基于知识图谱的海关知识库平台建设研究

朱金连[1] 法 勇[1] 吕 健[1] 沈 卓[2] 张 雨[1]
（1. 南京海关，江苏 南京 210001；
2. 南通海关，江苏 南通 226006）

摘 要 海关知识库平台建设是适应海关治理体系和治理能力现代化发展要求的具体体现。平台以海关内外部各条线业务数据为基础，重点解决知识图谱的构建。本文主要研究在海关业务领域，通过对本体的概念模型、约束定义、本体的形式化以及领域公认的概念集等四个方面进行领域化处理，尝试通过对领域本体对象根类型定义来促进动态本体的知识图谱构建，并对建立动态本体指标评价体系方面提出研究意见。

关键词 海关知识库平台 知识图谱 动态本体 对象根类型
中图分类号：F752.5；TP3 **文献标识码**：A **文章编号**：1007-0745(2022)11-0094-03

海关治理体系和治理能力现代化是国家治理体系和治理能力现代化的重要组成部分。“十四五”海关科技发展规划指出，“将法律法规、技术标准、工作规范、业务基础知识汇聚、融合，建设海关通用和业务重点领域知识图谱。引入智能规则引擎等工具，应用语音识别、视频识别、图像识别、自然语言处理等技术，建设海关业务知识库平台”。

1 海关知识库平台功能设计

海关知识库平台需要以海关内外部各条线业务数据为基础，重点解决领域知识图谱的构建，并图像识别、自然语言处理、文字识别、多语种翻译、语音识别等人工智能技术为辅助，以海关业务知识资源持续开发和运营为手段，实现“前台综合执法后台知识支撑”的智能化服务，满足海关内外部用户不同群体的知识获取需要，实现海关领域知识的“业务百科”。

1.1 汇聚海关各业务条线的数据资源

作为知识来源，各业务条线的数据资源汇聚，既是海关知识库的平台特色，也是平台建设的关键点。各业务条线的数据主要包括海关各类政策法规、海关辞库、数字图书馆、动植标本库、贸易单据无纸化数据、国外证书样本数据、商品库数据等内容，具有对接系统众多、数据量大、数据时效性差异显著、数据结构及对接环境复杂等特点[1]。基于此，数据采集应采用一次性采集、增量采集、定期采集等多种方式，接口应采用数据库接口、文件接口、系统接口等多种对接方法，逐步汇聚起海关知识库底层数据资源。

1.2 强化基于知识图谱的海关知识专业化

主要体现在以下两个方面：一方面是通过对海关长期积累的大量文档、业务数据等资源进行图谱构建，逐步建立领域的基础知识图谱平台和可视化场景，并通过对接智能搜索等应用，辅助海关对积累的知识进行管理，提升内、外用户对知识的获取效率。另一方面，知识图谱的专业性体现在特定领域的知识图谱构建方面，逐步实现特定领域的专项知识图谱构建[2]。比如建设洋垃圾专项图谱、濒危物种专项图谱等，可以广泛应用于海关业务领域的知识发现、关联关系挖掘、风险分析控制等场景。

1.3 建设支撑海关业务的知识应用体系

“图库、研库、学库及关搜、关问、智识、智用”等“三库四用”应用体系开创性地提出了知识库的领域应用解决方案。作为一种应用框架，“三库四用”通过各具特色的应用场景发挥各自的应用价值。“图库”重点面向图像检索、识别，“研库”侧重于各类政策法规、文献期刊等应用，“学库”面向于微学习场景；“四用”方面，“关搜”是典型的多模态搜索应用，“关问”则侧重专家系统、问答系统，“智识”为业务场景提供智能识别支持，“智用”则贴合具体业务场景，以现场操作、执法依据、作业流程为核心，辅助用户的实际作业，切实发挥出海关知识库应用价值。

1.4 搭建共建共享的海关基础 AI 服务中心

海关知识库平台 AI 服务中心，同样可为其他业务系统提供相应的 AI 服务。作为基本能力平台，通过汇集各其他业务场景中涉及的 AI 能力，可支撑海关建立起规模化、体系化、共建共享的海关 AI 服务资源池。比如：标签识别能力可以作为“智能审图”的基础支撑，为进出口食品监管等场景提供智能识别服务；智能翻

表 1 本体定义的演变过程

范畴	提出人 / 提出时间	定义
哲学		客观存在的一个系统的解释或说明，客观现实的抽象本质
计算机科学	Neches 等 /1991	给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义
	Gruber/1993	概念模型的明确的规范说明
	Borst/1997	共享概念模型的形式化规范说明
	Sduder/1998	共享概念模型的明确的形式化规范说明
	Fensel.D/2000	特定领域中重要概念的共享的形式化的描述
	G.L.Zuniga/2001	本体是用来描述一个特定领域中的知识的形式化语言

译可通过服务开放，为舆情分析、缉私办案及监管等场景提供文本翻译服务等。通过建设共建共享的 AI 服务中心，可有效提高服务能力本身，并降低全国海关在类似服务能力建设方面的成本投入。

1.5 促进多种 AI 技术的场景化融合

多种 AI 技术在海关知识生产和应用过程中还应加快场景化融合。比如，平台支持文字、语音、图像等多种识别技术的多模态搜索服务，能够在语义理解的基础上对语音输入内容、文字输入内容或者是拍照图片内容进行精准识别及需求理解；融合自然语言理解、情感分析、智能问答等能力，精准理解用户问题中提及的业务、服务领域和意图，并基于此对意图、指代等进行准确分析，提供以自然语言对话、语音合成对话等形式，使关员有更良好的知识检索、知识应用体验 [3]。

2 海关领域知识图谱构建

知识图谱是当前很受热捧的人工智能研究方向。从概念上讲，知识图谱是由各类本体相互连接而成的语义网络，它基于图数据库，本质上是一张具有 N 个节点、M 条边的图。它能在现有 web 基础之上构建一层覆盖网络，在 web 表达上建立概念之间的语义链接关系，从而将网络上各种信息组织起来，成为可以被利用的知识。在海关知识库平台中，可利用动态本体知识图谱构建技术，基于对现有数据的再加工、结构化，逐步形成一个统一的、逻辑上全局的海关行业性知识库。

本体最早起源于哲学上的一个概念：本体是客观存在的一个系统的解释或说明，是客观现实的抽象本质。目前业界关于“本体”的定义，已被人工智能赋予新的内容。（见表 1）

在上述定义中，人们引用最广泛的是 1998 年 Studer 提出的：本体是共享概念模型的明确的形式化规范说明。这个定义体现了本体 O 的四层含义：概念模型 M、明确 U、形式化 F 和共享 R，可以表达为：$O=\{M(),U(),F(),R()\}$[4]。

在海关业务领域，我们认为“概念模型”是指通过抽象出海关业务中的具体概念而得到的模型 MC，比如法人机构、口岸、商品等；“明确 U”是指所使用的业务概念及使用这些业务概念的约束都有明确的定义 U_C，比如针对商品，应有海关领域的具体约束，区别于其他行业对商品的不同定义；“形式化”是指本体 O_C 是能被计算机处理的 F_C，即能够结构化描述；“共享”是指海关业务本体体现的是行业认可的知识，反映的是领域中公认的概念集 R_C，以下是海关业务领域的本体定义：

$$O_C=\{M_C(\cdot),U_C(\cdot),F_C(\cdot),R_C(\cdot)\}$$

通常本体所展示的逻辑或概念是相对稳定的，甚至是固定的。但在实际行业应用中，本体的概念和外延并非一成不变，往往需要随着时间、政策调整、业务变更等因素进行动态调整。因此，我们在海关知识库项目中，运用动态本体理论，将通过基于动态本体的知识图谱构建技术，来实现领域知识图谱的构建。

动态本体是指动态的本体结构，它可以在本体部署应用之后，仍然可以不断地进行修改。基于动态本体的知识构建是目前本体研究的热点课题，不少机构对动态本体的知识建模、构建流程、本体构建标准等进行了较为深入的研究，但目前国内外学者尚未形成统一的认识 [5]。

本文主要从海关业务领域，通过对本体的概念模型 MC、约束定义 UC、本体的形式化 FC 以及领域公认的概念集 RC 进行领域化处理，尝试通过对领域本体对象根类型定义来促进动态本体的知识图谱构建。

通过对海关行业的海量数据分析，我们首先归纳出两种基本的数据对象：实体对象和事件对象。其中，实体对象一般是作为主体存在的，和我们现实世界中

有着明确的实体对应关系。在海关业务中，实体是各业务环节中的关键要素，比如进出口食品监管环节，实体主要包括企业、食品、国家 / 地区、口岸等，数据来源基本以特定业务系统数据为主；事件对象则通常是某实体的行为集，或者某几个实体之间的事件关系集，在海关业务中，事件对象是对业务行为的具体描述，比如针对某个物品的查验业务，查验环节就是事件描述，它发生在物品、企业和查验机构等实体关系之间，具有查验事件、查验过程、查验结果等属性，其数据来源也以业务系统数据采集为核心。

对于海关行业领域，文本、图片、视频等数据所占的比例非常高。比如众多的海关相关政策法规、业务指南、图书文献等，基本以文本格式存在的数据为主，用户往往需要从大量文本中进行业务知识的提取、分析，这些业务知识对充实实体或事件对象的描述非常重要。因此，我们认为应将文本对象当作海关行业的一种基础对象类型，研究将主要围绕基于 NLP 技术的海关领域智能分词、文本智能识别等方面展开，其数据来源包括海关内部各管理系统产生的文本文档、资源库中存储的大量非结构化文本数据或者从互联网爬取的部分补充描述数据等。

除此之外，在有害生物监管、进出口食品安全监管等领域，会产生大量的图片、音频、视频等数据，在 AI 分析技术能力不断提高的基础上，用户也需要从大量现场图片、样本图片、监控视频、语音录音等数据中进行业务知识提取，比如从现场拍摄的物品照片中对商标进行识别，从而验证该物品的归属，并与特定企业进行关联管理。因此，我们认为应将图片对象、音视频对象也作为重要对象类型加以研究 [6]。

基于以上分析，海关知识图谱领域的本体对象基本就可以总结为五种基本类型：实体对象 C_E、事件对象 C_V、文本对象 C_T、图片对象 C_P 和音视频对象 C_M，它们同样具有继承性、封装性、多态性等对象特征。比如实体对象作为父类，可以扩展出机构类实体、人员类实体、商品类实体等子类，并可通过继承关系进行约定和描述。

为了在知识库平台中形成对知识的统一规范性描述，我们为五种对象设立了一个知识本体根对象 C_R，这五种类型的对象都从该本体对象 C_R 向下进行扩展、继承，从而构成海关知识图谱的动态本体表达 D。以下是海关业务领域的动态本体表达：

$$D=\{\{O_C:M_C(\cdot),U_C(\cdot),F_C(\cdot),R_C(\cdot)\};\{C_R:C_E,C_V,C_T,C_P,C_M\}\}$$

3 平台建设展望

海关知识库平台应用动态本体技术进行知识图谱构建，还需重点考虑动态本体的构建标准问题。目前，关于动态本体的构建标准大多使用本体评估方法，我们认为应该系统性地提出动态本体构建的指标评价体系。

动态本体指标评价体系的建立，需要综合考虑海关知识库平台的定位，以及整个海关业务知识的运营体系构建。通过组织、制度、管理和技术等措施，从知识应用、知识资源、能力服务三个层面逐步推进海关动态本体指标评价体系的建设：

3.1 知识应用是抓手

“三库四用”的特色应用体系应进一步专业化，为业务提供知识应用的“业务中台”组件，不断丰富知识应用的场景，增强平台的业务价值和用户黏性，以知识应用为抓手推动动态本体的指标评价标准。

3.2 知识资源是关键

通过组织、制度手段，确立知识库平台对各业务条线数据的汇聚职责和权利，从根本上保证各业务条线数据对知识库的数据供给，促进知识库的“数据中台”组件建设，并通过与知识应用良性互动，推动建立、完善海关知识图谱的动态本体评价指标集。

3.3 知识运营是根本

不同于普通的业务应用软件，知识库不仅需要技术上的运维保障，更需要进行业务上的运作。建立专门的知识运营管理机构、工作机制和配套制度规范，梳理、构建海关知识体系，协调海关知识资源基础数据，保障知识图谱等的顺利构建，并有力推动动态本体评价指标体系的落地。

参考文献：

[1] 罗钧旻，王蕾 . 基于互表性的动态本体体系结构研究 [J]. 微电子学与计算机 ,2013,30(02):124-127.
[2] 王茜 . 基于文本挖掘的动态本体构建方法研究 [D]. 北京 : 中国农业大学 ,2007.
[3] 中国中文信息学会 , 语言与知识计算专委会 . 知识图谱发展报告 (2018):VI.
[4] 王美琴，吴庆斌 . 基于本体的医学知识库构建方法综述 [J]. 医学信息学杂志 ,2017,38(03):73-76.
[5] 樊小辉，石晨光 . 本体构建研究综述 [J]. 舰船电子工程 ,2011,31(06):15-18,53.
[6] 郑姝雅，黄奇，张戈，等 . 面向用户生成内容的本体构建方法 [J]. 情报科学 ,2019,37(11):43-47.