多对多：艺术史与摄影研究中的实体对齐可用性挑战

摘要

目的
本文探讨文化机构之间异构记录对齐所面临的挑战，重点关注 PHAROS 联盟中的艺术史照片档案。

设计／方法／途径
本研究通过案例研究分析对齐工作流与编目传统，并关注机构语境、数据粒度和建模策略。

研究发现
对齐很少是一对一的操作。歧义、数据不完整、归属变更以及实践差异都会影响结果。观察到的策略包括创建匿名实体或集体实体、使用上位术语、添加不确定性限定词，以及在歧义无法消除时选择信息隐匿。

原创性
本文强调需要明确地对不确定性进行建模，并提出一个将技术对齐方法与机构实践相连接的框架。

实践意义
PHAROS 的见解为设计更稳健、可互操作且可持续的文化遗产基础设施提供了指导。

关键词：艺术史；语义网；数据对齐

1. 引言[1]

在画廊、图书馆、档案馆和博物馆（GLAM）领域，关联数据已成为构建大规模元数据聚合器的核心策略，以支持跨藏品发现和研究（Hyvönen，2012年；Isaac 等，2012年）。机构通过将数据作为关联开放数据发布，并将其与 VIAF、盖蒂词表或维基数据等共享规范资源对齐，便能以可互操作且可在本地目录之外重复利用的方式展示其馆藏。此类聚合器的创建在很大程度上依赖数据对齐，以匹配描述同一人物、地点或艺术品的统一资源标识符；而这些标识符往往在不同机构间存在重复、分散或表述各异的情况。因此，对齐这些实体对于确保聚合器呈现连贯、丰富且易于使用的文化遗产信息网络至关重要。

---

[1] 已发表版本：Marilena Daquino、Francesca Mambelli 和 Artem Kozlov。2026年。多对多：艺术史与摄影研究中的实体对齐可用性挑战。《文献学报》 [在线先行出版], ı–19。https://doi.org/10.1108/JD-09-2025-0284。作者责任声明。所有作者均参与了该项工作的概念设计和设计。Marilena Daquino 负责第2节和第4节，Francesca Mambelli 负责第3节，Artem Kozlov 负责第5节。所有作者均同等参与了第1、5和6节。

然而，据我们所知，尽管对账方法日益增多（Zeng 等，2021年），且文化遗产领域已开展实证研究，文化机构之间数据对齐的流程与政策仍鲜有记录。以往工作主要聚焦于技术方法，而跨机构对账的组织和政策层面尚未得到充分探讨。考虑到机构编目实践各有特点，单一通用框架难以构想；但机构数据比较分析所揭示的许多问题在不同情境中反复出现且具有共性。因此，记录这些分析至关重要，因为它们能够为对账政策和最佳实践的设计提供依据，助力未来艺术与人文学科领域的大规模倡议。

本文探讨了艺术品记录中缺乏明确实体对齐政策的问题。本文借鉴 PHAROS 艺术史照片档案协会的协作实践，报告了对齐艺术家、摄影师和艺术品记录的工作，重点介绍了协会遇到的挑战及其制定的解决方案。研究结果包括一些有效策略，可推广至面临类似问题的文化遗产与人文学科领域；这些领域往往需要处理相同的信息。本文提出的框架整合了识别、建模、工作流程和界面策略，因此为任何希望在共享的语义互操作环境中整合异构数据集的 GLAM 机构提供了可重复使用的结构。

2. 现状

画廊、图书馆、档案馆和博物馆通常会先考虑哪种规范控制最适合其数据范围，再开始对账任务。规范控制确保名称、标题和主题采用标准化形式，从而在机构的组织与发现工具中作为连贯的馆藏检索点（Joudrey 等，2015年）。名称的规范形式收录于规范档中。VIAF 汇集了图书馆的国家规范档，其中描述了作者、艺术家和历史人物。盖蒂词表（ULAN/AAT/TGN）以关联开放数据的形式提供有关艺术家、材料、艺术品类型和历史地点的信息，许多机构都会与之进行对齐。如今，维基数据也是实体对齐和规范数据匹配的理想候选，因为其中丰富的传记学和地理数据注释包含指向主要文化机构标识符以及第三方权威库（如前述 VIAF 和盖蒂词表）的直接链接。

本体规范了对齐的对象。CIDOC-CRM（CIDOC 2021）提供了用于描述文化遗产数据的核心事件中心模型。博物馆聚合平台采用关联艺术（Delmas-Glass and Sanderson, 2020）等应用配置文件；关联艺术提供了应用程序接口（API）[2]，简化跨机构对齐及后续对齐工作。同样，Europeana 的实体 API 提供人物、地点和主题的规范记录，以供链接[3]。CIDOC-CRM 越来越多地与 IFLA 图书馆参考模型结合使用，其面向对象的表述形式（LRMoo）已整合进 CIDOC-CRM，以协调书目与博物馆文献实践（Riva and Žumer 2018）。

在实践中，对齐的半自动策展已是标准流程，例如使用 OpenRefine 或盖蒂的服务（Delpeuch 2019），并且常与本地规则结合，用于数据规范化、变体处理、出生／死亡日期管理、角色或职业管理等。数据集以资源描述框架（RDF）发布时（Baker et al. 2011），LIMES（Ngomo et al. 2021）等框架允许策展人员声明链接规范（属性、相似度指标、阈值），并运用剪枝规则，在大型图上高效运行匹配器。监督机器学习以及日益普及的深度学习可增强（而非取代）匹配规则（Di Cicco et al. 2019）。这些技法与领域约束及权威档案结合时，可应用于人文学科语料库。

尽管如此，在 GLAM 语境下——名称、归属和历史身份往往存在噪声——策展人监督至关重要。OpenRefine 对账 API 支撑着许多权威档案映射工作流，但其评分／用户体验局限已得到确认，亟待改进（Knoblock 等 2017；Delpeuch 2019）。即使拥有强大的权威档案和方法，仍会面临多对多的现实情况，包括同名异人、归属变更（确定的艺术家与归属的艺术家之间的区别）、不确定的生卒日期，以及角色和法律身份粒度问题（摄影师与工作室之间的区别）；即便是专家自己，也会

[2] https://linked.art/about/1.0/

[3] https://europeana.atlassian.net/wiki/spaces/EF/pages/2324561923

难以清晰识别这些情况（Doerr 2025）。值得注意的是，参与对账工作的文化遗产机构专家愿意投入大量时间和精力，以确保最终数据准确无误（Knoblock 等 2017）。然而，公开发布管理歧义策略（聚类、证据聚合、周期性重新处理）的机构寥寥无几，这说明所需的不仅是算法，还包括政策和流程（Hickey 和 Toves 2014）。例如，RDA（资源描述与检索）对变体身份的处理，尤其是“作品组”这一构造：当来源信息或作者身份无法被唯一确定时，它会将异质的表现形式聚合到一个共同的概念框架之下；这是解决重要识别和对账问题的一次显著尝试（Coyle 和 Hillmann 2007）。同样，馆藏及其目录分析模型（Heaney 2000）对馆藏与子馆藏的建模，后来在都柏林核心馆藏描述应用纲要（DC 2024）中得到正式确立，也揭示了数据建模作为一种在形式语言中暴露歧义的代理手段的重要性。总的来说，鉴于 PHAROS 机构所持文献及摄影收藏本身的特殊性，PHAROS 遇到的对账挑战并非摄影档案所独有，而是反映了图书馆、档案馆和博物馆中普遍存在的结构性问题：这些机构的描述传统和规范来源各自独立发展，因此需要有原则的对齐策略。

3. 背景

PHAROS 协会由来自欧洲和北美洲的 13 家艺术史照片档案机构组成。其馆藏合计包含 3,100 万份摄影文献，记录了各个时期的艺术品（Caraffa 等 2020）。目前，七家机构（赫茨阿纳图书馆—马克斯·普朗克艺术史研究所、马尔堡图片档案馆、弗里克艺术研究图书馆、费代里科·泽里基金会、佛罗伦萨艺术史研究所、荷兰艺术史研究所（RKD）、沃伯格研究所）正在试运行在线研究平台 artresearch.net。该平台目前收录了约 100 万件艺术品的 170 万张照片。

平台 artresearch.net 是一项基于 ResearchSpace（Oldman 和 Tanase 2018）构建的协作计划，支持导入转换为关联开放数据的数据。为确保互操作性，PHAROS 协会采用 CIDOC-CRM 应用规范作为标准，以统一元数据描述，并尽可能提供统一的馆藏视图。对齐活动是 artresearch.net 提取/转换/加载（ETL）流程不可或缺的一部分。由于持续导入新的数据源，以及随之而来的数据管理政策修订（例如扩展本体、受控词表和匹配规则），这些活动仍在进行中。

原始数据采用多种元数据标准来描述艺术品的摄影文档，例如用于描述艺术品（Scheda OA）[4] 和照片（Scheda F）的意大利 ICCD 国家编目标准[5], 、马尔堡信息、文献与管理系统（MIDAS）元数据标准[6], 、MARC21（美国国会图书馆 2021年）以及机构标准。这些标准都涵盖一组相似的核心元数据字段，用于描述艺术品（如艺术家、材料与技法、日期、主题、保管者、来源信息）和摄影文档（如摄影师、技法、拍摄与印制日期、版权持有者）。

艺术史照片档案中描述的数据具有显著的异质性，其形成不仅受到摄影对象本身多样性的影响，也受到塑造其整理工作的编目实践的影响。从历史上看，编目是一项成本高昂、劳动密集的工作，并受不断变化的机构优先事项、学科兴趣和技术限制影响。在许多档案馆中，早期编目工作反映了当时的偏见：描述哪些内容、如何描述，以及优先采用何种粒度，这些选择都受当时知识框架的制约。尤其是在20世纪90年代启动的编目活动中，描述完全聚焦于被再现的对象，而对与摄影媒介相关的信息关注甚少，这些信息往往仅以图注或自由文本描述的形式记录。因此，照片档案中的元数据通常参差不齐、不完整或前后不一致。

[4] http://www.iccd.beniculturali.it/it/ricercanormative/29/0a-opere-oggetti-d-arte-3_00
[5] http://www.iccd.beniculturali.it/it/ricercanormative/62/f-fotografia-4_00
[6] https://d-nb.info/1212592956/34

自21世纪初以来，随着摄影收藏的系统化编目逐渐兴起，人们对摄影来源的关注也日益增强，但档案工作者开展工作时仍缺乏既有的规范档和标准化词表。尽管人们曾尝试在不同机构间建立共享术语，但这些努力当时仍处于起步阶段，尚不具备后来在艺术史等领域中形成的稳健性。

的确，在艺术史领域，学术兴趣和方法上的进步推动了编目实践的改进。艺术史照片档案中编目实践的修订，源于学术界和艺术市场群体的兴趣变化；这些群体日益要求以更严谨、更一致的方式访问数据集，例如按艺术家或体裁检索，这就需要专业介入以提升数据质量。尽管付出了这些努力，数据往往仍保留着源于地方传统和编目员个人判断的独特之处。

语义网技术的兴起，为重新审视以往的编目实践、应对长期存在的术语和互操作性挑战带来了前所未有的机遇。然而，这些技术也暴露了底层数据的脆弱性：数据中存在不确定性、模糊性和不一致性，这些问题既承袭自早期的编目实践，也源于无法唯一且无可争议地识别实体（例如艺术家的追随者）。因此，记录并不总能与外部规范档一一对应，甚至无法在不同机构之间进行对齐。全面修订遗留数据的成本高得令人难以承受，而且很可能也无法解决不同学术观点所带来的不确定性（例如，学者对身份不明的艺术家有不同认定）。

对于照片档案的数据而言，这些挑战尤为突出。如前所述，摄影研究作为一个领域的出现远晚于艺术史研究，而且直到最近，照片才被视为独立的研究对象。缺乏关于照片和摄影师的综合术语体系，使机构可用的描述框架存在重大缺口。与逐步发展出规范控制和受控词表的艺术品不同，照片至今仍缺乏标准化描述词。这种缺失不仅使跨机构互操作性更加复杂，也有可能延续早期编目工作中固有的不均衡与偏见。因此，如今备受学者重视的元数据必须置于历史脉络中加以审视，并结合此前未曾提出的精确性要求重新诠释。

编目员、元数据专家和开发人员不得不应对这种异质性，以及实现精确对齐这一目标本身固有的不可能性。此外，结构化数据系统所要求的精确程度，与艺术史本身存在的认识论上的不确定性相矛盾。因此，任何转化过程都必须承认绝对准确性无法实现，并围绕这一现实进行设计。

4. 内容数据协调中的挑战

所选取用于摄取和对齐的 PHAROS数据源，鲜明地体现了机构偏见和编目传统的深刻影响。这些影响表现为：随着时间推移，编目工作的优先事项和关注点不断变化，导致对同一信息的描述准确度各不相同；同时，详略程度也有所不同，反映出人们对数据质量和内容的预期在不断变化。综合来看，这些模式与编目流程中已知的若干现象相呼应（Barabucci 等，2022；Vitali 和 Pasqual，2025），具体包括：

- 信息隐匿。由于时间紧迫或缺乏政策，信息未被记录。- 扁平化。使用同一个元数据字段同时描述不同的信息。- 强制归类。找不到合适字段的信息被强行填入其他字段。- 倾倒。找不到合适字段的信息以纯文本形式置于描述字段中。- 可靠性差异。与某个人工制品相关的各项信息并非都具有相同的可靠性，但有关动机、来源以及对不确定性的承认等信息可能缺失。- 假定确定性。即使对某个字段的取值没有分歧，人们对该取值真实性的置信度也可能各不相同。

这类问题可能导致记录的检索点不一致（扁平化、信息隐匿），阻碍数据和实体的自动处理（强制归类、倾倒），并使人无法区分主观信息与事实信息

（可靠性差异、假定确定性），从而使读者不自觉地认为记录具有事实性和确定性。

设计 artresearch.net 平台，进而设计对齐流程，需要在呈现不确定性与确保可用性之间不断权衡。虽然保留并传达数据中蕴含的不同程度的确定性和可靠性至关重要，但将如此细致的粒度直接呈现给终端用户，可能会令他们不堪重负，并削弱界面的清晰度。无论是学者还是普通用户，如果在用户体验中将不确定性作为首要特征突出呈现，都可能感到无所适从（Battisti and Daquino 2025）。因此，可用性考量不可避免地会影响设计决策。信息必须以易于理解且有意义的方式呈现，既不掩盖底层的复杂性，也不让用户承担过多的解读工作。 在 PHAROS，我们随着时间推移采用了多种解决方案，有时还会并行测试几种方法，以评估其效果（如第5节所述）。

然而，对齐难题不仅源于编目流程引入的偏见，也源于数据本身的固有特征。例如，同一位摄影师在不同机构的记录中可能使用不同的姓名变体——有时日期不完整，有时归属相互矛盾——这使得利用规范档或跨数据源进行自动对齐尤为困难。此处我们列举了若干此类对齐问题；解决这些问题需要元数据专家介入，并与机构共同设计政策，以建立一致的实践、协商用户界面（UI）中可接受的折衷方案，并确保不同的编目传统能够实现对齐，同时保留其历史特性。

艺术家

PHAROS 中的艺术家主要与 ULAN 进行对齐；ULAN 是有史以来最权威、最完整的艺术家名单。在 ULAN 中不存在匹配项时，则会次要地与维基数据进行对齐。然而在实践中，直接对齐常常会受到复杂案例的阻碍，而这些案例需要预先制定政策，以指导对候选匹配项的审核。我们重点介绍三种反复出现的情况。

归属限定语。许多记录将艺术品归属于匿名艺术家，而这些艺术家的身份是以某位已知人物为参照来表述的，例如“列奥纳多·达·芬奇画派”“追随者”或“圈子”。虽然保留与已知艺术家的关联对于检索十分重要，但在 PHAROS 中，此类情况会被建模为与已识别艺术家相关联的独立匿名实体。这种方法使我们能够区分已知个人与未知个人——此前二者被“扁平化”到同一字段中——同时保留关联关系。随后，这些关系可以通过层级分面呈现给 artresearch.net 的用户。

姓氏。当无法进行具体归属，或艺术品是家族合作的成果时，档案工作者往往只记录姓氏。例如，“贝利尼”这一标签可能指雅各布、乔瓦尼、真蒂莱（他们都是同一家族的艺术家）中的某一人，可能指其中一部分人，也可能指整个家族，或指某位身份不明的其他亲属。不同机构的解释也各不相同：在一个馆藏中，“贝利尼”可能始终指乔瓦尼，而在另一个馆藏中，它可能在不同记录中被不一致地使用。为应对这种歧义，我们会为字符串“贝利尼”所代表的匿名群体创建一个独立实体，但不主张其与已识别的家庭成员之间存在正式关系。该链接仅保留在用户界面层面，即分面中；在那里，匿名群体与已识别个人可以在视觉上并列呈现，从而尊重原始编目中已有的信息隐匿。

身份分歧。在某些情况下，不同学者对化名归属的识别并不一致。例如，“Pseudo Ambrogio di Baldese（1375–1425）”被分别认为与Ventura di Moro（1399–1486，ULAN 500012920）和Lippo d’Andrea（1371–1451，ULAN 500082343）有关。由于ULAN本身并未认可某一种识别方式，而是将该化名列为两个候选对象的别名，PHAROS便为Pseudo Ambrogio di Baldese创建了一个新的统一资源标识符，供各机构达成共识。该统一资源标识符通过rdfs:seeAlso链接到两个ULAN候选对象，而非语义上含义更强的owl:sameAs，从而在实现对齐的同时保留不确定性。用户界面中会显示艺术家姓名及其与ULAN统一资源标识符的所有链接。界面既不记录，也不显示某一归属相较于另一归属所获得的机构认可，因为即使在同一机构内部也可能存在分歧

。例如，当档案馆的资料源自某位特定艺术史家（如Federico Zeri）时，档案工作者应尊重创作者的假设，而不重新评估。
一个类似的案例是 Ercole Grandi。ULAN 将其认定为 Ercole de' Roberti（ULAN 500124891），而一些机构则将两者视为不同的人物。为避免支持任何一方，同时防止 PHAROS 权威档案中出现重复记录，我们采用了相同的策略：为 Ercole Grandi 新建一个统一资源标识符，并通过 rdfs:seeAlso 将其链接到 ULAN 中的候选实体。同样，我们不保留机构认可，从而在缺乏学术共识时保持中立。

匿名群体。当无法通过名称或化名识别某位艺术家或某个群体时，编目员会将其记录为匿名实体。一些机构可能会采用惯例为匿名实体命名，例如记录其所属画派、地区和世纪，如“16世纪佛罗伦萨匿名艺术家”。尽管无法辨别这一标签是否以及何时指代同一位匿名人物或同一匿名群体，该标签仍会在多条记录中重复使用。因此，即使在同一机构内，更不用说跨机构，对这类记录进行对齐也会在语义层面产生不一致性。在 PHAROS 中，出于可用性考虑，我们忽略这一粒度差异，仅创建一个实体来代表具有相同时空范围的匿名群体。

材料、技法、艺术品类型

最引人注目的对齐异常之一，源于某些术语同时指涉不同的概念层级，而这些层级在 AAT 中看似彼此分离，例如材料、技法和作品类型。比如，照片记录中使用的术语“蛋清”，可能指原材料本身（AAT 300011802）、使用蛋清的摄影工艺（AAT 300133274），也可能指特定的产品“蛋清印相”（AAT 300127121）；而“蛋清印相”又常常与更广泛的“纸”类混为一谈。这种术语重叠使对齐变得复杂，因为在受控词表中，同一个字符串可能含糊地指向三个不同的实体。这会导致机构之间以及机构内部出现不一致性：编目员在对齐术语时，可能会选择不同的 AAT 术语。因此，若没有明确区分这些层叠概念的政策作为支持，自动对齐规范记录就有引入语义噪声的风险。在 PHAROS 中，我们采取最全面的方法，请编目员识别可能参与定义该术语的所有概念层级对应的术语。

来源信息：保管者与地点

当前和历任艺术品保管者的对齐，在建模和对齐层面都存在复杂问题。原则上，保管者是代理主体——可以是人物，也可以是组织——可能与某处住所相关联；该住所位于一座建筑物内，而这座建筑物又坐落于特定城市。

代理主体或建筑物。首先，在解释实体的语义性质时会出现歧义：博物馆应被视为组织（概念上的代理主体），还是建筑物（实体对象）？与维基数据进行自动对齐时，往往会发现不一致性，因为维基数据贡献者并不总是遵循统一的原则——有时将文化机构登记为组织，有时登记为其建筑物，也有时两者皆有（例如 Q3330350）。在 PHAROS 中，我们通过与组织类实体对齐来统一处理，因为这更能体现机构随时间迁址的可能性。

作为代理主体的收藏与混合结构。元数据字段扁平化的另一个例子，是记录保管者的方式未能遵循连贯的语义区分。例如，收藏常被记录为保管者，尽管它们只是对象分组，并非能够获取或持有这些对象的代理主体。为处理这类情况，只要能够明确识别代表收藏品集合所有者的实体，PHAROS 就会对其进行对齐；若无法识别，则为匿名所有者创建一个统一资源标识符。此外，博物馆及其保存的收藏常被交替记录，例如“Frick Art Research Library”和“Frick Collection”。同样，我们对齐被记作保管者的机构，而非收藏，并制定建模方案，将艺术品与特定收藏品集合关联起来。

除馆藏外，PHAROS数据还包含种类繁杂、被标记为保管者的实体：拍卖行（代理主体）、艺术经销商（代理主体）、博物馆（代理主体）和建筑物（物理对象）。这些差异造成了特殊的

层级安排，需要谨慎对齐，以免错误表示定位。将馆藏、建筑物和机构合并在分层描述中的情况尤其具有挑战性，因为语义角色的扁平化会模糊行动者、其所在地以及其所包含对象之间的区别。例如，可移动和不可移动实体包括：属于另一建筑物一部分的建筑物（“Inferior basilica, Basilica of S. Clemente”）、位于某建筑物内的博物馆（“Museum of Porziuncola, Basilica of S. Maria degli Angeli”）、位于另一博物馆内的博物馆（“The Fine Arts Museums of San Francisco, M.H. de Young Memorial Museum”）、博物馆内的馆藏（“Ajmanov grad, collection J. Demsar”）、建筑物内的馆藏（“Allington Castle, collection Lord W.M. Conway”）、建筑物内的拍卖行（“Hotel George V, Auction Tajan”），以及建筑物内的古董商（“The Manor House, Phillips of Hitchin (Antiques) Ltd”）。对每个实体的对齐，均依据相关实体的语义性质进行：拍卖行将与维基数据中对应的代理主体对齐，而其所在地则与相应的行政所在地对齐。为保留两个实体之间的关系——这是正确定位艺术品的必要条件——我们采用建模方案，根据具体艺术品将可移动实体（如馆藏）与不可移动实体（如建筑物）关联起来。

庄园，例如 Broomford Manor 或 Chatsworth House，是语义扁平化的另一个例子，这一次源于其行政属性本身含有歧义。它们介于地名与建筑物之间，因此难以明确判定其所属的行政区或城镇。这种歧义还因庄园中保存的收藏名称常与庄园本身的名称相同而加剧，模糊了保管实体的语义。我们采用与维基数据相同的做法，在适用情况下优先将它们视为地点。

最后已知位置。将拍卖行或古董商记录为保管者，通常意味着艺术品在某个日期最后一次被发现由其持有（该日期记载于同一元数据字段或其他元数据字段中）。然而，艺术品之后很可能被转移到了别处。虽然对齐会将这些代理主体（拍卖行和古董商）视为稳定实体，但实际记录的事件是：在特定时间和地点对人工制品的一次临时观察——这在概念上不同于由某个机构永久保管。

地点的粒度。 对地理实体进行对齐——例如确定博物馆当前位于某个城市或城镇的何处——又增加了一层复杂性。维基数据为地理数据对齐提供了全面的数据，但自动对齐往往会产生地理具体程度不一的层级结构。例如，维基数据中 National Gallery of Art 在伦敦的地理位置将其总部标注为威斯敏斯特市（属于伦敦），这一粒度过于精细，不适合展示给终端用户，因此需要进行后处理。在 PHAROS 中，我们接受这些自动对齐结果，但会应用协调规则，将地点规范化到城市层级。

来源重建。 所有这些问题都使重建来源事件序列尤其困难。合并不同机构的描述时，这些描述采用的粒度不一致（例如，提及人物而非馆藏，或反之），且往往缺少步骤或相互矛盾，因此需要大量工作来进行数据规范化、冲突消解，以及管理模糊性和不确定性。此外，学界缺乏一套共享的艺术品来源事件权威数据，可用于对齐销售、购藏和所有权变更。更复杂的是，这类信息通常仅以自由文本描述的形式记录（这属于倾倒），而各机构采用的编目规范差异很大，使得自动提取实体和事件序列十分困难。

目前，PHAROS 采用渐进式方法向终端用户提供这些信息。尽管知识提取和对齐流程仍在进行中，文本来源描述仍以原样显示在 artresearch.net 的记录中，让用户能够在一处获取来自不同提供方的所有可用信息——即使这些信息尚未得到充分解读、去重或合并。

艺术品

尽管 GLAM机构在制作关联开放数据（LOD）方面取得了显著进展，但要将艺术品对齐到一份共享且全面的规范名单中，仍是一项艰巨任务。Getty CONA 可视为一个潜在的选择，但遗憾的是，它并未像其他 Getty 标准那样受到广泛关注，而且目前也没有

这样一份全面的名单存在于 LOD 中。因此，PHAROS 致力于对齐合作机构的艺术品，从而在 artresearch.net 上对其目录形成连贯、统一的视图。这种对齐可视为该流程的主要目标之一，依赖于多种策略的结合，主要包括元数据对齐和视觉相似性分析。

元数据（错位）对齐。仅基于元数据进行对齐面临着诸多挑战，尤其是当艺术品条目采用不同惯例描述，且其中包含相互竞争、彼此矛盾或不准确的信息时。例如，艺术品的标题可能会随着时间推移而发生变化；照片档案可能采用不同的历史标题，也可能在没有现成标题时自行拟定标题，从而造成潜在的错配。日期的记录粒度也可能各不相同（世纪、十年、年份范围或模糊估计），不同机构给出的归属也可能有所出入。艺术家信息也存在同样的问题：某家机构可能将作品列为某位大师工作室中的无名成员所作，而另一家机构则可能注明该工作室中已确认身份的实际艺术家。这些差异十分普遍，因此，仅靠元数据对齐并不是一种可靠的全面对齐方法。

视觉相似性。为解决这一问题，PHAROS首先通过图像相似性算法识别候选匹配项，然后再由艺术史学家和编目员参与比对并审查相关元数据。然而，这种方法也会带来自身的难题。例如，摹写者创作的精确副本常常会被算法匹配到一起；如果只考察元数据，这些匹配可能看起来正确，从而误导非专家，使其将结果解读为学术观点存在差异的证据。只有通过专家视觉比对，才能确认或否定此类匹配。在这些情况下，虽然严格的等价链接并不恰当，但仍应保留关联关系。同样，习作图或复合艺术品的单个组成部分也可能与完成后的作品匹配。在此情况下，一对一对齐同样会产生误导，但保留关联关系仍然具有价值。

摄影师

PHAROS 中的摄影师可以是人物、工作室，也可以是委托拍摄照片的文化机构。在适用情况下，我们会将这些实体与维基数据进行对齐。选择与通用知识库而非专业知识库进行对齐，源于 PHAROS 所收录照片的来源信息种类繁多；专业规范档案列表只能涵盖其中的小群体（例如，摄影师身份目录（PIC）[7] 主要涵盖美国摄影师），因此我们需要采用多个专业规范档案列表。相比之下，维基数据覆盖范围更广，即便对较为小众的专业人士而言，其数据仍不完整，也能让我们获得更高的覆盖率。

摄影师相关数据——主要是姓名和日期——通常来自对照片上题字或印章的转录。如前所述，早期的编目活动往往会记录这些姓名，却没有明确意图去识别一个稳定实体，以便日后将其用作数据访问点。在这类情况下，摄影师姓名的作用更像是图注，而非规范数据。随着时间推移，一些机构（如 Zeri 照片档案）修订并简化了这些命名规范，建立了自己的规范档，但大多数机构并未如此。同样，少数采用 MIDAS标准的机构创建了内部统一的规范名单，但该名单并未与外部来源全面互联。这导致不同馆藏之间存在显著差异，包括命名规范、归属实践和准确度水平各不相同。与将艺术家与 ULAN 对齐相比，PHAROS 中摄影师的对齐所面临的挑战，很大程度上源于姓名沿袭而来的模糊性和不稳定性；所采用的解决方案则是为了简化用于访问数据的分面。这些问题可概括如下。

与某个名称相关联的不确定法律地位。当一个名称可能指代人物，也可能指代工作室，而又缺乏足够证据加以区分时，就会反复出现一个问题。例如，*“Böhm”*可能指摄影师*OsvaldoBöhm*，也可能指他的工作室*Foto Böhm*。机构通常会记录这两种形式中的一种，从而明确其法律地位，但并非总能做到。在含义模糊的情况下，人们会不加区分地使用字符串*“Böhm”*（有时后面会加上“？”），因此无论通过程序还是实证方法，都无法确定其所指的实体。为处理这些情况，

[7] https://pic.纽约公共图书馆.org/

我们创建了一个名为“Böhm（集合名称）”的新实体，用以表示一个匿名群体。与上述艺术家姓氏的情形类似，由于边界模糊，我们不会将不确定实体“Böhm（集合名称）”与确定实体“Osvaldo Böhm”和“Foto Böhm”之间的关系具体化。相反，我们会另行创建一个代表常用标签“Böhm”的统一资源标识符，并将其关联到所有上述实体。我们非正式地将这个新实体称为“伞形术语”。它并不代表现实世界中的实体（如人物或群体），而是一个具体化标签，用来将具有共同字符串的实体汇集起来并建立松散关联。在用户界面中，*“Böhm”*会显示为一个顶层分面，可展开查看底层的确定实体（“Osvaldo Böhm”和“Foto Böhm”）以及不确定实体（“Böhm（集合名称）”）。

实体构成方面的变化。摄影工作室或工作室与其他组织一样，往往会更改名称、所在地或所有权，因此引发了一个问题：这些变化应被建模为不同的实体，还是仅视为姓名变体。维基数据对此反映出不一致性，提供了粒度各异的统一资源标识符：有时一个统一资源标识符涵盖一家公司的多种表现形式，而有时每种表现形式都有自己的记录。例如，Braun & Cie (WD Q79493769) 包含 Maison Adolphe Braun & Cie 和 Braun, Clément et Cie 等实体，并将它们视为同一家公司的不同形式。然而，在 PHAROS 中，数据中发现的每种表现形式都对应一个新创建的统一资源标识符，且仅与最匹配的维基数据实体建立对齐关系。随后还会创建一个伞形术语，将所有这些表现形式归入同一标签之下。另一个例子是“Armoni Moretti Raffaelli”公司。该工作室由 Luigi Armoni 创办，随后由 Luigi Raffaelli 继承，并更名为“Armoni e Raffaelli”；后来又由 Mario Moretti 接手，并更名为“Armoni Raffaelli & Moretti”。维基数据中有一条描述摄影师“Luigi Armoni Raffaelli”的不精确记录，并将其标识为人物。在 PHAROS 中，我们为每种机构形式分别创建统一资源标识符，再将它们汇总到一个名为“Armoni Raffaelli Moretti”的伞形术语之下。
姓氏变更。个人姓名的变更也使对齐变得复杂。摄影师兼艺术史家 Hilde Lotz-Bauer 在 PHAROS 中以多个姓名记录，这些姓名反映了婚姻状况和机构惯例：“Lotz-Bauer, Hilde”、“Degenhart-Bauer, Hilde”以及“Bauer Lotz, H.”。虽然将其与维基数据条目（WD Q1618235）对齐不存在语义问题，但用户界面只能显示一种规范形式，从而掩盖了这些变体在时间和传记层面的重要意义。为保留这些信息，我们采用 CIDOC-CRM，在照片与记录编目员所记录的确切形式的语言对象之间建立关系，再将这些语言对象关联回人物实体。

委托摄影的机构。当摄影师身份不明，但记录了委托机构——该机构通常同时作为版权所有者——时，编目员常常会将机构名称填入同一个元数据字段（这是另一种强制归类）。在 PHAROS 中，我们选择了一种简化的建模方法：出于可用性考虑，我们不增加大量匿名实体，而是将机构保留在摄影师规范名单中。例如，阿姆斯特丹的国立博物馆（WD Q190804）既可能作为摄影师出现，也可能作为艺术品的保管方。当记录的是下属部门或相关办公室时，则会出现特殊情况，例如 Fotocommissie Rijksmuseum Amsterdam。由于这类实体没有专门的维基数据记录，且无法与上级机构对齐，PHAROS 会创建单独的实体，并将其归入伞形术语“Rijksmuseum”之下。我们不在这些实体之间声明直接关系，因为它们的法律地位可能随时间发生变化。这样既保留了必要的模糊性，又简化了用户界面分面。

然而，当机构和摄影师都已知时，这种简化便显现出明显的局限。应当只将个人认定为摄影师，还是也应承认其所属机构？虽然分别对这些实体进行规范处理并不困难，但若将它们纳入同一份摄影师规范名单进行管理，就可能引入不准确之处。例如，Mario Sansoni 曾先后独立工作，也曾为多家工作室和机构工作（如 Alinari、Frick Art Research、Nesti、Bencini）。如果只将 Sansoni 记录为摄影师，却忽略雇用过他的机构信息，就无法呈现其职业生涯的演变，也无法反映照片拍摄时的具体背景。要解决这一问题，需要进行大量建模工作，以呈现个人与机构之间随时间变化的合作关系，从而为规范名单增添语义关系。

多个或替代摄影师。在 PHAROS 数据中，我们有时会看到诸如“Anderson/Brogi”这样的条目。这里，两位摄影师——或各自的工作室——被一并记录，但尚不清楚这表示的是

实际合作，还是替代性的归属。法律地位不确定的问题再次出现，因此需要使用匿名实体和“上位术语”来表示这种歧义。与此同时，照片与其可能的作者（一位、两位或其中任一位）之间的关系，也带来了建模方面的挑战：如何表示对字符串的解释存在不确定性，并由此体现作者身份的归属，而编目员并不认可这种归属。在 PHAROS 中，我们采用定制的本体模式来表示替代陈述，并使用命名图保留其来源信息。

不确定的姓名。除了字符串所代表实体的法律地位存在不确定性外，我们还遇到了一些案例，其中歧义影响了对齐本身的语义。例如，标签“Beyer, Constantin?”暗示摄影师可能确实是 Constantin Beyer（WD Q95218985），但也可能指另一位毫不相关的人。在此类情况下，我们将该字符串与已识别的维基数据记录对齐，并在具体照片层面明确对归属的不确定性进行建模。另一些情况则更为棘手，例如“Beyer, ?”，无法判断所指的是 Constantin、Günther 还是 Klaus Beyer。由于此时无法作出合理推测，我们会创建一个未与任何候选对象关联的匿名实体，有意避免作出归属，因此也谨慎避免提出一个或多个可能的识别结果。

5. 文化遗产数据中管理歧义与异质性的挑战

案例研究表明，文化遗产数据中的不确定性并非边缘性的例外，而是编目实践的一种结构性条件。基于在 PHAROS 中观察到的挑战，我们提出了一个由四部分组成的不确定性管理框架，涵盖建模、工作流程和界面方面的考量。该框架明确了对齐不同机构所属异构记录所需的步骤，同时保留来源信息和机构视角。

1. **识别策略**。用于检测实体之间细微或显著差异的方法，包括同名异人、笔名、匿名群体、文字变体和不完整的历史归属。这些策略依托规范数据（如 VIAF、ULAN、维基数据）和机构元数据，在自动对齐开始之前标记潜在的歧义。

2. **建模策略**。用于编码歧义而非消除歧义的一系列表征技术，包括：

    a. 区分可能的等价关系与确定的等价关系（必要时使用 rdfs:seeAlso 而非 owl:sameAs）；b. 当相关实体之间无法定义上述相似或等价关系时，描述其关联关系；c. 运用更高层级的概念节点（如上位术语），将异质或相互冲突的归属归为一组，同时保留机构差异；

3. **工作流程策略**。 结合自动对齐、基于规范数据的匹配、冲突检测和人工验证的多步骤流程。该框架强调机构间的分工，使每个档案馆都能审核与自身材料相关的候选对象，同时在集中式图中共享合并结果。4. **界面与交互策略**。关于如何向终端用户传达不确定性的决策，包括以可视化方式归组变体标题、在过细的区分妨碍导航时予以隐藏，以及展示来源轨迹以说明每条陈述如何进入图中。这些策略将用户体验视为对齐的构成部分，而非事后考虑的事项。

这一框架不仅适用于摄影档案，也提供了概念工具，可供面临类似跨机构元数据对齐挑战的博物馆、图书馆和档案馆使用。

5.1 命名与身份歧义（人物、群体、笔名 ）

PHAROS 的整体策略是尽可能将实体名称（人物、组织、地点）与第三方权威库对齐，仅在无法找到匹配项时才创建新的统一资源标识符，之后再寻求跨机构对齐。ULAN、VIAF、GND 和维基数据等权威数据集提高了覆盖范围，但也引入了相互竞争的规范，进一步增加了身份解析的复杂性。

我们的方法将自动化相似度度量与权威库交叉核验相结合，以生成候选匹配项，再由机构专家进行验证。当不确定性依然存在时，我们会通过以下方式编码关系：1）使用非承诺性谓词（例如 rdfs:seeAlso），而非 owl:sameAs（例如针对艺术家）；或 2）使用上位术语，定义现实世界实体与共享标签之间的关联关系。

第一种情况适用于艺术家：我们通过结合基于规则的算法与图嵌入的两步流程，将其与 ULAN 和维基数据进行对齐，最后由人工审查员核验。最初，我们尝试使用大语言模型（LLM，具体为 Gemini 2.5 Pro）自动执行对齐。我们向大语言模型提供了来自不同机构的艺术家结构化数据（包括姓名和生卒日期），以及一份与姓名部分匹配的 ULAN 候选对象清单，并要求其剔除字符串相似度较低或日期不相符的对象。然而，这种方法会产生难以预测的错误和幻觉，例如输出不完整、字段缺失（如日期）、重复记录、假阴性（漏掉已有匹配项）和假阳性（提出输入清单中不存在的匹配项），而这些问题很难诊断或纠正。

因此，我们转而采用基于规则的方法，比较字符串相似度和日期。尽管这种方法的召回率较低，但它能生成精度很高的匹配结果，从而开展可控的第一轮人工审核，避免让审核人员被可能不正确的匹配项淹没。对于所有剩余的未解决案例，我们采用基于图嵌入的第二种方法。这种方法利用本地目录中的姓名和日期，以及 ULAN 中的传记数据。在此，LLM 被用于一种限制更多的角色：筛选嵌入模型提出的候选匹配项，并返回最有可能的匹配结果供人工审核。这种组合方法在基于规则的系统无法有把握地识别匹配项时，能够得出更可靠的结果。然而，当不同机构对同一位艺术家的描述与 ULAN 候选对象进行匹配时，可能会出现差异。例如，一家机构可能使用艺术家的全名“Gavasio, Giovanni Giacomo”，并在 ULAN 中找到正确的匹配项；而另一家机构使用不同的姓名形式“Gavazzi Giovanni”，却无法在 ULAN 中找到可信的匹配项。在这些情况下，对候选匹配项进行专家审核有助于我们避免最终列表中出现重复项，并在有相应记录时最终找到正确的 ULAN 匹配项。

值得注意的是，一些机构已经提供了与现有权威库的对齐结果。弗里克艺术研究图书馆包含指向美国国会图书馆（LOC）艺术家 ID 的链接，赫兹iana图书馆包含指向 GND 和维基数据的链接，而 Zeri 照片档案则为部分实体提供了指向 ULAN、VIAF 和维基数据的链接。为统一实体对齐结果，并获取所有艺术家与 ULAN 的匹配项，我们进行跨权威库对齐，自动检索 LOC、GND、VIAF、ULAN 和维基数据记录之间的链接。令人意外的是，我们发现许多情况下，数据集之间存在种种分歧。例如，某个维基数据统一资源标识符链接到 VIAF 中的一个统一资源标识符，而这个相同的 VIAF 统一资源标识符又指向另一个维基数据统一资源标识符。在 PHAROS 中，此类情况占跨权威库对齐统一资源标识符的 27%。图1展示了我们用于统一跨权威库链接的工作流程示例。

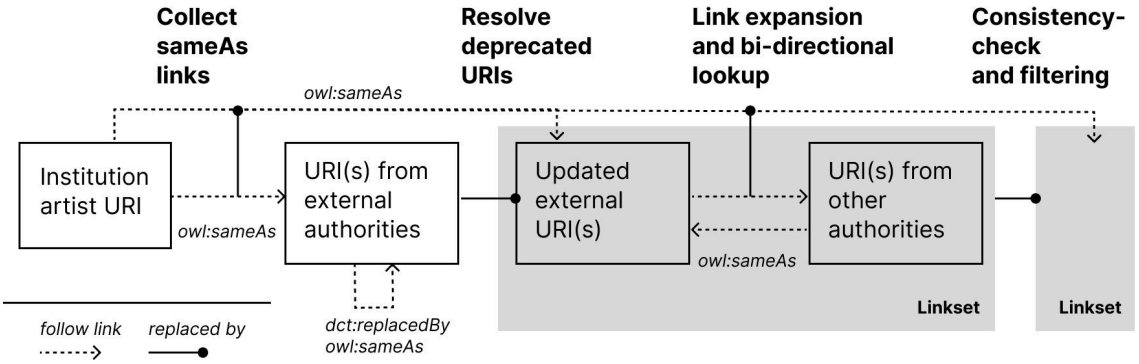

图 1。跨权威机构协调流程

具体而言，我们首先收集各机构为相关实体提供的所有 sameAs（或等效）链接，并先解析任何已弃用或过时的权威库 URI。例如，在 Zeri 照片档案中，我们会沿着艺术家对应的 ULAN 和维基数据链接进行检索，检查 AAT/ULAN 中的 dct:replacedBy 和维基数据中的 owl:sameAs，并在必要时替换过时的 URI。然后，我们尽可能扩展每个链接集。例如，弗里克艺术参考图书馆提供了指向美国国会图书馆（LoC）记录的链接。我们会沿着 LoC 链接，确认相应的 LoC 记录是否还包含指向维基数据、ULAN 及其他相关权威库的链接。我们也会进行反向检查：维基数据记录是否包含指回同一 LoC URI 的链接。通过这一流程，每个机构提供的实体都会关联到一个全面的 sameAs 链接集，其中每条链接都标注了来源信息。接下来，我们评估这些链接集在内部是否一致。理想情况下，exactMatch/sameAs 链接应具有传递性——例如，LoC URI → 维基数据 URI → LoC URI 应回到原始标识符。但实际上，情况往往并非如此：某个 LoC URI（A）可能指向一个维基数据 URI（B），而 B 又指向另一个 LoC URI（C）。出现此类不一致性时，我们会剔除相互冲突的实体对齐结果。为选出最可靠的映射，我们优先采用美国国会图书馆、德国国家图书馆、RKD 艺术家数据库和 ULAN 作为最具权威性的来源，并将维基数据作为后备来源，因为它拥有活跃的编辑社群。我们有意排除 VIAF 等聚合平台，不将其用于实体对齐决策，因为这类平台更新记录的频率较低，因此更可能包含过时或失效的外部链接。

另一个例子是机构层面进行跨机构实体对齐时继承的匹配冲突。例如，Zeri 照片档案将“乔瓦尼·安东尼奥·巴齐”与 ULAN (500015183)、VIAF (311436515 和 76586951) 及维基数据 (Q8506) 对齐。然而，在维基数据中，该统一资源标识符关联到上述 VIAF 记录 76586951，以及另一条 VIAF 记录 125158790735238852393（重复记录）。此外，VIAF 记录 311436515 也是一条重复记录，信息较少且不够精确，但 Zeri 照片档案并未舍弃任何一个最终可能正确的匹配。这又带来了一个问题：在一个显然无法假定记录具有唯一性的权威数据源中，如何从多个候选匹配项中进行选择。在 PHAROS 中，我们会协调并仅存储现有的一致链接。

第二种情况适用于摄影师：首先，由人工在各机构之间进行实体对齐，共同设计“伞形术语”并共享首选名称形式。其次，将标记为人物或工作室的实体以类别为约束条件与维基数据进行对齐，以消除假阳性匹配（例如，某个人物与维基数据中的某个组织对齐）。候选匹配项由人工审核。伞形术语及相关实体会在界面中显示为分层列表的根名称，用于筛选艺术品（图1）。

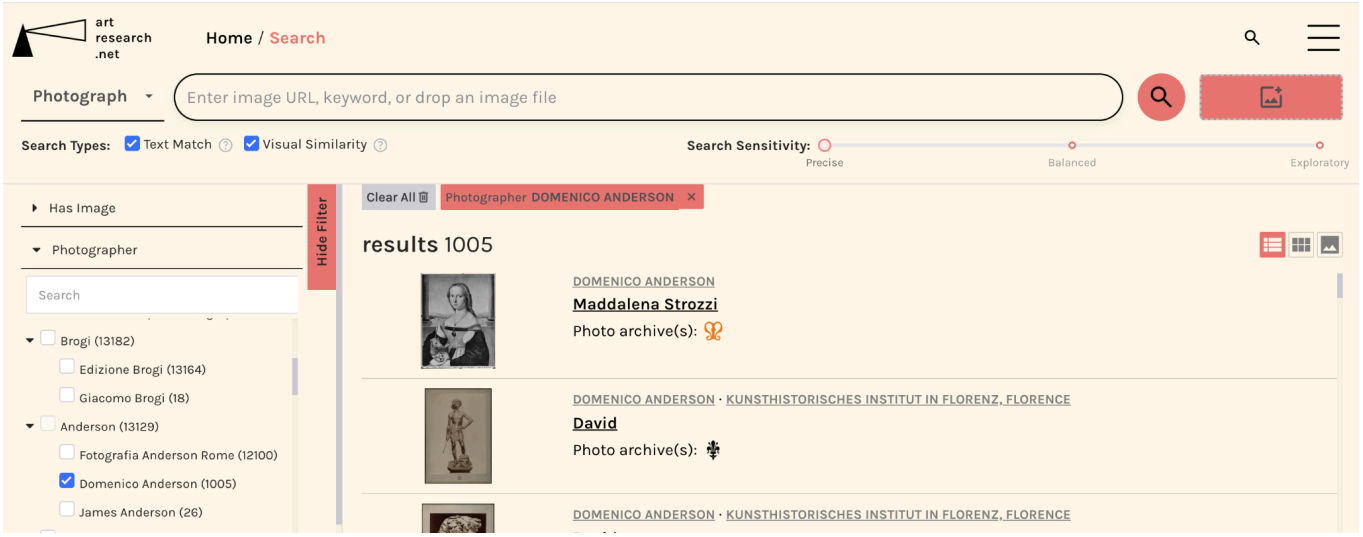

图2。artreseach.net 层级分面（左侧）展示了用于归类实体的上位术语：尽管这些实体通常以共享标签为人所知，但由于其本体定义不同，无法实现一对一匹配。

GLAM领域也存在类似模式：例如，RDA 的“作品组”机制会在无法唯一确定作者身份时，将不同的身份归入一个概念性统称之下——这与我们处理身份不明的摄影师时采用的策略相似。

5.2 结构与语义歧义（收藏、建筑物、地点）

上述策略也用于对保管者和地点进行实体匹配：系统会使用字符串匹配算法将其与维基数据自动匹配，之后再由人工审核候选对象，以检查类别声明是否一致。

为解决与收藏相关的歧义，PHAROS 采用了类似于 Heaney 2001 所提出的收藏级描述原则。我们的对齐流程为收藏创建明确的节点，在保留机构来源的同时，实现不同结构层级之间的对齐。

5.3 粒度与变体传统（标题、日期、材料）

描述粒度的差异——无论源于编目政策、机构传统还是历史不确定性——都会反复带来对齐挑战。一张照片可能有多个标题（原始标题、补充标题、译名）、不同程度的描述，或相互冲突的日期（对于正片而言，可能分别指拍摄、印制、出版和再版日期）。因此，不建议在不同机构之间或与第三方权威库进行此类信息的自动对齐，而需要人工策展。

PHAROS 并不将某个单一的“正确”值置于优先地位，而是通过明确的来源信息对所有变体进行建模，记录其来源（机构目录）和确定性类型（例如传统的、归属的）。当异质性妨碍导航时，我们会采用更高层级的分组。例如，在建立两条艺术品记录之间的链接时，相关作品的相互竞争的标题会与一个共享标题关联；该共享标题由各机构共同设计（详见下文）。系统要求审核人员将现有标题中的一个标记为“首选”；或者，如果现有标题都不令人信服，则用英语创建一个新标题。材料与技法由人工对齐至 Getty AAT 术语，以便专家处理这些术语的多层语义；随后，这些分类层级会通过层级结构呈现给用户。这种方法类似于 RDA 对作品家族和变体载体表现的处理方式，在保留细节的同时，使用户能够浏览概念一致的实体。

5.4 组织治理与工作流分配（艺术品）

对齐不仅是建模方面的挑战，也是治理方面的挑战。各机构都有自己的编目传统、对准确性的要求以及项目优先事项。由于没有任何单一权威档案可以被视为最终依据，PHAROS 采用分布式工作流：各档案馆为自身记录验证候选匹配项，随后将合并后的分类体系整合到共享知识图谱中。

主要目标之一是实现跨机构的记录对齐，以匹配等价艺术品。PHAROS 采用图像匹配算法，提出艺术品之间的跨机构对齐。候选匹配项通过用户友好界面（图 3）呈现给专家，由专家逐一接受或拒绝。不同于基于权威的对齐——其候选对象来自单一外部来源——艺术品在不同合作机构中可能存在多个潜在匹配项，每个匹配项都需要单独验证。为此开发的用户界面在不确定性的管理方式上发挥着决定性作用。通过展示相互竞争的候选对象、不同名称和来源轨迹，该界面确保了透明度，同时仍使策展人能够作出明智决策。这些设计选择反映了语义网技术领域普遍认可的更广泛现实：在开放世界假设和 AAA原则（“任何人都可以对任何事物说任何话”）下，目标并不是强行实现绝对准确，而是维护一种灵活、可扩展且具备来源信息意识的表示方式。资源描述框架的三元组结构和命名图的使用支持了这种原子化，使每个机构的贡献都能保持明确且机器可读。

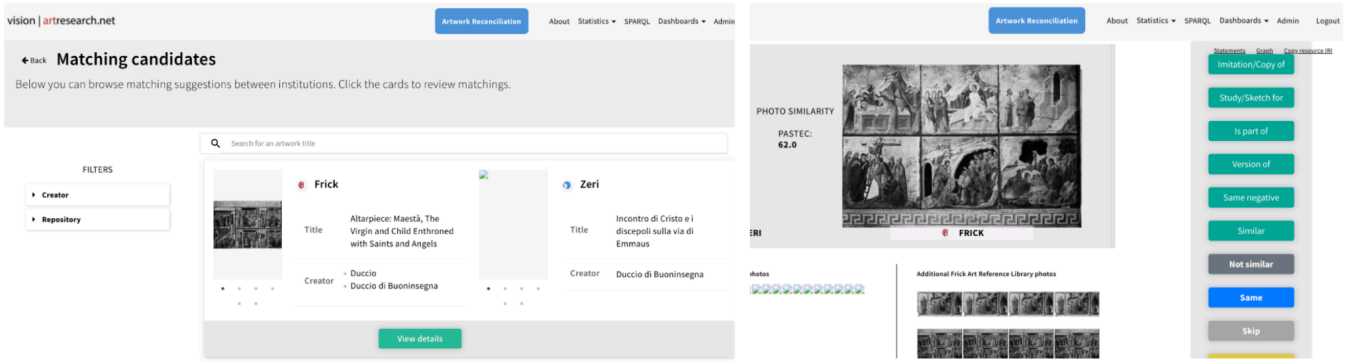

图 3. vision.artresearch.net 是 PHAROS 合作机构使用的艺术品匹配界面。审核人员可以从不同机构选择成对的艺术品，并定义它们之间的等价关系／关联关系。

管理这一对账流程会带来两个主要的社会技术挑战：协调工作分工，以及处理记录的合并（连同其继承的所有权威记录链接）。
目前，有七家机构正在试用 artresearch.net基础设施。每家机构都被视为有责任将其自身记录中提及的实体与第三方权威库进行对齐。然而，艺术品级匹配往往同时涉及多个机构，因此难以明确划分责任。虽然某家机构自然会更倾向于审核包含其馆藏艺术品的匹配，但这样做也会减轻其他机构在多机构匹配中承担的共同工作量。为避免不公平，界面会向每家机构分配数量相同的候选匹配项，不论其来源如何。这项基础工作为未来的协会成员奠定了基础，他们将在这些实体对齐结果之上，审核涉及各自艺术品的新匹配项。

完成一件艺术品的跨机构对齐后，最突出的挑战在于如何实际合并信息，并将其呈现在统一记录视图中。所有信息片段都与其来源信息相关联（在界面中以负责机构的标识概括显示），包括那些最初未受对齐影响的信息。不过，这项操作还需要专家进一步介入，他们希望统一冗余的文本描述。例如，对齐后的艺术品会继承各机构记录的所有标题，因此需要作出决定，向最终用户仅展示一个所有机构均认可的标题，而不是一串（可能很长且十分相似的）标题。在 PHAROS 中，已对齐的艺术品记录既提供机构间一致认可的主要归属标题，也保留最初记录的所有标题，以便在不那么醒目的位置保留各机构的偏好和语言变体。另一个值得关注的问题是人工制品的层级组织结构：这些人工制品分别加以描述，但合在一起构成一件完整艺术品（例如壁画组画）。在 PHAROS 中，我们保留复合作品各部分的记录，并在记录中展示其层级关系。艺术品主题的呈现也采用了类似方案；在 PHAROS 中，这些主题与 ICONCLASS 对齐。记录中保留了 ICONCLASS 的分类结构，以展示各机构针对具体案例所采用的不同粒度层级。

6. 讨论

我们的案例研究表明，文化遗产领域中的实体对齐——尤其是在艺术史照片档案中——不能被视为一项纯粹的技术任务。这是一个具有深刻阐释性、由政策驱动的过程，受到参与机构的知识传统和编目实践所塑造。尽管算法方法、规范控制和关联开放数据框架提供了至关重要的基础设施，却无法消除文化数据固有的歧义。PHAROS 的实践表明，对齐需要接纳而非抹去这种歧义，同时仍要构建易于使用、便于浏览的知识图谱；而这又要求领域专家参与工作流程中多个阶段的决策过程。

一项关键发现是可用性与语文学准确性之间的张力。语义网技术鼓励创建非常复杂的数据；然而在实践中，现有工具限制颇多，并倾向于扁平化和规范化数据，使其能够纳入整齐的层级结构、筛选器和可视化之中。然而，艺术史数据充斥着模糊性、相互竞争的归属以及不断演变的学术观点。严格强制统一，可能会抹去这种复杂性；而将其完全呈现出来，又可能令用户不堪重负。PHAROS 不得不寻求折中：创建上位术语，将不确定或异质的实体归入共享标签之下，并采用分层用户界面分类面，默认隐藏复杂性，同时允许用户按需探索。这些设计选择表明，对齐不仅关乎数据建模，也同样关乎用户体验。

本研究发现的对齐挑战，与语义网的基本假设密切相关。开放世界假设（OWA）确保信息的缺失不能被解读为矛盾，因此可以整合不完整、不确定或彼此不同的陈述，而不必过早定论。AAA原则（“任何人都可以对任何事物说任何话”）有助于解释为何外部规范数据集偶尔会包含相互冲突或不兼容的断言——这是一种预期情况，而非错误。资源描述框架（RDF）的三元组结构旨在供机器处理原子陈述，既能捕获细粒度的来源信息，又不规定应如何向用户呈现这些陈述。这使建模开放性与由界面驱动的简化之间自然产生张力，文化遗产机构必须通过对齐政策和界面设计明确处理这种张力。

本项目还表明，鉴于数据在历史和机构层面的异质性，完全自动化的对齐并不可行。同样，人工审核不仅是质量控制，更主要是一项认识论工作。文化遗产专家不会满足于概率性的“最佳猜测”，而是期待由人类参与判断且有充分依据的断言。这对人工智能和大语言模型（LLM）在对齐中的作用有着重要影响。大语言模型和嵌入可以帮助找出候选匹配项，或缩小大量可能性的范围，但无法取代专家判断，也不应被视为真相来源；尤其是在连人类都难以确认身份的情况下（Doerr 2025）。它们的概率性本质与人文学者对证据的要求相冲突，实验也表明，它们可能生成损害信任的幻觉链接。

另一个重要之处在于，对齐并非一次性任务，而是一个持续进行的过程。新的数据摄取、不断演变的规范记录以及变化中的学术观点，都会持续对现有对齐关系提出挑战。这要求采取可持续的组织策略：迭代式摄取管道、对匹配结果进行版本管理和来源追踪，以及制定明确的政策，说明何时以及如何修订先前接受的链接。文献中很少涉及这些治理层面的问题，但它们对于长期维持可靠性，以及建立人们对聚合型文化遗产平台的信任至关重要。

归根结底，对齐必须接受一定程度的不完美。追求绝对精确既不现实，也会适得其反；关键在于设计能够透明地管理不确定性、并明确说明取舍的流程。PHAROS 表明，与其强行采用一对一或多对一匹配，不如接受多对多关系，这是一条可行的前进路径，前提是通过审慎的建模和用户界面设计来处理这种复杂性。这意味着，对齐不应被视为最终目标——即生成单一权威视图，而更应被视为一种基础设施，用以支持在异质文化数据之间持续协商意义。

7. 结论

本研究考察了艺术史照片档案中异质数据的对齐，重点关注 PHAROS 项目中形成的挑战与策略。我们的分析表明，对齐并非纯粹的技术任务，而是一项阐释性、协作性的工作，受到机构遗产、不断演变的编目实践以及学科认识论的影响。缺乏全面的规范档案列表、元数据存在冲突或不完整，以及艺术史知识本身的模糊性，使得完全自动化无法实现，因此需要在算法支持与专家主导的决策之间谨慎平衡。

我们强调，创建上位术语、对不确定性进行建模，以及谨慎地区分（或归组）实体等策略，既能保留编目实践的历史特殊性，又能为终端用户提供连贯且易于使用的界面。这种方法将对齐重新界定为一个持续且可修订的过程，而非一次性地试图生成一个确定、权威的数据集。

我们如今的重点并非追求文化遗产数据集之间完全一致——这一目标与国际书目控制的早期诠释相关，如今已普遍被认为既不现实，也不可取——而是实现可互操作且具备来源信息的表示，同时保留多元声音。在此背景下，对齐并非将数据规范化为单一真相的过程，而是一种以透明且可复用的方式管理异质性的方法。采用这一视角，PHAROS 有助于重新界定对齐：它不是消除异质性，而是提供一个调和异质性并帮助人们理解异质性的框架，为围绕摄影档案及借助摄影档案开展研究开辟新的可能性。未来，PHAROS 将通过定制的 OpenRefine 服务，公开提供相关方法和数据，包括已对齐的实体；这项服务将承袭艺术史照片档案中记录的丰富诠释多样性，并将其传递给更广泛的艺术史界。

PHAROS 扎根于摄影档案领域。然而，本文所述的对齐策略和管理挑战远不止于此。图书馆在依据图书馆参考模型、RDA 和旧有编目规则对齐名称规范档时，面临类似问题；档案机构在 EAC-CPF记录和传记实体中也会遇到相似的歧义；博物馆则必须在 ULAN、AAT 和本地叙词表等系统之间对齐异构的对象描述、归属和馆藏历史。因此，本文提出的框架——整合识别、建模、工作流程和界面策略——为任何寻求在共享的语义互操作环境下整合异构数据集的 GLAM 机构提供了可复用的结构。

致谢

作者谨向所有参与人工对齐工作，以及参与为决策过程提供依据的讨论的同事致谢。特别感谢 Ute Dercks、Dagmar Keultjes、Kurt Scharenberg（KHI）；Pietro Liuzzo（Hertziana）；John McQuaid、Kerri A. Pfister（Frick）；Spyros Koulouris（I Tatti）；Christoph Glorious（马尔堡）和 Louisa Wood Ruby 提供的细致审阅和建设性意见。他们的专业知识和对细节的关注，对于小组讨论、解决分歧以及完善最终决策都至关重要。

参考文献

1. Baker, T., Bermès, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., Panzer, M., Schneider, J., Singer, R., Summers, E., Waites, W., Young, J. and Zeng, M. (2011), 图书馆关联数据孵化组最终报告：万维网联盟孵化组报告，*2011年10月25日*. https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ 2. Barabucci, G., Tomasi, F. and Vitali, F. (2022), “公共史学与文化遗产中的数据复杂性建模”，Noiret, S., Tebeau, M., and Zaagsma G. (eds.),数字公共史学手册，德格鲁伊特·奥尔登堡出版社，柏林，波士顿，第459-474页。https://doi.org/10.1515/9783110430295-041 3. Battisti, T., and Daquino, M. (2025), “数字人文信息可视化项目中的不确定性、叙事性与批判性方法”，载于第*21*届信息与研究科学：数字与图书馆科学连接会议论文集，CEUR，第1-10页。4. Caraffa, C., Pugh, E., Stuber, T. and Wood Ruby, L. (2020), “Pharos：照片档案数字研究空间”，《艺术图书馆学刊》，第45卷第1期，第2-11页。https://doi.org/10.1017/alj.2019.34 5. CIDOC CRM特别兴趣小组 (2021), *CIDOC*概念参考模型定义，*7.1.1*版。编者：Bekiari, C., Bruseker, G., Doerr, M., Ore, C-E., Stead, S., Velios, A. https://cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf

6. Coyle, K., and Hillmann, D. (2007), “资源描述与检索（RDA）”，*D-Lib* 杂志，第13卷第1/2期，第1082-9873页。 https://mirror.dlib.org/dlib/january07/coyle/01coyle.html 7. Delmas-Glass, E. and Sanderson, R. (2020), “通过采用开放标准促进PHAROS学者社群发展”《艺术图书馆学刊》，第45卷第1期，第19–23页。https://doi.org/10.1017/alj.2019.32. 8. Delpeuch, A. (2019) “OpenRefine 对齐服务调查”，载于第十五届本体匹配国际研讨会论文集，CEUR, 第1-5页。9. Di Cicco, V., Firmani, D., Koudas, N., Merialdo, P. and Srivastava, D. (2019), “解读用于实体消歧的深度学习模型：使用LIME的经验报告”，载于第二届利用人工智能技术进行数据管理国际研讨会论文集，计算机协会，第1-4页。https://doi.org/10.1145/3329859.3329878 10. Doerr, M. (2025), “可识别个体与现实：通过形式命题描述过去”，《计算与文化遗产期刊》，第18卷第2期，第1-32页。https://doi.org/10.1145/3715011 11. 都柏林核心馆藏描述工作组 (2004), 都柏林核心馆藏描述应用纲要：数据模型。
https://www.dublincore.org/specifications/dublin-core/collection-description/collection-model/ 12. Heaney, M. (2001), 馆藏及其目录分析模型。英国图书馆与信息网络办公室。https://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf 13. Hickey, T. B., and Toves. J. A. (2014), “管理VIAF中的歧义。” *D-lib* 杂志，第20卷第7/8期。https://doi.org/10.1045/july2014-hickey 14. Hyvönen, E. (2012), “在语义网上发布和使用文化遗产关联数据”，Morgan & Claypool 出版社，圣拉斐尔（加利福尼亚州）。15. Isaac, A., and Haslhofer, B. (2013), “Europeana 关联开放数据–data. europeana. eu”，语义网，第4卷第3期，第291-297页。https://doi.org/10.3233/SW-120092 16. Joudrey, D. N., Taylor, A. G., and Miller, D. P. (2015), “编目与分类导论”，布鲁姆斯伯里美国出版社。17. Knoblock, C. A., Szekely, P., Fink, E., Degler, D., Newbury, D., Sanderson, R., Blanch, K. et al. (2017), “为美国艺术协作组织构建关联数据的经验教训”，载于语义网国际会议，施普林格国际出版社，查姆，第263-279页。https://doi.org/10.1007/978-3-319-68204-4_2618. 美国国会图书馆。(1999年), *MARC 21* 书目数据格式。网络开发与 MARC 标准办公室，美国国会图书馆，华盛顿哥伦比亚特区。19. Ngonga Ngomo, A., Sherif, M. A., Georgala, K., Hassan, M. M., Dreßler, K., Lyko, K., Obraczka, D., and Soru, T. (2021), “LIMES：语义网上链接发现框架”，《人工智能》，第35卷第3期，第413-423页。https://doi.org/10.1007/s13218-021-00713-x 20. Oldman, D., and Tanase, D. (2018), “通过在ResearchSpace中连接研究者、数据与实践，重塑知识图谱。”载于语义网国际会议，施普林格国际出版社，查姆，第325-340页。https://doi.org/10.1007/978-3-030-00668-6_2021. Riva, P., and Žumer, M. (2018), “FRBRoo、IFLA图书馆参考模型以及LRMoo：发展的循环”，载于国际图联世界图书馆与信息大会——吉隆坡，马来西亚——变革图书馆，变革社会。http://library.ifla.org/id/eprint/2130 22. Vitali, F., and Pasqual, P. (2025), “将历史不确定性表征为相互竞争且不相容的确定性的结果”，载于早期现代私人政治档案关系数据库中的数据提取与架构模型，威尼斯大学出版社，第61-77页。https://doi.org/10.30687/978-88-6969-919-1//003 23. Zeng, K., Li, C., Hou, L., Li, J., Feng, L. (2021), “知识图谱实体对齐综合综述”，《*AI Open*》，第2卷，第1–13页。https://doi.org/10.1016/j.aiopen.2021.02.002