arXiv:2305.06088v1 [cs.AI] 10 May 2023

构建以目标为导向的互操作电子健康记录知识图谱*

Simone Bocca、Alessio Zamboni、G´abor Bella、Yamini Chandrashekar、Mayukh Bagchi、Gabriel Kuper、Paolo Bouquet 和 Fausto Giunchiglia

特伦托大学（意大利）信息工程与计算机科学系（DISI）

摘要。 在构建新应用时，我们越来越常遇到需要重用和整合现有知识的情况。然而，事实上，这些先验知识几乎不可能被重用 *as-is*。即使在某些领域中也是如此，例如 *eHealth*，尽管人们已投入大量精力来开发高质量的标准和参考本体，例如 FHIR[1]。本文提出了一种集成方法，称为 *iTelos*，旨在通过重用数据和知识来构建 *Interoperable Electronic Health Records* (iEHR)。其核心理念是，应用的 *data level* 和 *schema level* 应独立开发，从而最大限度地提高重用先验知识的灵活性，同时始终以待满足的需求为总体指导，并将这些需求形式化为 *competence queries*。这一理念通过将所有需求（包括与 *reuse*相关的需求）编入由 *a priori*定义的 *purpose* 中来实现，随后利用该 *purpose* 驱动 *middle-out* 开发流程，在这一过程中，应用模式与数据持续保持一致。本文通过将所提方法应用于一项大规模案例研究，对其进行了验证。

关键词： 互操作电子健康记录 · 知识与数据复用 · 知识图谱。

1 引言

曾几何时，人们设计应用时会采用自上而下的方式，从需求出发，一直推进到实现，而不考虑数据：这些数据会在系统投入生产后生成，最多只需用被替代的遗留系统中的数据对其进行初始化。如今，我们越来越多地设计这样的系统：它们不仅在初期，而且在投入生产后，都必须与来自其他系统的数据集成，这些系统也可能属于第三方。例如，健康系统会集成来自多个机构的个人

* 本文是对短篇论文 [9] 的重大扩展。第 1、2、6、7 节均为全新内容，而第 3、4、5 节则是在 [9]，中所述工作的基础上进行了大幅扩展，该工作主要侧重于 *iTelos* 在医疗健康领域的应用。[1] https://hl7.org/fhir/.

数据；还有 B2C（企业对消费者）应用会利用 Web 上可获取的大数据，例如开放数据或流式数据。

这一复用问题的关键在于如何处理语义异构性，这种异构性会在任何需要跨多个来源进行数据集成的情况下出现 [12]。这一问题过去已得到广泛研究，并提出了两种主要方法。第一种方法是使用本体，在不同应用程序之间就一种固定的共享语言或模式达成一致[6]。第二种方法是使用知识图谱（*KGs*），并利用其固有的灵活性和可扩展性 [15], ，以促进对现有异构数据的适配和集成。然而，这一问题在很大程度上仍未得到解决。在开发应用程序时，无论采用本体还是知识图谱，都不可能原样复用现有知识。总会存在某些特定性，导致当前问题需要专门开发，而由此产生的应用程序也再次难以复用。

我们提出了一种通用方法，称为 $iTelos$，其主要目标是尽可能最大限度地减少上述现象带来的负面影响。$iTelos$ 充分利用了此前的所有研究成果，尤其是以本体和知识图谱为基础。与此同时，$iTelos$ 更进一步，对应用程序的开发流程进行了精确规范，重点关注如何有效复用来自多个来源的数据。$iTelos$ 建立在三个关键假设之上：

– 应用的数据层和模式层应独立开发，从而能够最大限度地灵活复用已有的数据和模式，例如本体，但同时应以待满足的需求为总体指导，并将这些需求形式化为能力查询；– 待复用的数据和模式，以及能力查询，应在开发开始之前尽可能精确地确定，并事先定义为应用目标的一部分。此外，假定目标还应规定一组约束，用以明确其三个组成要素各自的满足程度在多大程度上可以影响其他两个要素的满足程度；– 应以目标为驱动，采用中间向外的开发过程，在此过程中，将应用模式和数据（两者均建模为知识图谱）的逐步演进，持续与待复用的参考模式进行自下而上的对齐，并与待复用的数据进行自上而下的对齐。

上述三项假设是$iTelos$方法论的核心，与电子健康领域中的应用程序非常契合，尤其适用于互操作电子健康记录（简称 iEHRs）的生成、集成和适配。首先，数据与模式之间的明确分离是所有医疗健康应用程序所固有的，这至少有两个原因。第一个原因是，在模式层和数据层都存在丰富的标准（例如，参见 [2,3]）。将二者分离有助于实现渐进式的

用于对 iEHR 建模的知识图谱（KG）的对齐。第二，它允许逐步集成，首先是模式，然后是数据。后者比前者更为关键，因为与模式不同，数据属于高度敏感数据，受到极其严格的隐私规则约束。关于上述列出的第二个假设，电子健康领域对正确性和精确性的需求，通过$iTelos$ 对应用程序目标的定义得到了支持。目标的定义能够精确确定在构建知识图谱（KG）时必须考虑哪些数据和模式资源，从而使该知识图谱能够表示并利用 iEHR。最后，第三个假设仅为实现上述提及的 iEHR 渐进式生成提供了指导方针，而这种生成方式得益于数据与模式之间的分离。

本文的主要目标是报告在欧洲项目 InteropEHRate[2] 中，将 $iTelos$ 方法应用于医疗健康领域，特别是应用于互操作电子健康记录生成过程中所获得的经验教训。本文的组织结构如下。第 2 节介绍目标及其组织方式。第 3 节重点介绍 $iTelos$ 自中向外的过程。第 4 节详细说明 $iTelos$ 如何实现对现有数据的可复用性。第 5 节介绍 $iTelos$ 如何实现应用知识图谱的共享性和未来可复用性。第 6 节介绍用于确保$iTelos$ 自中向外的过程始终符合目标所规定约束的主要对齐度量指标。第 7 节介绍$iTelos$ 如何在 InteropEHRate 欧盟项目中得到应用。最后，第 8 节总结全文并给出结论。

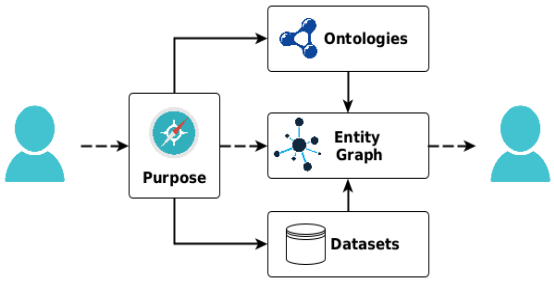


图 1：$iTelos$ 方法。

2 目标

$iTelos$ 流程在图 1 中进行了高度概括的展示。由虚线表示的逻辑顺序表明，用户 输入她希望解决的问题的非正式规范，即目标，并接收输出的知识图谱（KG），该知识图谱在图 1 中被命名为实体图（EG）。具体的

[2] 项目的详细信息可在以下网址找到：https://www.interopehrate. eu/。

该过程由四条实线表示，说明目标如何促使先验知识的重用：在数据层以数据集的形式表示，在模式层以本体的形式表示，旨在构建实体图。

目标作为该过程的主要输入，由三个要素构成：

– 最终应用程序的功能需求，换言之，即最终实体图必须能够提供以满足用户需求的信息。在 IEHR 项目的具体示例中，这些需求列出了需要纳入 IEHR 的医疗数据。我们假设此类需求被形式化为一系列能力问题（*CQ*） [14]。

– 一组可重复利用的现有知识数据集，进而将其整合到最终的实体图中。作为整体$iTelos$ 战略的一项关键假设，是将此类数据集作为知识图谱处理，从而促进其未来的重复利用。在 IEHR 项目中，四家不同的医院合作伙伴提供了用于实现这一目标的数据集。

– 一组现有的知名参考模式，即本体，但不仅限于本体，可重复使用，以开发实体图（EG）的特定目标模式，随后可将其共享以供未来的应用程序使用。值得注意的是，已有许多知名模式可供使用；例如，LOV、LOV4IoT 和 DATAHUB,[3] 是其中最相关的三个存储库。然而，对于 IEHR 项目，主要考虑了医院提供的数据集的参考模式，并大量采用了 FHIR 参考本体。按照重用方法，为了支持$iTelos$ 流程，目前正在构建一个名为$LiveSchema$的新存储库，[4] 其中的参考模式均使用非常丰富的元数据集进行注释，参见例如 [5,11], ，其目标是尽可能实现$iTelos$ 流程的自动化。

在目标的结构设计中，一个关键的设计决策是将数据层与模式层分离，使二者保持彼此独立，同时将其建模为两种不同类型的知识图谱。这一设计反映了图 1 所示的$iTelos$ 流程结构。这一重要特征使得我们能够将集成过程中重用现有数据集的问题，与为最终实体图开发一个统一模式的问题分开处理，而该统一模式可以在未来的应用程序中轻松重用。我们将数据层中的知识图谱称为实体图（*EGs*）。它们是以实体 作为节点的图（例如，我的狗$Pluto$）。实体由用于描述它们的数据属性值组成。数据层知识图谱中的链接是对象属性，用于描述任意两个实体之间的关系。相比之下，模式层中的知识图谱称为实体类型（*etype*）图（*ETGs*），即定义实体图模式的图。因此，每个实体图都有一个与之对应的 ETG，用于定义其模式。ETG 的节点是实体类型，即实体的类（例如，dog)类）。每种实体类型都由一组数据属性和一组对象属性来描述（实际上也由它们来表示），分别用于定义每种实体类型自身的模式以及这些实体类型之间可能存在的关系。

[3] 分别参见：https://lov.linkeddata.es/、http://lov4iot.appspot.com/、https://old.datahub.io/。[4] 参见：http://liveschema.eu/。

数据集 和本体 分别是图 1 中所示的实体图和实体类型图的示例。$iTelos$ 使用实体图来表示：(i) 整个流程最终生成的实体图（见图 1），以及在特定形式下可用的 (ii) 输入数据集。另一方面，ETG 则用于表示：(i) 最终实体图的模式，(ii) 可供重用的输入参考模式，以及经过形式化处理后，(iii) 从初始能力问题中提取的功能需求。$iTelos$ 保持这种表示方式的统一性，以便在数据层和模式层充分利用知识图谱相互组合的强大能力，同时支持扩展并适应不同的目标。与其底层方法一致，这些知识图谱特性使$iTelos$ 不仅能够重用现有知识，还能够生成可重用且可互操作的数据，进而减少未来构建应用程序所需的工作量。

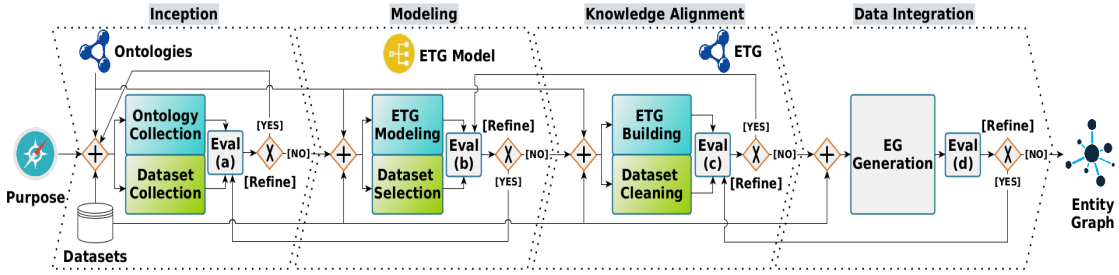


图 2：$iTelos$ 流程。

3 $iTelos$ 作为一个过程

图 2 根据 [9]，自由改编，描述了 $iTelos$ 流程（图 1 中的虚线）的关键逻辑阶段。图 2 所示的流程分为四个阶段，从初始目标出发，逐步生成实体图。每个阶段的目标可简要概括如下：

– 构思：用户在输入中提供的目标被形式化为一组能力问题，同时用于识别和收集一组现有的、可重复使用的数据集和本体；– 建模：通过考虑能力问题和数据集，构建一个特定于目标的 ETG 模型；– 知识对齐：基于先前设计的模型，通过重用选定的参考本体，构建一个可重复使用的 ETG；– 数据集成：通过将输入数据集集成到 ETG 中，构建最终的实体图。

让我们详细分析这些阶段。初始阶段 从用户处获取目标，该目标最初以所需目标的自然语言描述形式指定。功能需求从目标中提取，并形式化为能力问题列表。在图 2 所示的 本体收集 和 数据集收集 活动中，将作为目标一部分的可复用数据集和本体与能力问题进行匹配，以便选择最适合用于构建最终实体图的资源。最终的数据和模式资源集合可以扩展，纳入初始目标提供的列表中未考虑的资源。关键在于，能力问题与数据集（的模式）的匹配对于项目的成功至关重要。覆盖度不足意味着没有足够的数据来实现实体图，因此需要修订能力问题，或进一步完善数据收集。如果属于前一种情况，我们遵循的流程受到文献 [13]中工作的启发。

建模阶段接收此前收集的本体、数据集以及能力问题列表。该阶段的主要目标是从能力问题中提取一组实体类型，并通过其相应属性对其进行描述，随后利用这些实体类型和属性构建最合适的实体类型图（ETG）模型，作为最终实体图（EG）的模式，该模型在图 2 中称为*ETG* 模型。根据从能力问题中提取的实体类型和属性设计的 ETG 模型，还可以通过数据集所建议的额外实体类型和属性进行扩展。这种扩展是可选的，但建议采用，以便支持未来的扩展，并提高现有数据集的复用率。值得注意的是，数据的可用性能够降低这一步骤的成本，尤其是在早期阶段就将其纳入考虑的情况下。*ETG* 建模活动旨在通过将 ETG 模型构建为增强实体关系图（EER diagram）的形式来设计该模型，换言之，即将能力问题转化为 ETG 模型。与此同时，数据集选择活动通过筛除与所生成的 ETG 模型不匹配的数据集，最终确定此前收集的数据集。

知识对齐 阶段以先前生成的 ETG 模型，以及选定的数据集和参考本体集合作为输入。该阶段的主要目标是为最终实体图（EG）创建 ETG。关键在于，ETG 的构建应尽可能具有可复用性，从而进一步提高整个 EG 的可共享性。为此，输入的 ETG 模型本身可以作为一种可行的解决方案，但它与能力问题（CQs）的契合程度过高，因为其中包含的实体类型和属性定义对于不同目标而言可复用性较低。为了达到预期的可共享性水平，我们使用输入的参考本体中复用程度更高的实体类型和属性来构建 ETG（在图 2 中称为 ETG），同时保留 ETG 模型中设计的目标特定语义结构。此步骤通过 实体类型图对齐 活动完成，该活动由 [11] 中描述的机器学习算法实现。该算法以参考本体集合和 ETG 模型作为输入，并生成最终的 ETG 作为输出。在获得最终批准之前，需要验证生成的 ETG 是否符合输入数据集的要求。与此同时，数据集清洗 活动对数据集执行清洗和格式化操作，以使数据类型和数据格式与生成的 ETG 保持一致。

在最后一个阶段，即数据集成阶段，所考虑的输入包括 ETG 以及此前经过清洗和格式化的数据集。本阶段的目标是根据最初设定的目标，将前几个阶段处理的模式和数据资源整合在一起，从而构建实体图（EG）。本阶段包含一项活动，称为实体图生成，其目标是利用一种名为$KarmaLinker$的数据映射工具，将 ETG 与数据集进行合并。该工具由$Karma$ 数据集成工具 [17] 扩展而来，能够对短句执行自然语言处理（即我们通常所说的数据语言）[1]。借助该工具，数据集中的数据值被映射到 ETG 的实体类型和属性上。在此步骤中，将生成最终实体图（EG）中的实体；当发现多个实体实际上是同一现实世界实体的不同表示时，则将它们合并为一个有效的表示形式。Karmalinker 为这些活动提供了全面支持。本阶段所述的流程将针对上一阶段选定的数据集列表迭代执行，并按顺序逐一处理。该流程最终以将实体图（EG）导出为 RDF 文件结束。

关键观察是，正如引言所述，期望的自中而外收敛过程已在两个独立的子流程中实现；每个阶段内，这两个子流程并行执行，一个在模式层运行，另一个在数据层运行（图 2 中的蓝色和绿色方框）。在这一过程中，初始目标不断演进，逐步搭建起能力问题、数据集与参考本体之间的桥梁。为确保这一过程收敛，并避免代价高昂的错误，每个阶段都以一项评估活动结束（图 2 中的 评估 方框）。第 6 节将详细介绍这项活动。这里值得指出两点。第一，这项评估由目标所提出的非功能需求驱动。第二，在每个阶段内，评估旨在验证该阶段的目标是否达成，具体而言：在阶段 1 中，使能力问题与数据集和本体对齐，从而最大限度地重用现有资源；在阶段 2 中，使 ETG 模型与数据集对齐，从而确保项目成功（即实现目标的特定要求）；在阶段 3 中，使 ETG 与本体对齐，从而最大限度地提高实体图的可共享性。阶段 4 的评估旨在检查最终实体图是否满足目标所规定的要求。如图 2 所示，如果在预期进行评估的任何步骤中，评估结果未达到可接受的阈值，流程就会退回到上一阶段的评估步骤。若出现重大早期设计错误这一极端情况，也可以从阶段 4 退回到阶段 1。

4 在$iTelos$中的数据再利用

在构思阶段和建模阶段，iTelos 旨在增强对现有资源的复用，以减少构建实体图（EG）所需的工作量。从这一角度来看，前两个阶段的主要目标是将输入目标所提供的规范转换为 ETG 模型。该过程通过贯穿这两个阶段的五个不同步骤来实现这一目标：

1. 将目标初步形式化为一组自然语言句子，每个句子都以非正式的方式定义一个能力问题；2. 从每个能力问题中提取相关实体类型列表及对应属性，从而进一步形式化初始目标；3. 选择模式与上一步从能力问题中提取的实体类型相匹配的数据集；4. 通过与从能力问题中提取的实体类型及属性进行匹配，生成从先前选定的数据集中提取的实体类型及相关属性列表；5. 使用上一步中的实体类型和属性构建 ETG 模型。

大部分工作都在启动阶段完成，该阶段涵盖步骤 1–4，而步骤 5 则在建模阶段进行，在此阶段，将选择并利用前一阶段所做的选择来构建 ETG 模型。然而，如果建模阶段的评估未能产生预期结果，则可以通过回溯来纠正启动阶段所做的错误选择。为了提高可复用性，核心思想是将前两个阶段处理的三类资源（能力问题、本体、数据集）划分为三个类别，以定义这些资源的可复用程度。此外，在这两个阶段的执行过程中，资源将通过一系列三次迭代执行来处理，每次执行对应一个特定类别，并按照可复用性逐渐降低的顺序进行。这些类别定义如下（有关这一三级分类的初步描述，请参见 [9] ）：

通用：归入此类别的资源用于表达所有领域（包括相关特定目的领域之外的领域）共有的方面。通常，这些知识资源对应于上层本体 [7], 中规定的抽象实体类型，例如人员、组织、事件、位置，或来自非常通用领域的实体类型；这些领域通常为大多数应用程序所需，例如空间 和时间。$iTelos$归类为通用的实体类型和属性，对应于知识组织中所谓的通用孤立项 [18]。此外，$iTelos$ 将开放数据网站提供的数据资源归类为通用资源。

核心：归类为核心类别的资源体现了相关特定目的领域中最核心的方面。鉴于这些资源所承载的信息与目标密切相关，因此这些信息具有基础性作用；缺少这些信息，就无法构建实体图（EG）。例如，考虑以下目标：

”有必要通过开辟新的途径，在需要时支持欧洲公民获取其健康数据（不受特定国家医疗保健体系的限制）。为此，需要通过整合不同国家的本地数据来生成可互操作的健康数据，因为这些数据采用不同的医疗标准和语言进行表示。”[5]

[5] 本示例以及下文所述的所有后续材料均摘自 InteropEHRate 欧盟项目。

在本例中，核心资源可以是报告患者健康信息的数据值，例如用药详情、药物、医学代码（如健康领域、各种类型的互操作性标准）。相比之下，常见资源的示例包括有关患者的一般信息，如姓名、姓氏和出生日期，以及可以在上述存储库中找到的上层本体。值得注意的是，通常情况下，数据比本体更难获取，尤其是在医疗保健等敏感领域，因为这些领域对个人数据的保护尤为重视。

上下文类：归类于此类别的资源包含了相关特定目的领域的信息，这些信息可能具有独特性。核心资源有助于构建有意义的应用程序，而上下文类资源则通过明确体现与竞争对手之间的差异，为所开发的实体图（EG）创造附加价值。在上述示例中，归类为上下文类的资源可以是健康数据的翻译，这些翻译需要纳入输出的互操作健康数据中，以便在不同国家得到利用。在模式层， 上下文实体类型及其属性是指那些能够区分不同本体的实体类型和属性；这些本体虽然覆盖相同的领域，但实际上存在显著差异。 [10] 对如何比较这些本体进行了详细的定量分析。鉴于数据级上下文资源具有特定性和内在属性，通常并不容易找到。在过去开发的各种应用程序中，这类数据最终都成为一组全新的资源，需要专门为正在开发的应用程序生成，有时甚至需要在生产过程中生成。在上述示例中，一些上下文数据资源是不同医学标准代码之间的映射。这些映射并非总是现成可用，因此必须专门通过人工方式创建。

总体而言，可以得出的结论是，资源的可获取性及其可复用性，尤其是数据的可获取性及可复用性，会从通用类别向上下文类类别逐渐降低，换言之，数据会从基础设施层面的数据逐渐转向更具应用特定性的数据。此外，如 [9]，所述，可复用性的降低与创建和处理更多上下文类数据所需预处理工作的增加是同步发生的。

表1：按三个可复用性类别划分的能力问题。

编号	问题	行动	类别
1	患者的基本信息是什么? 患者摘要中报告的内容?	返回互操作患者摘要	共同
2	患者药物的医疗信息有哪些?	返回用药信息 ， 以及所涉及药品的代码	Core
3	哪一种是意大利药物的国际版本?	返回多语言互操作药物信息 上下文类	

我们来看一个例子，了解本节开头列出的五个细化步骤如何与上述三个类别相互交织。表1展示了从上文介绍的案例研究中提取的一些能力问题（步骤1）。请注意，在

这一步中，能力问题已被归类为通用、核心或上下文类（最后一列）；同时，左列中目标的描述经过转换后，变得更接近实体图的需求（中间列）。表2则列出了从能力问题中提取的实体类型和属性（步骤2）。由于篇幅所限，表中没有呈现这一点，但这些实体类型和属性继承了相应能力问题的类别。请注意，实体类型可以属于核心、通用或上下文类，而其属性却同时涵盖这三类；这是因为某个实体类型可能出现在多种（类型的）能力问题中。

表2：能力问题中的实体类型及属性。

CQ 编号实体	类型	属性
1	患者，　生命体征，护理计划	患者标识符　　, Name, Surname, 出生日期　　, 血压, 护理计划类别 _　_
2	药物，药品	药物主体、药物日期、药品标识符　　_ 编码系统，代码值　_
3	药物　　，翻译	目标语言，源语言，　_ 药物剂量说明，药物文本备注，　_　_

下一步是选择数据集（步骤 3）。假设用于（部分）回答表1中查询的数据集生成了与能力问题相匹配的属性，如表3所示（步骤 4）。作为匹配示例，请比较表3中的属性 *CD-ATC* 与表2中能力问题 n°2 的属性 代码值 。请注意，分析顺序（从通用到核心，再到上下文类）如何形成依赖关系，而这些依赖关系可能影响选择某个数据集而非另一个数据集。例如，药物剂量说明 可能作为上下文类属性而被需要，但要对其进行恰当定义，我们还需要核心实体类型 药物。按照上述顺序，依据三个可复用性类别分析所需资源，意味着更多地使用可复用性较高的资源。至此，前两个*iTelos* 阶段的流程结束，步骤 5 构建 ETG 模型，如图 3 所示，其中包含上述各表中的实体类型和属性。

表 3：根据可复用性类别划分的数据集属性。

属性	描述	Type	类别
患者标识符	患者标识符纳入数据集	字符串	Common
firstname	将患者姓名纳入数据集	string	Common
姓氏	将患者的姓氏纳入数据集	string	Common
CD-ATC	数据集中指定的药品医疗代码	string	Core
beginmoment/date	数据集中指定的药物日期时间		Core
content/text	有关一种药物的文本信息	string	上下文类

5 在$iTelos$中的数据共享

在知识对齐阶段，主要目标是通过构建一个可供未来不同目标共享的 ETG，增强待生成实体图（EG）的可共享性。$iTelos$ 为实现这种可共享性所采用的方法是：在先前构建的 ETG 模型的整体指导下，尽可能多地重用来自规范良好的标准参考本体的实体类型和属性来构建 ETG。ETG 模型使 ETG 的生成始终聚焦于初始目标，而参考本体的应用则提高了其在不同领域中的共享可能性，因为这些领域已经涉及此类本体。一个关键观察是，对齐工作主要涉及通用类型，以及可能涉及的核心类型，而对上下文实体类型的要求则低得多。值得注意的是，从事后角度来看，与最合适的本体进行对齐能够实现对所生成数据的重用。例如，将 FHIR [6] 选为医疗数据的参考本体，可以确保所生成的 ETG 符合大量已采用 FHIR 进行结构化的医疗保健信息。这类决策应在启动阶段作出；如果直到此阶段才发现需要作出此类决策，则可能导致回溯。

$iTelos$ 通过调整 [11]中提出的实体类型识别（*ETR*）流程，实现了可共享 ETG 的构建。该流程分为三个步骤。第一步是本体选择。此步骤旨在选择最适合 ETG 模型的一组参考本体。根据[11], ，此选择步骤通过依据两个度量指标对每个参考本体进行测量来完成，这两个度量指标能够：

– 用于识别参考本体中有多少种实体类型与 ETG 模型中定义的实体类型相同，以及– 用于衡量每种本体实体类型的属性可共享性值，以表示其与 ETG 模型实体类型共享的属性数量。

第一步的输出是一组经过筛选的本体，这些本体能够最大程度地覆盖 ETG 模型，并且已验证其符合数据集模式，包括实体类型和实体类型属性两个层面。第二步是实体类型识别(ETR)。这一步的主要目标是针对 ETG 模型中的每种实体类型，逐一分析先前选定的本体，预测其中哪种实体类型与 ETG 的匹配程度最高。实际上，ETG 模型的实体类型被用作分类任务的标签。如 [11], 中所述，这种执行方式利用了与本体匹配中所采用的技术非常相似的技术（例如，参见 [8]）。这一步的最终结果是一个预测值向量，用于返回 ETG 模型的实体类型与所选本体中的实体类型之间的相似度分数。第三步是 *ETG*生成。这一步利用上一步生成的预测向量，识别参考本体中的实体类型及属性，并将其纳入 ETG 的最终版本。在这一过程中，保留与数据集模式之间的映射至关重要；每当这一点难以实现时，就可能需要进行回溯。

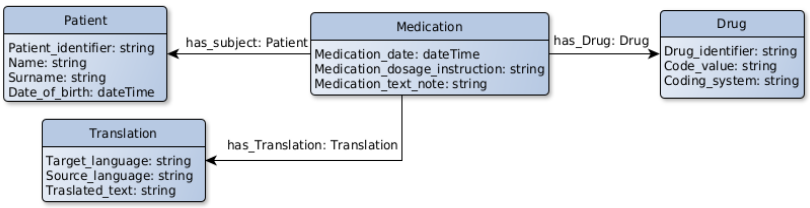


图 3：ETG 模型的局部示意图。

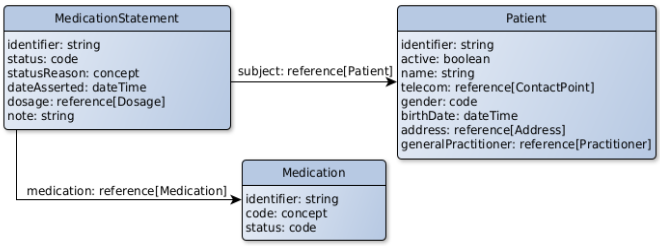


图 4：*FHIR* 参考本体的一个局部。

常见实体类型、核心实体类型和上下文实体类型及其属性之间的区别，以及它们的处理顺序，在这一阶段发挥着重要作用，具体可表述如下：

– 应尽可能多地从参考本体中采用通用实体类型，采用比例应尽可能接近 100%。这通常会通过重用现有的通用跨领域实体类型表示（例如，事物、产品、事件、位置），丰富 ETG 模型的顶层。这些实体类型通常因其抽象性质而较少受到开发者关注，但对于构建 ETG 至关重要：它们能确保所有属性都置于恰当的位置，从而可供不同领域和不同目标共享。此外，这还能实现对这些通用孤立项 的跨领域对齐（见第 4 节）；这些孤立项通常有大量公开的（开放）数据可用（例如街道）；– 核心实体类型暂且按与通用实体类型相同的方式处理，但结果在很大程度上取决于可用的参考本体。不妨想想上文的 FHIR 示例；– 上下文实体类型，尤其是上下文属性，是现有参考本体中最难找到的。因此，这类实体类型和属性主要用于从可用本体中进行高层级筛选，而不是用于在这些本体内部挑选单个实体类型 [11]。

[6] http://hl7.org/fhir/

作为上述流程的一个示例，请比较图 3 中的 ETG 模型片段与图 4 中的$FHIR$ 参考本体片段。[7] 可以看出，$FHIR$ 实体类型用药记录、药物 和患者 分别可与 ETG 模型中的实体类型药物、药品和患者匹配。因此，$iTelos$ 采用 FHIR 实体类型和属性来构建最终版本的 ETG。值得注意的是，ETG 模型中名为翻译 的上下文实体类型，在 FHIR 参考本体中没有任何可重用的对应项。因此，这种特定目标的实体类型取自 ETG 模型，并纳入最终的 ETG。

6 对齐评估

$iTelos$ 流程中实施的评估活动（见第 3 节）基于一组度量指标，这些指标应用于不同阶段的中间产出，例如能力问题、ETG 和数据集，以及最终的实体图。由于篇幅有限，下面我们仅介绍三项度量指标在其具体应用阶段中的使用情况。为了描述这些评估度量指标，设 α 和 β 为两个通用的元素 集合，其中一个元素可以是实体类型或属性。我们有如下定义：

覆盖度。该指标用于衡量两个元素集之间的重叠程度。具体而言，分析两个集合中共同包含多少种实体类型和属性。该评估通过 α 和 β 的交集与整个 α 的比值计算得出。

$$Cov = (\alpha \cap \beta)/\alpha \tag{1}$$

例如，在启动阶段（图 2 中的 *Eval(a)* ）， Cov 在评估潜在数据集（通过数据集模式）的<escape> 可复用性 </escape>方面发挥着核心作用，评估依据是能力问题。对于每个数据集， Cov的高值（无论应用于实体类型还是属性）都意味着该数据集非常适合这一目标。 Cov 的低值则意味着目标与数据集之间的重叠程度极低，其结果是该数据集不会被考虑用于复用，同时可能需要修改（规范不充分的）能力问题。

扩展度。该指标用于量化任意元素集（例如 β）所提供的知识占比，以实体类型集合或属性集合的形式表示，并相对于所考虑的全部知识（此处为 α 和 β）进行衡量。

$$Ext = (\beta - (\alpha \cap \beta))/((\alpha + \beta) - (\alpha \cap \beta)) \tag{2}$$

在建模阶段（评估（*b*）图 2 中），评估利用 Ext 来衡量 ETG 模型扩展能力问题集合的程度，目标是构建最适合该目标的模型。为此， Ext 的高值表明 ETG 扩展了能力问题的范围，

[7] 在本文中作为一个用例进行报告的 InteropEHRate 项目中，与 ETG 模型的匹配工作是手动完成的。

这表明能力问题在生成 ETG 模型方面的贡献有限。另一方面， Ext 的较低值则表明，能力问题对 ETG 模型的构建作出了显著贡献。

稀疏度。该指标量化任意数量的相似元素集之间的元素级差异，其定义为不属于 β 的 α 的百分比与反向情况的百分比之和。

$$Spr = ((\alpha + \beta) - 2(\alpha \cap \beta))/((\alpha + \beta) - (\alpha \cap \beta)) \tag{3}$$

在知识对齐阶段（*Eval(c)* 图 2 中），我们的主要重点是利用 Spr 来确保 ETG 的可共享性 。我们通过确保 ETG 与每个参考本体之间达到所需的 Spr 阈值，逐步实现可共享性。该阈值表明，ETG 包含反映上下文知识的公理。然而，这一评估旨在最大限度地采用参考本体中关于通用和核心元素的知识。

上述所有度量指标均作用于模式层。但它们并未涉及由于数据集之间存在语义异质性而导致的数据集集成结果。为了描述 $iTelos$ 所采用的数据层评估标准，我们假设以下情形：使用 $Karma\text{-}Linker$，我们将一个新的数据集 D_1 集成到一个部分构建的实体图（EG）中。此外， D_1 具有一个实体类型 E_1，其属性集为 A_1 ，而实体图（EG）中已经包含一个实体类型 E_2，其属性集为 A_2。那么，我们将面临以下几种可能的情况:

– ($E_1 = E_2$) D_1 中的实体类型 E_1 已集成到实体图（EG）中。将 D_1 集成到实体图中，会增加通过实体类型 E_1表示的实体数量，而集成后的实体图将通过新实体之间的连接得到丰富。然而，这种情况包含两个不同的子情况：(i) 当 $A_1 = A_2$时，即两个实体类型具有相同的属性集合， A_1 和 A_2的值集合之间可能存在冲突（同一属性具有不同的值）。在这种情况下，$iTelos$ 旨在识别集成过程中出现的此类冲突的数量，以及集成后仍未赋值（值为空或无意义）的属性数量。(ii) 在第二种子情况下，即 $A_1 \neq A_2$，两个实体类型具有不同的属性集合。因此， A_1 和 A_2的值集合之间不存在冲突，并且我们能够实现对 E_1中实体更大程度的集成。

– ($E_1 \neq E_2$) D_1 中的实体类型 E_1 尚未存在于实体图（EG）中。因此，将 D_1 整合到实体图（EG）中，会增加最终实体图（EG）中的实体类型数量（以及此类实体类型的实体数量）。同样，在这种情况下，我们可以区分两种子情况：(i) 当 E_1 和 E_2 至少通过一个对象属性相互关联时，整合D_1后得到的实体图（EG）将是一个连通图。在这种子情况下，$iTelos$ 旨在通过识别 E_1中有多少实体的对象属性值不为空，且这些对象属性将 E_1 与实体图（EG）中的其余实体连接起来，从而评估图的连通程度。在第二种子情况下，不存在连接.....的对象属性。

E_1 和 E_2 （或者 E_1 与实体图（EG）中的任何其他实体类型相连）。因此，集成 D_1之后的实体图（EG）将不会保持连通，且无法通过遍历实体图（EG）获取 D_1 所承载的信息。因此，集成D_1 并不会提高实体图（EG）的连通性。

上文简要介绍的数据驱动标准对于评估最终实体图的质量至关重要。目前，我们在处理过程中以及针对最终实体图，依据上述标准对其特征进行了评估。不过，用于评估最终实体图质量的一套更精确的度量指标仍在制定中。

7 InteropEHRate 案例研究

$iTelos$ 方法已应用于 InteropEHRate 欧盟项目。该项目旨在让欧洲公民能够完全掌控自己的医疗数据。为实现这一目标，合作方重点关注两个关键方面：

1. 通过跨国数字医疗基础设施为公民和医疗保健从业者提供支持。该基础设施由移动应用程序、医院数字服务和第三方数字服务组成 [16]。不过，仅有这样的基础设施还不足以完全实现项目目标。 2. 项目考虑的第二个关键方面是互操作电子健康记录（iEHR）的生成与利用。iEHR 是一种医疗数据资源，公民和医疗保健从业者（以及研究人员）均可加以利用，不受所在欧洲国家的限制，即使他们身处的国家与惯常居住地不同也不例外。具体而言，IEHR 旨在将目前不同国家（甚至同一国家内部）医疗数据的异构性整合为一种统一的多语言数据格式（采用 FHIR 医疗标准），借助项目提供的基础设施，使其能够在欧洲得到便捷利用。

在上述背景下，$iTelos$ 已被用于整合来自不同欧洲国家的本地医疗数据。具体而言，该方法以分别来自意大利、比利时、罗马尼亚和希腊四家不同医院的若干本地数据测试集作为输入，并结合各自的参考模式进行处理。处理这些本地数据时需要解决的主要问题是所提供数据之间存在高度异构性。一个具体的例子是意大利医院提供的数据与比利时医院提供的数据之间的差异：前者采用 HL7-CDA 标准 [8], 表示，而后者则采用一种名为$SumEHR$[9]的比利时医疗标准表示。CDA 数据采用 XML 格式进行结构化，而 SumEHR 数据则采用 JSON 文件格式，并使用不同的属性来表达相同的医疗信息。

[8] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7[9] https://www.ehealth.fgov.be/standards/kmehr/en/transactions/ 汇总电子医疗保健记录-v20

这两类数据之间的另一个主要区别在于不同标准所采用的医学编码系统。CDA 支持 LOINC [10] 代码，而 SumEHR 使用的是一种在比利时以外不被认可的本地编码系统。此外，如 [3], 所述，在这种背景下还需要考虑另一个问题，即自然语言异构性，这导致了构建多语言 iEHR 的需求。在模式层面，FHIR 医学本体被采用为参考本体，作为主要输入，用于构建一个 ETG，以构建能够维护互操作电子健康记录的知识图谱。$iTelos$ 已在该项目中实施，采用半自动化流程构建知识图谱，以维护由参与医院提供的本地信息的互操作多语言版本。从该知识图谱中提取的数据构成了满足项目要求的互操作电子健康记录。

在 IEHR 的最后阶段，已实施该项目的试点项目，以便在真实场景中检验项目成果是否能够满足最初的需求。我们收集了参与试点项目的合作方所提供的反馈，这些反馈帮助我们了解了 IEHR 背景下所采用方法的一些优势和不足之处。以下报告了这些反馈的具体描述：

– *(优势)* 该方法侧重于复用现有数据，因此有助于发现医疗数据在不同国家之间的异质性，并找到解决方案，使此类数据在欧洲范围内实现互操作。在项目期间，已向 HL7 协会提交了一些关于如何改进 FHIR 标准以实现更高互操作性的建议。

– *(优势)* 本地与国际知识以及数据之间的映射，以及现有资源的整合，初始构建过程较为繁重，但后续维护则相对轻量。换言之，从方法论的角度来看，良好的建模阶段难以实施，但能够构建出易于维护和演进的知识图谱。这进一步强化了 $iTelos$ 在构建知识图谱时所采用的复用与共享理念。

– (弱点) $iTelos$提供的输入标准知识和工具支持是必要的，但并不充分。在医疗保健领域，为了精确、正确地生成数据并将其从一种情境转换到另一种情境（从本地到国际），始终需要人工监督。

– (弱点) 所需的人工干预限制了构建知识图谱的系统化方法的可扩展性。为此，需要改进$iTelos$ 方法，以实现最高程度的（半）自动化，同时将所需的人工干预降至最低。

[10] https://loinc.org/

8 结论

本文介绍了$iTelos$，这是一种采用循环式开发流程、用于创建特定用途知识图谱的新方法。我们的意思是，$iTelos$ 的隐含目标是通过复用现有资源来推动知识图谱的开发，从而减少构建新数据（基于知识图谱）的工作量；这些数据今后又可被其他应用程序大量复用。此外，我们还介绍了$iTelos$ 在欧盟 InteropEHRate 项目中的应用，其目标是生成多语言 iEHR。

致谢

本文所述研究得到了 InteropEHRate 项目的支持。该项目属于欧盟委员会“地平线2020”计划，资助协议编号为 826106。我们感谢特伦托大学所有支持我们开展这一项目的人员，特别是：Danish Asghar Cheema、Ronald Chenu Abente。本文对 InteropEHRate 项目中的缩写 IEHR 进行了自由改编，将其写作 iEHR，代表互操作电子健康记录。

参考文献

1. Bella, G., Zamboni, A., Giunchiglia, F.：多语言结构化数据中的基于领域的词义消歧。载于：ECAI 2016 的 DIVERSITY 研讨会（2016）2. Bella, G., Elliot, L., Das, S., Pavis, S., Turra, E., Robertson, D., Giunchiglia, F.：利用多层分布式知识开展跨境医学研究。载于：智能系统杰出应用论文集（PAIS@ECAI）。西班牙圣地亚哥-德孔波斯特拉（2020）3. Bocca, S., Bella, G., Giunchiglia, F.：多语言领域数据中的互操作词汇表。载于：企业词汇表、本体与规则国际研讨会（VORTE）（2022）4. Bouquet, P., Giunchiglia, F.：关于理论充分性的推理：限定问题的一种新解决方案。Fundamenta Informaticae **23**(2, 3, 4), 247–262（1995）5. Dutta, B., Nandini, D., Shahi, G.K.：Mod：用于本体描述与发布的元数据。载于：都柏林核心与元数据应用国际会议。第 1–9 页（2015）6. Ekaputra, F., Sabou, M., Serral Asensio, E., Kiesling, E., Biffl, S.：多学科工程环境中的基于本体的数据集成：综述。开放信息系统期刊 **4**(1), 1–26（2017）7. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.：利用 DOLCE 丰富本体。载于：知识工程与知识管理国际会议。第 166–181 页。Springer（2002）8. Giunchiglia, F., Autayeu, A., Pane, J.：S-Match：用于匹配轻量级本体的开源框架。语义网 **3**(3), 307–317（2012）

9. Giunchiglia, F., Bocca, S., Fumagalli, M., Bagchi, M., Zamboni, A.：人气驱动的数据集成。载于：伊比利亚美洲知识图谱与语义网会议。第277–284页。Springer（2022） 10. Giunchiglia, F., Fumagalli, M.：目的论：对象、动作与功能。载于：ER—概念建模国际会议。第520–534页（2017） 11. Giunchiglia, F., Fumagalli, M.：实体类型识别——应对知识的多样性。载于：知识表示会议（KR）。希腊罗得岛（2020） 12. Giunchiglia, F., Zamboni, A., Bagchi, M., Bocca, S.：分层数据集成。载于：第二届知识图谱构建国际研讨会（KGCW），与2021年扩展语义网会议（ESWC）同期举行。线上（2021） 13. Gr¨uninger, M.：作为对话的本体验证（2019） 14. Gr¨uninger, M., Fox, M.S.：能力问题在企业工程中的作用。载于：基准测试——理论与实践，第22–31页。Springer（1995） 15. Kejriwal, M.：领域特定知识图谱构建。Springer（2019） 16. Kiourtis, A., Mavrogiorgou, A., Mavrogiorgos, K., Kyriazis, D., Graziani, A., Symvoulidis, C., Bella, G., Bocca, S., Torelli, F.：欧洲民众触手可及的电子健康记录：interopehrate 协议。载于：pHealth 2022，第145–150页。IOS Press（2022） 17. Knoblock, C.A., Szekely, P.：利用语义实现大数据集成。《AI Magazine》**36**(1)，25–38（2015） 18. Ranganathan, S.R.：图书分类法导论。Asia Publishing House，伦敦，第3版（1967）