FAIR 中的“I”：将原则上的互操作性转化为实践中的数据互操作

Evan Morris[1], Gaurav Vaidya[1], Phil Owen[1], Jason Reilly[1], Karamarie Fecho[2], Patrick Wang[3], Yaphet Kebede[1], E. Kathleen Carter[1], Chris Bizon[1]

[1] 北卡罗来纳大学教堂山分校 Renaissance Computing Institute，美国北卡罗来纳州，邮编 27599[2] Copperline Professional Solutions, LLC，美国北卡罗来纳州皮茨伯勒，邮编 27312[3] 杜克大学电气与计算机工程系，美国北卡罗来纳州达勒

摘要BSTRACT

FAIR（可发现、可访问、可互操作和可重用）数据原则 [1] 通过鼓励使用持久标识符、标准化词汇表和形式化元数据结构，促进科学数据的互操作性。许多资源采用符合 FAIR 原则且注释完善的词汇表创建，但这些资源构成的整体生态系统在实践中往往无法有效实现互操作。这一持续存在的挑战主要源于这些资源所使用的标识符模式和数据模型存在差异。我们创建了两种工具，以弥合原则上的互操作性与实践中的数据互操作之间的鸿沟。**Babel**通过生成经过整理的标识符映射集合，创建等价标识符的团，并通过高性能应用程序接口提供这些映射，从而解决多种标识符方案并存的问题。**ORION** 通过摄取知识库并将其转换为通用的社区管理的数据模型，解决多种数据模型并存的问题。本文介绍 Babel 和 ORION，并展示它们支持数据互操作的能力。通过应用 Babel 和 ORION 构建的、完全可互操作的知识库集合可在 https://robokop.renci.org 下载和使用。

1 I引言

当代研究日益受到大型复杂数据集可用性的影响。在生物医学科学领域，这一点尤为明显：高通量技术、电子健康记录、群体规模生物样本库和分子分析等领域的进展，带来了前所未有的数据规模与多样性。这些资源为发现带来了非凡机遇，使研究人员能够提出更具雄心的问题，并以新的分析能力应对长期存在的难题。然而，这些数据的科学价值受到集成和再利用方面挑战的限制。在许多生物医学领域，研究成果经过整理后以知识库（KB）的形式呈现。这些知识库是结构化的代码库，以正式的机器可读框架表示生物医学实体、它们之间的关系及相关元数据。由不同研究团队、机构或联盟在不同背景下生成的知识库，通常采用不同的标识符系统、数据模型和访问机制。若不进行谨慎协调，这些差异便会阻碍生物医学知识的集成、比较以及统一表示的构建。随着公开可用的生物医学知识库数量不断增加、类型日益多样，缺乏可扩展的集成机制，正日益限制这些资源的科学价值。

FAIR（可发现、可访问、可互操作和可重用）数据原则 [1] 已成为促进科学知识库重用的广泛认可的框架。FAIR 鼓励使用持久标识符、标准化词汇表和正式元数据，以提升可访问性和互操作性。尽管在可发现性和可访问性方面取得了显著进展，但在促进互操作性和可重用性方面，实际挑战依然存在[2,3]。要实现真正的互操作性，不仅需要采用共享标准，还需要建立机制来协调不同知识库之间的语义和结构差异。缺少此类机制，仍在持续制约大规模集成。

知识库集成与互操作面临的两项关键挑战是标识符模式各异以及数据模型存在差异。知识库通常使用不同的标识符或命名规范来表示同一底层生物学概念，例如基因、疾病或化学物质。要实现有效集成，用户和计算系统都必须能够识别不同名称或标识符是否指向同一实体。除此之外，数据模型的差异（例如与各实体关联的信息类型，以及用于编码这些信息的结构）也带来了额外的复杂性。虽然标识符标准化已被广泛视为一项核心挑战，[4]，协调不同的数据模型同样是大规模集成的一大障碍。知识库在其模式、本体和粒度层级方面存在很大差异，这些不一致可能会掩盖有意义的关系，并阻碍统一框架的构建。

为应对这些挑战，知识库创建者通常会将其系统设计为遵循 FAIR 原则，以实现互操作性。这些原则鼓励采取包括以下内容在内的做法：

I1。（元）数据使用一种正式、可访问、共享

例如，I2 可以通过采用成熟公共资源中的基因标识符来满足，例如 EntrezGene [4] 或 Ensembl [5], ，这些资源使用户能够明确识别知识库中的基因。然而，目前尚无普遍认可的标识符标准。此外，不同知识库往往根据各自的优先事项，对同一概念空间进行划分。一个知识库可能将水归类为环境特征，另一个则将其归类为药品，还有一个可能将其归类为生化代谢物。因此，许多知识库虽然使用符合 FAIR 原则且注释完善的词汇表构建，但这些资源构成的整体生态系统在实践中仍无法实现有效互操作。

要将原则上的互操作性扩展为实践中的互操作，需要能够灵活应对相互竞争的标准的工具。在许多情况下，知识库之间的差异反映了优先事项或基本假设的合理差异。然而，其他差异仅仅源于模型的独立开发，这带来了本可避免的复杂性，阻碍集成，却没有增加实质价值。解决这种差异对于真正实现大规模、多来源的知识库集成至关重要。

为此，我们开发了一套工具，旨在跨越一系列知识库<styles id='1'>实现实际互操作；这些知识库各自符合 FAIR 要求 [5,6] ，但集体整合时仍然困难重重。这些工具直接应对数据集成的核心障碍，即需要：**(1)** 解析实体等价关系，以及 **(2)** 协调数据模型之间的差异。具体而言，Babel 及其相关服务节点规范化器和名称解析器通过对标识符进行稳健的标准化来应对第一个挑战。网络摄取与输出操作流程（ORION）则通过使用社区维护的顶层本体和数据模型 Biolink 模型 [7]，将多样化的知识库转换为<styles id='4'>知识图谱（知识图谱），从而应对第二个挑战。这些组件共同构成了一个协调一致的框架，可自动化执行原本需要手动、特定于来源且难以在大规模场景下推广的过程。该框架支持摄取异构知识库，将其映射到统一的标识符体系中，在共享数据模型内进行结构化，并通过一组一致的接口访问。通过在上述各个层面落实 FAIR 原则，我们的框架促进了集成式高质量知识图谱的构建，为严谨的科学发现提供支持。

2 结果

2.1 Babel：标识符等价性的人工整理框架

Babel 是一组经过人工整理的处理流程，用于下载、筛选和合并来自多个来源的映射，覆盖 23 个广义的生物实体类别，范围从基因和蛋白质到过程和通路。Babel 系统地处理来自广泛的生物医学数据库、本体、叙词表和受控词表的等价信息。它应用经过整理的规则，将这些映射综合为结构化的组或“团”，每个团代表被认为指向同一底层概念的标识符。在每个团中，按照 Biolink 模型 [7] 定义的来源优先顺序，选定一个标识符作为“团领导标识符”或代表性标识符（见下文“方法”部分“标识符等价：Babel”一节）。下游来源可以使用团领导标识符来代表该概念，从而对标识符进行<styles id='2'>标准化。

鉴于生物医学词汇表的规模和多样性，逐一整理每种等价关系并不可行；因此，我们优先汇集来自可信来源、且有充分依据的等价断言集合。通过人工整理流程，我们决定接受哪些映射来源中的哪些等价关系分组，从而建立一个受控框架，以解决不同领域间的标识符等价问题。此流程明确聚焦于标识符对齐，并不试图协调源本体中蕴含的完整逻辑结构或关系层级。这一取舍体现了有意优先考虑可扩展的跨领域互操作性，而非追求详尽的语义粒度；大多数源数据库本身也并未维护这种粒度。Babel 将等价标识符团作为统一概念发布，而非一组单独映射，从而提供供下游应用使用的共享词汇表，促进实际互操作。

虽然实现完美的语义等价是一个合理的目标，但知识的实际集成往往需要采取更务实的方法。在某些情况下，团的使用者需要合并标识符，而其他使用者则倾向于将它们保留为不同标识符。我们将这一过程称为合并：一种受控机制，用于将密切相关但并非严格相同的概念归为一组。例如，基因蛋白质合并归并提供交叉引用，以便将代表蛋白质编码基因的团与代表这些基因所编码蛋白质的团合并，从而简化基因数据库与蛋白质数据库之间的协调。当区分基因及其蛋白质产物几乎无法带来额外洞见，却会增加数据库维护和查询成本时，这种方法尤其有用，例如研究它们在某条生化通路中的共同作用时。类似地，药物化学物质合并归并会将代表仅含一种活性成分的药物的团，与代表该活性成分的化学物质合并。Babel

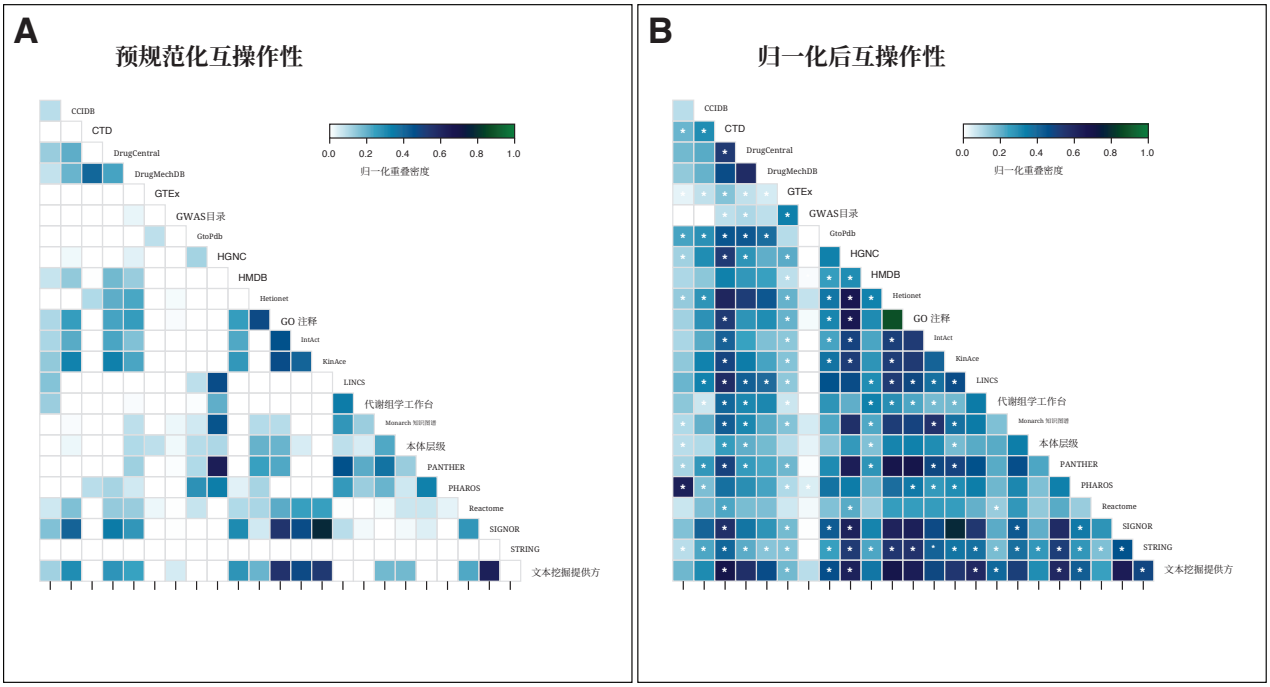

图1。标识符规范化对跨源互操作性的影响 知识库重叠热图显示了源对之间经规范化的共现密度（**A**）规范化前的重叠矩阵包含 138 个连接。（**B**）规范化后的重叠矩阵包含 267 个连接。每个单元格表示源 A 和源 B 之间经缩放的重叠密度。规范化产生了 129 个新的源对连接，使总重叠增加了 93.5 %；这里的总重叠定义为至少共享一个规范化实体的唯一源对数量。 面板 **B** 中的白色星号（*）表示规范化前不存在的新连接。

流水线会根据定义的合并策略生成合并列表，其中列出在相应策略下应视为等价的团代表标识符。

下游系统（包括节点规范化器，见下文“Babel 在标识符解析中的应用”）可以利用这些信息，在查询时根据应用需求调整或扩展等价类。该能力确保框架同时支持严格和宽松的等价性定义，从而在分析精度与大规模实际数据集成之间取得平衡。

除了合并之外，Babel 还能为团补充信息，以提高其分析实用性。主要的增强步骤是为每个团分配一个 Biolink 类。对于大多数处理流程而言，这一步很直接，但有时仍需进一步消歧，尤其是在统一医学语言系统和化学数据集中；在这些数据集中，等价标识符可能关联不同概念类别的实体，例如代谢物和一般化学物质。Ubergraph 根据每个本体类的子类数量和存在关系，计算其归一化信息含量值，并将数值范围设为 0（表示子类众多的高度一般化类）至 100 （表示没有子类的高度特异类）。Babel 通过为每个团赋予其成员标识符中观测到的最低值，纳入这些度量，从而提供语义一般性的统一指标。基因和蛋白质团还会进一步分别标注来自 NCBI Gene 数据库和 UniProtKB 的分类学信息，确保集成过程中保留物种语境。这些增强使团的用途超越了标识符解析，还能支持依赖特异性度量和生物学语境的分析。

Babel 的大规模等价性集成将标识符整合为结构化的团，从而实现协调一致的跨源引用，为下游知识图谱构建和分析提供基础。在全部十四条处理流程中，Babel 构建了超过 4.23 亿个不同的团，其团代表标识符来自 44 个词汇表（由 CURIE 前缀区分），其中包括来自 66 个词汇表的近 6 亿个唯一 CURIE （补充表 1 ）。

Babel 对跨资源连通性的影响如**图1**所示。标准化前后的热图展示了一组常用知识库之间共享标识符的密度；标准化使观测到重叠的来源数量几乎翻了一番（从138个增至267个）。为比较不同规模知识库之间的重叠强度，我们将归一化重叠密度定义为：

$$density(A,B) = overlap_count(A,B)/\sqrt{|A| \cdot |B|} \qquad (1)$$

其中， $overlap_count(A,B)$表示来源 A 与B之间共享标识符的总数，而$|A|$ 和$|B|$ 分别表示

各自唯一标识符的数量。平方根缩放采用基于几何平均数的调整方式，以控制来源规模的影响，从而得出不受规模影响的跨来源实体重叠度量。标准化凸显了通过 Babel 进行标识符协调，如何大幅提升知识库之间的语义连通性，并揭示了此前彼此孤立的来源之间新的互操作性连接点。如**图1 B面板**所示，白色星号标记了标识符规范化后新出现的跨来源重叠。除了唯一重叠数量有所增加外，平均节点度也从 11.5 上升至 22.25，表明每种资源都与更多不同来源建立了连接。总之，这些结果表明，通过 Babel 系统地解析等价关系，既能建立新的连接，也能强化已有连接，从而增强生物医学资源之间的互操作性。

2.2 Babel 在标识符解析中的应用

Babel 大规模的等价团通过一对公共服务提供，这些服务旨在满足知识库集成中的两项基本需求：1）将标识符规范化为定义明确的标识符团；2）将自由文本字符串解析为标准化标识符，并提供足够的信息以区分不同选项。虽然 Babel 的团可以导入通用数据库，但生物医学数据的规模和复杂性要求知识库集成采用专用基础设施，以精准、快速地提供这些功能。为满足这一需求，我们开发了两个互补服务——节点规范化器和名称解析器，并通过使用 OpenAPI [8] 描述的 Web 服务公开 Babel 的输出。

节点规范化器。 节点规范化器提供一个 HTTP API，可接收待标准化的标识符列表，并返回这些标识符对应的完整团，包括团领导标识符、Biolink 类型，以及信息量分数等其他属性。该服务可将异构标识符解析为规范团领导标识符，并支持可选的合并，从而消除标准化在大规模工作流中长期存在的瓶颈。节点规范化器通过在内存中维护等价信息并通过 API 将其公开，支持大规模实时标准化。这一功能对于 NCATS Translator [9]，等系统至关重要，因为这类系统的查询会遍历多个异构知识库。对 2025年6月一周内提交到某个节点规范化器实例的 19,043 条查询进行分析发现，查询响应时间介于 0.83 毫秒和 10.89 秒之间，平均每个请求的响应时间为 352.33 毫秒。该数据集中的每条查询最多包含 2,769 个待标准化的 CURIE，平均每个 CURIE 耗时 5.68 毫秒。其 GitHub 仓库中提供了一个展示节点规范化器用法示例的 Jupyter Notebook。

名称解析器。名称解析器提供一个 HTTP API，接受文本字符串，并搜索首选标签或同义词中包含该字符串的团。该服务还支持解析文本输入，通过排序后的搜索结果将标签和同义词映射到规范标识符。这使自由文本术语能够无缝集成到结构化知识图谱查询中。例如，搜索“aspirin”会返回一组按相关性排序的概念，其中可能包括 CHEBI:15365、PUBCHEM.COMPOUND:2244 以及相关药物制剂，并提供足够的上下文信息以便区分这些概念。同样，搜索“ADA”可以区分基因 腺苷脱氨酶（NCBIGene:100）、疾病腺苷脱氨酶缺乏症（MONDO:0011130）和兰科属名 *Ada Lindl*（NCBITaxon:125078，现归类于 *Brassia*）。名称解析器提供 Biolink 类型、首选标签和同义词等上下文元数据，使自动化系统和用户都能够消歧术语，并将其映射到可计算标识符。其 GitHub 仓库中提供了一个展示名称解析器用法示例的 Jupyter Notebook。

Node Normalizer 和 Name Resolver 共同使 Babel 协调后的等价集合能够应用于自动化流水线和面向用户的探索环境。通过以 Babel 团为标识符解析奠定基础，这些服务将标识符协调转化为可运行的基础设施，从而在实践中实现生物医学数据资源之间真正的互操作性。

2.3 模型等价性：ORION

即使知识库使用共享的标识符命名空间，它们也很少遵循相同的语义模型或数据模式。生物实体之间的关系通过各种术语体系和本体来描述，而这些术语体系和本体在粒度和特异性方面各不相同。数据以多种格式提供，而这些格式往往彼此不兼容，例如通过未标准化的基于 Web 的 API 提供，或以 JSON、XML 或 CSV 文件的形式编码。此外，许多重要的生物学知识来源所提供的数据格式完全缺乏形式化数据模型，例如统计分析的结果。

网络摄取与输出操作流程（ORION）是一种知识图谱流水线，旨在应对这些挑战，具体方式是

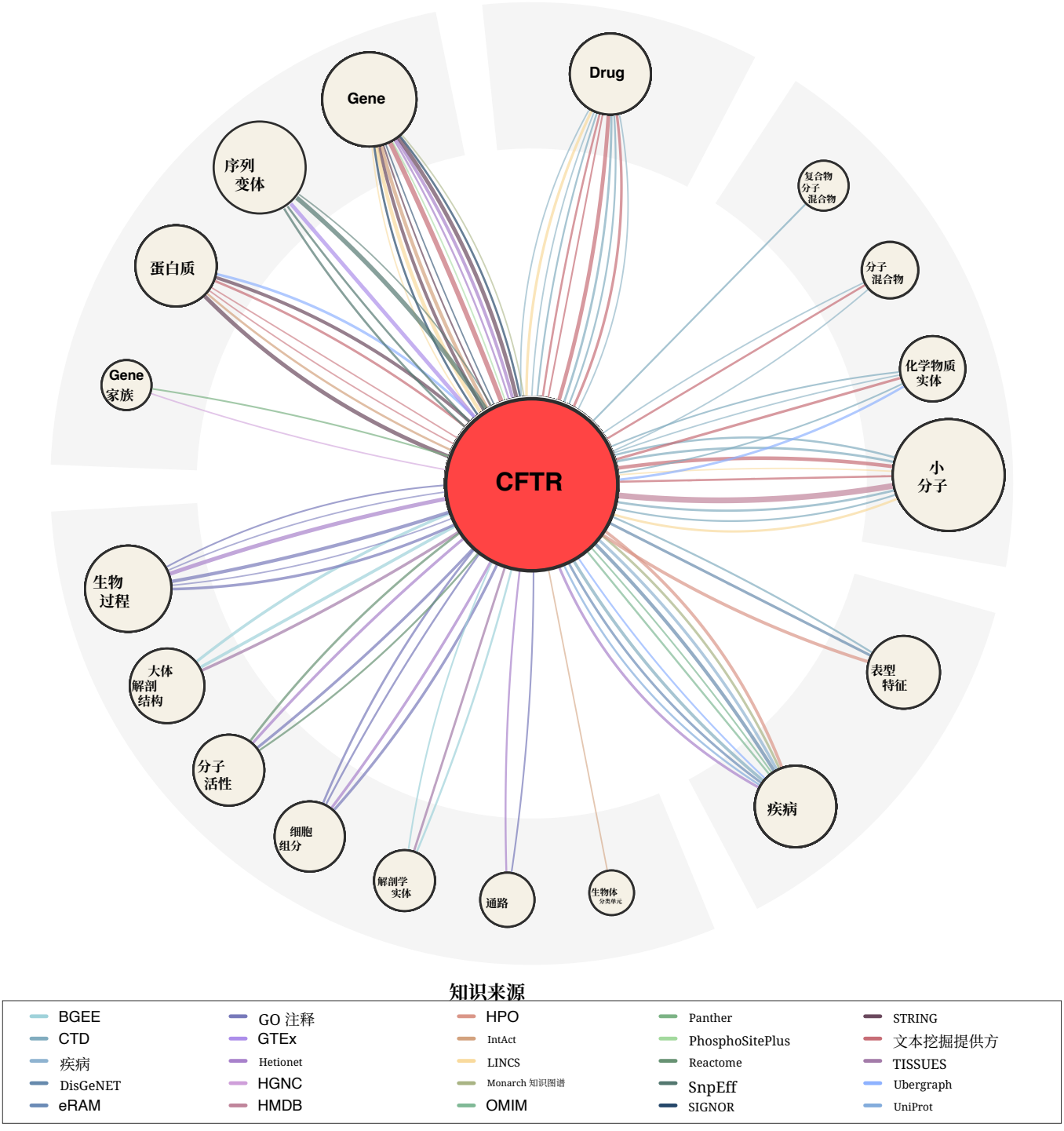

图 2. ORION 生成的 ROBOKOP 知识图谱中与 CFTR 相关组件的网络表示。在此示例中，ORION 集成的 25 个来源中的知识通过 35 种不同的关系类型，将*CFTR*与 18 种不同的生物节点类型关联起来。所有实体和边均采用 Biolink 模型模式建模，并使用节点规范化器进行规范化，从而能够在不同知识库之间无缝探索。*CFTR*显示为中心节点，周围的节点则表示响应单次数据库查询而检索到的不同生物实体类型。节点按更高层级的类别分组，包括化学物质、药物、基因、蛋白质、生物学背景、疾病和表型。此处的可视化中，节点大小反映了与 CFTR 相连的该类型实体数量，并使用对数转换后的计数。边将 CFTR 与每种实体类型相连，并按关系来源着色。边宽表示 CFTR 与每个来源–组件对之间的边数，同样经过对数转换。

使用共享的语义框架 Biolink 模型，将不同的生物数据表示转换为模块化且可互操作的知识图谱。

ORION 流水线已应用于 40 多个公共知识库。它将异构数据模式中表达的关系转换为语义三元组，即严格遵循 Biolink 模型定义的主语-谓语-宾语关系。ORION 使用节点规范化器，将知识库使用的标识符规范为共享且一致的标识符命名空间，并为每个实体分配 Biolink 模型类别，例如基因或疾病。因此，使用 ORION 生成的知识图谱具有完全的互操作性，因为它们在标识符用法和语义结构上均保持一致，并且可以使用一致且可预测的模型轻松集成或查询，从而支持各种下游应用。这些模块化知识图谱既可独立使用，也可通过 ORION 组合起来构建集成网络，在保留来源级别区分的同时实现无缝互操作。

为使知识图谱能够立即投入使用，ORION 会生成可直接部署的输出格式。每个知识图谱都以 KGX 格式 [10] 序列化，并附有元数据文件，记录版本信息、来源信息、汇总统计信息以及有关所执行转换的具体信息。除 KGX 文件外，ORION 还可选择生成 Neo4j 数据库转储文件，这些文件可导入为独立的 Neo4j 实例并供查询。使用 ORION 生成的知识图谱可借助具备模型感知能力的知识图谱查询工具进行部署，例如 Plater [11], 。Plater 是我们开发的一项 Web 服务，可通过应用程序接口提供 Biolink 模型知识图谱，支持多种查询类型，包括 TRAPI [12], 、Cypher，以及用于从知识图谱中提取信息的自定义端点。由于采用了一致的数据模型，Plater 等工具可以通过统一的过程和应用程序接口部署并开放 ORION 生成的知识图谱。此外，采用标准化数据模型还能支持可复用的分析流水线，而无需额外转换或重新格式化。

ORION 的一个应用示例是面向知识路径的生物医学对象推理（ROBOKOP）知识图谱，这是一个完全由 ORION 生成的组件构建而成的大规模生物医学知识图谱。ROBOKOP 知识图谱包含约 1,000 万个节点和 1.3 亿条边，汇集了 40 多个生物数据源的数据。由于纳入 ROBOKOP 知识图谱的每个来源都经过语义对齐和标识符规范化，因此无需定制协调即可使用整个图谱，支持从化学物质-基因-疾病分析到跨生物学尺度的语义子图探索等多种应用。该图谱以 KGX 格式和 Neo4j 数据库转储文件的形式提供下载，并附有元数据。这些资源使研究人员能够直接使用 ROBOKOP 知识图谱，或以其为基础创建特定用途图。

为了说明互操作如何助力发现，不妨设想一位研究人员正在探索一个感兴趣的基因，例如 *CFTR*，以考察它与表型、化学物质、分子通路及其他生物特征之间的关系。借助 ROBOKOP 知识图谱，研究人员可以通过一次数据库查询检索所有相关关系。如**图 2**所示，查询结果包含 *CFTR* 与 18 种不同类型生物实体之间的关联，这些实体来自 25 个不同的知识库。

在第二个使用案例中，研究人员希望汇编一个数据集，其中包含直接关联化学物质与疾病或表型的关系，以及通过分子靶标建立的此类关系。这项任务可能涉及集成比较毒理基因组学数据库（CTD）中的化学物质-疾病关联 [13], 、BindingDB 中的化学物质-靶标相互作用[14], , 以及 Monarch Initiative 中的基因-表型关系[15]。尽管这些来源均可公开获取，且名义上具有互操作性，但在实践中将它们集成起来仍面临重大挑战。CTD 使用 MeSH 标识符表示化学物质，并使用 MeSH 或 OMIM 术语表示疾病，同时以自由文本注释标示关系类型。BindingDB 使用 InChIKey 或其他化学物质标识符，并以 UniProt 登录号报告蛋白质靶标；关系则由数据结构隐式推断。与此不同，Monarch Initiative 提供语义显式且符合 Biolink 模型的基因-表型关联，使用 HGNC 和 HPO 标识符。

即使是经验丰富的信息学专家，要合并这些数据集，通常也需要熟悉每个来源的数据模式、利用外部对照表，并编写自定义脚本。对于缺乏计算专业知识的研究人员而言，这些障碍可能难以逾越。相比之下，ORION 让用户只需通过一条 Python 命令或命令行界面，指定所需的知识库，即可自动生成标识符一致、采用共享数据模型的连贯图。生成的统一图可以载入数据库、进行可视化，或借助 LinkML 工具导出为 RDF 等其他格式。

图 3 展示了这种集成的效果。左侧面板显示了来自 BindingDB、CTD 和 Monarch Initiative 的源特定网络，每个网络都涵盖了核心 Biolink 实体类型（如基因、蛋白质、化学物质和疾病）之间不同的关系子集。右侧面板展示了 ORION 生成的集成图，其中的实体标识符已使用节点规范化器进行规范化，关系也已在语义上与 Biolink 模型对齐。合并后的网络揭示了新的跨源连接和共享节点，展现了如何通过跨模式协调，将原本彼此孤立的知识库转变为连贯、可互操作的结构，从而支持统一的探索与分析。

通过对齐标识符、协调关系语义并生成模块化且可互操作的成果，ORION 降低了生物医学生态系统中集成和二次利用的技术门槛。ORION 生成的知识图谱遵循 Biolink 模型并采用规范化标识符，从而确保不同来源之间的结构与语义一致性。这种

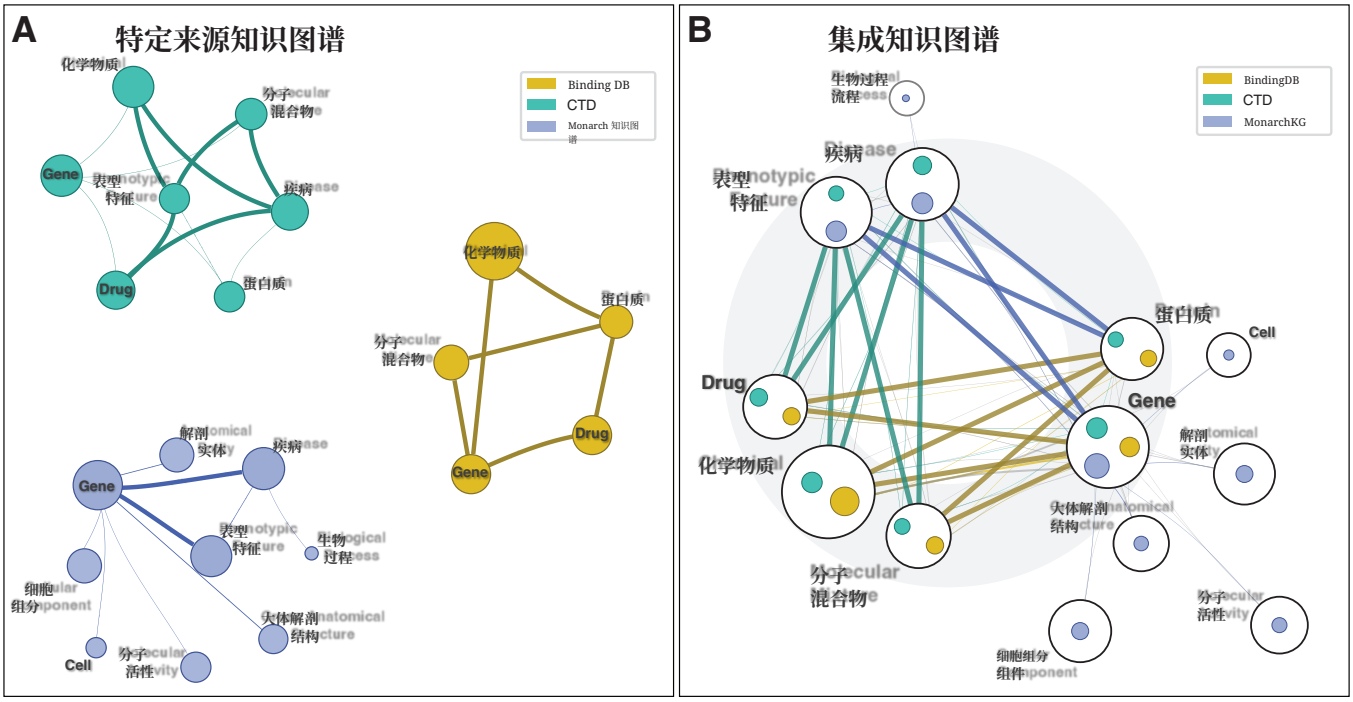

图 3。ORION 生成的特定来源知识图谱与集成知识图谱的比较。（**A**）网络表示展示了三个生物学知识库如何连接关键的 Biolink 实体类型：BindingDB（金色）、CTD（青绿色）和 Monarch 知识图谱（蓝色），这些类型包括基因、蛋白质、疾病和化学物质。左侧面板展示了从各资源中提取的源内网络，突显其覆盖范围和关系结构的差异。（**B**）右侧面板展示了 ORION 生成的集成图，其中的实体和关系已使用 Biolink 模型进行规范化和语义对齐。共享节点体现了不同来源之间的互操作性，而边的密度则反映了协调过程中新建立的跨源连接。

一致性使用户能够专注于提出具有生物学意义的问题，并深入探索跨学科知识，而不是解决知识库之间的不兼容问题。因此，ORION 推动了更广泛的目标：将异构数据转化为标准化、可互操作的资源，促进透明且高效的发现，从而使生物医学知识真正符合 FAIR 原则。

3 方法

3.1 标识符等价：Babel

生物医学数据源集成面临的一项基础性挑战，在于解决实体等价问题。同一底层生物学概念可能由不同知识库使用不同的标识符来表示，这反映了各知识库特定的优先事项和使用场景。例如，在异构来源之间进行集成时，需要理解生物学兴趣化学实体（ChEBI） [16] 标识符“15377”和 PubChem 标识符 “962” 均指向同一概念 —— 水。在本文中，标识符以紧凑统一资源标识符（CURIE）的形式表示，例如 CHEBI:15377 和 PUBCHEM.COMPOUND:962。

现有资源通常只能提供部分解决方案。为促进集成，许多数据库和本体都包含交叉引用，将其标识符与其他资源所使用的标识符关联起来。例如，ChEBI 中关于水的记录包含指向许多其他数据库中水这一化学物质概念的交叉引用，包括 PubChem [17] (PUBCHEM.COMPOUND:962)、化学文摘社（CAS）注册库 [18] (CAS:7732-18-5)、人类代谢组数据库 [19] (HMDB:HMDB0002111) 以及京都基因与基因组百科全书 [20] (KEGG.COMPOUND:C00001)。然而，其中明显缺少 UNII 和 RxNorm 标识符，而这两者用于标识药物成分和制剂。映射可以由统一医学语言系统（UMLS） [21], TogoID [22,23], 或 EMBL-EBI 本体交叉引用服务（OxO） [24], 等集成资源综合生成，也可以通过 Mapping Commons 等项目在外部维护

[25], 这些资源使用本体映射共享简单标准 [26] 发布现有资源之间的映射。尽管如此，二者有时并不兼容，从而造成缺口和冲突，这凸显了仅依赖任何单一资源的局限性。

许多生物医学概念也存在于丰富的层级体系或部分—整体结构中。例如，美国国家医学图书馆（NLM）的医学主题词表（MeSH）词库 [27] 将水的一般概念（MeSH:D014867）与更具体的下位概念区分开来，例如碳酸水（MeSH:D061545）、饮用水（MeSH:D060766）、冰（MeSH:D007053）和蒸汽

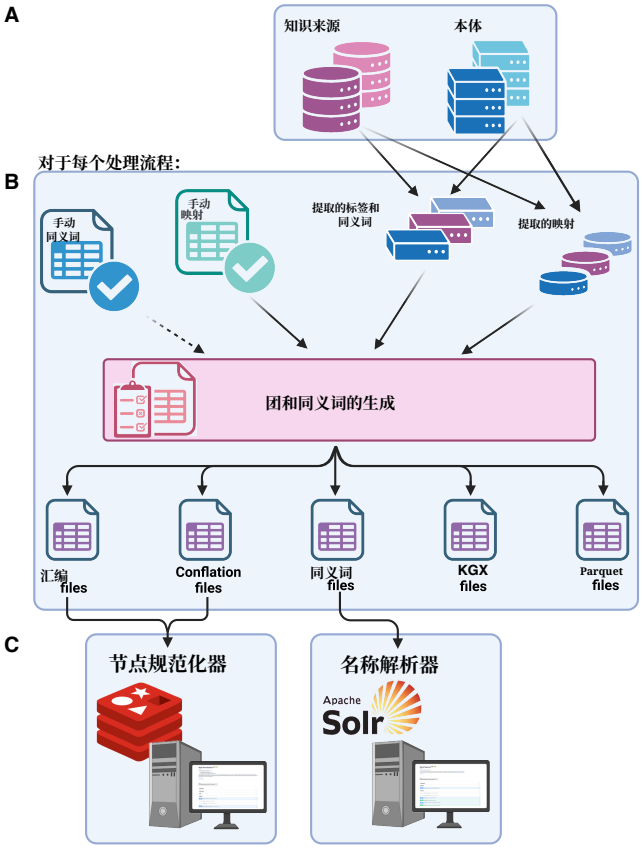

图 4。 Babel。Babel 中的整体数据流。（**A**）从知识库和本体中提取标签、同义词和映射。（**B**）十八个特定输出流水线中的每一个都会选取这些映射的某个子集，将其与人工映射和同义词相结合，并生成等价标识符的团，以及与这些标识符相关的所有同义词。这些团用于生成汇编和融合文件（供节点规范化器前端使用），以及同义词文件（供名称解析器前端使用）（**C**）。它还会以其他输出格式（KGX 和 Apache Parquet）生成团。

要实现尊重这些区别的全面集成，需要明确编码子类或部分-整体关系，以厘清每个标识符的适用范围。虽然 Ubergraph [28] 等基于图的项目已尝试纳入这些复杂情况，但大多数生物医学资源并未提供如此精确的语义。此外，对于两个标识符是否应被视为等价，不同词汇表往往意见不一，这反映了各领域视角的多样性。

Babel 是一个基于 Python 的流水线，旨在应对生物医学数据整合中普遍存在的标识符等价难题：不同资源经常使用不同的标识符或命名规范来表示同一生物学概念。该系统使用 Snakemake [29], 实现，Snakemake 将工作流构建为由相互依赖的规则组成的有向无环图，以自动从各类生物医学数据库、本体词库和受控词表中检索、处理并整合等价信息（来源见补充表 1）。Babel 并不试图协调更广泛的

（MeSH:D013227）；而其他资源则对水的所有形态仅使用一个概念（例如 CHEBI:15377）。

嵌入源本体中的逻辑层级，而是明确专注于构建一组组标识符，称为团，以表示严格等价。这种设计选择避免了简单进行本体合并可能引发的矛盾，并为需要统一标识符的下游应用提供了一致的基础。

Babel 采用单个 Snakemake 工作流构建，其中包含 18 个针对特定输出的工作流，每个工作流都由各自的 Snakemake 文件表示（**图 4**）。其中大多数工作流对应单独的 Babel 流水线，负责生成若干密切相关的生物医学类别中的团，这些类别按 Biolink 类型组织。例如，解剖学流水线会生成与四个 Biolink 类相关的团：解剖实体、宏观解剖实体、细胞组分和细胞（完整列表见补充表 2）。两个工作流会生成可选的基因蛋白质和药物-化学物质合并归并，将其作为首选标识符集合，并在启用相应合并归并时将其组合（如前文“结果”部分所述）。另有五个辅助工作流不属于任何特定的 Biolink 类型，而是提供用于下载和转换生物医学资源、生成报告或导出文件等共享规则，以及其他共用步骤。这种模块化布局有多项优势：可以独立于其他类型生成或调试各个 Biolink 类型；可以并行运行互不相关的规则；还支持在需要复杂逻辑来为每个团选择 Biolink 类型的场景中使用模块，例如化学物质的分类——区分分子混合物、小分子和其他化学实体时，可能需要检查化学式，或比较多个来源的意见。最后，所有团和同义词都会载入 DuckDB [30] 数据库，以便标记重复项或相同标签。

每条 Babel 流水线都从一组专用的 Snakemake 规则开始，用于下载并提取相关内容，并将其转换为标准化中间表示。例如，对于化学实体，ChEBI 流水线会协调从 ChEBI 数据库获取完整的数据集导出文件 [16], ，其中包括结构文件和登录号表，随后对文件进行解析，以提取标签（概念的主要文本名称）、同义词（次要文本名称）以及其他元数据（例如描述、次要标识符等）。之后，特定领域的处理流程会处理这些下载内容，提取标识符、等价映射和同义词信息，作为等价关系构建阶段的输入。

映射的提取并不包含某个来源的所有映射。这样做会导致错误不断累积，并形成包含无关条目的超大团。相反，人工策展人会从每个来源中选取一组最小映射，使该来源中的每个标识符都能与另一个来源中的标识符关联起来。像 MONDO [31] 或

UniChem [32] 这样经过精心策展的来源，可以充当映射中心，人工策展人可以选择从中导入多种类型的映射。映射策展并非在单个标识符层面进行，因为大规模开展这项工作并不切实际。相反，策展人会在整个词汇表之间做出更高层级的决策，例如决定 Babel 是否应接受 MONDO 到疾病本体 [33]的映射。补充表3概述了 Babel 整合的交叉引用范围、采用的筛选策略，以及从各来源提取的交叉引用数量。可以通过在 Babel GitHub 仓库中创建 GitHub 问题来申请添加新的标识符来源；该仓库还提供了其他有关如何为 Babel 作贡献的指南。

提取映射后，Babel 会将这些信息整合为结构化的等价团。每个团代表一组被判定为指向同一生物医学概念的标识符。在每个团内，标识符会依据 Biolink 模型 [7], 定义的策展层级进行筛选和排序；该层级为每个 Biolink 类指定有效前缀，并由 Biolink 模型工具包 [34]实现。系统会选定一个以 CURIE 表示的标识符作为该团的首选标识符，选择依据是 Biolink 模型的首选前缀顺序。同样，系统也会选定一个标签作为首选标签，其选择主要依据该团 Biolink 类型所规定的首选前缀顺序，同时结合若干质量检查进行调整（例如已知的优质标签来源，以及抑制过长标签）。这种方法可确保不同知识库中的标识符解析稳定且标准化、标签清晰，同时有意避免超出严格等价关系的更广泛逻辑蕴含。

这两个合并文件通过使用特定映射生成，用于关联基因与蛋白质（UniProt [35], 的 ID 映射下载文件，其中包含 UniProtKB 标识符到 NCBIGene 标识符的映射），以及药物与化学物质（使用 RxNorm [36] 和 UMLS 中基于 RXCUI 的映射）。对于基因-蛋白质合并归并，基因始终列在蛋白质之前。对于药物-化学物质合并归并，算法会按照自定义的 Biolink 类型顺序排列标识符，优先选择更具体的类型（例如小分子），而非更通用的类型（例如化学实体）。在每种 Biolink 类型内，标识符按信息量、团大小和 CURIE 数字后缀排序。

团和合并构建完成后，Babel 会生成多种类别的输出。团汇编以按行分隔的 JSON 文件形式写入 [37], ，每个文件表示一个团及其排序后的等价标识符列表。相应的同义词文件会记录每个团的首选标识符，以及所有已识别的文本同义词，从而支持稳健的基于名称的解析。合并输出会列出在特定且明确定义的融合策略下应视为等价的一组团代表标识符。所有输出都会加载到 DuckDB [30] 并导出为 Apache Parquet 文件 [38], ，以便执行跨索引检查，例如识别出现在多个团中的标识符，或识别首选名称相同的团，并将结果写入机器可读报告。此外，Babel 还会生成 KGX 格式的导出文件 [10,39] ，将等价关系表示为明确的节点和边结构，并遵循知识图谱规范。

3.2 Babel 在标识符解析中的应用

节点规范化器。 Babel 生成的团可以直接导入任何基于文档的数据库（如 ElasticSearch 或 MongoDB）进行查询。不过，由于标识符规范化可能成为查询时集成的关键瓶颈，我们构建了一个专用的高速系统，以确保快速完成规范化。

节点规范化器是一个高性能系统，它将内存中的 Redis 数据库与符合 OpenAPI 规范的 FastAPI [40] 应用程序相结合，用于标识符规范化。Python 脚本将 Babel 团以 JSON 格式加载到 Redis 中。存储的内容包含 Babel 团的大部分数据，但不包括同义词信息。系统创建了七个 Redis 表：(1) 标识符到其各自标签的映射；(2) 标识符到其规范化标识符（团代表标识符）的映射；(3) 团代表标识符到其 Biolink 类型的映射；(4) 团代表标识符到其他属性的映射，例如描述、信息含量值和分类单元；(5) 包含数据库元数据的表，包括每种 Biolink 类型的前缀分布；以及 (6–7) 两个表，其中包含根据基因蛋白质和药物化学物质合并归并策略应进行合并的标识符列表。

Node Normalizer 的用户可以使用任意数量的标识符进行查询，以获取它们的 Biolink 类型、等价标识符列表、首选标识符和标签，还可以选择获取团的信息量及其所属的分类单元。运行这些查询时，可以启用 GeneProtein 合并和 DrugChemical 合并，或同时启用两者。其他 API 端点还提供有关 Node Normalizer 数据库中数据的信息，例如其中包含哪个版本的 Babel 输出、支持哪些合并，以及涵盖哪些 Biolink 类型和 CURIE 前缀。其 GitHub 仓库中提供了一个展示 Node Normalizer 用法示例的 Jupyter Notebook。

名称解析器。 名称解析器的用户可以通过文本字符串进行查询，从数据库中获取可能匹配项的列表。搜索范围既包括每个团的首选名称，也包括所有已知同义词。这样，用户就可以查找与特定生物医学术语对应的生物医学标识符，既可以由用户控制（例如，为用户输入查询所用生物医学术语时提供自动补全建议），

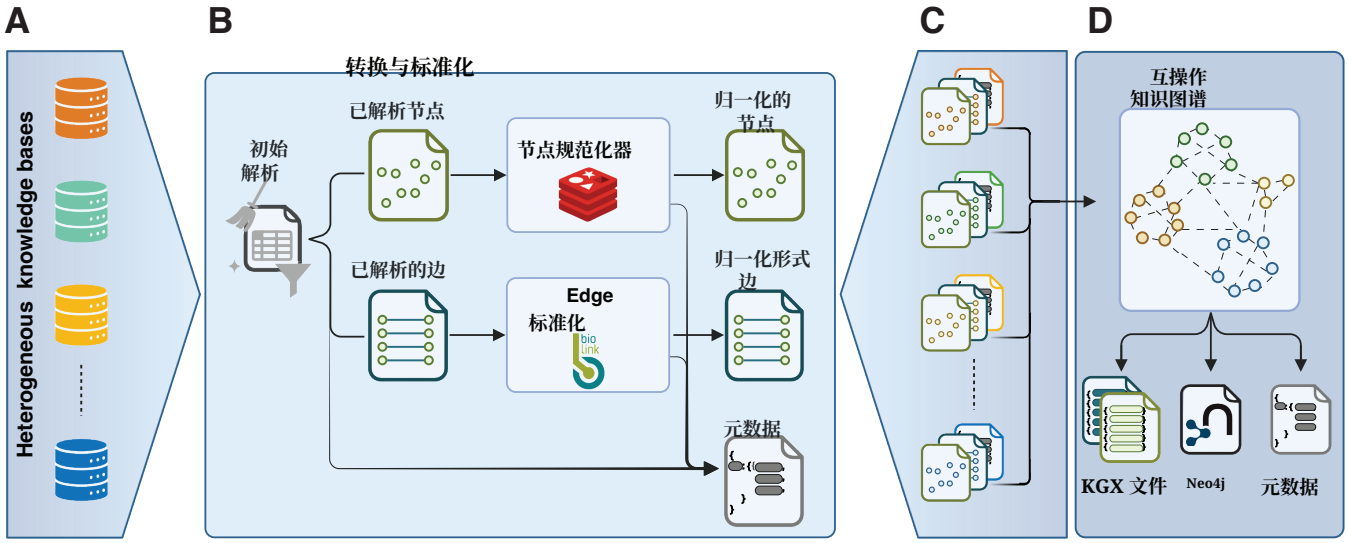

图 5. ORION 流水线。ORION 使用源特定解析器摄取异构源知识库（**A**），这些解析器获取带版本的数据发布，并将其转换为共享的中间表示（**B**）。这些经过解析的节点文件和边文件无需完全符合 Biolink 模型，但必须满足最低约束，以支持统一的下游处理。源无关规范化在所有输入之间对齐标识符和语义，通过节点规范化器或 ClinGen 等位基因注册库解析节点等价性，并使用 Biolink 模型工具包统一谓词。规范化后的图产物以 KGX JSON Lines 格式序列化，同时在整个流水线中生成记录版本信息、来源信息和质量指标的元数据。通过并集和去重，跨来源整合规范化输出，以生成互操作知识图谱（**C**），该图谱支持多个下游数据产品和访问方式（**D**）。

也可以批量处理（例如，为一系列生物医学术语执行实体链接）。各种参数可用于按 Biolink 类型、CURIE 前缀或特定分类单元集合筛选结果。高亮和调试选项还可返回有关匹配项的其他信息。第二个 API 端点可返回特定首选 CURIE 的所有已知同义词。其 GitHub 仓库中提供了一个展示名称解析器用法示例的 Jupyter Notebook。

名称解析器为 Babel 团中包含的标签和同义词提供基于 Apache Solr 的搜索引擎。该服务包括一个符合 OpenAPI 的 FastAPI 应用程序，该应用使用 Python 编写，支持单个查询和批量查询，以及同义词查询；在同义词查询中，会返回某个首选标识符的所有标签和同义词。一个轻量级摄取模块通过以 JSON 格式加载 Babel 团数据来准备 Apache Solr 数据库。为减小数据库大小，该索引仅包含首选标识符、Biolink 类型、分类群信息和同义词。等价标识符、描述和信息含量值则通过随后在节点规范化器中的查询来获取。

Solr 后端使用 Apache Lucene 以两种并行方式为标签和同义词建立索引。它使用标准的 Solr 分词器将每个同义词拆分为单个词，同时还使用 KeywordTokenizer 将每个同义词作为单个词元建立索引，从而支持不区分大小写的精确匹配。查询通过 eDisMax 解析器执行。在自动补全模式下，解析器还会识别输入字符串不完整时的匹配结果。提升参数会优先显示精确匹配而非关键词匹配，并优先显示首选标签而非同义词。

3.3 模型等价性：ORION

ORION 是一个基于 Python 的软件包，用于构建和操作互操作知识图谱（KG）。ORION 的核心功能是一条流水线，可将原始知识库中的数据转换为节点规范化的 Biolink 模型兼容知识图谱（**图 5**）。由于知识库很少遵循标准模式，因此必须针对每个知识库的特定格式和语义，使用相应的自定义代码模块和配置来处理。对于 ORION 所摄取的每个知识库，专用解析器都会检索一个带版本的数据发布，读取其原始结构，并将其转换为标准化知识图谱表示。这些初步表示无需完全符合 Biolink 模型，也无需使用标准化标识符。不过，节点标识符必须能够由节点规范化器或 ClinGen 等位基因注册库解析 [41]，，且边谓词必须与 Biolink 模型兼容，或能够借助现有工具映射到该模型[34]。这组最低限度的约束，加上统一的文件格式，可确保所有来源的数据都能以统一方式进行下游标准化和转换。当前来源解析器的列表见补充表 4 和 ORION 的 GitHub 仓库[42]。

完成初始解析后，ORION 会应用一系列独立于来源的转换，以规范化标识符并统一语义。大多数实体标识符都使用上文所述的

节点规范化器进行标准化，该工具会将每个输入解析为首选 CURIE、规范标签、同义词以及一组指定的 Biolink 模型类别。序列变异不在节点规范化器的处理范围内，因此会通过 ClinGen 等位基因注册库 [41] 进行处理，并被赋予 ClinGen 标准等位基因 ID 作为节点标识符。谓词使用 Biolink 模型工具包 [34], 进行协调，该工具会将其映射为符合 Biolink 模型的表示形式，并在适用时附加方向、方面或上下文等限定符。最终生成的输出是一个经过规范化、与 Biolink 模型对齐的知识图谱，在保留来源特定内容的同时实现跨源互操作性。图会序列化为 KGX JSON Lines 格式 [10,39], , 可供各种下游应用使用。

由于每个 ORION 图都遵循一致的标识符方案和共享的语义模型，因此集成多个知识库十分简单。用户提供一个输入规范 YAML 文件，称为 图规范，用于定义要纳入的知识库集合。ORION 通过合并指定组件图中的节点和边来构建集成图。在此过程中，会采用边合并策略：如果两条边具有相同的主语、宾语、谓词、限定符和主要知识库，就将它们视为等价。该策略会合并重复的断言，同时保留来自不同来源的独立贡献。

图规范还可用于以声明式方式指定图的其他属性及其构建方式。用户可以选择为每个图指定源数据、Biolink 模型或规范化配置参数的特定版本。规范化配置可用于各种标识符合并和回退行为（详情见上文）。如果未指定版本，ORION 会自动检测并使用每个知识库和 Biolink 模型的最新可用版本。

图规范还允许用户配置边合并策略。例如，用户可以选择只纳入给定知识库中与其他来源所含实体相连的关系。ROBOKOP 知识图谱 [43], 采用了这种选择性方法：它纳入 Ubergraph [28] 中的子类关系，但仅限于涉及集成图中已有实体的关系。这样既避免了不必要的扩展，又保留了相关的层级信息。图规范还支持在不同来源之间传播属性。例如，在 ROBOKOP 知识图谱中，MONDO [31] 的子类注释和化学实体生物学兴趣数据库 [16] 中的角色指定会被转换为节点属性，并应用于所有具有同义标识符的节点，从而利用经过整理的元数据丰富图，而无需完整复制来源数据。

在整个 ORION 流水线中，元数据会被系统地生成和使用，以支持来源追踪和质量控制。每个图产物都有版本信息，且这些版本标识符都会记录在元数据文件中。

ORION 还会采集过程元数据，包括转换过程中所用工具的版本，以及用于概述生成的图的关键特征的质量指标。这些指标包括按类型划分的节点和边的数量，以及成功完成标准化的实体所占比例。跨连续构建跟踪这些指标，有助于用户发现意外偏差，例如标准化成功率下降或节点数量突然增加；这些情况可能表明某个来源的模式发生变更，或数据获取存在问题。最后，每个由 ORION 生成的知识图谱（KG）都会附带一份可分发的元数据文件，其中包含来源信息、版本信息和许可信息，以确保遵循 FAIR 原则。

ORION 利用与每个流水线制品相关联的元数据和版本，实现了一套自定义的资产依赖管理系统。这可以确保流水线不会重复执行已完成的过程，例如下载或解析源数据，或执行标准化。当有新版本的数据或 Babel 可用时，ORION 能够以编程方式确定流水线中哪些阶段需要运行，并且只执行必要的更新。

由于 ORION 的所有输出都遵循统一的数据模型，因此该系统还为下游的模型感知型图转换提供了基础。具体而言，ORION 支持利用 Biolink 模型的层级结构，对逻辑上蕴含的关系进行实体化。例如，当下游应用需要一个完全冗余的图，其中所有子类关系都被显式包含时，可以通过 Biolink 模型工具包提供的工具实例化这些边 [7]。此功能使用户能够根据特定用例的推理需求定制集成图的结构，无论是使用紧凑表示，还是使用包含推断关系的扩展图。

ORION 的设计充分考虑了计算资源管理，通常只需较低的资源配置即可使用（例如 16GiB 内存）。大型数据源和通用流水线组件（例如标识符规范化）的处理采用流式处理技术实现，因此任意时刻只需将极少量数据保留在内存中。ORION 使用 JSON Lines 文件格式存储知识图谱，以便进行流式处理。在某些无法进行流式处理的情况下，系统还实现了其他节省内存的技术。例如，系统提供了一种边合并算法，可将排序后的边批次写入磁盘并按顺序处理，而不是一次性将大量边保存在内存中。

4 数据可用性

Babel 的输出包括汇编、同义词文件、合并映射和 KGX 导出文件，托管在 RENCI 基础设施上（https://stars.renci.org/var/babel/2025dec11- umls-level-0/），可在此基础上将其用于更广泛的转化信息学应用，例如将其加载到本地 SQLite 数据库以便快速查询。在

出版之前，将从这些输出文件中整理出最新的 Babel 版本，上传至 UNC Dataverse 并存档，以便长期保存。

也可以使用 RENCI 托管的公开 Node Normalizer 实例对标识符进行标准化，访问地址为 https://nodenormalization-sri.renci.org/docs。或者，可以通过 Name Resolver 在 https://name-resolution-sri.renci.org/docs 对所有标识符进行文本搜索。这两项服务都提供 `/status` 端点，可用于查询这些实例当前加载的 Babel 版本。NCATS Translator 在 https://nodenorm.transltr.io/（Node Normalizer）和 https://name-lookup.transltr.io/docs（Name Resolver）分别维护着这些工具的生产实例。Node Normalizer 用法示例可见其 GitHub 仓库中的 Jupyter Notebook。您可以通过我们的邮件列表联系这些服务的其他用户，地址为 https://lists.renci.org/mailman3/lists/babel-tools.lists.renci.org/。

使用 ORION 为 ROBOKOP 构建的图可供下载、通过基于浏览器的用户界面进行探索，或在 https://robokop.renci.org/ 以编程方式查询。

5 代码可用性

完整的 Babel软件栈，包括所有 Snakemake工作流、Python模块、Dockerfile文件和 Helm配置，均以 MIT许可证发布，并在公开代码库中维护（ https://github.com/NCATSTranslator/Babel）。节点规范化器和名称解析器服务同样是开源的，其各自 GitHub 项目中提供了部署说明和运维文档。由于 Babel 会处理规模极大的等价空间，尤其是蛋白质、小分子和基因等类别，因此完整运行需要大量计算资源。典型运行需要约 500 GB 内存，以支持在内存中构建和验证团。

节点规范化器的源代码以 MIT许可证开源，网址为 https://github.com/NCATSTranslator/NodeNormalization。其 GitHub 仓库中提供了展示节点规范化器用法示例的 Jupyter Notebook。

名称解析器的源代码以 MIT许可证开源，网址为 https://github.com/NCATSTranslator/NameResolution。其 GitHub 仓库中提供了展示名称解析器用法示例的 Jupyter Notebook。

ORION 的整个软件栈均为开源软件，并以 MIT许可证发布在公开的 GitHub 仓库（ https://github.com/RobokopU24/ORION）中。该 GitHub 仓库包含 Dockerfile文件、Docker Compose配置和 Helm图表，可用于多种部署方案。仓库还包含 GitHub Actions工作流，用于自动运行 pytest测试，并在每次发布新版时将镜像发布到 GitHub容器注册表。一个 Jupyter Notebook 展示了 ORION 用法示例，可在其 GitHub 仓库中找到。

6 讨论

按照 FAIR 的定义，互操作性是某个知识库自身的一种属性，而不是描述来源 A 如何与来源 B 配合使用的成对属性。换言之，互操作的数据集 A 本身描述得足够充分，使得有意愿的用户能够将其与其他描述同样充分的数据集连接起来。这种方法对于推动公众资助的生物医学数据整体再利用，具有极其重要的意义和深远影响；然而，从彼此独立且具备互操作性的数据输入中生成集成数据集，仍然给下游用户带来负担，从而降低了整体再利用率。

此外，生物医学界不太可能完全统一采用一套通用的实体标识符或单一数据模型，以使数据集能够原生互操作。无论是基于原则还是任意作出的数据模型或标识符方案选择，始终都会存在于独立生成的数据中。因此，生物医学界需要一些工具，填补理论上的互操作性与实践中的互通之间的鸿沟。

Babel / ORION 系统是一套用于创建互操作数据集的务实工具集。从根本上说，该系统承认所有模型都有缺陷、所有抽象都存在疏漏，但要实现互通，就必须对模型和标识符作出选择。事实上，将这些选择分享给更广泛的群体是有益的，即便它们并不完美——而它们始终不会完美。Babel 的输出已不仅用于资助其开发的 NCATS 生物医学数据转换器项目，也被 ROBOKOP（Bizon 等，2019）、Dug 语义搜索引擎（Waldrop 等，2022）、DrugMechDB（ Gonzalez-Cavazos 等，2023）、MeDIC 项目（DeLuca 等，2025）、Every Cure 机器学习/人工智能辅助扩展治疗再利用（MATRIX）项目（正在开发中），以及 NCATS 本体与数据集成网络（ODIN）项目（正在开发中）采用。

认识到系统存在缺陷，并不意味着听天由命；我们相信，信息学界会发现这些工具以目前的形式颇为有用，而且持续使用这些工具将推动底层 Biolink 模型和 Babel 的标识符映射不断改进。此外，我们预计会增强底层工具的灵活性。具体而言，合并这一概念体现了标识符团形成中的一种灵活性，我们计划将其扩展到其他领域，例如合并高度相似的化学结构，或可能的直系同源基因。我们还计划允许用户配置 Babel，例如根据特定应用的需要关闭处理流程或添加新的映射。同样，ORION 的灵活性将使用户能够创建特定的、由使用场景驱动的集成和数据集，朝着实现真正可互操作、可重复使用的生物医学数据这一愿景迈进。

最后，我们欢迎社区就本文所述工具和资源的改进提出意见，这些工具和资源包括 Babel、节点规范化器、名称解析器以及 ORION 数据集成与协调流程。所有这些工具均已根据开源许可证发布。运行这些处理流程或使用这些服务时遇到的任何问题，以及任何以拉取请求形式提交的贡献，都应在“代码可用性”部分列出的 GitHub 代码库中报告。Babel 代码库中包含一份“贡献者”文档，详细说明了如何尽可能方便地将您的贡献纳入源代码。您也可以通过 Babel Tools 邮件列表与其他 Babel 、节点规范化器和名称解析器用户交流。

致谢

我们感谢 Jon Michael Beasley（AbbVie）、Olawumi Olasunkanmi（RENCI）、Hina Shah（RENCI）、Shalki Shrivastava（LBNL）和 Hong Yi（RENCI）为项目代码库的开发与维护所作的贡献。他们在软件开发、文档编写、测试和数据整理方面付出的努力，对推动这项工作至关重要。

我们还要感谢 NIH NCATS生物医学数据转换器联盟的成员，他们的讨论、问题报告和概念性反馈，为这一框架的设计与完善提供了重要参考。他们的协作参与有助于识别集成方面的关键挑战，并确定对实现不同知识库间互操作性至关重要的功能优先级。尤其是 Babel 流水线的开发，需要许多人为这一项目进行规划、讨论和测试，报告问题，并贡献想法与工具。我们特别感谢 Kent Shefcheck（Helix）、Marian Mersmann、Eric Deutsch（系统生物学研究所）、Colleen Xu（斯克里普斯研究所）、Amy K Glen（Phenome Health）、Andrew Su（斯克里普斯研究所）、Mark D. Williams（NCATS）、Richa Kanwar（RENCI）、Kamileh Narsinh（系统生物学研究所）、Stephen Ramsey（俄勒冈州立大学）和 Gwênlyn Glusman（系统生物学研究所）。

除了 Translator Consortium，我们还要感谢更广泛的研究与开源社区。该社区提供的工具和资源，为本文所述的计算框架奠定了重要基础。

经费UNDING

本研究由美国国家环境卫生科学研究所（NIEHS）与美国国立卫生研究院下属的数据科学战略办公室联合资助，资助编号为 **U24ES035214**；此外，还获得了美国国立卫生研究院下属国家转化科学促进中心（NCATS）的资助，资助编号为 **OT2TR005712**（DOGSLED）。本文内容完全由作者负责，并不一定代表美国国立卫生研究院的官方观点。

参考文献

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, 等. 科学数据管理与 stewardship 的 FAIR指导原则. 载于: Nature Publishing Group UK [互联网]. 15 2016 [20 2022] 10 1038 2016 18 3月 引用日期: 2 4月. DOI: . /sdata. . García-Closas M, Ahearn TU, Gaudet MM, Hurson AN, Balasubramanian JB, Choudhury PP, 等. 推动流行病学研究走向可发现、可访问、可互操作和可复用的实践.《美国流行病学杂志》. 2023 040 3 192 995 1005 10 1093 ; : – . DOI: . /aje/kwad . Hutchison VB, Norkin T, Zolly LS, Hsu L. 数据现状：评估美国地质调查局数据的 FAIR 程度.《数据科学杂志》. 2024;23: 22. DOI:10 022 4 5334 2024 . /dsj- - . Krauthammer M, Nenadic G. 生物医学文献中的术语识别.《生物医学信息学杂志》. 2004;37: 512–526. DOI:10 004 5 1016 2004 08 . /j.jbi. . . . Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, 等. FAIRsharing：以社区为基础的标准、代码库和政策方法.《自然生物技术》. 2019;37: 358–367. DOI:10.1038/s41587-019-0080-8 6. Clarke DJB, Wang L, Jones A, Wojciechowicz ML, Torre D, Jagodnik KM, 等. FAIRshake：评估研究数字资源 FAIR 程度的工具包.《细胞系统》. 2019;9: 417–421. DOI:10 011 7 1016 2019 09 . /j.cels. . . . Unni DR, Moxon SAT, Bada M, Brush M, Bruskiewich R, Caufield JH, 等. Biolink 模型：临床、生物医学和转化科学知识图谱的通用模式.《临床与转化科学》. 2022;15: 1848– 1855. DOI:10.1111/cts.13302 8. OpenAPI 倡议组织 . OpenAPI. 2025. 可访问: https://github.com/OAI/OpenAPI-Specification 9. Fecho K, Glusman G, Baranzini SE, Bizon C, Brush M, Byrd W, 等. 生物医学数据转换器正式发布：首次公开发布.《临床与转化科学》. 2025 70284 10 18 70284 10 1111 ; : e . DOI: . /cts. . Caufield JH, Putman T, Schaper K, Unni DR, Hegde H, Callahan TJ, 等. KG-Hub：构建和交换生物知识图谱.《生物信息学》. 2023 418 11 39 418 10 1093 ; : btad . DOI: . /bioinformatics/btad . Plater：Plater 可为符合 Biolink 标准的 neo4j图自动创建 TRAPI接口. GitHub；可访问：https://github.com/TranslatorSRI/Plater 12. ReasonerAPI：NCATS生物医学数据转换器推理器标准API. GitHub；可访问：https://github.com/NCATSTranslator/ReasonerAPI 13. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, 等. 比较毒理基因组学数据库：2019年更新.《核酸研究》. 2019 868 14 47 948 954 10 1093 ; : D –D . DOI: . /nar/gky . Chen X, Liu M, Gilson M. BindingDB：可通过网络访问的分子识别数据库. Comb Chem High Throughput Screen. 2001;4: 719– 725. DOI:10.2174/1386207013330670 15. Putman TE, Schaper K, Matentzoglu N, Rubinetti VP, Alquaddoomi FS, Cox C, 等. Monarch Initiative 2024年报告：整合跨物种表型、基因和疾病的分析平台.《核酸研究》. 2024 1082 16 52 938 949 10 1093 ; : D –D . DOI: . /nar/gkad . Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, 等. ChEBI 2016年报告：改进的服务和不断扩大的代谢物集合.《核酸研究》. 2016;44: D1214–9. DOI:10 1031 17 1093. /nar/gkv . Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, 等. PubChem 2025年更新.《核酸研究》. 2025;53: D1516–D1525. DOI:10 1059 18 1093. /nar/gkae . Leiter DP, Morgan HL, Stobaugh RE. 化合物注册库的安装与运行. J Chem Doc. 1965;5: 238–242. DOI:10.1021/c160019a010 19. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, 等. HMDB：人类代谢组数据库.《核酸研究》. 2007;35: D521–6. DOI:10 923 20 1093. /nar/gkl . Kanehisa M, Goto S. KEGG：京都基因与基因组百科全书.《核酸研究》. 2000 27 21 28 27 30 10 1093 28 1 ; : – . DOI: . /nar/ . . . Bodenreider O. 统一医学语言系统（UMLS）：整合生物医学术语.《核酸研究》. 2004;32: D267– 70 061 22 10 1093. DOI: . /nar/gkh . Ikeda S, Ono H, Ohta T, Chiba H, Naito Y, Moriya Y, 等. TogoID：用于连接生物数据集的探索性标识符转换器.《生物信息学》. 2022 491 23 38 4194 4199 10 1093 ; : – . DOI: . /bioinformatics/btac . Ikeda S, Aoki-Kinoshita KF, Chiba H, Goto S, Hosoda M, Kawashima S, 等. 通过引入

多语义和标签特征，扩展 TogoID 中标识符转换的概念.《生物医学语义学杂志》. 2025;16: 1. DOI:10.1186/s13326-024-00322-1 24. Simon Jupp, Thomas Liener, Sirarat Sarntivijai, Olga Vrousgou, Tony Burdett 和 Helen Parkinson. OxO——本体映射提取项的浓缩汇集. 载于：Matthew Horridge, Phillip Lord, Jennifer D. Warrender，编辑. 2017. 25. Mapping Commons. GitHub；可访问：https://github.com/mapping- commons 26. Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, 等. 共享本体映射的简单标准（本体映射共享简单标准）.《数据库》（牛津）. 2022 035 27 2022 10 1093 ; . DOI：. /database/baac . 医学主题词表——主页. 载于：美国国家医学图书馆 [互联网]. 7月23日 2020 [引用日期：10月29日 2025]. 可访问：https://www.nlm.nih.gov/mesh/meshhome.html 28. Balhoff JP, Bayindir U, Caron AR, Matentzoglu N, Osumi-Sutherland D, Mungall CJ. Ubergraph：将 OBO 本体整合为统一的语义图. Zenodo；2022 7249758 29 10 5281. DOI：. /ZENODO. . Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch C H, Sochat V, 等. 使用 Snakemake 实现可持续的数据分析. F1000Research. 2021;10: 33. DOI:10.12688/f1000research.29032.2 30. Raasveldt M, Mühleisen H. DuckDB：可嵌入式分析型数据库. 2019年国际数据管理会议论文集. 美国纽约州纽约：美国计算机协会；2019. 第1981–1984页. DOI:10.1145/3299869.3320212 31. Vasilevsky NA, Toro S, Matentzoglu N, Flack JE, Mullen KR, Hegde H, 等. MONDO：整合不同社区的疾病术语.《遗传学》. 2025 215 32 215 10 1093 ; iyaf . DOI：. /genetics/iyaf . Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, 等. UniChem：统一的化学结构交叉引用和标识符追踪系统.《化学信息学杂志》. 2013 3 33 5 3 10 1186 1758 2946 5 ; : . DOI：. / - - - . Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, 等. 人类疾病本体2018年更新：分类、内容和工作流扩展。《核酸研究》。2019年；47：D955–D962。DOI:10 1032 34 1093. /nar/gky . biolink-model-toolkit：一组用于查询信息和使用 Biolink 模型的实用 Python 函数。GitHub ；可获取：https://github.com/biolink/biolink-model-toolkit 35. UniProt联盟。UniProt：2025年版通用蛋白质知识库。《核酸研究》。2025年；53：D609–D617。DOI:10 1010 36 1093. /nar/gkae . Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. 临床药物的标准化名称：RxNorm 六年。《美国医学信息学会杂志》。2011年；18：441 448 10 1136 2011 000116 – 。DOI: . /amiajnl- - . 37 11 2025 [Ward I. JSON Lines。9月，引自10月15日 2025]。可获取：https://jsonlines.org/ 38. Apache软件基金会。Apache Parquet。2025年。可获取：https://parquet.apache.org/ 39. biolink/kgx：KGX 是一个用于交换知识图谱的 Python 库。载于：GitHub [Internet]。[引自12月3日 2024]。可获取：https://github.com/biolink/kgx 40. fastapi：FastAPI 框架，性能出色、易于学习、编码快速，可用于生产环境。GitHub；可获取：https://github.com/fastapi/fastapi 41. Pawliczek P, Patel RY, Ashmore LR, Jackson AR, Bizon C, Nelson T, 等。ClinGen 等位基因注册库关联遗传变异信息。《人类突变》。2018 23637 42 39 1690 1701 10 1002 ; : – 。DOI: . /humu. . ORION：用于解析来自不同来源的数据集并将其转换为可加载到图数据库中的代码。GitHub；可获取：https://github.com/RobokopU24/ORION 43. Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, Morton K, 等。来自联邦数据源的 ROBOKOP 知识图谱和 KGB：集成知识图谱。《化学信息与建模杂志》。2019年；59：4968–4973。DOI:10.1021/acs.jcim.9b00683

补充数据

Jupyter 笔记本

随附的 Jupyter 笔记本提供了节点规范化器和名称解析用法示例。

- 节点规范化器：
  https://github.com/NCATSTranslator/NodeNormalization/blob/master/documentation/NodeNormalization.ipynb
- 名称解析器：
  https://github.com/NCATSTranslator/NameResolution/blob/master/documentation/NameResolution.ipynb

补充表

补充表 1。Babel 中按 CURIE 前缀和 Biolink 类型统计的标识符数量。 Babel 2025dec11 中的标识符数量，按 CURIE 前缀分组；CURIE 前缀表示标识符所属的词汇表。已知有些标识符会在 Babel 中错误地重复出现，因为每条 Babel 流水线都会并行地独立为其文件获取标识符：因此，本表提供了 CURIE 的两种计数，即标识符总数（“CURIE 数量”）和标识符去重后的数量（“不同 CURIE 数量”）。此外，还按 Biolink 类型提供了细分统计。

前缀	CURIE 数量	不同 CURIE 数量	按 Biolink 类型细分
UniProtKB	199,651,241	199,651,241	- biolink:蛋白质：199,651,241 个 CURIE
PubChem 化合物	125,139,588	125,139,588	- biolink:小分子: 111,547,805 CURIE - biolink:分子混合物: 11 ,716,398 CURIE - biolink:化学实体: 1,875,385 CURIE
InChIKey	115,976,745	115,976,745	- biolink:小分子: 104,165,630 CURIE - biolink:分子混合物: 9 ,801,481 CURIE - biolink:化学实体: 2,009,633 CURIE - biolink:化学混合物: 1 CURIE
NCBIGene	63,742,099	63,742,099	- biolink:基因：63,742,099 个 CURIE
PMID	39,682,478	39,682,478	- biolink:出版物：39,682,478 个 CURIE
ENSEMBL	38,268,117	38,261,517	- biolink:蛋白质：21,279,080 个 CURIE（21,272,480 个不同 CURIE） - biolink:基因：17,002,237 个 CURIE（16,995,637 个不同 CURIE）
doi	30,758,070	30,758,070	- biolink:出版物：30,758,070 个 CURIE
PMC	10,658,639	10,658,639	- biolink:出版物：10,658,639 个 CURIE
CAS	4,127,675	4,127,675	- biolink:小分子：3,813,016 个 CURIE - biolink:分子混合物：307 ,963 个 CURIE - biolink:化学实体：6,689 个 CURIE - biolink:ComplexMolecularMixture：4 个 CURIE - biolink:Drug：3 个 CURIE
NCBI分类单元	2,709,586	2,709,586	- biolink:OrganismTaxon：2,709,586 个 CURIE
ChEMBL 化合物	2,480,347	2,480,347	- biolink:小分子：2,233,709 CURIE - biolink:化学实体：135,913 CURIE - biolink:分子混合物：110 ,725 个 CURIE

UMLS	2,459,955	2,459,378	- 生物体分类单元：769,724 CURIE - 医疗程序：244,699 CURIE - 药物：225,570 CURIE - biolink:化学实体：200,799 个 CURIE（200,798 个不同项） - biolink:蛋白质：160,925 个 CURIE - biolink:表型特征：120,794 个 CURIE - biolink:解剖实体：119,934 个 CURIE（119,933 个不同项） - biolink:临床属性：109,956 CURIE - biolink:疾病：107,657 CURIE - biolink:小分子：72,846 CURIE - biolink:基因：46,912 个 CURIE - biolink:出版物：45,262 个 CURIE - biolink:信息内容实体：44,812 个 CURIE - biolink:基因组实体：41,922 个 CURIE - biolink:生物过程：39,927 个 CURIE - biolink:分子活性：28,396 CURIE - biolink:设备：21,460 CURIE - biolink:活动：9,860 CURIE - biolink:细胞组分：8,141 CURIE - biolink:分子混合物：6,326 个 CURIE - biolink:细胞：6,027 个 CURIE（6,026 个不同项） - biolink:地理位置：4,262 个 CURIE - biolink:个体生物种群：4,106 个 CURIE - biolink:现象：3,940 个 CURIE - biolink：队列：3,580 个 CURIE - biolink:行为：2,870 个 CURIE - biolink：物理实体：2,845 个 CURIE - biolink：施事者：2,704 个 CURIE - biolink:宏观解剖结构：2,380 个 CURIE - biolink：核酸实体：454 个 CURIE - biolink：多肽：376 个 CURIE - biolink：事件：135 个 CURIE - biolink：生理过程：111 个 CURIE - biolink：生物实体：109 个 CURIE - 复杂分子混合物：87 个 CURIE - biolink：人类：46 个 CURIE - biolink：疾病或表型特征：27 个 CURIE - biolink：命名实体：23 个 CURIE - biolink：分子实体：13 个 CURIE - biolink:化学混合物：6 个 CURIE
MESH	446,681	346,827	- biolink:化学实体：170,767 个 CURIE（170,743 个不同项） - biolink:蛋白质：105,470 个 CURIE - biolink:小分子：78,742 个 CURIE - 生物体分类单元：70,718 个 CURIE（70,653 个不同） - biolink:疾病：11,267 个 CURIE - biolink:分子混合物：6,608 个 CURIE - biolink:表型特征：982 个 CURIE - biolink:宏观解剖结构：779 个 CURIE - biolink:解剖实体：644 个 CURIE（607 个不同） - biolink:细胞：340 个 CURIE（331 个不同） - biolink:细胞组分：284 个 CURIE（280 个不同） - biolink:ComplexMolecularMixture：183 个 CURIE（182 个不同项） - biolink:Drug：29 个 CURIE - biolink:ChemicalMixture：8 个 CURIE
RGD	317,397	317,397	- biolink:Gene：317,397 个 CURIE
PR	253,470	253,470	- biolink:Protein：253,470 个 CURIE
HMDB	217,920	217,920	- biolink:小分子：217,469 个 CURIE - biolink:分子混合物：362 个 CURIE - biolink:化学实体：89 个 CURIE

化学实体 生物学兴 趣数据库	202,960	202,960	- biolink:小分子：172 ,720 CURIE - biolink:化学实体：26,121 CURIE - biolink:分子混合物：3 ,582 CURIE - biolink:化学混合物：537 个 CURIE
UNII	139,647	139,647	- biolink:小分子：67 ,158 个 CURIE - biolink:化学实体：55,679 个 CURIE - biolink:分子混合物：16 ,809 个 CURIE - biolink:化学混合物：1 个 CURIE
RxNorm 概念唯一标 识符	124,874	124,874	- 药物：102,786 个 CURIE - biolink:化学实体：13,214 个 CURIE - biolink:分子混合物: 4 ,529 CURIE - biolink:小分子: 4 ,345 CURIE
REACT	112,754	112,754	- biolink:分子活性：89,878 CURIE - biolink:通路：22,358 CURIE - biolink:生物过程：518 CURIE
FMA	98,633	98,633	- biolink:解剖实体：98,633 CURIE
MGI	94,277	94,277	- biolink:基因：94 ,277 CURIE
RHEA	71,132	71,132	- biolink:分子活性：71,132 CURIE
NCIT	59,588	59,531	- biolink:疾病：27 ,359 个 CURIE - biolink:表型特征：21,890 个 CURIE - biolink:解剖实体：4,224 个 CURIE - biolink:细胞：2 ,577 个 CURIE - biolink:宏观解剖结构：2 ,081 个 CURIE - biolink:细胞组分：1,457 个 CURIE
WormBase	48,778	48,778	- biolink:基因：48 ,778 个 CURIE
HGNC	44,741	44,741	- biolink:基因：44 ,741 个 CURIE
GO	39,354	39,354	- biolink:生物过程：24,568 个 CURIE - biolink:分子活性：10,143 个 CURIE - biolink:细胞组分：4,058 CURIE - biolink:通路：585 CURIE
CLO	38,810	38,810	- 细胞系：38 ,810 CURIE
ZFIN	37,990	37,990	- biolink:基因：37 ,990 CURIE
FB	30,256	30,256	- biolink:基因：30 ,256 CURIE
SMPDB	30,248	30,248	- biolink:通路：30,248 CURIE
OMIM	28,018	27,969	- biolink:基因：17 ,759 个 CURIE - biolink:疾病：10 ,259 个 CURIE
MONDO	26,333	26,333	- biolink:疾病：26 ,333 CURIE
PANTHER.FAMILY	26,140	26,140	- biolink:GeneFamily：26,140 CURIE
MedGen	21,139	21,138	- biolink:疾病：21,139 个 CURIE（21,138 个不同项）
HP	19,076	19,076	- biolink:表型特征：16,707 个 CURIE - biolink:疾病：2 ,369 个 CURIE
DrugBank	17,678	16,547	- biolink:小分子：11 ,571 个 CURIE - biolink:化学实体：3,147 个 CURIE - biolink:蛋白质：2 ,447 CURIE - biolink:分子混合物：509 CURIE - biolink:复杂分子混合物：3 CURIE - biolink:药物：1 CURIE

KEGG 化合物	16,066	16,066	- biolink:小分子：14 ,039 CURIE - biolink:化学实体：1,502 CURIE - biolink:分子混合物：468 CURIE - biolink:化学混合物：57 CURIE
UBERON	14,608	14,608	- biolink:宏观解剖结构：10 ,507 CURIE - biolink:解剖实体：4,090 CURIE - biolink:细胞组分：8 CURIE - biolink:细胞：3 CURIE
dictyBase	13,892	13,892	- biolink:基因：13,892 个 CURIE
GTOPDB	13,510	13,510	- biolink:小分子：13 ,231 个 CURIE - biolink:化学实体：211 个 CURIE - biolink:分子混合物：68 个 CURIE
DOID	12,012	12,012	- biolink:疾病：12 ,012 个 CURIE
SNOMED CT	11,953	11,953	- biolink:疾病：9 ,828 CURIE - biolink:表型特征：2,125 CURIE
EFO	11,586	11,586	- biolink:表型特征：7,859 CURIE - biolink:疾病：3 ,727 个 CURIE
Orphanet	11,048	11,048	- biolink:疾病：11,048 个 CURIE
EC	8,813	8,813	- biolink:分子活性：8,813 CURIE
SGD	7,167	7,167	- biolink:基因：7 ,167 CURIE
DrugCentral	4,995	4,995	- biolink:小分子：4,302 个 CURIE - biolink:化学实体：388 个 CURIE - biolink:分子混合物：305 个 CURIE
CL	3,160	3,160	- biolink:细胞：3 ,159 个 CURIE - biolink:细胞组分：1 个 CURIE
ICD-10	2,520	2,520	- biolink:疾病：2 ,520 个 CURIE
ICD9	2,233	2,233	- biolink:疾病：2 ,233 CURIE
HGNC 基因家族	1,999	1,999	- biolink:GeneFamily：1,999 CURIE
EMAPA	966	966	- biolink:解剖实体：966 个 CURIE
监管活动医学词典	732	732	- biolink:疾病：534 个 CURIE - biolink:表型特征：198 个 CURIE
ComplexPortal	631	631	- biolink:大分子复合物：631 个 CURIE
ZFA	606	606	- biolink:解剖实体：606 个 CURIE
OMIM.PS	572	572	- biolink:疾病：572 个 CURIE
PANTHER 通路	175	175	- biolink:通路：175 个 CURIE
FBbt	117	117	- biolink:解剖实体：117 个 CURIE
KEGG 疾病	40	40	- biolink:疾病：40 个 CURIE
MP	33	33	- biolink:表型特征：30 个 CURIE - biolink:疾病：3 个 CURIE
TCDB	21	21	- biolink:分子活性：21 个 CURIE
WBbt	18	18	- biolink:解剖实体：18 个 CURIE
KEGG.REACTION	9	9	- biolink:分子活性：9 CURIE
ICD-11	5	5	- biolink:疾病：5 CURIE

补充表 2。 构成 Babel 一部分的十四个输出处理流程。 每个处理流程都会为一个或多个 Biolink 文件生成标识符。表中还列出了团的数量、标识符数量和不同同义词的数量，以及 CURIE 前缀（按标识符数量从多到少排列）。

处理流程	Biolink 类型	数量 CURIE	数量 不同的 CURIE	团领导标识符 前缀	CURIE 前缀
解剖学 所有尺度上的解剖学实体 尺度，从大脑到 内皮到胰腺 β细胞	解剖实体	227,885	227,885	医学主题词表、国家癌症研究所术语, UBERON , UMLS	解剖学基础模型、ZFA、统一医学语言系统、医学主题词表, WBbt, EMAPA , UBERON , NCIT, FBbt
	Cell	12,088	12,088	CL, MESH, NCIT, UMLS	CL, MESH, UMLS, UBERON , NCIT
	细胞组分	13,945	13,945	基因本体、医学主题词表、国家癌症研究所术语, UMLS	细胞本体、医学主题词表、统一医学语言系统、基因本体, UBERON , NCIT
	大体解剖结构	15,747	15,747	UBERON	NCIT, MESH, UMLS, UBERON
细胞系 来自不同 物种	细胞系	38,810	38,810	CLO	CLO
化学物质 各种化学物质 , 包括药物、小分子 分子．分子 混合物 ，等等	分子混合物	21,976,133	21,976,133	化学实体生物学兴趣数据库, CHEMBL.COMPOUND、HMDB、PUBCHEM.COMPOUND, RxNorm 概念唯一标识符 , UNII	ChEMBL 化合物 , KEGG 化合物 , HMDB, GTOPDB、UNII、医学主题词表、统一医学语言系统、DrugCentral, PubChem 化合物 , 化学实体生物学兴趣数据库 , DrugBank , CAS, InChIKey , RxNorm 概念唯一标识符
	小分子	222,416,583	222,416,583	化学实体生物学兴趣数据库 , CHEMBL.COMPOUND, DrugCentral, GTOPDB, HMDB, PUBCHEM.COMPOUND, UNII	ChEMBL 化合物 , KEGG 化合物 , InChIKey, GTOPDB, UNII, 医学主题词表, 统一医学语言系统, DrugCentral, PubChem 化合物 , 化学实体生物学兴趣数据库 , DrugBank , CAS, HMDB , RxNorm 概念唯一标识符
	多肽	163	163	UMLS	UMLS
	复杂分子混合物	275	275	DrugBank, MESH	医学主题词表、统一医学语言系统、DrugBank, CAS
	化学实体	4,488,540	4,488,540	化学实体生物学兴趣数据库, CHEMBL.COMPOUND, DrugBank , DrugCentral, KEGG 化合物 D, MESH, PubChem 化合物 UND, RxNorm 概念唯一标识符 , UMLS, UNII	ChEMBL 化合物, KEGG 化合物 , HMDB, GTOPDB , MESH, UMLS, UNII、DrugCentral, PubChem 化合物 , 化学实体生物学兴趣数据库 , DrugBank , CAS, InChIKey , RxNorm 概念唯一标识符
	化学混合物	610	610	化学实体生物学兴趣数据库	KEGG 化合物 , MESH, UMLS, UNII, 化学实体生物学兴趣数据库 , InChIKey
	Drug	314,727	314,727	RxNorm 概念唯一标识符，统一医学语言系统	医学主题词表、统一医学语言系统、DrugBank, RxNorm 概念唯一标识符 , CAS
疾病和表型	疾病	248,905	248,904	疾病本体、实验因子本体、医学主题词表, MONDO , NCIT, OMIM, UMLS, Orphanet	哺乳动物表型本体, KEGG 疾病, SNOMED CT , 监管活动医学词典 , ICD-11 , ICD9, MESH, UMLS, OMIM.PS, MONDO, MedGen, ICD-10, Orphanet, 实验因子本体, 国家癌症研究所术语, OMIM, HP, DOID
	表型特征	170,585	170,585	实验因子本体、人类表型本体、医学主题词表, NCIT, UMLS	哺乳动物表型本体、SNOMED CT、监管活动医学词典, MESH, UMLS, EFO, NCIT, HP

药物化学物质 药物与 其活性成分作为 化学物质	9,098 组合并归并，其中每组包含 2 至 729 个由其首选 CURIE 表示的团。总共有 106,199 个无重复的首选 CURIE。				
Gene 所有物种的基因	Gene	81,396,905	81,396,905	ENSEMBL、果蝇数据库、 HGNC、MGI、 NCBIGene ，OMIM, RGD, SGD, UMLS, WormBase ，ZFIN	dictyBase、NCBIGene、HGNC, ENSEMBL、MGI、统一医学语言系统、SGD, RGD, ZFIN, WormBase ，FB, OMIM
基因家族 基因家族	基因家族	28,139	28,139	HGNC 基因家族, PANTHER.FAMIL Y	HGNC 基因家族, PANTHER.FAMILY
基因蛋白质 基因与其编码的 蛋白质的合并	15,051,450 组合并结果，其中包含 2 到 609 个团，并以其首选 CURIE 表示。总计 32,781,254 个首选 CURIE，且没有重复项。				
剩余的统一医学语言系统 一种特殊的流水线，用于添加 每个统一医学语言系统概念不 到 Babel 中 Babel	umls	569,764	569,764	UMLS	UMLS
大分子 **复合物**	大分子复合物	631	631	ComplexPortal	ComplexPortal
过程、活动和通路 生物过程， 活动和通路	通路	53,366	53,366	GO, PANTHER.PATHW AY、REACT、 SMPDB	SMPDB、REACT、 PANTHER 通路、基因本体
	生物过程	65,013	65,013	基因本体、REACT、统一医 学语言系统	REACT、统一医学语言系统、 基因本体
	分子活性	208,097	208,097	EC、基因本体、REACT, RHEA, UMLS	TCDB、统一医学语言系统、基因本体、RHEA、EC, KEGG.REACTION ，REACT
蛋白质 所有物种的蛋白质	蛋白质	221,445,86 1	221,445,86 1	ENSEMBL ，PR, 统一医学语言系统, UniProtKB	PR, ENSEMBL ，UniProtKB ， 医学主题词表，统一医学语言系统, DrugBank
出版物 所有来自 PubMed	出版物	81,099,187	81,099,187	PMID	PMID, PMC, doi
分类单元 分类实体 ， 包括物种、属、 科 ，等等，来自 NCBI分类法	生物分类单元	3,549,962	3,549,962	医学主题词表、NCBI分类单元、 UMLS	医学主题词表、统一医学语言系统、 NCBI分类单元

补充表3：Babel 用于合并来自不同来源的标识符的映射来源列表。在某些情况下（例如 UBERON本体），我们会纳入映射来源中的所有交叉引用；而在其他情况下（例如 Wikidata），我们只提取部分映射，有时还会进行额外筛选。在这种情况下，“映射数量”列列出的是成对 CURIE 前缀之间的交叉引用数量（例如，以“CL:”开头的细胞本体标识符与以“UMLS:”开头的统一医学语言系统标识符之间有 586 个交叉引用）。在某些情况下（例如“HGNC 基因家族标签”），会纳入所有标识符，不进行任何交叉映射。

Biolink 类型	**映射来源**	**映射数量**
解剖实体 Cell 细胞组分 大体解剖结构	UBERON本体	来自 UBERON:0001062“解剖实体”及其其子类。
	GO本体	来自 GO:0005575“细胞 组分”及其子类的 305 条交叉引用。
	Wikidata	查询 *Wikidata*，找出同时具有 *wd:P7963*（“细胞本体标识符”）和 *wd:P2892*（“*UMLS* 概念唯一标识符”）的实体，然后生成它们之间的交叉引用。 CL, UMLS：596
	UMLS	FMA, UMLS：103 ,205 GO, 统一医学语言系统：5 ,412 MESH, 统一医学语言系统：1 ,794 NCIT, 统一医学语言系统：8 ,902
生物过程 分子活性 通路	RHEA	EC，RHEA：7,858
	GO	26,来自 GO:0007165 的 512 个交叉引用 , GO:0008150 和 GO:0003674 及其子类。
	UMLS	GO, 统一医学语言系统：53 ,274
细胞系	细胞系本体	38,810 个标识符
Gene	MIM2Gene MedGen	NCBIGene , OMIM：17 ,461
	NCBIGene gene2ensembl.gz	ENSEMBL，NCBIGene：13,467,445
	NCBIGene 基因_info.gz	果蝇数据库，NCBIGene：17,872 HGNC, NCBIGene：43 ,795 MGI, NCBIGene：58 ,200 NCBIGene , RGD：303 ,125 NCBIGene , SGD：6 ,471 NCBIGene , WormBase：46 ,890 NCBIGene , ZFIN：27 ,274 NCBIGene，dictyBase：13,892
	UMLS-UniProtKB 映射	NCBIGene，统一医学语言系统：23,036
	UMLS	HGNC, 统一医学语言系统：44 ,066
基因家族	PANTHER 序列分类：人类	所有标识符
	HGNC 基因家族标签	所有标识符
大分子复合物	生物体 559292 的 ComplexPortal（酿酒酵母）	631个标识符
生物分类单元	MeSH	医学主题词表，NCBI分类单元：70,106
	UMLS	MESH, 统一医学语言系统：69 ,247
出版物	PubMed 基线数据及更新	所有标识符
	来自 PMC 的 PMC/PMID 映射	PMC，PMID：12,593,920 PMID, DOI：35 ,216,885
化学实体 化学混合物 复杂分子混合物 Drug 分子混合物 多肽 小分子	DrugCentral	化学实体生物学兴趣数据库、DrugCentral：4,302 ChEMBL 化合物、DrugCentral：7,129 DrugBank、DrugCentral：4,368 DrugCentral、GTOPDB：2,106 DrugCentral、医学主题词表：2,014 DrugCentral，PubChem 化合物：5,013 DrugCentral，统一医学语言系统：5,244 DrugCentral，UNII：5,185

	MeSH	CAS, MESH：48 ,440
	MeSH	MESH, UNII：18 ,077
	Wikidata	化学实体 , MESH：6 ,208 生物学兴 趣数据库
	化学实体 生物学兴 趣数据库	化学实体 , KEGG 化合物：15 ,638 生物学兴 趣数据库
	GtoPDB	GTOPDB , InChIKey：11 ,177
	PubChem 化学文摘 社	CAS, PubChem 化合物：4 ,220,873
	PubChem 医学主题词	MESH, PubChem 化合物：77 ,747
	UMLS	DrugBank , 统一医学语言系统：8 ,003 MESH, 统一医学语言系统：148 ,791 , 统一医学语言系统 116 ,638
	UMLS	DrugBank , RxNorm 概念唯一标识符: 5 ,212
疾病 表型特征	EFO 本体	28,559 个交叉引用
	DOID	38,566 个交叉引用
	HP 本体	17,962 个交叉引用
	MONDO 本体	来自 MONDO:0000001 后代的 107,021 条交叉引用，以及 MONDO:0042489，包括 Orphanet 前缀
	UMLS	HP, 统一医学语言系统：20 ,664 MESH, 统一医学语言系统：12 ,082 NCIT, 统一医学语言系统：48 ,081 OMIM, 统一医学语言系统：11 ,616
	手动一致性匹配	DOID, MONDO：3 EFO, MONDO：1 EFO, 统一医学语言系统：7 HP, 国家癌症研究所术语：1 HP, 统一医学语言系统：1 MONDO , 统一医学语言系统：63 UMLS, 统一医学语言系统：8 HP, MONDO：2 MESH, MONDO：1
药物化学物质合并归并	手动一致性匹配	32 个交叉引用
	来自 PubChem 的 RxNorm 映射	4,246 个交叉引用
	RxNorm 映射	310,002 个交叉引用（尽管并非全部都会使用）
蛋白质	NCIt-SwissProt_映射	国家癌症研究所术语、 UniProtKB：6,277
	PR	205,来自 PR:000000001 的子类和交叉引用的 386 个交叉引用
	UniProtKB idmapping.dat	ENSEMBL，ENSEMBL：9,673,501 ENSEMBL , UniProtKB：9 ,731,073
	统一医学语言系统 基因蛋白 质	UMLS, UniProtKB：23 ,036
	UMLS	NCIT, 统一医学语言系统：15 ,478 DrugBank , 统一医学语言系统：2 ,447 MESH, 统一医学语言系 ,470 统：105

补充表 4. 提供 ORION 解析器的知识来源。此表列出了已在 ORION 中实现专用解析器的上游生物医学知识源。每个解析器都会从来源获取带版本的数据发布，并将其转换为适合下游规范化和集成的标准化中间表示。列入此表表示相应解析器可用，但不一定意味着该来源已纳入某个特定的集成知识图谱版本。

可用解析器	描述
BindingDB https://www.bindingdb.org/	一个可通过网络公开访问的数据库，收录实测结合亲和力，主要侧重于 被视为候选药物靶点的蛋白质与小分子配体之间的相互作用， 这些配体具有类药物特性。
CAM KP https://github.com/ExposuresProvider/cam-kp-api	CAM（因果活动模型）是使用网络本体语言构建的小型知识图谱 （OWL）。CAM 数据库整合了许多 CAM 图以及一个大型 合并后的生物本体，包含文中引用的 各个 CAM 中涉及的全部概念词汇。每个 CAM 描述了在 特定情境下对其中某些概念的实例化，并将这些实例之间的相互作用建模为相互关联的 复杂生物过程或环境过程的表征。
化学实体生物学兴趣数据库 https://www.ebi.ac.uk/chebi/	生物学兴趣化学实体（ChEBI）是一个免费提供的分子 实体词典，专注于“小分子”化合物。
临床试验知识图谱 https://aact.ctti-clinicaltrials.org/	由多组学提供方创建和维护的临床试验知识图谱，提供 有关临床试验的信息，这些信息最终源自研究人员提交至 clinicaltrials.gov，通过临床试验汇总分析（AACT）数据库）。 部分试验的信息包括该试验的 NCT 标识符 , 使用的干预措施 , 与该试验相关的疾病/状况 , 不良事件 , etc.
COHD https://cohd.io/	哥伦比亚开放健康数据（COHD）提供计数和频率（即 电子健康记录中的患病率），包括状况、医疗程序、药物暴露和患者 人口统计信息在住院患者和门诊患者中的数据，以及它们之间的共现频率。 这些数据源自哥伦比亚大学欧文医学中心的观察性医疗结局合作组织（OMOP） 哥伦比亚大学欧文医学中心的数据库。
DrugCentral https://drugcentral.org/	DrugCentral 提供有关活性成分、化学实体的开源数据， 以及已批准药物的药品、药物作用机制、适应症和药理作用信息， 这些数据源自美国食品药品监督管理局、 欧洲药品管理局和日本医药品医疗器械综合机构提供的信息。 注意：DrugCentral 关于已停用药物和已批准药物的信息有限 美国境外 , 但用户应注意，这些信息尚未经过验证。
DrugMechDB https://sulab.github.io/DrugMechDB/	一个用于表示药物在某种适应症下对疾病的作用机制路径的数据库 。
GTEx https://www.gtexportal.org/	基因型-组织表达（GTEx）门户网站开放提供组织 特异性基因表达和调控数据，这些数据来自分子检测（例如，全基因组测序、全外显子组测序、 RNA测序），涵盖近 1000 名个体的 54 个非患病组织部位。
GWAS目录 https://www.ebi.ac.uk/gwas/	全基因组关联研究（GWAS）目录提供了一个经过整理、可搜索、 可视化且开放获取的单核苷酸多态性（SNP）-性状 关联数据库，数据源自策展人员识别的所有 GWAS 出版物；策展人员随后提取其中 报告的性状、显著的 SNP-性状关联和样本元数据。
基因组资源联盟 https://www.alliancegenome.org/	由 7 个模式生物数据库（MOD）和基因本体（GO）组成的联盟 该联盟旨在为所有生物学家提供其数据的整合视图， 以及临床医生和其他感兴趣的各方。
HGNC https://www.genenames.org/about/	HUGO 基因命名委员会（HGNC）数据库开放提供 经 HGNC 批准的人类基因、基因组及相关对象的唯一符号和名称 资源，包括指向基因组学、蛋白质组学和表型信息的链接。
HMDB https://hmdb.ca/	人类代谢组数据库（HMDB）是一个开放访问的数据库，包含 人体内小分子代谢物的详细信息，并提供 化学数据、临床数据与分子生物学/生物化学数据之间的链接，包括 蛋白质序列（例如酶和转运蛋白）。
Hetio https://het.io/about/	Hetionet 是一个开源生物医学异构信息网络（异构网络）或图 型资源，描述了数百万项生物医学研究发现的关系 这些研究发表于过去五十年间。
基因本体注释 https://www.ebi.ac.uk/GOA/	基因本体（GO）联盟的人类基因本体注释（Human GOA） 资源提供对 UniProt 知识库（UniProtKB）中的蛋白质、来自 RNACentral 的 RNA分子以及蛋白质复合物的 GO 术语人工整理分配结果的开放访问 这些蛋白质复合物来自 Complex Portal。
IntAct https://www.ebi.ac.uk/intact/	IntAct分子相互作用数据库提供对分子相互作用数据的开放访问 这些数据源自文献整理或直接用户提交。
KinAce https://kinace.kinametrix.com/	KinAce 网络门户汇总并可视化蛋白质之间的相互作用网络- 人类基因组中的激酶及其底物。
Monarch 知识图谱 https://monarchinitiative.org/	Monarch知识图谱是 Biolink 模型规范的参考实现 。它包含来自 Monarch Initiative 汇总数据库的数据，以及 若干 OBO 本体。这些数据来源提供了疾病、表型之间的关联， 人类和非人类生物体的基因组特征。
PANTHER http://pantherdb.org/	蛋白质进化关系分析（PANTHER）分类系统 提供一个开放获取的基因家族系统发育树注释库，其中 树中的所有节点均附有持久标识符，并对每个蛋白质成员进行注释

	该家族的家族和蛋白质类别、亚家族、直系同源基因、旁系同源基因、GO系统发育注释项目功能和Reactome通路。
PHAROS https://pharos.nih.gov/	Pharos是可成药基因组照明计划（IDG）项目知识管理中心（KMC）面向公众开放的用户界面，旨在开发一个针对可成药基因组（DG）的全面整合型知识库，以阐明DG中尚未充分表征和／或注释不充分的部分，重点关注最常被用作药物靶点的三类蛋白质家族：G蛋白偶联受体、离子通道和激酶。
Reactome https://reactome.org/	Reactome 是一个免费、开源、经过整理并经过同行评审的通路数据库。
STRING-DB https://string-db.org	相互作用基因/蛋白质检索搜索工具（STRING）数据库提供已知和预测的蛋白质-蛋白质相互作用（直接和间接）的信息这些信息来源于基因组背景预测、高通量实验室实验、保守共表达、自动化文本挖掘，以及来自原始数据源的汇总知识。
文本挖掘提供方知识图谱 https://github.com/NCATSTranslator/Translator-All/wiki/Text-Mining-Provider	文本挖掘提供方知识图谱包含从以下内容中提取的主语-谓语-宾语断言将自然语言处理（NLP）算法应用于 PubMedCentral Open访问出版物合集，以及来自 PubMed 的其他标题和摘要。
Ubergraph https://github.com/INCATools/ubergraph	Ubergraph 是一个开源图数据库，包含多个集成本体，包括基因本体（GO）、化学实体生物学兴趣数据库（CHEBI）、人类表型本体（HPO）和 Uberon 的解剖学本体。
ViralProteome https://www.ebi.ac.uk/GOA/proteomes	基因本体（GO）联盟的病毒蛋白质组资源向公众开放经过整理的 GO 术语与蛋白质的对应关系，以及从UniProt知识库中为所有被 NCBI分类群视为病毒的分类单元提取的蛋白质组关系。