### 创新与改进

在本节中，我们将讨论创建知识图的主要技术，并随后从各种遗留数据源（从纯文本到结构化格式（以及介于两者之间的任何内容）中丰富知识图。创建知识图时应遵循的适当方法取决于所涉及的参与者、领域、设想的应用程序、可用的数据源等。然而，一般来说，知识图表的灵活性使其能够从一个初始核心开始，该核心可以根据需要从其他来源逐步丰富（通常遵循敏捷[243]或“现收现付”[448]方法）。在我们的运行示例中，我们假设旅游局决定从头开始构建一个知识图，旨在最初描述智利的主要旅游景点——地点、事件等，以帮助来访的游客确定他们最感兴趣的景点。委员会决定推迟以后再添加更多数据，如交通路线、犯罪报告等。

#### 人类合作

创建和丰富知识图表的一种方法是征求人类编辑的直接贡献。此类编辑可以在内部找到（例如，旅游局的员工），使用众包平台，通过反馈机制（例如，游客添加对景点的评论），通过协作编辑平台（例如，开放给公众编辑的景点维基），等等。尽管人工参与会产生高成本[376]，一些著名的知识图表主要基于人类编辑的直接贡献[214515]。然而，根据征集贡献的方式，这种方法有一些关键的缺点，主要是由于人为错误[382]、分歧[535]、偏见[255]、故意破坏[218]等。成功的协作创作进一步提出了有关许可、工具和文化的挑战[382]。有时，人们更倾向于验证和策划通过其他方式[382]（例如，通过有目的的视频游戏[259]）提取的知识图的添加，以定义来自其他来源的高质量映射[113]，定义适当的高级模式[267，287]，等等。

#### 6.2文本源

文本语料库——比如来自报纸、书籍、科学文章、社交媒体、电子邮件、网络爬虫等——是丰富信息的丰富来源[220422]。然而，为了创建或丰富知识图，以高精度和高召回率提取此类信息是一个不小的挑战。为了解决这个问题，可以应用自然语言处理（NLP）[258322]和信息提取（IE）[192317522]的技术。虽然不同的文本提取框架的过程有很大差异，但在图28中，我们展示了一个示例句子中文本提取的四个核心任务。我们将依次讨论这些任务。

**6.2.1预处理。**预处理任务可能涉及对输入文本应用各种技术，图28展示了标记化，它将文本解析为原子术语和符号。应用于文本语料库的其他预处理任务可能包括：词性标记[258，322]，以识别代表动词、名词、形容词等的术语。；依存分析，它为一个句子提取一个语法树结构，其中叶节点表示单个单词，这些单词共同构成短语（例如名词短语、动词短语），最终形成从句和句子[258，322]；词义消歧（WSD）[350]用于识别单词使用的意义（又称词义），将单词与词义词典（例如WordNet[331]或BabelNet[351]）联系起来，例如，术语flights可能与WordNet的意思“乘飞机旅行的实例”有关，而不是“一层楼和下一层楼之间的楼梯”。适用的适当预处理类型通常取决于后续任务的要求。

**6.2.2命名实体识别（NER）**。NER任务识别文本中提到的命名实体[347，409]，通常针对提到的人、组织、地点和潜在的其他类型[300，348，537]。存在多种NER技术，许多现代方法基于学习框架，利用词汇特征（例如，词性标签、依赖关系解析树等）和地名录（例如，常见名字、姓氏、国家、知名企业等的列表）。监督方法[42,154,288]要求手动标记训练语料库中的所有实体提及，而基于自举的方法[94,143,202,348]则需要一小组实体提及的种子示例，从中可以学习模式并应用于未标记的文本。远程监控[300412537]使用知识图中的已知实体作为种子示例，通过这些示例可以检测到类似的实体。除了基于学习的框架之外，手动起草的规则[84275]有时仍然被使用，因为它们的行为更可控、更可预测[85]。由NER标识的命名实体可用于为知识图生成新的候选节点（称为新兴实体，如图28中的虚线所示），或可根据以下描述的实体链接任务链接到现有节点。

**6.2.3实体链接（EL）**。EL任务将文本中提到的实体与目标知识图的现有节点相关联，目标知识图可能是正在创建的知识图或外部知识图的核心[526]。在图28中，我们假设节点Santiago和Easter Island已经存在于知识图中（可能是从其他来源提取的）。EL然后可以将给定的提及链接到这些节点。EL任务提出了两个主要挑战。首先，提到同一实体可能有多种方式，比如拉帕努伊岛和复活节岛；如果我们创建了一个节点Rapa Nui来表示该提及，我们会将两个提及下的可用信息分割到不同的节点上，因此目标知识图捕获各种别名和多语言标签非常重要，通过这些别名和标签可以引用实体[340]。第二，不同语境中的同一提及可以指不同的实体；例如，圣地亚哥可以指智利、古巴、西班牙等国的城市。因此，EL任务考虑了一个消歧阶段，其中提及与知识图中的候选节点相关联，对候选节点进行排序，并选择最有可能提及的节点[526]。在这个阶段可以使用上下文；例如，如果复活节岛有可能与圣地亚哥并列，我们可能会提高提及智利首都的可能性，因为这两个候选人都位于智利。消除歧义的其他启发式方法考虑。

**6.2.4关系提取（RE）**。RE任务提取文本中实体之间的关系[22545]。最简单的情况是在封闭环境中提取二元关系，其中考虑了一组固定的关系类型。虽然传统方法通常使用手工制作的模式[215]，但现代方法更倾向于使用基于学习的框架[421]，包括手动标记示例的监督方法[74545]。其他基于学习的方法再次使用引导[75143]和远程监督[2333343415462485532]来放弃手动标签的需要；前者需要手动标记种子示例的子集，而后者则在一个大型文本语料库中找到句子，其中提到了一对具有已知关系/边缘的实体，这些实体用于学习该关系的模式。二进制RE也可以在开放环境中使用无监督方法应用——通常被称为开放信息提取（open Information Extraction，OIE）[28、144、145、320、321、335]——其中目标关系集不是预定义的，而是基于依赖关系解析树从文本中提取的。

人们提出了多种RE方法来提取n元关系，从而进一步捕捉实体之间的关联。在图28中，我们看到n元关系如何捕捉额外的时间背景，表示拉帕努伊何时被命名为世界遗产；在这种情况下，将创建一个匿名节点来表示有向标记图中较高的算术关系。n元RE的各种方法都基于框架语义[153]，对于给定的动词（例如，“命名”），框架语义捕获了所涉及的实体以及它们之间的相互关系。然后，FrameNet[24]等资源定义了单词的框架，例如，确定“命名”的语义框架包括说话人（命名某物的人）、实体（命名某物）和名称。可选的框架元素包括解释、目的、地点、时间等，可以为关系添加上下文。其他重建方法则基于话语表征理论（DRT）[262]，该理论认为文本的逻辑表征基于存在事件。例如，根据这一理论，将复活节岛命名为世界遗产被认为是一个（存在主义）事件，其中复活节岛是患者（受影响的实体），从而得出合乎逻辑的（新戴维森）公式：∃e:？命名（e），患者（e，复活节岛），姓名（e，世界遗产）？这种公式类似于物化的概念，如前3.3节所述

最后，虽然在封闭环境中提取的关系通常直接映射到知识图，但在开放环境中提取的关系可能需要与知识图对齐；例如，如果OIE过程提取出圣地亚哥复活节岛has flights to的二元关系，则可能是知识图没有标记has flights to的其他边，如果知识图中使用了flight，则对齐可能会将这种关系映射到圣地亚哥复活节岛flight的边。有多种方法可用于此类目的，包括用于对齐n元关系的映射[98169]和规则[423]；基于分布和依赖的相似性[339]、关联规则挖掘[130]、马尔可夫聚类[131]和语言技术[318]来调整OIE关系；其中包括。

**6.2.5联合任务**。在介绍了从文本构建知识图的四个主要任务之后，需要注意的是，框架并不总是遵循这一特定的任务序列。例如，一种常见的趋势是组合相互依赖的任务，联合执行WSD和EL[340]，或NER和EL[310，360]，或NER和RE[413，542]等，以相互提高多个任务的性能。关于从文本中提取知识的更多细节，请参阅梅纳德等人[322]的书和马丁内斯·罗德里格斯等人[317]最近的调查。

##### 6.3标记源

Web建立在相互链接的标记文档上，其中标记（又名标记）用于分隔文档的元素（通常用于格式化目的）。网络上的大多数文档都使用超文本标记语言（HTML）。图29显示了一个关于智利世界遗产的HTML网页示例。其他标记格式包括Wikipedia使用的Wikitext、用于排版的TeX、内容管理系统使用的Markdown等。从标记文档中提取信息的一种方法——为了创建和/或丰富知识图——是去除标记（例如HTML标记），只留下纯文本，可以应用上一节中的技术。然而，标记可用于提取目的，其中上述文本提取任务的变体已被修改以利用此类标记[304、307、317]。我们可以将标记文档的提取技术分为三大类：与特定格式中使用的标记无关的通用方法，通常基于将文档元素映射到输出的包装器；以文档中特定形式的标记为目标的集中方法，最典型的是web表（但有时也包括列表、链接等）；以及基于表单的方法，根据deepweb的概念，提取网页底层的数据。这些方法通常可以从给定网站的网页共享的规则中获益，无论是由于关于如何在网页上发布信息的非正式约定，还是由于重复使用模板来自动。

**6.3.1基于包装的提取**。许多通用方法都基于包装器，包装器可以直接从标记文档中定位和提取有用的信息。虽然传统的方法是手动定义这种包装器——这是一项定义了各种声明性语言和工具的任务——但这种方法对网站布局的更改很脆弱[152]。因此，其他方法允许（半）自动诱导包装[155]。现代的这种方法——用于丰富LODIE[174]等系统中的知识图表——是应用远程监控，其中，EL用于识别网页中的实体并将其链接到知识图中的节点，从而可以提取、排列标记中连接已知边的成对节点的路径，并将其应用于其他示例。例如，以图29为例，远程监督可能会使用EL将拉帕努伊岛和世界遗产地与知识图中的节点复活节岛和世界遗产地联系起来，并给出知识图中命名的复活节岛世界遗产地的边缘（根据图28提取），确定候选路径（x，td[1]− · tr− · 桌子− · h1，y）作为形式x ynamed的反射边，其中t[n]表示标记t，t的第n个子项− 它的倒数和t1·t2串联。最后，可以使用高置信度的路径（例如，知识图中许多已知边“见证”的路径）来提取新边，例如Qhapaqña世界遗产网站，无论是在该页面上还是在具有类似结构的网站的相关页面上（例如，对于其他国家）。

**6.3.2 Web表提取**。其他方法针对特定类型的标记，最常见的是web表，即嵌入HTML网页中的表。然而，web表的设计是为了增强人类的可读性，这往往与机器可读性相冲突。许多web表用于布局和页面结构（例如导航栏），而那些包含数据的web表可能采用不同的格式，如关系表、列表、属性值表、矩阵等[78，104]。因此，第一步是对表进行分类，以找到适合给定提取机制的表[104134]。接下来，web表可能包含列跨距、行跨距、内部表，或者可以垂直拆分以改善人类的审美观。因此，需要一个表规范化阶段来识别标题、合并拆分表、取消嵌套表、转置表等[78、104、121、140、292、393]。随后，方法可能需要确定主角[104345]——表中描述的主要实体——在网页的其他地方可以找到；例如，尽管世界遗产是图28表格的主角，但表格中并未提及。最后，可以应用提取过程，可能将单元格与实体[297343]、列与类型[121297343]以及列对与关系[297345]相关联。为了丰富知识图，最近的方法再次应用远程监控，首先将表单元链接到知识图节点，这些节点用于生成类型和关系提取的候选对象[297343,3] 。

**6.3.3深网爬行**。Deep Web提供了丰富的信息源，只有通过搜索Web表单才能访问，因此需要使用Deep Web爬行技术来访问[312]。有人提出了从DeepWeb源中提取知识图的系统[93172290]。这些方法通常试图生成合理的表单输入（可能基于用户查询或参考知识生成），然后使用上述技术从生成的响应（标记文档）中提取数据[9317229]。

6.4结构化来源

组织内部和网络上可用的许多遗留数据都是以结构化格式表示的，主要是表格——以关系数据库、CSV文件等的形式——但也包括JSON、XML等树结构格式。与文本和标记文档不同，结构化源通常可以映射到知识图，从而根据映射（精确地）转换结构，而不是（不精确地）提取结构。映射过程包括两个步骤：1）创建从源到图形的映射，2）使用映射将源数据具体化为图形或虚拟化源（在遗留数据上创建图形视图）。

**6.4.1表格映射**。表格数据源非常普遍，例如，许多组织、网站等的结构化内容都存放在关系数据库中。在图30中，我们展示了一个关系数据库实例的示例，我们希望将其集成到正在构建的知识图中。然后，有两种方法可以将内容从表映射到知识图：直接映射和自定义映射。

直接映射自动从表生成图形。我们在图31中展示了标准直接映射[18]的结果，该映射为表的每个（非标题、非空、非空）单元格创建一个边x zy，这样x代表单元格的行，y代表单元格的列名，z代表单元格的值。具体而言，x通常对一行的主键值进行编码（例如，claimer.id）；否则，如果没有定义主键（例如，根据报告表），x可以是匿名节点或基于行号的节点。节点x和边标签y进一步对表的名称进行编码，以避免在具有不同含义的相同列名的表之间发生冲突。对于每一行x，我们可以根据其表的名称添加一个类型边缘。值z可以基于源域映射到相应图形模型中的数据类型值（例如，日期类型的SQL列中的值可以映射到RDF数据模型中的xsd:Date）。如果该值为null（或空），通常会忽略相应的边。28关于图31，我们强调了节点XY12SDA和XY12SDA之间的区别，其中前者表示由后者主键值标识的行（或实体）。如果两个表之间存在外键，例如Report。索赔人引用索赔人。id–我们可以链接，例如，发送到claimer-XY12SDA而不是XY12SDA，其中前一个节点还具有索赔人的名称和国家。为了将关系数据库映射到RDF[18]，沿着这些路线的直接映射已经标准化，Stoica等人[475]最近提出了一种类似的属性图直接映射。为CSV和其他表格数据定义了另一种直接映射[489]，进一步允许指定列名、主键/外键和数据类型，作为映射本身的一部分，这些数据格式中经常缺少。

尽管直接映射可以自动应用于表格数据源，并保留原始数据源的信息，即允许从输出图[446]重建表格数据源的确定性逆映射，但在许多情况下，需要定制映射，例如将边缘标签或节点与正在增浓的知识图对齐，按照这些思路，声明式映射语言允许手动定义从表格源到图形的自定义映射。这些方面的标准语言是RDB2RDF映射语言（R2RML）[113]，它允许从表的单个行映射到一个或多个自定义边，节点和边定义为常量、单个单元格值，或使用模板将一行中的多个单元格值和静态子字符串连接到一个术语中；例如，模板{id}-{country}可以从索赔人表中生成诸如XY12SDA-U.S.之类的节点。如果无法从一行中定义所需的输出边，R2RML允许（SQL）查询生成可以从中提取边的表，例如，可以通过定义一个查询的映射来生成U.S.2crimes等边，该查询在claimer=id上连接报告和claimer表，按国家分组，并应用计数。然后，可以在结果表上定义映射，使源节点表示国家的值，边缘标签表示持续犯罪，目标节点表示计数值。还有一个类似的标准，用于将CSV和其他表格数据映射到RDF图，再次允许选择键、列名和数据类型作为映射的一部分[490]。

定义映射后，一个选项是使用它们按照提取转换负载（ETL）方法将图形数据具体化，从而使用映射将表格数据转换并显式序列化为图形数据。第二种选择是通过查询重写（QR）方法使用虚拟化，从而将图形上的查询（使用SPARQL、Cypher等）转换为表格数据上的查询（通常使用SQL）。通过比较这两个选项，ETL允许使用图形数据，就像它们是知识图中的任何其他数据一样。然而，ETL要求将对底层表格数据的更新显式地传播到知识图中，而QR方法只维护要更新的数据的一个副本。然后，基于本体的数据访问（OBDA）[528]领域涉及支持本体蕴涵的QR方法，如第4节所述。尽管大多数QR方法只支持可表达为单个（非递归）查询的非递归蕴涵，但一些QR方法通过重写递归查询来支持递归蕴涵[447]。

**6.4.2从树木绘制地图**。许多流行的数据格式都基于树，包括XML和JSON。尽管人们可以想象——抛开树中子节点的排序等问题不谈——通过简单地为源树中x的子节点y的每个节点y创建x ychild形式的边，从树到图的简单直接映射，但这种方法通常不使用，因为它代表源数据的文字结构。相反，树结构数据的内容可以更自然地表示为使用自定义映射的图形。按照这些思路，GRDLL标准[95]允许从XML映射到（RDF）图，而JSON-LD标准[469]允许从JSON映射到（RDF）图。相比之下，XSPARQL[43]等混合查询语言允许以集成方式查询XML和RDF，从而支持在传统数据的树结构源上实现图形的实体化和虚拟化。

**6.4.3其他知识图的映射**。构建或丰富知识图的另一个途径是利用现有的知识图作为源。例如，在我们的场景中，智利旅游局的大量兴趣点可能存在于现有的知识图中，如DBpedia[291]、LinkedGeoData[472]、Wikidata[515]、YAGO[232]、BabelNet[351]等。然而，根据正在构建的知识图，并非所有实体和/或关系都有兴趣。提取相关数据子图的标准选项是使用SPARQL构造查询，生成图形作为输出[353]。知识图之间的实体和模式对齐对于更好地集成（部分）外部知识图可能是进一步必要的；这可以使用基于外部标识符[382]的图形链接工具[356513]完成，也可以手动完成[382]。例如，Wikidata[515]使用Freebase[50382]作为源；Gottschalk和Demidova[190]从Wikidata、DBpedia和YAGO中提取以事件为中心的知识图；而Neumaier和Polleres[353]则利用地理名称、Wikidata和PeriodO[186]（以及表格数据）构建了时空知识图。

##### 6.5模式/本体创建

到目前为止，讨论的重点是从外部来源提取数据，以创建和丰富知识图。在本节中，我们将讨论基于外部数据源（包括人类知识）生成模式的一些主要方法。有关从知识图本身提取模式的讨论，请参阅第3.1.3节。一般来说，该领域的大部分工作都集中在使用本体工程方法和/或本体学习创建本体上。我们依次讨论这两种方法。

**6.5.1本体工程**。本体工程（Ontology engineering）是指开发和应用构建本体的方法，提出原则性的过程，通过这些过程，可以以较少的努力构建和维护质量更好的本体。早期的方法学[151195366]通常基于瀑布式的过程，在开始用逻辑语言实现本体之前，需求和概念化是固定的，例如使用本体工程工具[187267269]。然而，对于涉及大型或不断演化的本体的情况，已经提出了构建和维护本体的更加迭代和灵活的方法。

DILIGENT[389]是敏捷方法论的一个早期例子，提出了本体生命周期管理和知识进化的完整过程，并将局部变化（知识的局部视图）与本体核心部分的全局更新分开，使用审查过程授权将变化从局部传播到全球。这种方法类似于维护和发展大型临床参考术语SNOMED CT[249]（也可作为本体）的方式，其中（国际）核心术语根据全球需求进行维护，而国家或地方对SNOMED CT的扩展则根据当地需求进行维护。然后由一组作者决定向核心术语传播哪些国家或地方扩展。更现代的敏捷方法包括极限设计（XD）[46395]、模块化本体建模（MOM）[227282]、简化敏捷本体开发方法（SAMOD）[384]等。此类方法通常包括两个关键要素：本体需求和（最近）本体设计模式。

本体需求根据作为其模式的本体，指定生成的本体的预期任务，或者实际上是知识图本身。表达本体需求的一种常见方式是通过能力问题（CQ）[196]，这是一种自然语言问题，说明需要本体（或知识图）提供的典型知识。如果本体还应该包含用于推断新知识或检查数据一致性的限制和一般公理，那么这样的CQs可以通过额外的限制和推理需求来补充。测试本体（或基于本体的知识图）的一种常见方法是将CQ形式化为对一些测试数据集的查询，并确保包含预期结果[49268]。例如，我们可以考虑CQ“圣地亚哥发生的所有事件是什么？”哪个可以表示为图形查询事件？事件类型圣地亚哥。以图1的数据图和图11的公理为例，我们可以检查预期结果EID15是否被本体和数据所包含，并且由于它不是，所以我们可以考虑扩展公理来断定位置转移VyType。

本体设计模式（Ontology Design Patterns，ODP）是现代方法学的另一个常见特征[48167]，它指定了可推广的本体建模模式，这些模式可以作为类似模式建模的灵感，作为建模模板[135460]，或作为可直接重用的组件[157454]。一些模式库已经在网上提供，从精心策划的[17454]到开放和社区主持的[157]。例如，在为我们的场景建模本体时，我们可能会决定遵循Krinadhi和Hitzler[281]提出的核心事件本体模式，该模式指定了事件的时空范围、子事件和参与者，进一步建议能力问题、正式定义等，以支持该模式。

**6.5.2本体学习**。之前的方法概述了可以手动构建和维护本体的方法。相比之下，本体学习可以用于（半）自动地从文本中提取对本体工程过程有用的信息[73,89]。早期的方法侧重于从可能代表相关领域类别的文本中提取术语；例如，从一组关于旅游业的文本文档中，一个术语提取工具可以识别n-gram，如“游客签证”、“世界遗产”、“非高峰费率”等，该工具使用统一性度量来确定n-gram作为统一短语的内聚性，以及确定该短语与某个领域的相关性的术语[318]。，作为对旅游领域特别重要的术语，这可能值得包含在这样一个本体中。 公理也可以从文本中提取，其中子类公理通常以修改名词和形容词为目标，这些名词和形容词逐渐专门化概念（例如，从名词短语“Visitor Visa”中提取Visitor Visa Visasubc.of，并在其他地方单独出现“Visa”），或者使用赫斯特模式[215]（例如，根据模式“X，如Y”，从“许多折扣，如非高峰价格可用”中提取非高峰价格折扣Subc.）。还可以从大型文本中获取文本定义，以提取超名关系，并从头开始归纳分类法[506]。最近的一些作品旨在从文本中提取更具表现力的公理，包括不相交公理[512]；以及涉及类的并集和交集的公理，以及存在性、普遍性和限定的基数限制[387]。本体学习过程的结果可以作为更通用的本体工程方法的输入，使我们能够验证本体的术语覆盖范围，识别新的类和公理等。