CHICO：连接文化遗产馆藏中的稀疏元数据与多模态证据

Vagner Inácio de Oliveira，Paula Dornhofer Paro Costav208939@dac.unicamp.br, paulad@unicamp.br

坎皮纳斯州立大学（UNICAMP）电气与计算机工程学院（FEEC）坎皮纳斯，圣保罗州，巴西

arXiv:2611.30923v1 [cs.DL] 23 Nov 2026

摘要

文化遗产聚合平台提供对数百万件数字化藏品的访问。然而，其目录记录通常包含稀疏且异构的元数据，关联媒体链接也可能无法指向可访问的图像。对这些资源进行结构化和验证，有助于多模态人工智能研究实现可复现实验。为此，我们推出CHICO（文化遗产图像与目录对象），这是一个源自 Europeana 的多模态数据集，汇集了整理后的目录字段、图像和上下文信息，用于支持基于证据的描述生成，同时保留原始 Europeana记录。该数据集包含超过 60,000 条带有关联图像的文化遗产记录。尽管它最初是为这一目的开发的，但也可用于多模态信息检索、元数据丰富和基于证据的评估。

1 引言

博物馆、图书馆和档案馆已经将数百万件文化遗产藏品数字化，使其能够通过 Europeana（Isaac and Haslhofer，2013）和 Brasiliana Museus（Siqueira etal.，2022年b）等聚合平台更加便捷地获取。这些藏品汇集了来自不同机构的记录，体现出多样化的编目实践、元数据结构和完整性水平。在此背景下，目录描述有助于访客发现并理解文化遗产藏品，而其可用性和详细程度则反映了藏品与编目实践的多样性。除目录元数据外，关联图像和上下文注释还可以提供补充视角。综合这些来源，可以采用多模态方法，包括根据与每件藏品相关的信息生成文本描述。

这些信息来源的可用性和结构在不同文化遗产记录中差异很大。目录信息（包括标题、日期和创作者）在不同藏品中的记录详细程度各不相同。同样，媒体链接也并非总能提供对可用图

。机构描述在可用时，其形式从简短陈述到详细的策展文本不等，且可能包含其他元数据字段或图像中未体现的信息。因此，将这些描述视为生成或评估的真实依据，可能会引入无法通过现有证据验证的信息。此外，Europeana记录遵循 Europeana 数据模型（EDM），结合了嵌套结构、多语言值以及基于 URI 的引用（Isaac and Haslhofer，2013）。将这些信息源整合到多模态方法中，需要选择相关目录字段、处理异构元数据，并验证关联图像的可用性。

CHICO（文化遗产图像与目录对象）是一个源自 Europeana记录的多模态数据集，旨在支持以证据为依据的描述生成。该数据集汇集了经过整理的目录字段、关联图像和上下文信息，同时保留原始 Europeana记录。其结构为这些互补来源提供了一致的表示形式，支持可复现实验，并有望用于其他多模态方法。该数据集包含来自74个提供方的60,397条文化遗产记录，其中83,852张图像已确认可解码。构建与分析代码已公开发布；[1] 数据集制品已发布于 Zenodo（Oliveira 和 Costa，2026年）。[2]

图 1 展示了 CHICO 的构建过程，从Europeana馆藏到最终的数据集。该过程包括整理记录及其相关媒体、整理目录信息，并将其与图像关联

[1]https://github.com/AI-Unicamp/heritage-data-hub[2]https://doi.org/10.5281/zenodo.23954726

，以及通过检索和排序实体注释来纳入上下文信息。最终的记录将机构描述单独保留，作为后续使用的参考。

本研究的三项主要贡献是：整理并构建了一个多模态数据集，包含 60,397 条Europeana记录及其关联图像、目录字段和上下文注释，同时保留原始记录及其来源信息；提供了有文档记录的构建流程，并划分训练集、验证集和测试集，以避免相关记录之间发生重叠；以及刻画了该数据集在不同藏品中的组成、语言和目录字段覆盖率。

以下章节回顾相关工作（第 2 节），介绍 CHICO 及其构建过程（第 3 节），说明其获取方式（第 4 节），讨论局限性（第 5 节），并给出总结性评论（第 6 节）。

2 相关工作

文化遗产领域的数据集与多模态方法。 现有资源为文化遗产物品的视觉—语言研究提供支持，涵盖不同任务和标注格式。DEArt 提供欧洲艺术图像及其物体与姿态标注，可用于检测和场景理解（Reshetnikov 等人，2023年）。EMA 推出了巴西文化遗产图像数据集，其中包含来自博物馆藏品的标注图像，可通过图像分类支持自动化元数据标注（de Oliveira 等人，2022）。在艺术品图像描述方面，Cioni 等人（2023年）探索了基于扩散模型的数据增强，而Castellano 等人（2024年）结合视觉—语言模型与基于语言的方法，生成并优化文本描述。Bongini 等人（2023年）研究了视觉问答，通过结合视觉信息与有关艺术品的上下文知识来实现。其他方法则利用 Iconclass 分类系统探索图像学概念的多模态检索（Banar 等人，2021；Santini 等人，2023年）。

元数据聚合与数字编目。
数字聚合计划汇集了来自不同机构的文化遗产记录，反映出多样化的编目实践和元数据结构。Europeana 和 Brasiliana Museus 展示了聚合并提供数字文化遗产藏品访问的不同方式（Isaac and Haslhofer，2013；Siqueira et al.，2022年b）。Europeana 使用 Europeana 数据模型（EDM）表示聚合记录，并纳入都柏林核心元数据术语和关联实体（Isaac and Haslhofer，2013；都柏林核心元数据倡议，2020）。在巴西，Tainacan 平台支持数字博物馆藏品的管理与发布，而 Brasiliana Museus 则通过元数据映射和聚合，将来自不同机构的藏品汇集起来（Siqueiraet al.，2022年a）。

针对这些资源的研究探索了语义丰富化和计算编目的方法。 Koch 等人（2021）研究了运用语义词汇表和图像分析来改进历史图像的描述与检索；而Domingo 等人（2020）则将视觉识别与语义整合相结合，并应用于 Europeana。 Abele 等人（2026年）探讨了视觉语言模型在档案图像编目中的应用，并比较了生成式描述与人工撰写文本。机构描述提供了宝贵的策展知识，尽管其内容可能超出关联元数据或图像所能提供的证据范围（Malevé，2021；Hanley 等人，2021）。

CHICO 以文化遗产记录为分析单位，将整理后的目录字段、关联图像和上下文证据汇集起来，同时保留其原始 Europeana 来源信息。

3 CHICO 数据集

目录元数据、视觉信息和上下文来源的整合，构成了 CHICO 多模态结构的基础。该数据集源自 Europeana，在保留原始记录的同时纳入整理后的信息，以呈现这些互补来源。其构建过程包括筛选和处理记录、验证关联图像，以及整理信息以供后续实验使用。

3.1 数据集构建

该数据集基于对约 3,350 万条 Europeana记录的本地采集构建而成，采集过程不设主题限制，并筛选出开放许可内容以及已声明可提供媒体的记录。每条记录均以完整的 Europeana 数据模型（EDM）形式获取，并经过规范化处理，以便于数据管理和后续处理。在此基础上，随机抽取了 100,000 条记录，且未按提供方、语言或项目类型进行分层抽样—

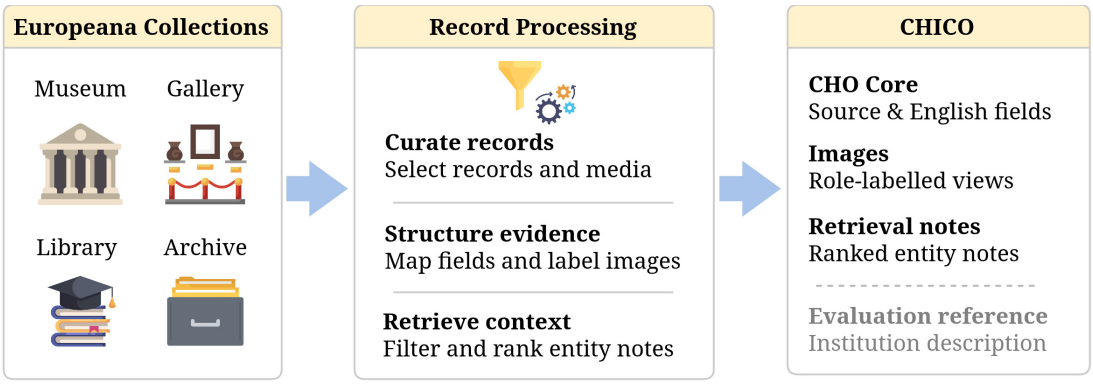


图1：从 Europeana 馆藏构建 CHICO，形成目录字段、角色标注图像和排序检索注释，并保留机构描述作为评估参考。

按提供方、语言或项目类型进行筛选。

最终研究样本还通过验证关联图像的可用性进行筛选。我们下载媒体文件，并检查其是否能够成功解码为栅格图像。仅保留至少有一张图像成功解码的记录，最终得到来自 74 个提供方的 60,397 条记录和 83,852 张图像（84.8% 为单图像记录；15.2% 为多图像记录）。这一步验证也影响了样本构成：媒体链接指向可解码图像的藏品比链接到查看器页面或打包扫描件的藏品更有可能被保留。

为描述数据集的构成，我们根据目录中记录的类型，将记录划分为若干宽泛的项目类别，并同时考量源语言和英语表述（图2）。自然史标本是数量最多的类别，在初始随机样本中已约占 40%；图像验证后，其占比进一步上升，因为这些提供方更常提供指向可解码图像的媒体链接。因此，最终形成的是一个固定的、经过图像验证的研究样本，而非经过平衡处理或具有代表性的 Europeana 馆藏总体样本。

目录记录为多语言，反映了参与藏品的多样性。英语是最常声明的目录语言，同时也包含大量非英语材料。目录字段保留其源语言，并在可用时提供英语表示。

为尽量减少实验子集之间的信息泄漏，我们将相关记录归入连通分量，并将每个连通分量整体分配到训练集（80%）、验证集（10%）和测试集（10%）划分中。

3.2 多模态记录结构

CHICO 的多模态结构汇集了每件文化遗产项目的目录元数据、关联图像和上下文信息。该结构定义为

$$E = (M,\ I,\ R,\ d^{\text{ref}}), \tag{1}$$

其中，M 包含整理后的目录字段，I 表示关联图像及其角色，R包含从记录所引用的相关实体中选取的上下文注释，d^{ref} 表示可用时的机构描述。原始 Europeana 记录也会予以保留，从而维持对源元数据的访问。对于以证据为依据的描述生成，M、I和 R 构成指定的支持性证据，而d^{ref} 则单独保留以供比较，并从生成输入中排除。

目录字段的选取依据是对随机抽取的 1,000 条 Europeana 记录进行的考察，重点关注字段的可用性、多语言表示形式，以及这些字段对于描述文化遗产物品的相关性（Isaac 和 Hasl- hofer，2013）。这项分析为精选文化遗产对象（CHO）字段的定义提供了依据，优先纳入与物品本身相关的信息，而非管理元数据。字段选取兼顾描述相关性与可用性；即使来源信息、材质和范围等信息的记录频率较低，仍予以保留。M 物品层级字段包括标题、类型、创作者、日期、时期、地点、主题、材质、尺寸和所有权历史。不同文化遗产类别的目录信息分布差异显著（图3a）。

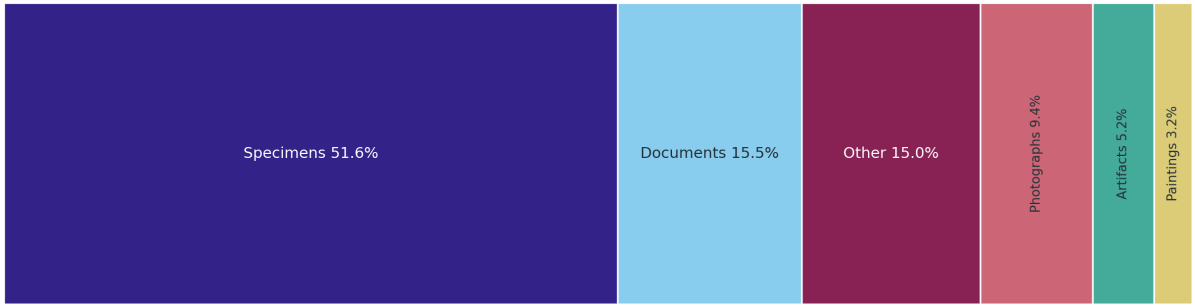


图 2：60,397 条记录中物品类别的分布。

关联图像（I）会被赋予不同角色，以区分物品视图、细节、标签、装裱物和模糊帧。仅包含一张图像的记录会被视为物品视图；包含多张图像的记录则会自动添加标签，以便后续使用时将外观证据与包含文本的帧或非物品帧区分开来。

上下文信息（R）来自记录所引用的相关实体（概念、地点、行动者和时间跨度）的描述性注释，而非 M中整理后的目录字段。在每条记录中，使用 BGE-M3 检索候选注释，并通过 BGE-reranker- v2-m3 重新排序（Chen 等人，2024年），并以 M为指导，使相关性更高的注释排在前面。随后对排序列表进行筛选，以避免重复这些字段中已陈述的信息；如果没有合适的注释剩余，则将 R 留空。

机构描述（d^{ref}）在不同物品类别中的出现频率和篇幅各不相同（图3b）。因此，这些描述不会被视为物品描述的真实依据：该字段经常为空，即使有内容，也无法可靠地完整反映物品情况。

4 可用性

CHICO 可公开获取，用于研究。构建与分析代码已发布在 heritage-data-hub 仓库中。[3] 该数据集存储于 Zenodo（Oliveira 和 Costa，2026年）。存储内容包括目录记录、训练/验证/测试集划分、附带校验和以供验证的图像URL，以及文档。可选的归档文件还提供上下文补充注释。图像文件不再分发；可通过记录的URL获取，并根据提供的校验和进行核验，但须遵守各条记录的版权和再利用条款。

5 局限性

作为从附带关联媒体的开放许可 Europeana 记录中提取的固定研究样本，该数据集不应被视为能够代表范围更广的 Europeana 馆藏。成功解码图像这一要求会影响样本构成，使其偏向媒体链接可解析出可解码图像的提供方，而非链接到查看器页面、打包扫描件或不可用内容的提供方。

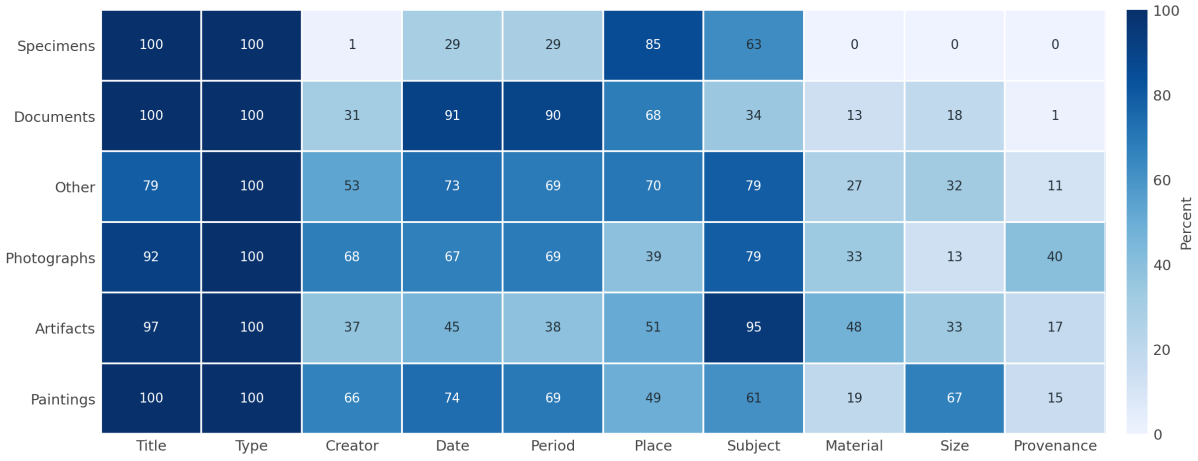
数据集划分通过标识符、图像URL和加密哈希值的精确匹配来关联相关记录。馆藏中已有的视觉相似图像通常具有不同的哈希值，因此不会被视为相关项，这意味着此类近似重复项仍可能出现在不同的数据划分中。此外，数据集分析采用的宽泛项目类别源自异质的目录类型值，可能无法充分体现原始分类。上下文丰富仅依据每条记录所引用的相关实体的备注。

尽管 CHICO 最初旨在支持基于证据的描述生成，但此版本侧重于数据集构建与特征描述。其在其他应用中的潜在用途，包括多模态检索和元数据丰富，仍有待评估。

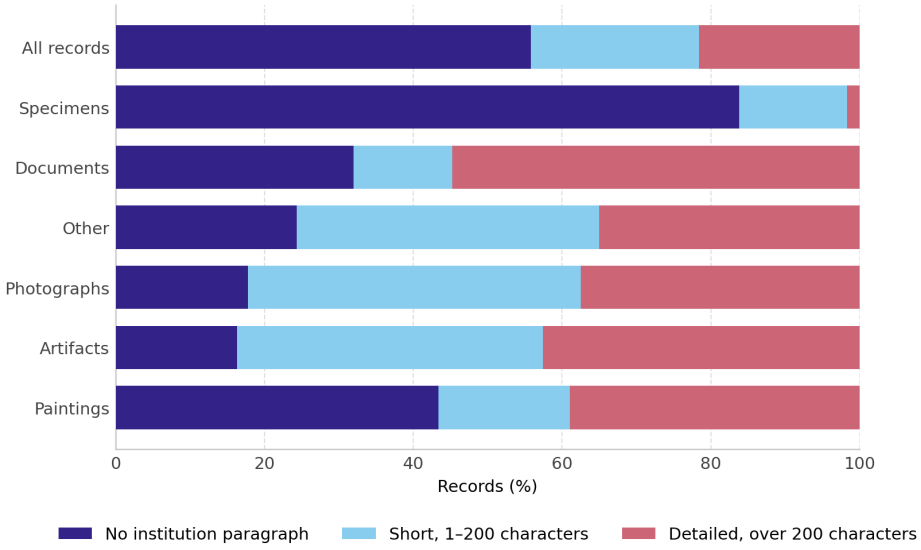
6 结论

文化遗产聚合平台中的目录记录稀疏且异质，难以汇集多模态证据以开展可复现的研究。本研究通过构建一个包含 60,397 条 Europeana记录的固定研究样本来弥补这一不足。该样本汇集了整理后的目录字段、经核实的图像和上下文注释，同时保留

[3] https://github.com/AI-Unicamp/heritage-data-hub



(a)



(b)

图3：（a ）按对象类别划分的整理后的目录字段覆盖情况。（b）按对象类别划分的机构描述（d^{ref}）。

原始记录及其来源信息。构建过程中形成了预定义划分，将相关记录归为一组；同时还构建了多模态结构，将机构描述单独保留，以供比较。本文从数据集的组成、语言和各藏品的目录字段覆盖率等方面对其进行介绍，并记录了数据集的构建过程。代码和数据均已公开，为基于证据的描述生成及文化遗产馆藏相关多模态研究提供了可复用的基础。

致谢

本研究部分由巴西高等教育人员提升协调机构（CAPES）– 财务代码001资助。作者也隶属于人工智能实验室 Recod.ai。

参考文献

Line Abele、Gerrit Anders、Tolgahan Aydın、Jürgen Buder、HelenFischer、Dominik Kimmel 和 Markus Huff。2026年。ArchiveGPT：以人为本的视觉语言模型图像编目使用评估。人文与社会科学传播，13(1):1241。

NikolayBanar、Walter Daelemans 和 MikeKestemont。2021年。视觉艺术的多模态标签检索：以 Iconclass 为例。载于第*13*届国际智能体与人工智能会议（*ICAART*）论文集，第1卷，第622–629页。

Pietro Bongini、Federico Becattini 和 Alberto Del Bimbo。2023年。GPT-3 是否足以胜任文化遗产领域的视觉问答？载于计算机视觉——*ECCV 2022* 研讨会，《计算机科学讲义》第13801卷，《计算机科学讲义》，第268–281页。

Giovanna Castellano、Nicola Fanelli、Raffaele Scaringi 和 Gennaro Vessio。2024年。探索视觉-语言预训练与 ChatGPT 在艺术品图像描述中的协同作用：一项初步研究。载于图像分析与处理——*ICIAP 2023* 研讨会，《计算机科学讲义》。

Jianlv Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2024年。M3-embedding：通过自知识蒸馏实现多语言性、多功能性和多粒度文本嵌入。载于计算语言学协会论文集：*ACL 2024*，第2318–2335页。Dario Cioni、Lorenzo Berlincioni、Federico Becattini 和 Alberto Del Bimbo。2023年。用于文化遗产图像描述和检索的基于扩散的增强方法。载于*IEEE/CVF* 计算机视觉国际会议研讨会论文集（*ICCVW*），第1699–1708页。Vagner de Oliveira、Dalton Martins 和 Paula Costa。2022年。EMA：巴西文化遗产图像数据集——面向数字馆藏的基于人工智能的元数据标注。载于*iPRES2022*。Juan Duque Domingo、Jaime Gómez-García-Bermejo 和 Eduardo Zalama。2020年。利用深度神经网络和与 Europeana 的语义集成识别埃及沙布提俑。《应用科学》，10(18):6408。都柏林核心元数据倡议。2020年。DCMI 元数据术语。
h
t
t
p
s://
w
w
w.
d
u
b
lincore.org/specifications/dublin-core/dcmi-terms/。访问日期：2026年10月5日。Margot Hanley、Solon Barocas、Karen Levy、Shiri Azenkot 和 Helen Nissenbaum。2021年。计算机视觉与价值冲突：使用自动生成的替代文本描述人物。载于*AAAI/ACM* 人工智能、伦理与社会会议论文集（*AIES*），第543–554页。Antoine Isaac 和 Bernhard Haslhofer。2013年。Europeana 关联开放数据：data.europeana.eu。语义网，4(3):291–297。Gerda Koch、Amelie Dorn、Renato Rocha Souza、Yalemisew M. Abgaz 和 Ramiro Osvaldo Ortiz。2021年。使用 ChIA 词汇表描述和检索历史图像中的食物内容。载于*SWODCH*。Nicolas Malevé。2021年。在数据集的废墟之上。《人工智能与社会》，36(4):1117–1131。Vagner Oliveira 和 Paula Costa。2026年。CHICO：文化遗产图像与目录对象。数据集。

Artem Reshetnikov、Maria-Cristina Marinescu 和 Joaquim More Lopez。2023年。DEArt：欧洲艺术数据集。载于计算机视觉 – *ECCV 2022* 研讨会，第13801卷，收录于《计算机科学讲义》，第218–233页。

Cristian Santini、Etienne Posthumus、Tabea Tietz、Mary Ann Tan、Oleksandra Bruns 和 Harald Sack。2023年。使用视觉-语言预训练模型搜索 Iconclass。载于*ACM/IEEE*数字图书馆联合会议（*JCDL*）论文集，第 285–287 页。

Joyce Siqueira、Dalton Lopes Martins 和 Vinícius Nunes Medeiros。2022年a。Tainacan 中的元数据映射：面向巴西博物馆研究所关联数字博物馆的新功能。载于《信息科学高级札记》，第 2 卷，第 182–191 页。Pro-Metrics。

Joyce Siqueira、Dalton Lopes Martins 和 Daniela Lucas da Silva Lemos。2022年b。Brasiliana Museus：巴西博物馆研究所数字馆藏聚合信息的搜索与检索服务。*Enancib*，2(1)。