据的来源信息与消歧。 *PLOS ONE*. 6, 1-13 (2011,10)，https://doi.org/10 0025513 14 1371. /journal.pone. . 化学方法本体（CHMO）化学方法本体（CHMO）。（FAIRsharing,2020），https://doi.org/10.25504/FAIRsharing.9j4wh2，访问日期：2025年9月26日。15. Moxon, S., Solbrig, H., Unni, D., Jiao, D., Bruskiewich, R., Balhoff, J., Vaidya, G., Duncan, W., Hegde, H., Miller, M., Brush, M., Harris, N., Haendel, M., Mungall, C.:《关联数据建模语言（LinkML）：以机器可读语义为基础的通用数据建模框架》，第148-151页。2021年国际生物医学本体会议论文集，博尔扎诺（2021）。https://ceur-ws.org/Vol-3073/paper24.pdf 16. Group, D. 用于研究数据及其他研究成果发布与引用的 DataCite 元数据模式。（DataCite e.V.,2024），https://doi.org/10.14454/mzv1-5b55，访问日期：2025年9月26日。17. 都柏林核心元数据倡议（DCMI）《都柏林核心元数据元素集》，版本1.1。（2008），https://www.dublincore.org/specifications/dublin-core/dces/，访问日期：2025年9月26日。18. 美国国会图书馆，METS 编辑委员会《元数据编码与传输标准（METS）》，版本1.13。（2016），https://www.loc.gov/standards/mets/，访问日期：2025年9月26日。19. Cox, S., Perego, A., Browning, D., Beltran, A., Winstanley, P. & Albertoni, R. 数据目录词汇表（DCAT）——版本3。（W3C,2024,8），https://www.w3.org/TR/2024/REC-vocab-dcat-3-20240822/ 20. Cox, S., Lefran¸cois, M., Janowicz, K., Haller, A., Phuoc, D. & Taylor, K. 语义传感器网络本体。（W3C,2017,10），https://www.w3.org/TR/2017/REC- vocab-ssn-20171019/ 21. Gregory, A., Bell, D., Brickley, D., Buttigieg, P., Cox, S., Edwards, M., Doug, F., Gonzalez Morales, L., Heus, P., Hodson, S., Kanjala, C., Le Franc, Y., Maxwell, L., Molloy, L., Richard, S., Rizzolo, F., Winstanley, P., Wyborn, L. & Burton, A. WorldFAIR（D2.3）跨域互操作性框架（CDIF）（汇总学科与跨学科研究领域建议的报告）。（Zenodo,2024,5 11236871 22 10 5281），https://doi.org/ 。/zenodo. . Guha, R., Brickley, D. & Macbeth, S. Schema.org：网络结构化数据的演进。 *Commun. ACM*. 59, 44-51 (2016) 23. Magagna, B., Schindler, S., Stoica, M., Moncoiffe, G., Devaraju, A., Pamment Haller, A. I-ADOPT 框架本体，https://w3 3 24 1 0 3 3 id.org/iadopt/ont/ . . . W3C 数据目录词汇表（DCAT）——版本 .（2023），https://www.w3 2023 25 3 3 4 .org/TR/vocab-dcat- /，W3C 建议，4月。欧盟委员会 / 互操作欧洲《DCAT 欧洲数据门户应用规范实施指南——第1版》。（2016），https://interoperable-

europe.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-implementation-guidelines/release-1，访问日期：2025年9月26日。26. Zeginis, D., Kalampokis, E., Palma, R., Atkinson, R. & Tarabanis, K. 面向精准农业与畜牧养殖领域的数据集成与开发的语义元模型。 *Semantic Web*. 15, 1165-1193 (2024), https://journals.sagepub.com/doi/abs/ 10 3233 233156 . /SW- . Klımek, J., Cock, J., 27 Escriu, J., Fragkou, P., Perego, A., Robberecht, J., Schiltz, A. & Nuffelen, B. GeoDCAT-AP 3.0.0：面向欧洲数据门户的 DCAT 应用配置文件的地理空间扩展。(https://semiceu.github.io/GeoDCAT- AP/930/releases/3.0.0/,2024)，由 SEMIC 在欧盟委员会互操作性计划下发布。访问日期：2025年9月26日。28. Sofou, N. & Dragan, A. StatDCAT-AP——用于描述统计数据集的 DCAT 应用配置文件，版本 1.0. 1. *European Union ISA2*. (2019), https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/statdcat-application-profile-data-portals-europe 29. SEMIC / 欧盟 MLDCAT-AP 2.0.0 — 机器学习 DCAT-AP。(2024), https://semiceu.github.io/MLDCAT-AP/releases/2.0.0/，访问日期：2025年9月26日。30. HealthDCAT-AP。(FAIRsharing,2024), https://doi.org/10.25504/FAIRsharing.b2399f ,

访问日期：2025年9月26日。31. 欧盟委员会 / 互操作欧洲应用配置文件：应用配置文件是什么，以及如何正确地对其建模和复用？通过 DCAT-AP 示例进行考察。（2024），https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/application-profiles-what-are-they-and-how-model-and-reuse-them-properly-look-through-dcat-ap，访问日期：2025年9月26日。32. Brickley, D. & Miller, L. FOAF词汇规范 0.97。（FOAF项目，2010,1），http://xmlns.com/foaf/spec/20100101.html，访问日期：2025年9月26日。33. Groth, P. & Moreau, L. PROV概述。（W3C，2013,4），https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/ 34. QUDT – 量、单位、量纲和类型。（FAIRsharing，2024），https://doi.org/10.25504/FAIRsharing.d3pqw7，访问日期：2025年9月26日。35. Arp, R., Smith, B., Spear, A. *Building Ontologies with Basic Formal Ontology*；麻省理工学院：美国马萨诸塞州剑桥，2015。36. 信息人工物本体 信息人工物本体（IAO）。（OBO Foundry，2025），https://doi.org/10.25504/FAIRsharing.gq1xtx，访问日期：2025年9月26日。37. Sansone, S., Rocca-Serra, P., Field, D. & Al. 生物医学调查本体（OBI）。（FAIRsharing，2014），https://doi.org/10.25504/FAIRsharing.284e1z，访问日期：2025年9月26日。38. Prudhomme, T., De Colle, G., Liebers, A., Sculley, A., Xie, P., Cohen, S. & Bever-ley, J. 将溯源本体映射至基本形式本体的语义方法。 *Sci. Data*。12页，第282号（2025），https://doi.org/10.1038/s41597-025-04580-1，访问日期：2025年9月26日。39. Batchelor, C. 化学反应本体（RXNO）。2012。在线 उपलब्ध：https: // github.com/rsc-ontologies/rxno，访问日期：2025年9月26日。40. 开放知识基金会：CKAN——开源数据管理系统。https://ckan.org/ 41. FAIRsharing.org：MARGARITAS；用于报告小分子液态核磁共振实验的 MIChI 标准，

https://doi.org/10.25504/FAIRsharing.c29400，最后编辑时间：2025年1月24日，星期五，14:03；最后访问时间：2025年9月27日，星期六，17:26；最后审阅时间：2025年1月23日，星期四，9:38。42. FAIRCORE4EOSC 元数据模式与交叉对照注册表（MSCR）。（2025），https://faircore4eosc.eu/eosc-core-components/metadata-schema-and-crosswalk-registry-mscr，访问日期：2025-09-27