arXiv:2301.07676v1 [cs.DB] 18 Jan 2023

定量档案研究中的整体数据管理与语义互操作性工作流模型

Pavlos Fafalios[1],Yannis Marketakis[1], Anastasia Axaridou[1],Yannis Tzitzikas[1,2],和 Martin Doerr[1]

[1] 信息系统实验室，FORTH-ICS，希腊伊拉克利翁[2] 计算机科学系，克里特大学，希腊伊拉克利翁{fafalios, marketak, axaridou, tzitzik, martin}@ics.forth.gr

摘要。 档案研究是一项复杂的任务，涉及从一组档案文献中提取证据和知识的多种不同活动。这些活动通常在数据连接与流转方面彼此脱节，导致难以对其进行递归修订和执行，也难以在数据元素层面检查来源信息。本文提出了一种用于档案研究整体数据管理的工作流模型，涵盖从转录并记录一组档案文献，到整理转录数据、将其整合到丰富的语义网络（知识图谱）中，再到对集成数据进行定量探索。该工作流具备来源感知能力、高度递归，并着重于语义互操作性，旨在产出高价值且长期有效的可持续数据。我们详细介绍了工作流各步骤的实现，并展示其在海事史研究中的应用。我们还讨论了相关的质量方面以及在实际情境中应用该工作流所获得的经验教训。

1 引言

档案研究是一种通过调查并从通常保存在图书馆、博物馆或其他组织中的档案记录中提取证据的研究。在最经典的意义上，档案研究涉及对历史文献的研究，因此它是原创历史研究的核心（Ventresca 和 Mohr，2017年）。

大量研究关注对具有重复性 结构的档案文献进行研究，例如登记册、航海日志、工资册、人口普查等；这些文献提供有关一种或多种实体类型的信息，例如人物、地点、对象、组织等。在这种情况下，研究通常首先收集一组与某个领域相关的档案文献，

然后对其进行转录和整理，以便对实证事实进行定量（也包括定性）分析，并描述和解释其可能的成因、影响及演变趋势（Petrakis等，2020年）。

在此背景下，常见的数据管理问题包括：应转录哪些数据，以及如何转录？如何整理转录数据，以便开展有效的定量分析并提供更有效的探索服务？如何在通用模式下集成数据，以支持对信息需求的研究，而这类需求需要结合多个来源的数据？如何支持数据的长期保存与再利用？如何维护所有数据溯源信息？这些信息对于使用这些数据的研究发现的验证及其长期有效性至关重要。

例如，SeaLiT 项目[1]（欧洲研究理事会海事史领域启动基金项目）是一个真实案例。该项目研究从帆船航行向蒸汽航行的转变及其对地中海和黑海航海人口的影响（19世纪50年代至20世纪20年代）（Delis，2020年）。该项目的历史学家从五个国家的多个主管机构收集了大量不同类型、不同语言的档案文献，并对其进行了研究，例如船员名册、工资册和水手登记册。同一关注实体（如一艘船或一名水手）的补充信息可能存在于不同的档案文献中。例如，针对同一艘船，一个来源（账簿）可能提供其所有者的信息，另一个来源（海军舰艇登记名录）可能提供船舶的建造详情和特征（长度、吨位、马力等），而其他来源（船员名册）则可能提供有关该船航行及船员的信息。此外，还可能存在另一个来源（民事登记册），提供船员的补充信息，例如婚姻状况和既往职业。在此背景下，数据集成至关重要，因为它有助于历史学家找到需要结合多个来源信息才能回答的问题，例如“查明抵达特定港口的大型船舶上水手的国籍”。

此外，由于拼写错误、语言不同、字符无法辨认或使用缩写，同一实体（例如某个人）的姓名在不同来源中可能有所不同（例如“G. Schiaffino”“Gaetano Schiaffino”“Gaetano Schi- afino”）。此外，同一术语（例如职业或船舶类型）在不同来源中也可能有不同的名称（例如“brigantine”“brigantino”）。在这种情况下，数据整理，尤其是实体（实例）匹配和术语对齐，对于开展有效的定量分析（例如按职业对检索到的水手列表进行分组）至关重要。然而，同时，这类整理不得更改原始转录数据，因为原始数据对于核验至关重要，也关系到研究发现的长期有效性。

为应对这些问题，本文介绍了一种用于档案研究整体数据管理的工作流模型（如图1所示）。该工作流依赖研究人员（领域专家）与数据工程师（建模专家）之间的紧密协作，并侧重于语义互操作性，即计算机系统能够以明确且共享的含义交换数据的能力

[1] https://sealitproject.eu/

（Ouksel 和 Sheth，1999年），因为这种方法有助于生成高价值的可持续数据，使其能够在特定研究活动或项目之外得以扩展和重复利用。

该工作流根据真实用户的需求设计，并具备来源意识，也就是说，它保留每条数据的完整溯源链。为此，它将数据录入与数据整理及整合解耦。研究者可以回到转录文本或原始来源，检查某条信息的初始形式。该工作流还具有很强的递归性，支持修订转录、整理和整合步骤，例如因研究过程中获取了新知识而进行修订。与相关研究相比，我们以整体视角处理相关活动，并特别注重在微观层面（数据元素）维护来源信息；在开放科学时代，这对于可复现研究至关重要（Vicente-Saez and Martinez-Fuentes，2018年）。

我们展示了该工作流模型在海事史领域一个真实案例中的实现，并报告了其应用的实证结果，说明该模型如何满足一大群历史学家的真实信息需求。我们还讨论了相关的数据质量方面和经验教训。

本文其余部分安排如下：第 2 节介绍所需的背景知识并描述相关工作。第 3 节概述所提出的工作流模型及其主要特征。第 4 节详细说明如何实现工作流模型的各个步骤。第 5 节介绍工作流的自动化。第 6 节描述一个实际用例。第 7 节讨论质量方面及相关经验教训。最后，第 8 节总结全文并概述未来工作。

2 背景与相关工作

我们首先介绍语义技术的基本概念（第 2.1 节），并回顾这类技术在人文研究中的应用，其中很大一部分涉及档案研究（第 2.2 节）。接着，我们聚焦于档案研究中旨在实现语义互操作性的各项数据管理活动，并介绍相关工作（第 2.3 节）。最后，我们说明本研究的定位（第 2.4 节）。

2.1 基本概念

语义技术旨在帮助机器理解数据。RDF（资源描述框架）[2] 和 OWL（Web 本体语言）[3] 是关键的语义技术，可用于编码数据的语义，从而正式表述信息中蕴含的含义（Antoniou 和 Van Harmelen，2004年）。这种表述的形式是一种语义网络

[2] https://www.w3.org/TR/rdf11-concepts/[3]
https://www.w3.org/TR/owl2-overview/

（或知识图谱），它以图结构存储相互关联的“实体”（对象、人物、事件、概念等）的描述，其中顶点代表实体，边代表实体之间的语义关系。典型的标准化语义网络以 RDF 三元组（形式为主语-谓词-宾语的陈述）表示，并存储在语义存储库（RDF 三元组存储库）中（Ali等，2021年）。语义技术有助于实现语义互操作性，即计算机系统能够以明确且共享的含义交换数据的能力；这是实现机器可计算逻辑、推理、知识发现以及信息系统之间数据联邦的必要条件（Ouksel 和 Sheth，1999年）。

2.2 人文研究中的语义技术

人文学科领域越来越多地采用语义技术，重点在于如何生成并公开发布可互操作的关联数据（Heath 和 Bizer，2011），使其能够轻松查询，并与其他数据集整合（Hyv¨onen，2020；Hyv¨onen 等，2014年；Hawkins，2021年；Beretta，2021年；Fafalios 等，2021a）。

Oldman 等（2015年）深入探讨了语义技术和关联数据理念在人文研究中的应用，并介绍了推动更广泛采用这些技术的策略，以支持高质量的数字人文项目，生成能够更好地表达人类知识、更加符合人文学科研究人员需求的数据。Hawkins（2021年）考察了有关档案的关联数据如何惠及从事数字人文研究与学术工作的人员，并探讨了目前阻碍数字人文学者利用数字化档案和原生数字档案的一些障碍。

我们认为，我们提出的工作流模型，尤其是其在数据元素层面具备来源意识，是解决前述文献所述若干重大问题的第一步，例如“将知识的来源追溯至微观层面（同时保留其原始语境和视角）”（Oldman 等，2015年，第10页），或“防止档案被去语境化并失去细微差别”（Hawkins，2021年，第11页）。

对于历史研究而言，档案研究是其中的核心组成部分。Mero˜no-Pe˜nuela 等人（2015年）调查了历史学家与计算机科学家在运用语义技术方面的协作。该文全面分析了知识建模、文本处理与挖掘、搜索与检索以及数据集成方面的研究成果和系统。文中还探讨了语义技术在历史研究中可进一步发挥作用的方面。其中之一便是“非破坏性数据转换”（Mero˜no-Pe˜nuela等人，2015年，第22页）。将数据录入与数据整理和转换分离，并在这些流程之间维持循环工作流，是所提出的工作流模型的核心特征，有助于实现这一目标。

2.3 档案研究中实现语义互操作性的数据管理

档案研究中实现语义互操作性的常见数据管理活动包括：

– 档案文献的数字化／转录（扫描文献、文本识别、人工转录）– 文档编制／元数据记录（文献的来源是什么、文献的内容是什么、由谁转录等）– 数据整理／为统计分析准备数据（修正或规范化数据值、实例匹配、术语对齐等）– 使用通用表示语言进行数据集成（基于本体的建模、创建映射、数据转换）– 数据发布（例如，作为关联数据发布）– 数据分析与探索（定性和／或定量分析、构建查询、数据可视化等）

针对上述每项活动，都有大量软件工具和系统可供使用。下文将介绍一些与人文研究相关的工作。

数字化／转录。可以使用文本识别软件自动从历史文献中提取文本，也可以手动进行转录；这两种方法各有优缺点。例如，自动化方法通常需要大量训练数据，其有效性（结果质量）很大程度上取决于待提取文本的类型和质量，以及训练数据的数量。另一方面，人工转录能提供高质量的结果，但需要投入大量精力。一种混合方法是将自动提取与人工校正和数据录入相结合。在软件工具方面，Transkribus（Kahle *et al.*，2017年）是一个广受欢迎的历史文献数字化平台，提供人工智能驱动的文本识别。FastCat（Fafalios *et al.*，2021b）是一款基于模板、用于人工协作转录的网络应用。它以表格形式（表格）组织数据（及元数据），类似于电子表格，为数据录入提供快速、易用的方式。

文档记录 / 元数据记录。 实现语义互操作性的文档记录主要有两种方法：a）将文档记录流程与基于本体的集成及语义网络的构建分离；b）从一开始就构建语义网络，即在文档记录过程中构建。Synthesis（Fafalios 等，2021年a）是一个基于网络的系统，采用第一种方法，以协作且科学的方式记录文化实体（对象、事件、人物、组织等），并提供内嵌流程，将数据转换为基于本体的 RDF 数据集。ResearchSpace（Oldman 和 Tanase，2018年）和 WissKI（Scholz 和 Goerz，2012年）是采用第二种方法的平台，支持对（元）数据进行直接的本体表示。

Microsoft Excel 等电子表格软件以及 Microsoft Access 等关系数据库管理系统（RDBMS）仍然是（元）数据录入和分析的常用工具（可能也是占主导地位的工具），并被广泛用于手动文档记录和元数据记录。此外，还有一些基于关系数据库管理系统的系统，例如 HEURIST[4] 和 nodegoat[5], ，专为人文学科研究者设计，集成了一系列用于构建和管理研究数据集的功能，但并不侧重于语义互操作性。

数据整理。 当需要对转录数据进行定量（统计）分析时，通常会进行这一可选步骤。在这种情况下，数据整理非常重要，因为数据质量会影响分析结果的可靠性。OpenRefine[6] 是一款常用的数据清洗桌面应用程序。它以数据行作为处理对象，每行的单元格位于各列之下（类似于关系表）。Silk[7] （Volz 等，2009）是一个开源框架，用于查找相关数据项之间的链接，例如进行实例匹配。它提供了一种声明式语言，用于指定链接规则，并支持通过$owl:sameAs$ 或其他类型的链接生成资源描述框架链接。对于全自动实例匹配（实体消歧），有大量基于学习的方法，这些方法需要人工或自动生成的训练数据（ Christophides 等，2020年）。最后，FastCat 系统（Fafalios 等，2021b）提供了一个名为 FastCat Team 的网络环境，支持自动（基于规则）和人工实例匹配，以及词汇整理流程。所应用的数据整理不会更改原始（转录）数据，并会保留整理后数据指向原始数据的链接。

数据集成。 此处的目标是使用领域（形式）本体（作为通用表示语言）对所有数据和元数据进行语义化表示，从而实现语义互操作性，使数据能够用于特定研究问题或项目之外的用途。这项工作包括数据建模和数据转换流程。数据建模包括定义或选择领域本体并创建模式映射；数据转换则根据模式映射转换数据，并创建集成数据的语义网络。

在软件系统方面，Prot´eg´e 是一款流行的本体编辑器，提供图形用户界面以定义本体。它可用于针对给定领域创建新的 OWL 本体，或扩展现有本体。对于模式映射的创建与执行，R2RML[8] 是一项 W3C 标准，用于将关系数据库映射为资源描述框架（RDF）；而 Dimou 等（2014年）则介绍了一种名为 RML 的扩展，用于将异构数据源映射为 RDF。最后，X3ML 工具包（Marketakis 等，2017年）提供了一种声明式（基于 XML）的映射定义语言，以及一组用于创

[4] http://heuristnetwork.org/[5] https://nodegoat.net/[6] https://openrefine.org/[7] http://silkframework.org/[8] https://www.w3.org/TR/r2rml/

建和维护模式映射，并将数据实际转换为 RDF 的工具。

数据发布 现在可以将集成数据导入语义存储库（RDF 三元组存储库），该存储库可以公开访问，也可以设为私有，并提供应用程序编程接口（API），以便访问数据并使用 SPARQL[9] 协议和语言运行结构化查询。随后，可以基于此 API 构建易于使用的应用程序，帮助最终用户探索和分析集成数据。也可以遵循关联开放数据（LOD）原则（Heath 和 Bizer，2011）将数据发布为关联数据。Sampo 模型[10] （Hyv¨onen 等，2014年）为协作发布和使用 LOD 提供了一个框架；该框架已通过构建所谓的“Sampo 门户”在多个领域得到测试（Hyv¨onen 等人，2020年）。

数据探索与分析。 探索集成数据主要有两种通用方法：（a）自由文本搜索：用户提供一组关键词或一个自然语言问题，以临时信息检索的方式进行查询；（b）交互式界面：系统通过易于使用的交互式界面，帮助用户表达信息需求。在这两种情况下，结果通常都是一个实体排序列表，用户可以从中开始探索相关信息，例如通过浏览、分面搜索，或使用图表、地图、时间线等不同的可视化方式。

在 RDF 数据上实现关键词搜索的方法多种多样，例如使用以文档为中心的信息检索系统（Kadilier- akis 等，2020年），或将关键词查询转换为结构化查询（SPARQL）（Izquierdo 等，2021年）。对于关键词搜索结果的呈现，Nikas 等（2020年）提出了一种多视角方法，提供多种呈现方法（视角），让用户能够轻松切换视角，从而充分利用每种视角的附加价值。在交互式界面方面，A-Qub（Kritsotakis 等，2018年）和 ResearchSpace（Oldman 和 Tanase，2018年）提供了用户友好的环境，支持最终用户逐步构建复杂问题（对应于 SPARQL 查询），以关联不同类型的实体和信息。

2.4 定位

据我们所知，目前尚无相关研究以整体性的方式处理档案研究中的数据管理部分；这里所说的整体性，是指所提出的工作流模型能够表示并高效管理信息，运用语义数据集成工具构建丰富的档案数据知识图谱，同时保留完整的溯源链，使研究人员能够从最终的语义集成集合追溯至原始手稿和转录手稿，反之亦然。

[9] https://www.w3.org/TR/sparql11-overview/

[10] https://seco.cs.aalto.fi/applications/sampo/

3 工作流模型：概述与主要特征

我们首先概述工作流模型（第 3.1 节），然后着重介绍其独特特征（第 3.2 节）。

3.1 角色、输入/输出与流程

图1 展示了为支持档案研究中的整体数据管理而提出的工作流模型。

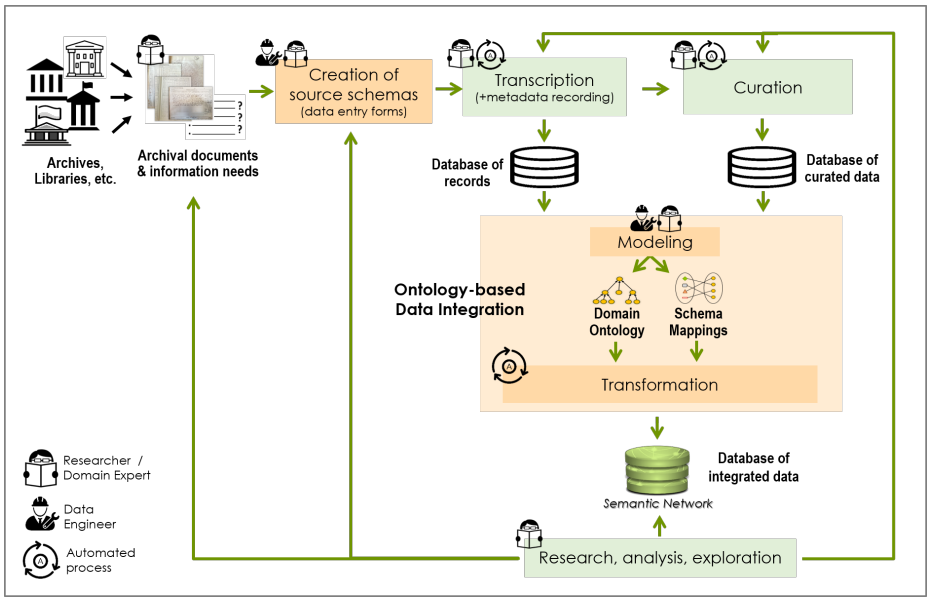


图1. 用于档案研究中整体数据管理与语义互操作性的工作流模型。

角色。 工作流中涉及两个主要角色：a) 研究者（领域专家/最终用户），负责收集和研究档案资料、提供领域知识并确定需求；b) 数据工程师（建模专家），负责设计和实施不同的工作流流程。

输入/输出。工作流的输入是一组由一名或多名研究人员从不同主管机构收集的档案文献，以及研究人员提供的信息需求。这些信息需求与其研究目标相关，且收集到的档案资料能够为其提供重要信息（证据）。在这一阶段收集信息需求非常有用，因为这有助于数据工程师更好地设计和实施后续工作流流程。工作流的输出是一个丰富的语义网络（一个知识图谱），其中包含集成信息，供研究人员用于

数据分析和探索，以及两个彼此独立的中间数据库：记录数据库（原始转录文本）和整理后数据数据库（整理后的实体实例和词汇术语）。

流程1：创建源模式。根据 Calvanese 等人（1998年）提出的基于描述逻辑的信息集成框架，我们首先需要创建源模式，每种不同类型的来源各创建一个，以便在软件系统中提供所需的数据录入表单，用于转录和记录原始档案文献。此步骤有助于根据源模型的语义进行数据整理和整合，也便于在通用本体下开展建模和集成；该本体可在研究过程中修改，而不会因此影响或延误转录流程。在此过程中，研究人员与数据工程师密切合作十分重要，这有助于正确设计模式，并避免数据录入时出现可能给后续步骤带来困难或限制的错误。此类错误的一个例子是使用同一个数据录入字段来记录计量单位和数值。这很可能会给最终用户在数据探索过程中进行比较时造成问题。

如果研究人员收集了不同来源类型的新档案文献，并需要对其进行转录，就必须创建新的源模式，或修改/扩展现有模式。这可能发生在整个流程的任何阶段，并且不会影响针对已收集材料并行运行的其他流程。

流程2：转录。为收集到的档案文献创建一个或多个源模式后，研究人员便可使用提供所需数据录入表单的软件系统开始对文献进行转录。除了转录文献的重要内容，此步骤还包括记录文献的元数据信息（档案馆/图书馆、年代等）以及转录过程的元数据信息（由谁进行转录等）。转录流程的结果是一个转录文本数据库。这项工作完全由研究人员小组负责，但可以借助软件工具来简化或自动化转录，例如文本识别软件。

流程3：整理。工作流的下一步是对转录数据进行整理。在此阶段，研究人员需要协调转录文本中出现的不同数据元素并解决身份歧义，使指代同一现实世界实体/概念的不同元素获得相同的标识符，并消除错误的共同指代关系。

数据元素可分为两大类：(a) 共相；属于特定词汇表或术语叙词表的概念实例，例如职业、对象类型等；以及 (b) 殊相；属于特定类别并附带特征/属性的实体实例，例如人物 （名字、姓氏、出生日期等）、地点 （名称、类型等）、组织 （名称、地点等）。整理还可以包括提供

更正后的/首选值（例如更正人物实例的名字）或实体丰富（例如为地点实例添加坐标）；这些任务通常有助于外部服务更好地利用和可视化经整理和集成的数据。

数据整理过程是由研究人员团队执行的一项任务，可能包括人工和自动化步骤。例如，实体的实例匹配或词汇术语的对齐，可能同时包含自动化步骤（基于规则）和人工步骤（用于验证有歧义的情况）。其结果是一个独立的整理后数据数据库，并与原始数据元素建立链接，这意味着数据整理步骤不会更改从原始来源转录的数据。

流程 4：基于本体的数据集成。下一步是对转录和整理后的数据进行基于本体的集成，其中包括建模和转换子流程。

在建模方面，良好实践建议使用成熟的领域模型（如果该应用领域有此类模型），或者创建一个与成熟的顶层本体兼容的新模型（专门化扩展）。这一过程通常需要领域专家与数据工程师进行深入讨论：领域专家了解数据，而数据工程师负责构建领域本体并创建映射。

建模过程中的一个重要环节是创建模式映射，用以描述如何将输入数据（转录文本和整理后的数据）映射到领域本体的类和属性。通常，当源模式数量众多且规模较大或结构复杂时，创建模式映射可能会耗费大量时间。不过，每种不同类型的数据源只需创建一次；若模式或目标本体发生变化，则可能需要修订。建议使用声明式语言定义映射，例如 X3ML（Marketakis 等人，2017年），因为数据源的局部变更只需对相应的映射规范进行局部修改，而且这些修改易于定位和执行。

转换过程以 i）数据库（转录和数据整理过程的输出）、ii）领域本体以及 iii）模式映射为输入，并生成丰富的集成数据语义网络。此步骤可以完全自动化；只要转录模式没有变化，每当有新的数据源完成转录和整理后，都可以重复执行。

流程 5：研究、分析与探索。研究人员通过一个或多个基于语义网络运行的服务，利用由集成数据构成的语义网络；这些服务提供用户友好界面，便于数据浏览、分析与探索。在此，最终用户能够回溯转录文本，甚至查看原始来源的扫描件，以检查某条信息的初始形式（即整理和转换之前的形式），或收集更多背景信息，这一点十分重要。此外，在研究过程中，用户可能会发现转录或整理后的数据需要更正，因此研究人员需要能够

重新执行转录和整理步骤、进行更正，然后（自动）再次转换数据，以更新语义网络。同样，任何时候都可能收集到新的档案文献，这意味着可能需要创建一个或多个新的源模式及相应的映射，以便对这些文献进行转录、整理和转换。

3.2 工作流的独特特征

下面我们重点介绍并说明所提出的工作流模型在设计和方法上的独特特征：

– 研究人员（领域专家）与数据工程师（建模专家）之间的紧密协作。这种协作有助于：a) 更好地设计源模式（以及相应的数据录入表单）；b) 更好地定义和设计目标（领域）本体并创建模式映射；以及 c) 更好地创建和配置数据探索服务的用户界面。

– 将数据录入与数据整理分开，并保留整理后数据与原始数据之间的关联。 这不仅对于维护数据溯源、核实信息，进而验证利用这些数据得出的研究发现非常重要，而且数据整理与整合可能存在歧义，需要在未来任何时候（由同一批或其他研究人员）开展进一步研究并反复修订。

– 将源模式创建与本体建模分开。 我们的目标是消除初始研究假设对目标（整合）模型造成的偏见。这是无偏研究中最棘手的哲学问题之一，也是科学实在论讨论的核心（Turner，2007年；Chapman 和 Wylie，2018年）。目标模型（本体）可以与数据录入过程并行开发，并可随时根据源数据带来的新见解进行调整，而不会使已录入的数据失效，也不会影响（或延误）转录和整理过程。

– 将数据库（存储转录文本和整理后的数据）与语义网络分离。将数据录入和整理与语义网络的创建解耦，有助于通过尽可能保留转录文本与原始（档案）文档的一致性（尽量维持其原有结构），来保持源模型的语义；同时也提供了一种熟悉的数据录入方式，可大幅加快这一耗时过程。此外，这还使得能够轻松生成不同版本的语义网络，以适配不同的本体，或同一本体的不同版本（这只需根据所需的目标模型创建模式映射即可）。

4 如何实现工作流

现在，我们将介绍实现该工作流的具体细节。

4.1 忠实、快速且协作式的数据转录

数据转录系统应满足的常见要求包括：

– 支持对档案文献中的信息进行忠实 且结构化 的转录（尽可能准确地还原原始信息），并记录元数据 信息。– 通过研究人员熟悉或能够快速熟悉的直观用户界面，支持快速 数据录入。– 支持多名研究人员协作式 转录，并使用相同的数据录入结构（源模式）。

这些特性会显著影响数据录入系统的易用性，进而影响最终用户（研究人员）对系统的接受程度。

为了开展后续的数据整理 流程，我们首先需要确定转录数据中出现且需要整理的主要实体类别（如人物、地点、对象等），以及主要词汇表或术语层级。为此，我们需要定义数据录入表单中的字段，以提供与实体或术语相关的信息。例如，数据录入字段名字 和姓氏 提供人物实例的信息，而职业 字段则提供一个词汇术语。这些字段的值必须复制（理想情况下自动复制）到一个新环境中，以便整理这些值，同时不改变其在转录文本中呈现的原始数据。随后，我们只需提供整理后的数据与原始数据之间的链接，和/或提供位置信息（例如记录名称、表格名称、行号），以保留来源信息。

4.2 来源感知的数据整理

需要专用软件系统支持的数据整理活动包括：

– 更正 实体的名称或其某项属性的值（通过设置首选标签）。– 实例匹配：匹配两个或多个指向同一现实世界实体的实体实例，这意味着它们必须具有相同的标识。– 实例取消匹配：将特定实体实例从一组自动匹配的实例中取消匹配，这意味着该实例将具有不同的标识。– 丰富化：为实体实例补充额外信息，例如为某个地点添加坐标。– 为词汇术语提供首选术语 （例如，来自固定叙词表的术语，或用英语为另一种语言中的术语提供对应术语）。

– 为词汇术语提供上位术语 （从而建立术语层级）。

在此语境下，实例匹配可以分为多个阶段。首先，可以通过自动化步骤，为一组具有某些共同特征的实体实例赋予相同的标识，例如，对于人物实例，可依据相同的名字、姓氏和出生日期（基于规则的方法），或者采用机器学习技术（监督式或半监督式方法）（Christophides 等，2020年）。随后，研究人员可以执行第二个人工步骤，将自动化步骤未能匹配的其他实体实例进行匹配，或者取消自动化步骤错误地将某个实体实例与其他实例进行的匹配。

实例匹配与取消匹配活动，以及为词汇术语提供首选术语，对于基于集成数据开展有效的定量（统计）分析至关重要。举例来说，研究海事史相关档案文献（如船员名册）的研究者，可能希望查找抵达特定港口的水手的出生地，或按其职业对他们进行分组。为所有代表同一现实人物的水手实例赋予相同的标识，并为对应于同一职业的不同职业提供相同的“首选”术语，可以确保生成的汇总信息准确无误。

4.3 基于本体的集成

转录数据和整理后的数据的基于本体的集成包括以下任务：

1. 使用领域本体进行数据建模。2. 创建模式映射，并定义如何生成实体标识符（统一资源标识符）。3. 执行转换，以生成集成数据的语义网络。

数据建模。CIDOC-CRM[11] (Doerr, 2003) 是一种描述时空中发生的人类活动、事物和事件的高层 ISO 标准本体（ISO 21127:2014）[12] ，因此可用于对转录数据进行建模，并支持语义互操作性和长期数据保存。根据应用领域的不同，可能需要扩展 CIDOC-CRM，以专门定义特定的关注概念。例如，在我们的用例中，我们创建了 SeaLiT 本体，这是 CIDOC-CRM 的扩展，用于对海事史相关数据进行建模和整合（详见第 6 节）。对于使用 CIDOC-CRM 进行语义数据管理，Tzitzikas 等人(2022) 分析了相关流程和任务，并回顾了应用机器

[11] http://www.cidoc-crm.org/[12]
https://www.iso.org/standard/57832.html

学习技术以降低基于 CIDOC-CRM 的合规与互操作性相关成本的文献。

标识符的映射与生成。此步骤定义如何转换转录文本和整理后的数据集，使其最终构建成语义网络。挑战在于保留完整的溯源链，从整理后的数据追溯至源转录文本的原始数据，以便研究人员轻松验证、进一步完善或查找更多信息。此步骤的第一部分是定义模式映射，确定输入模式中的哪些部分（例如特定表列）将映射到领域本体中的具体类和属性，从而确保原始数据的语义定义明确、无歧义，且不丢失任何数据。第二部分定义如何生成资源 URI 和标签。在此阶段，URI 将作为“黏合剂”，连接相关信息片段。

图 2 展示了一个如何使用统一资源标识符建立此类关联的示例。在这个例子中，有两条不同的转录记录，每条记录都描述了不同的人物。其中一条记录里有一位名为‘Agostino B??ndi’的人物（也就是说，问号表示无法从原始来源辨认出相应字符），另一条记录里则有一位名为‘A Brondi’的人物。由于这两个名字并不匹配，而且分别出现在不同的记录中，因此为这两个人物创建了两个不同的统一资源标识符。然而，在整理后的数据集中，历史学家一致认为这些指称指向同一个人。因此，系统创建了一个带有新统一资源标识符的人物实例，并将其与先前的实例关联起来。这个新实例称为‘主实例’，而与之关联的实例则称为‘本地实例’。

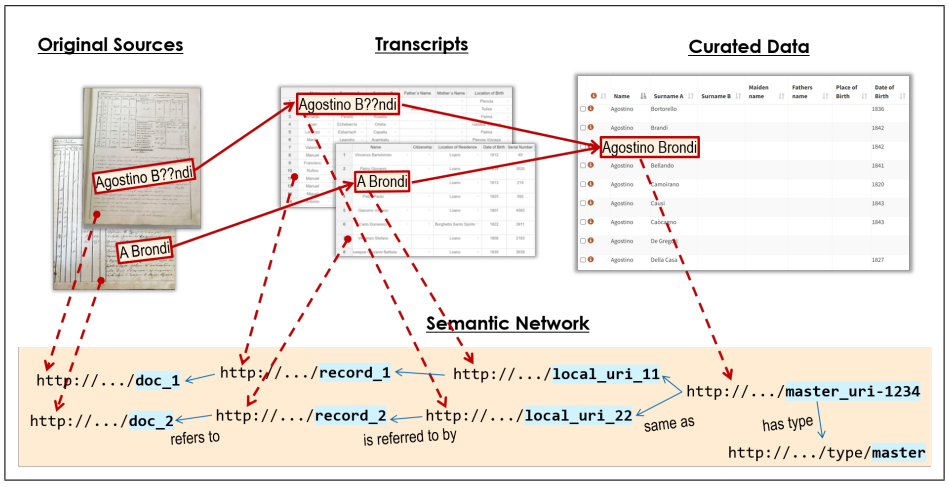


图 2. 标识（URI）管理与溯源链。

每个统一资源标识符由三部分组成：(a) 所有资源共用的统一资源标识符前缀，(b) 资源的类型或层级，(c) 资源的实际内容或哈希内容。一个示例统一资源标识符为：*https://rs.sealitproject.eu/kb/location/sardinia*。还应指出，在某些情况下并不采用上述策略。一个典型例子是在语义网络中构建中间节点，这类节点不需要统一资源标识符（例如‘E67 出生’事件）。在这种情况下，会为它们分配一个随机的通用唯一识别码。

转换。 此步骤以 (a) 已转录并整理的数据集，以及 (b) 模式映射和 URI 生成器的定义为输入，并根据领域本体生成本体实例（RDF 三元组），这些实例构成语义网络的核心内容。此步骤无需人工干预，可完全自动化。这种自动化的一个明显优势是，一旦有新数据完成转录，和/或研究人员整理了更多数据，就可以立即全部或部分刷新语义网络。

管理语义数据时，为灵活更新和版本控制，一个良好的做法是使用命名图（Carroll 等，2005），每条源记录对应一个命名图。当记录或其映射定义文件有新版本时，只需替换相应命名图中的 RDF 数据，即可轻松将新工作流周期生成的记录输出整合到语义存储库中。此外，术语层级和地点也可以在各自独立的命名图中得到有效管理和更新，语义推断语句的物化 过程结果（即生成作为捷径、用于表示长路径以提升查询性能的新 RDF 三元组）亦可如此管理。

4.4 语义网络的利用

现在可以将整合后的语义网络数据作为档案研究的主要来源加以利用。这包括为复杂的信息需求和分析查询寻找答案；这些查询需要综合来自不同来源的信息，并以表格、图表、时间线或地图等不同形式呈现结果，供研究直接使用。

实际的信息需求取决于应用领域，以及最终用户所需的探索或分析类型。这里的挑战在于，为研究人员提供用户友好、直观易用的界面，让他们能够放心地表达信息需求，并找到相关信息。因此，这类数据探索服务的关键成功因素是易用性和可信度。通过让用户直接检查所显示信息的来源，允许他们直接查看包含相关信息的转录文本，甚至查看原始档案文献的扫描件，可以提升可信度。

一些常见的信息需求类别包括：(i) 查找有关特定实体的信息，例如某个人的出生日期和出生地点；(ii) 根据这些实体的一项或多项属性检索实体列表（例如，

所有居住在特定地点的人物）；(iii) 根据某项属性或特征对检索到的实体列表进行分组（例如，按职业对所有检索到的人物分组）；(iv) 查找与某些实体相关的比较信息（例如，某组织在不同时间段雇用的人数）。

最后，制定一种处理数据中缺失值的策略非常重要，因为某些类型的档案文献中经常会出现缺失值。这样才能获得有效的汇总信息并得出可靠的结论。例如，原始文献中某些人物的居住地可能为空。按居住地对一组人物进行分组时，必须为这类信息缺失设置一个“未知”值。

5 工作流自动化

转录和数据整理过程中使用的系统需要相互通信，以便自动将需要整理的数据元素（实体、术语）从转录系统复制到数据整理系统。随后，只要建模过程已经完成，并且已为所有不同的源模式创建了映射（这些任务需要针对每种不同类型的档案文献各执行一次），工作流中的一个重要环节就可以实现完全自动化。在这种情况下，新转录和整理后的数据可以自动转换并导入集成数据的语义存储库，从而供终端用户直接通过数据探索应用程序进行探索。

具体而言，工作流场景如下：一组研究人员收集了第一批档案文献，并在专用系统中为每种不同类型的来源创建了数据录入表单。研究人员开始进行转录。当这批已收集档案文献的转录工作完成后，需要整理的数据元素会自动复制到整理环境中，研究人员随即开始整理这些数据。与此同时，数据工程师在研究人员提供的领域知识支持下，通过研究现有实物证据和专家需求，定义目标（领域）本体，并为每种不同类型的来源创建模式映射。当所有（或大部分）档案文献的转录和整理流程均已完成，且相应的模式映射也已创建后，研究人员便可以“发布”数据，也就是说，执行转换过程，并创建语义网络，将其载入语义存储库。随后，研究人员可以通过一个基于语义存储库运行的应用程序所提供的用户友好界面，开始探索集成数据。

研究人员可以随时转录并整理新的档案文献，也可以根据研究过程中获得的新知识，更正现有的（已整理）数据，然后重新执行转换过程，并自动更新语义存储库。语义存储库中的变更

会直接（并自动）反映在数据探索应用程序中。

研究人员所要处理的全部档案文献无需从一开始就确定，这意味着任何时候都可能收集新的文献进行转录。在这种情况下，如果这些新文献属于与现有类型不同的新来源类型，就需要创建新的源模式（数据录入表单）。因此，可能需要修改现有的数据录入表单（例如添加新列），以便转录某种重要的新信息；这类信息对于现有来源类型而言，原先并未计划或知晓。在这两种情况下，都可能需要修订或扩展领域本体，并创建新的模式映射或修改现有映射。

需要指出的是，即使转录模式和集成模型发生变化（项目进行过程中确实会出现这种情况），这些变化也独立于研究人员并行开展的其他转录和整理流程，因此不会影响或延误这些流程。此外，数据转换步骤的全面自动化将研究人员的额外工作量降至最低。

工作流中最耗时的两个步骤是转录和整理过程。正如前文所述，这两个过程中的若干子任务可以实现自动化或半自动化，例如使用最先进的文本识别软件（Kahle 等，2017年），或应用自动化实例匹配／实体消歧（Christophides 等，2020年）。这里的挑战在于，在任务完全自动化与获得高准确度结果以支持有效数据分析之间找到最佳平衡。我们建议采用半自动化解决方案，并纳入人在回路机制，以确保高质量（Wu 等，2022；Gurajada 等，2019）。

6 海事史研究中的用例

该工作流已在一个实际用例中全面实施，旨在帮助众多历史学家管理与海事史相关的多种档案来源。具体背景是 SeaLiT[13]，项目；在该项目中，海事史学家研究帆船向蒸汽航行的转变及其对地中海和黑海航海人口的影响（19世纪50年代至20世纪20年代）。

下文将详细介绍工作流中各个流程的实施方式，并举例说明如何利用集成数据满足历史学家提出的实际信息需求。

档案资料与信息需求。SeaLiT 研究的档案资料涵盖多种来源，使用五种语言（西班牙语、意大利语、法语、俄语、希腊语），包括船员名册和排水量名册、不同类型的登记册（水手、军舰、学生等）、航海日志、工资册、账簿，

[13] https://sealitproject.eu/

就业记录和人口普查。有关完整档案语料库及其来源的详情，可在项目网站上查阅。[14]

我们的首要任务是收集 SeaLiT 历史学家的信息需求，这些需求与其研究目标相关，且所研究的档案资料能够为之提供重要信息。这对于更好地设计源模式（数据录入表单）、集成模型以及数据探索服务至关重要。我们收集了约 100 项信息需求。以下是一些示例：[15]

– 特定时期内船舶的建造地点有哪些？– 从时间维度来看，来自黑海的船舶最常前往哪些欧洲目的地？– 在抵达某个特定地点（例如巴塞罗那）的人当中，有多少人的出生地距此地超过 X 英里？– 特定时期内，每艘船有多少位船主？

源模式的创建与转录。 我们使用 FastCat 系统（Fafalios等，2021b）创建源模式，并由 5 个国家的约 30 位用户（SeaLiT 历史学家）转录档案文献。FastCat 是一款可用的开源软件[16], 。在 FastCat 中，用户可以转录文献，并通过创建属于特定“模板”的“记录”来提供元数据信息。“记录”以一组表格的形式组织档案文献的数据和元数据，而“模板”则代表不同数据源的结构，即定义数据录入表、表中的列，以及每一列的类型（用于标识提供词汇术语或实体相关信息的列，其值将在转录后进行整理）。针对 SeaLiT 项目，共创建了 20 个模板，每种不同类型的档案来源各设一个。表 1 列出了这些模板，并概述了每个模板中可记录的信息。

SeaLiT 的历史学家目前已转录的记录总数超过 620 条。图 3 展示了一个真实记录的部分内容，该记录属于船员名册（*Ruoli di Equipaggio*）[17] 模板（共有 98 条记录属于此模板）。该模板由六个表格组成，使历史学家能够提供/转录以下信息：i）记录本身（创建日期、最后修改日期、转录者）；ii）来源（档案馆/图书馆、地点、登记号、签发机构等）；iii）船舶（名称、类型、吨位、建造地点等）；iv）船员名单（登船港及日期、下船港及日期、姓氏、名字、居住地、职业、薪酬信息等）；v）有记录的航行（日期、持续时间、首个计划目的地、船员总人数）；vi）航线（出发港及日期、抵达港及日期）。例如，在图 3 的记录中，转录者提供了二十

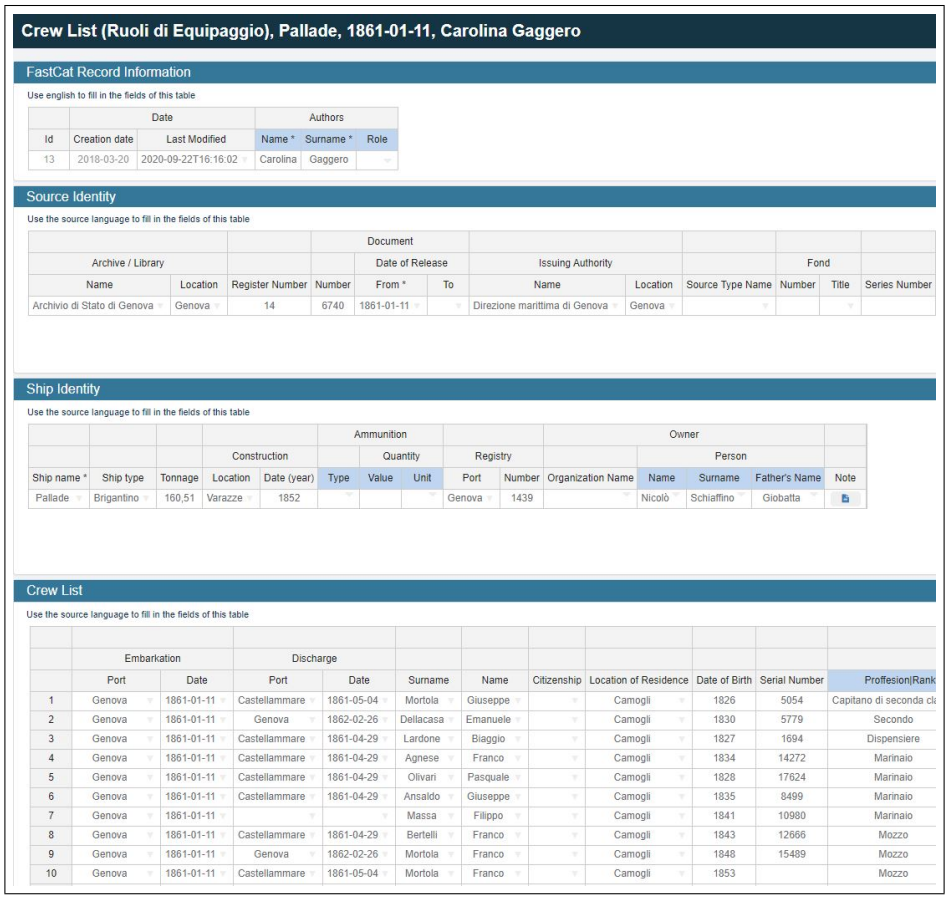
[14] https://sealitproject.eu/archival-corpus[15] 收集到的完整信息需求清单可在以下网址查阅：https://users.ics.forth. gr/~fafalios/SeaLiT_能力_问题_InfoNeeds.pdf[16] https://github.com/isl/FastCat[17] 完整记录可通过以下网址访问：https://tinyurl.com/2u35frya

表 1. 所考察的档案来源及所记录信息概览。

档案来源	所记录信息概览
船员及航行清单（Roll）船舶、船员和港口信息。	
船员名册（Ruoli di Equipaggio）	船舶、航行、船员和港口信息。
西班牙通用船员名册	有关船舶、船主、船员和航行的信息，港口。
水手登记册（Libro de registro de marineros）	有关水手的信息（包括职业和军事服组织地点）
海事人员登记册	人物信息（包括居住地、海事状态、以前的职业、兵役组织地点）。
海事工人登记册	有关海事工人、船舶、船长和港口的信息。
海上工作人员	有关人物（包括婚姻状况、职业、退役原因）、船舶和目的地的信息。
海军舰艇登记名录	有关船舶的信息（包括吨位、长度、建造地点、登记地）和船主。
船舶名录	船舶信息（包括曾用名、登记港和年份、建造地点和年份、吨位、发动机特性、所有者）。
民事登记册	人物信息（包括职业、原籍地、婚姻状况、死亡地点及死因）。
海事登记册，拉西奥塔	人物、登船和下船地点、船舶、船长、赞助人。
学生登记册	有关学生和课程的信息。
拉西约塔人口普查	有关居住者的信息（包括国籍、婚姻状况、宗教、职业、工作单位、家庭角色）。
俄罗斯帝国人口普查	有关居住者的信息（包括婚姻状况、社会等级、宗教、母语、家庭角色、职业）。
希腊船舶工资册	有关船舶、船长、航行、人物、雇佣——关系（包括工资）的信息。
（俄罗斯蒸汽航运与贸易公司工资册——tion and Trading Company)	有关船舶、人物、招聘的信息（包括薪水月薪）。
就业记录（造船厂：Messageries Maritimes，拉西奥塔）	有关工人的信息（包括婚姻状况、职业、公司任职状况）。
航海日志	有关船舶、船长、港口、航线行程和航行事件的信息。
账簿	有关船舶、航行、船长、港口和交易的信息。
公证文书	关于契据、公证人、证人、签约各方的信息，船舶。

六名水手和十三个航线港口，涉及该船*Pallade*（类型为双桅帆船）于1861年11月1日至1862年2月26日期间的航行。

FastCat中模板的创建和配置并非一次性完成的过程。新模板会根据历史学家收集的新档案资料定期创建；即使在记录创建之后，现有模板也可能被多次修改（例如在表格中加入额外的列），以纳入特定档案文献所提供的新型（且重要的）信息。



Crew List (Ruoli di Equipaggio), Pallade, 1861-01-11, Carolina Gaggero

FastCat Record Information

Use english to fill in the fields of this table

	Date		Authors		
Id	Creation date	Last Modified	Name *	Surname *	Role
13	2018-03-20	2020-09-22T16:16:02	Carolina	Gaggero	

Source Identity

Use the source language to fill in the fields of this table

			Document							
Archive / Library			Date of Release		Issuing Authority			Fond		
Name	Location	Register Number	Number	From *	To	Name	Location	Source Type Name	Number	Title	Series Number
Archivio di Stato di Genova	Genova	14	6740	1861-01-11		Direzione marittima di Genova	Genova				

Ship Identity

Use the source language to fill in the fields of this table

					Ammunition					Owner				
			Construction		Quantity			Registry		Person				
Ship name *	Ship type	Tonnage	Location	Date (year)	Type	Value	Unit	Port	Number	Organization Name	Name	Surname	Father's Name	Note
Pallade	Brigantino	160,51	Varazze	1852				Genova	1439		Nicolò	Schiaffino	Giobatta	

Crew List

Use the source language to fill in the fields of this table

	Embarkation		Discharge									
	Port	Date	Port	Date	Surname	Name	Citizenship	Location of Residence	Date of Birth	Serial Number	Proffesion|Rank
1	Genova	1861-01-11	Castellammare	1861-05-04	Mortola	Giuseppe		Camogli	1826	5054	Capitano di seconda cl
2	Genova	1861-01-11	Genova	1862-02-26	Dellacasa	Emanuele		Camogli	1830	5779	Secondo
3	Genova	1861-01-11	Castellammare	1861-04-29	Lardone	Biaggio		Camogli	1827	1694	Dispensiere
4	Genova	1861-01-11	Castellammare	1861-04-29	Agnese	Franco		Camogli	1834	14272	Marinaio
5	Genova	1861-01-11	Castellammare	1861-04-29	Olivari	Pasquale		Camogli	1828	17624	Marinaio
6	Genova	1861-01-11	Castellammare	1861-04-29	Ansaldo	Giuseppe		Camogli	1835	8499	Marinaio
7	Genova	1861-01-11			Massa	Filippo		Camogli	1841	10980	Marinaio
8	Genova	1861-01-11	Castellammare	1861-04-29	Bertelli	Franco		Camogli	1843	12666	Mozzo
9	Genova	1861-01-11	Genova	1862-02-26	Mortola	Franco		Camogli	1848	15489	Mozzo
10	Genova	1861-01-11	Castellammare	1861-05-04	Mortola	Franco		Camogli	1853		Mozzo

图3。 属于“船员名册（Ruoli di Equipaggio）”模板的一个真实FastCat记录示例。

整理。 转录数据（词汇术语和实体实例）的整理通过 FastCat 中一个名为 FastCat Team 的专用环境进行。具体而言，当历史学家完成一份或多份文档（记录）的转录后，便可以“发布”该记录，这意味着

所有涉及词汇术语和实体实例的数据都会被复制到 FastCatTeam，以便对其进行整理。

以 SeaLiT 为例，目前共有五十二个词汇表（例如：船舶类型、发动机类型、职业、婚姻状况）；可供整理的实体类型（及当前数量）包括 船舶 （约 2,400 个）、人物 （约 99,200 个）、地点 （约 9,800 个）、法人实体 （约 1,100 个）。对于词汇表中的每个术语，用户可以提供首选术语（英文）和上位术语，也可以查看包含该术语的记录。在整理实体实例时，用户可以更正取值，选择两个或更多实例进行匹配（表示它们指向同一个现实世界实体），取消某个实例与一组自动匹配实例之间的匹配，或查看包含该实体实例的记录。对于地点，用户还可以添加标识符（TGN/GeoNames ID）、坐标或次要地名（例如历史名称）。

图4展示了 FastCat Team 的用户界面，特别是用于整理船舶实例的页面。关于 FastCat（以及 FastCat Team）的更多信息，读者可参阅 Fafalios 等人（2021年b）。

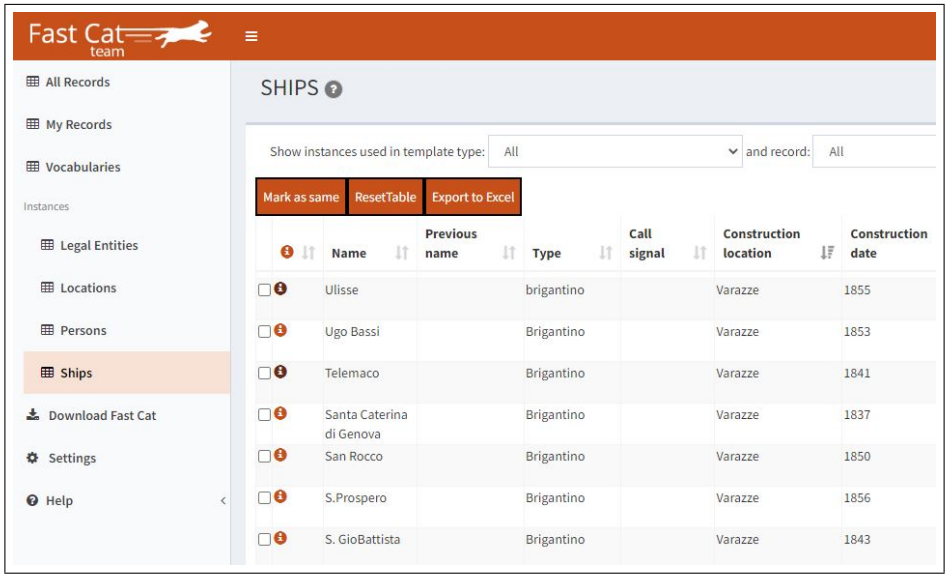


图4。FastCat Team 中船舶实例的整理。

基于本体的集成与转换。为了进行数据集成，我们创建了一个与 CIDOC-CRM 兼容的数据模型，称为“SeaLiT 本体”[18]。本体的当前版本（v1.1）包含四十六个类和七十九个属性，可用于描述有关船舶、航行的信息，

[18] https://zenodo.org/record/6797750

就业、支付、航海人员、教学课程以及其他相关活动。为了创建模式映射并将数据转换为资源描述框架，我们采用了 X3ML 框架（Marketakis 等，2017年）。具体而言，我们为 FastCat 中的每个模板创建了一个映射定义文件，为 FastCat Team 中的四类实体各创建了一个映射定义文件，并为所有词汇表创建了一个映射定义文件。

生成的语义网络包含超过 18.5M 个资源描述框架三元组，目前由数据探索应用程序（ResearchSpace；详见下文）使用，以帮助历史学家找到满足其信息需求的答案。完整的资源描述框架数据集已公开提供[19]。该网络包含数千名水手、船舶、地点、组织、航行及许多其他相关活动之间相互关联的信息，也与公开可用的资源（Geonames、盖蒂地名词典）相连接。

语义网络探索。为了让 SeaLiT 的历史学家和其他感兴趣的人士能够探索集成数据，并找到满足其信息需求的答案，我们使用 ResearchSpace（Oldman 和 Tanase，2018年）。ResearchSpace 是一个可配置的开源平台，可通过资源描述框架三元组存储库访问语义网络。它提供多种功能，包括查询构建界面，支持用户通过用户友好界面逐步构建并运行复杂查询。随后，用户可以浏览结果，并通过条形图等不同可视化方式进行定量分析。

该平台针对 SeaLiT 数据进行了配置，提供三项主要的数据探索功能：a）关键词搜索；b）语义搜索（通过辅助式查询构建界面）；以及 c）实体浏览（按档案来源类型浏览）。图5展示了语义搜索功能的屏幕截图。用户查看的是“1830年到1840年间建造的船舶的建造地点”。用户首先搜索了在1830年到1840年间建造的船舶（图5-A），然后选择按建造地点对检索到的船舶分组（图5-B），并以条形图展示结果（图5-C）。这一查询对应于 SeaLiT 的历史学家提出的实际信息需求，系统会立即向用户显示答案（不到一秒）。如果某艘船的建造地点未知，图表中也会显示这一信息缺失情况（见“未知”栏，图5-D）。用户还可以开始浏览检索到的船舶信息（例如查看某艘船的所有者，再查看同一所有者拥有的其他船舶），访问提供相应信息的 FastCat 转录文本（以进行验证或查看其他上下文信息），或将结果下载为 CSV 格式，以便进一步进行（外部）分析。

该应用的部署版本可公开访问。[20]

[19] https://zenodo.org/record/6460841

[20] http://rs.sealitproject.eu/

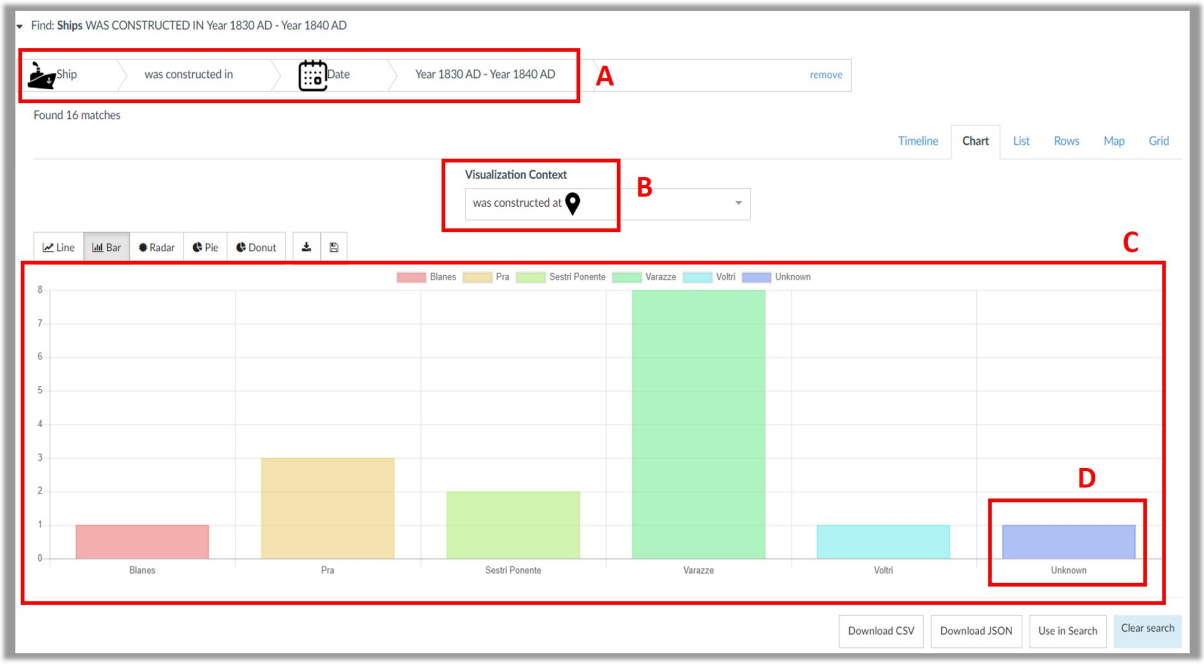


图5。 ResearchSpace 中的语义搜索与结果可视化。

7 质量方面与经验教训

我们探讨数据质量方面，以及将所提出的工作流模型应用于海事史研究时获得的相关经验教训。

7.1 质量方面

每个工作流周期结束时都会产生语义数据，而在某些情况下，这些数据可能存在质量特征较低的问题，使其在实际应用中难以满足研究需求。在相关文献中，数据质量通常被视为“适用性”，同时也是数据可用性的一个指标（Pipino 等，2002；Wang和Strong，1996）；此外，已有研究提出了用于衡量数据质量的多个维度和指标（Pipino 等，2002；Zaveri 等，2016）。尽管对质量因素进行详细研究超出了本文的范围，以下将重点讨论语义数据的三个主要质量维度，它们会显著影响定量分析过程：完整性、一致性和简洁性。

数据完整性。 这是一个质量维度，在模式/本体或特定用例场景的背景下很容易评估（Zaveri等，2016）。缺少必要信息，例如事件日期和地点缺失，或登记册中参与者的姓名和职业缺失，可能会影响研究分析，以及为历史主题作出判断所依据的证据。

数据一致性。 这一维度可以从多个角度来看（Zaveri 等，2016；Hassenstein 和 Vanella，2022）。我们的视角包括基于模式的一致性和基于值的一致性（或表征一致性）。基于模式的一致性可以依据特定模式/本体进行评估。它可以防止建模问题，例如实体之间不兼容的归属/关联，并避免潜在的推理失灵。例如，将“吨位”赋予一个人（而非船只）毫无意义；在特定推理前提下，这可能会导致不准确的推断，认为人被用于货物运输。基于值的一致性关注比较值（数字、日期、测量值）的格式和结构，以确保其可比性。若要有效比较量值、维度、数量、时间跨度、日期、地点坐标等，它们必须采用一致的参照点或计量单位。

数据简洁性。这一质量维度包含两个方面：模式级简洁性和实例级简洁性（Zaveri 等，2016；Mendes等，2012年）。模式级简洁性是指数据不包含名称不同但含义等同的属性（由数据建模工程师负责），而实例级简洁性是指数据不包含标识符不同但含义等同的对象（高度依赖数据整理过程的质量）。

7.2 经验教训

接下来，我们将介绍在实施工作流过程中遇到的数据质量相关问题，这些问题应当纳入考量。

信息缺失。缺失值十分常见，是研究人员需要了解的重要信息，因为它们会影响定量（统计）分析的准确性。这与上文所述的完整性质量方面有关。若原始来源未提供某项信息，则数据录入系统中的相应单元格留空。数据探索系统在汇总和展示信息时必须考虑这些空值。

数据录入错误。 在数据录入过程中，转录文本中常会出现错误，例如意外地将内容填入表格中的错误列，或因误解而将信息填错位置。这与上文所述的基于模式的一致性 这一质量方面有关。这类错误会直接反映在数据探索界面中，并损害用户体验。因此，应允许研究人员查阅原始转录文本，以进行验证或更正。此外，在用户界面中提供相应机制，帮助用户在数据录入过程中避免此类错误，也能减少这一问题。

不一致的比较值。不同类型的档案来源中，比较值（如日期、维度、数量、位置坐标）不一致的情况十分常见。这是由于参照点或计量单位不同，导致这些值难以用于比较、筛选等操作。这与上文所述的质量方面——基于值的一致性 有关。需要额外进行一个（自动、半自动或人工）步骤，

以对齐这些值，但不改变它们在原始来源中的呈现方式。这可以在数据整理或数据转换过程中完成。

数据整理成本高。 低质量的数据整理会降低用户满意度，并产生无效的分析结果。这与上述实例级简洁性质量方面有关。手动数据整理的成本取决于待整理数据的规模（实体实例数量、词汇术语数量）。对于 SeaLiT 这类实体和词汇表数量庞大的项目，整理过程可能会耗费研究人员大量时间。因此，需要借助工具尽可能实现整理自动化，同时不显著影响质量，例如采用半自动化流程、监督算法或特定应用的机器学习。

8 结论

我们提出了一种用于档案研究的整体数据管理工作流模型：从转录并编目一组档案文献，到整理转录数据、将其整合到丰富的语义网络中，再到探索和分析集成数据。该方法的优势在于加快数据录入；通过将数据录入与数据整理和集成解耦，实现对来源的感知；兼具交互性，并适用于语义互操作性，旨在产出高价值且长期有效的可持续数据。

我们展示了该模型在海事史研究中的可行性和有效性，并报告了其应用的实证结果（约30名用户、20种档案文献、600多条记录、50多个词汇表、11万多个实体实例，以及超过1850万条集成信息三元组）。

值得进一步研究的问题包括：(a) 加快数据整理的半自动化方法；(b) 探讨语义网络的演进需求，正如Marketakis 等人（2021年）所提出的；(c) 支持研究人员自行定义和更新源模式的方法与界面。

致谢
本研究获得了欧盟 Horizon 2020研究与创新计划的资助，资助依据为玛丽·斯克沃多夫斯卡-居里项目协议第890861号（ReKnow项目）；此外，还获得了欧洲研究理事会（ERC）项目协议第714437号（SeaLiT项目）的资助。

参考文献

Ali, W.、Saleem, M.、Yao, B.、Hogan, A.和Ngomo, A.-C. N. （2021）。关于用于查询知识图谱的资源描述框架存储库和 SPARQL 引擎的综述。*The VLDB Journal*，第1–26页。

Antoniou, G. 和 Van Harmelen, F.（2004）。《语义网入门》。MIT出版社。Beretta, F.（2021）。历史研究面临的一项挑战：利用协作本体管理环境（OntoME）使数据符合 FAIR 原则。《语义网》，12(2)：279–294。出版方：IOS Press。Calvanese, D.、De Giacomo, G.、Lenzerini, M.、Nardi, D. 和 Rosati, R.（1998年）。用于信息集成的描述逻辑框架。载于 KR，第2–13页。Carroll, J. J.、Bizer, C.、Hayes, P. 和 Stickler, P.（2005）。命名图。《网络语义学杂志》，3(4)：247–267。出版方：Elsevier。Chapman, R. 和 Wylie, A.（2018年）。《考古学中的证据推理》。布鲁姆斯伯里出版社。Christophides, V.、Efthymiou, V.、Palpanas, T.、Papadakis, G. 和 Stefanidis, K.（2020年）。大数据端到端实体消歧综述。《ACM计算调查》（CSUR），53(6)：1–42。出版方：ACM，纽约，纽约州，美国。Delis, A.（2020年）。《航海人生》与地中海海事史的交汇。《国际海事史杂志》，32(2)：464–478。出版方：SAGE出版公司，英国英格兰伦敦。Dimou, A.、Vander Sande, M.、Colpaert, P.、Verborgh, R.、Mannens, E. 和 Van de Walle, R.（2014年）。RML：一种用于异构数据集成资源描述框架映射的通用语言。载于 Ldow。Doerr, M.（2003）。CIDOC概念参考模型：实现元数据语义互操作性的一种本体方法。《人工智能杂志》，24(3)：75–75。Fafalios, P.、Konsolaki, K.、Charami, L.、Petrakis, K.、Paterakis, M.、Angelakis, D.、Tzitzikas, Y.、Bekiari, C. 和 Doerr, M.（2021a）。迈向历史研究中的语义互操作性：利用 Synthesis 编目研究数据与知识。载于国际语义网会议，第682–698页。Springer。Fafalios, P.、Petrakis, K.、Samaritakis, G.、Doerr, K.、Kritsotaki, A.、Tzitzikas, Y. 和 Doerr, M.（2021b）。FAST CAT：面向数字人文语义互操作性的协作式数据录入与整理。《计算与文化遗产杂志》（JOCCH），14(4)：1–20。出版方：ACM，纽约，纽约州，美国。Gurajada, S.、Popa, L.、Qian, K. 和 Sen, P.（2019）。用于实体消歧的人在回路学习方法。载于第28届ACM信息与知识管理国际会议论文集，第2969–2970页。Hassenstein, M. J. 和 Vanella, P.（2022）。数据质量——概念与问题。《百科全书》，2(1)：498–510。出版方：MDPI。Hawkins, A.（2021年）。档案、关联数据与数字人文：通过语义网提升数字化档案和原生数字档案的可访问性。《档案科学》，第1–26页。出版方：Springer。Heath, T. 和 Bizer, C.（2011）。关联数据：将网络演变为全球数据空间。《语义网综合讲座：理论与技术》，1(1)：1–136。出版方：Morgan & Claypool出版社。

Hyv¨onen, E.（2020）。在数字人文中运用语义网：从数据发布转向数据分析与偶然知识发现。《Semantic Web》，11(1)：187–193。出版方：IOS Press。Hyv¨onen, E. 等（2020）。“Sampo”模型与面向数字人文的语义门户：构建于语义网之上。载于《北欧国家数字人文第五届会议论文集》（DHN 2020）。CEUR-WS.org。Hyv¨onen, E., Tuominen, J., Alonen, M. 和 M¨akel¨a, E.（2014）。Linked Data Finland：用于发布和再利用关联数据集的七星模型与平台。载于《欧洲语义网会议》，第226–230页。Springer。Izquierdo, Y. T., Garc´ıa, G. M., Menendez, E., Leme, L. A. P., Neves, A., Lemos, M., Finamore, A. C., Oliveira, C. 和 Casanova, M. A.（2021）。通过 SPARQL 查询编译，在无模式 RDF 数据集上进行关键词搜索。《Information Systems》，102：101814。出版方：Elsevier。Kadilierakis, G., Fafalios, P., Papadakos, P. 和 Tzitzikas, Y.（2020）。使用以文档为中心的信息检索系统搜索 RDF 数据中的关键词。载于《欧洲语义网会议》，第121–137页。Springer。Kahle, P., Colutto, S., Hackl, G. 和 M¨uhlberger, G.（2017）。Transkribus——用于历史文献转录、识别和检索的服务平台。载于《2017年第14届IAPR国际文档分析与识别会议》（ICDAR），第4卷，第19–24页。IEEE。Kritsotakis, V., Roussakis, Y., Patkos, T. 和 Theodoridou, M.（2018）。面向语义数据的辅助式查询构建。载于第14届国际语义系统会议（SEMANTiCS）同期举办的海报与演示分会场论文集。Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samar-itakis, G., Theodoridou, M., Flouris, G. 和 Doerr, M.（2017）。用于文化遗产及其他领域信息集成的 X3ML映射框架。《International Journal on Digital Libraries》，18(4)：301–319。出版方：Springer。Marketakis, Y., Tzitzikas, Y., Gentile, A., Niekerk, B. V. 和 Taconet, M.（2021）。支持 RDF 语义数据仓库演进需求的工作流。《International Journal of Metadata, Semantics and Ontologies》，15(3)：220–232。Mendes, P. N., M¨uhleisen, H. 和 Bizer, C.（2012）。Sieve：关联数据质量评估与融合。载于《2012年联合 EDBT/ICDT研讨会论文集》，第116–123页。Mero˜no-Pe˜nuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S. 和 Van Harmelen, F.（2015）。用于历史研究的语义技术：综述。《Semantic Web》，6(6)：539–564。出版方：IOS Press。Nikas, C., Kadilierakis, G., Fafalios, P. 和 Tzitzikas, Y.（2020）。RDF 关键词搜索：单一视角是否足够？《Big Data and Cognitive Computing》，4(3)：22。出版方：Multidisciplinary Digital Publishing Institute。

Oldman, D., Doerr, M. 和 Gradmann, S.（2015）。《链接数据之禅与艺术：人文知识语义网的新策略》。载于《数字人文新指南》，第251–273页。John Wiley & Sons, Ltd.。章节：18。Oldman, D. 和 Tanase, D.（2018）。通过连接研究者、数据与实践，重塑 ResearchSpace 中的知识图谱。载于《国际语义网会议》，第325–340页。Springer。Ouksel, A. M. 和 Sheth, A.（1999）。全球信息系统中的语义互操作性。《ACM Sigmod Record》，28(1)：5–12。出版方：ACM New York, NY, USA。Petrakis, K., Samaritakis, G., Kalesios, T., i Domingo, E. G., Delis, A., Tzitzikas, Y., Doerr, M. 和 Fafalios, P.（2020）。数字化、整理并可视化海事史档案来源：以十九、二十世纪的船舶航海日志为例。《Drassana: revista del Museu Marıtim》，(28)：60–87。Pipino, L. L., Lee, Y. W. 和 Wang, R. Y.（2002）。数据质量评估。《ACM通讯》，45(4)：211–218。出版方：ACM New York, NY, USA。Scholz, M. 和 Goerz, G.（2012）。WissKI：面向文化遗产的虚拟研究环境。载于 ECAI 2012，第1017–1018页。IOS Press。Turner, D.（2007）。构建史前史：历史科学与科学实在论之争。剑桥大学出版社。Tzitzikas, Y., Mountantonakis, M., Fafalios, P. 和 Marketakis, Y.（2022）。CIDOC-CRM 与机器学习：综述与未来研究。《Heritage》，5(3)：1612–1636。Ventresca, M. J. 和 Mohr, J. W.（2017）。档案研究方法。载于《布莱克威尔组织研究指南》，第805–828页。John Wiley & Sons, Ltd.。章节：35。Vicente-Saez, R. 和 Martinez-Fuentes, C.（2018）。开放科学新篇章：面向综合定义的系统性文献综述。《Journal of Business Research》，88：428–436。Volz, J., Bizer, C., Gaedke, M. 和 Kobilarov, G.（2009）。Silk——面向数据网的链接发现框架。载于《WWW2009网络链接数据研讨会论文集》。Wang, R. Y. 和 Strong, D. M.（1996）。超越准确性：数据质量对数据消费者意味着什么。《Journal of Management Information Systems》，12(4)：5–33。出版方：Taylor & Francis。Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. 和 He, L.（2022）。人在回路机器学习综述。《Future Generation Computer Systems》，135：364–381。Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. 和 Auer, S.（2016）。关联数据质量评估：综述。《Semantic Web》，7(1)：63–93。出版方：IOS Press。