arXiv:2603.28108v1 [cs.DL] 30 Mar 2026

何为 VERITAS？档案文档分析的模块化框架

Leonardo Bassanini[1], **Ludovico Biancardi**[1], **Alfio Ferrara**[2], **Andrea Gamberini**[3], **Sergio Picascia**[4*], **Folco Vaglienti**[3]

[1]米兰大学，大学图书馆服务处，圣索菲亚街 7/9 号 - 20122 米兰（意大利）leonardo.bassanini@unimi.it, ludovico.biancardi@unimi.it

[2]米兰大学，文学研究、语文学与语言学系，赦罪节街 7 号 - 20122 米兰（意大利）alfio.ferrara@unimi.it

[3]米兰大学，历史研究系，赦罪节街 7 号 - 20122 米兰（意大利）andrea.gamberini@unimi.it, folco.vaglienti@unimi.it

[4]米兰大学，计算机科学系，切洛里亚街 18 号 - 20133 米兰（意大利）sergio.picascia@unimi.it

摘要

历史文献的数字化传统上被视为仅限于字符级转录的过程，所生成的纯文本缺乏开展实质性计算分析所需的结构与语义信息。我们介绍 VERITAS（档案来源的视觉增强阅读、阐释与转录），这是一个模块化、模型无关的框架，将数字化重新构想为涵盖转录、版面分析和语义增强的集成工作流。该处理流程分为四个阶段——预处理、抽取、精炼和增强——并采用模式驱动架构，使研究人员能够以声明式方式指定提取目标。我们在贝尔纳迪诺·科里奥的《米兰史》校勘本上评估了 VERITAS；这部文艺复兴编年史超过 1,600 页。结果表明，与商业 OCR 基线相比，该处理流程的词错误率相对降低了 67.6%；若将人工校正纳入考量，端到端处理时间则缩短至原来的三分之一。我们还通过检索增强生成系统查询转录语料库，展示该处理流程输出在下游应用中的价值，以及其支持历史研究的能力。

关键词： 视觉语言模型、文档版面分析、历史文献数字化、数字人文

1. 引言

长期以来，历史文献数字化一直被视为一个过程，其主要目标是将文本内容忠实地转录为机器可读形式。在这一范式下，产出是扁平文本：一串适合关键词检索、却缺乏开展实质性学术研究所需结构、语义和关系信息的字符序列。这一局限源于一种根深蒂固的架构假设，即语义增强属于独立的下游任务。实际上，这种分离意味着大多数数字化馆藏始终无法获得此类增强，因为重新处理已处理材料并执行增强流水线所需的资源往往难以获得。其后果是，数字化材料的数量与真正可检索、可互操作且适用于计算分析。

我们认为，要弥合这一差距，就必须重新构想数字化本身：不应将语义增强视为可有可无的后处理，而应将其作为数字化工作流程不可或缺的一部分，从而生成结构化、经过语义标注的数字对象。基础模型的最新进展，尤其是视觉语言模型（VLM）和大语言模型（LLM），使这一转变成为可能。传统 OCR系统仅在字符层面进行处理，并不理解文档语义；与之不同，视觉语言模型能够联合处理视觉信息和文本信息，在统一的推理过程中完成版面分析、转录和内容解读。反过来，大语言模型可用于执行以往需要专门的特定任务系统才能完成的下游语义任务。这些技术相结合，使构建处理流程成为可行之举，能够在单一自动化工作流程中贯穿从原始图像到语义

增强数字对象的完整流程。

本文介绍 VERITAS（档案资料的视觉增强阅读、阐释与转录），这是一个将这一集成方法付诸实践的模块化框架。VERITAS 由四个依次衔接的阶段构成，每个阶段都是一个独立模块，具有明确定义的输入和输出。该框架不依赖特定模型：随着新模型不断推出，各个组件均可替换，而无需修改架构。模式驱动的信息抽取流程使研究人员能够根据分析目标配置目标数据结构。我们以贝尔纳迪诺·科里奥的《米兰史》为对象评估 VERITAS。这部文艺复兴编年史篇幅超过 1,600 页；评估结果表明，与商业 OCR 基线相比，VERITAS 在转录准确率和处理效率方面均有显著提升。此外，我们还展示了其在后续任务中的实用价值：将完整转录文本提交给检索增强生成（RAG）系统，并提出具有历史依据的研究问题。

2. 相关工作

历史文献的自动化处理日益受到关注，尤其是在计算机科学与数字人文的交叉领域。我们从两个方面回顾相关文献：转录技术和集成数字化流水线。

2.1. 历史文献转录

历史文献的转录经历了多个技术范式的发展。商业光学字符识别引擎（例如 Tesseract（Smith, 2007年）长期以来一直是默认工具；然而，这类引擎依赖字符级模式匹配，因此难以应对历史材料中质量退化的图像、不规则的版面以及古老的排版惯例。深度学习显著提升了此类文献的处理能力。Transkribus（Kahle 等人，2017年；Muehlberger 等人，2019年）由欧盟资助的READ项目开发，成为人文学科中应用最广泛的自动文本识别平台之一。与此同时，eScriptorium（Kiessling 等人，2019年）将识别功能扩展到非拉丁文字和双向书写系统。尽管这两个平台显著降低了使用门槛，但它们仍主要是交互式工具，需要人工进行真值数据准备、模型训练和事后校正。

生成式人工智能为历史文献转录开辟了新途径。其中一条研究路线探讨了大语言模型作为 OCR 后校正器的应用：Thomas 等人（2024年）证明，经过指令微调的 Llama 2 模型可使 19 世纪英国报纸的字符错误率降低 54.5%，但后续研究在多语言环境中得出了不一的结果（Kanerva 等人，2025年；Boros 等人，2024年）。近期，另一条研究路线将重点转向借助视觉语言模型（VLM）直接转录；这类模型端到端地联合处理视觉和文本信息，无需单独的 OCR 阶段。多项评估表明，视觉语言模型的表现可媲美甚至超越专用 OCR 系统： Humphries 等人（2024年）报告称，在 18 至 19 世纪的英文手稿上，字符错误率为 5–7%；Kim 等人（2025年）发现，通用视觉语言模型在历史表格文献上的表现优于传统 OCR；而Levchenko（2025年）对 12 种多模态大语言模型处理 18 世纪俄文文本的表现进行了基准测试，指出其既有优势，也存在诸如过度历史化之类的隐患。在专业领域，*CHURRO*（Semnani 等人，2025年）是一种参数量为 30 亿、在涵盖 46 个语言簇的 155 个历史语料库上微调的视觉语言模型。研究表明，有针对性的训练能够以低得多的成本取得优于商业替代方案的准确率。VERITAS 沿用这一范式，将视觉语言模型作为主要抽取机制，同时保留基于大语言模型进行精修的灵活性。

2.2. 集成式数字化处理流程

尽管各个独立组件已取得显著进展，但针对如何将它们集成到端到端处理流程中的研究仍然较少。OCR-D 项目（Neudecker 等人，2019年）为德国图书馆的历史印刷文献光学字符识别开发了模块化框架，强调通过标准化格式实现互操作性，但其范围并未延伸至语义丰富化。DAHN 项目（Chiffoleau，2024年）提出了一种以 TEI 为中心的处理流程，包含六个阶段，利用 eScriptorium 进行手写文本识别，并通过 TEI Publisher 发布成果。这是重要的一步，但其方向仍侧重于数字学术版，而非灵活的数据提取。在更大规模上，SocFace（Boillet 等人，2024年）展示了针对法国人口普查的端到端处理流程，而 Constum 等人（2024年）则提出了一种处理巴黎手写婚姻记录的方法。然而，这些工作往往与特定文献类型紧密绑定。

VERITAS 在多个方面拓展了这项工作。不同于 Tran- skribus 和 eScriptorium 等交互式平台，它作为一个完全自动化且可配置的处理流程运行。不同于 OCR-D 等专注于 OCR 的框架，它超越了转录，涵盖语义增强、实体链接和结构化数据索引。其

与模型无关的架构提供了现有设计所缺乏的灵活性。最后，通过展示其对 LLM 辅助的学术探究的下游实用价值（第 5 节），我们填补了一个空白：评估通常在转录准确率处止步，却未评估其对实质性研究的可用性。

3. 方法论

VERITAS（视觉增强的档案来源阅读、解读与转录）旨在将原始历史文献转化为结构化、经过语义增强且机器可读的数据。通过利用视觉语言模型和大语言模型，该处理流程将数字化从字符级转录拓展至全面的文献理解。此架构与模型无关，使用者可替换为针对特定语言或文献类型优化的模型，而无需修改架构。结合模式驱动的信息抽取，这一设计确保该处理流程广泛适用于各类档案馆藏。

该处理流程包含四个依次进行的阶段——预处理、抽取、精炼和增强——每个阶段都是独立模块，具有明确定义的输入和标准化输出（插图 1）。可根据项目需求启用、跳过或替换各个组件。本节将使用第 4 节案例研究中的页面来演示相关操作：贝尔纳迪诺·科里奥的校勘本 《米兰史》（1978年），这是一部篇幅超过1,600页的文艺复兴编年史。

3.1. 预处理

预处理阶段通过三项操作，将异构的原始输入转换为适合自动化分析的标准化格式。

数据转换。处理流程接受档案研究中常见的输入格式，例如数字化 PDF、单页扫描件（TIFF、JPEG、PNG）以及使用手持设备拍摄的照片，并将其转换为统一的表示形式，确保后续所有处理流程阶段都能处理同质且高质量的图像格式。每项输入都会被渲染为高分辨率栅格图像，并归一化至一致的色彩空间和分辨率。

图像增强。可对归一化后的图像应用可选的计算机视觉技术：倾斜校正（纠正旋转错位）、去噪（去除噪声、污渍或扫描伪影）、二值化（转换为黑白图像，以提高文本背景对比度）以及页面检测（从多余的背景元素中分离出文档区域）。具体采用哪些增强操作，可根据源材料的质量和特征进行配置。插图2 展示了在拟议案例研究的一张图像上执行这些操作后的结果。

模式定义。此操作构成学术意图与机器可处理输出之间的关键接口：分析需求被形式化为结构化规范，以确定处理流程提取哪些信息，以及以何种形式提取。模式通常以 JSON Schema 表示，它定义了每个文档元素需要填充的语义字段，并决定后续阶段所采用的工具。模式设计是领域专家与技术人员协作完成的工作：领域专家阐明核心研究问题，技术人员则将其转化为正式规范。下面是一个示例。

```
{ "type": "object", "properties": { "bbox": { "type": "array", "
items": { "type": "integer" } }, "category": { "type": "string", "
enum": [ "标题", "文本", "页眉", "脚注", "插图", "表格" ]}, "text":
{ "type": "string" }, "speaker": { "type": "string" }, "date": { "
type": "string" }, "place": { "type": "string" }, "entities": { "type":
"array", "items": { "type": "object", "properties": { "mention": {
"type": "string" }, "type": { "type": "string", "enum": [ "人物", "
机构", "地点" ]}}}}}, "required": [ "bbox", "category", "text" ]}
```

3.2. 抽取

抽取阶段构成了处理流程的核心，负责将视觉信息转换为机器可读的结构化数据。该阶段经过精心设计，具备高度模块化的特点，提供三种不同的处理路径，可根据项目需求、可用计算基础设施以及目标提取任务的复杂程度进行选择。

- **专用视觉语言模型**。专为文档分析优化的紧凑型、高资源效率视觉语言模型，在版面检测、文本定位、阅读顺序判定和转录方面具有很高的准确性。其优势在于计算效率和转录保真度；然而，它们按照固定的输出模式运行，无法

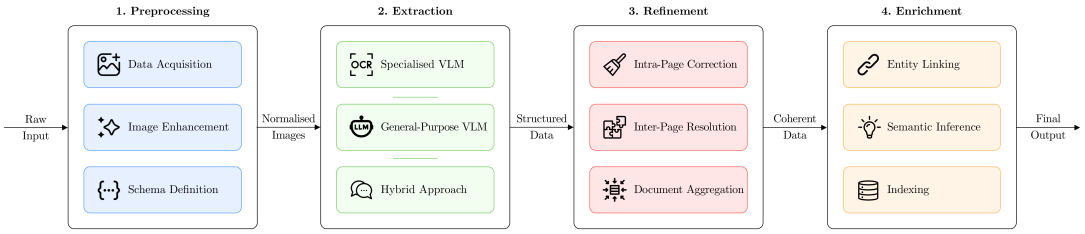

插图 1：VERITAS 架构图，展示了四个阶段及其输入和输出。

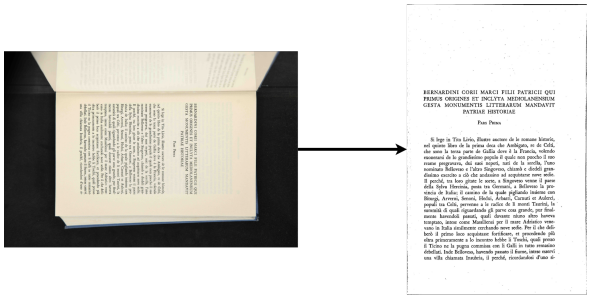

插图 2：对 Corio《米兰史》校勘本的一页所进行的预处理操作。左：原始彩色扫描图，以横向拍摄，可见相邻页面的部分内容。右：经过旋转校正、页面检测、灰度转换和自适应阈值处理后的结果。

满足任意用户自定义指令。

•**通用视觉语言模型**。能够处理视觉和文本信息的大规模视觉语言模型。与前述模型相比，它们在专业文档分析基准测试中的表现可能略低，但其优势在于灵活性：用户可以通过自然语言指令，引导抽取满足特定信息需求的内容，而不局限于传统转录。

• **多步混合方法**。通过两阶段流程结合两种方法：首先由专用视觉语言模型执行基础版面检测和转录，然后由通用大语言模型或视觉语言模型根据用户指定的指令完善并增强抽取效果。这种方法适用于既需要准确转录，又需要复杂语义解读的文档。

在这些路径中进行选择，取决于多种因素的共同作用，包括可用的计算资源、源文档的质量和复杂程度，以及所需语义分析的深度。对于优先考虑转录准确率的项目——

若希望在尽量降低计算开销的同时提高转录准确率，第一条路径是高效的解决方案。对于需要灵活的指令引导式抽取或语义推理的项目，第二条和第三条路径则更为合适。插图 3对比了专用视觉语言模型预定义的输出与通用视觉语言模型进行语义推理时更灵活的输出。

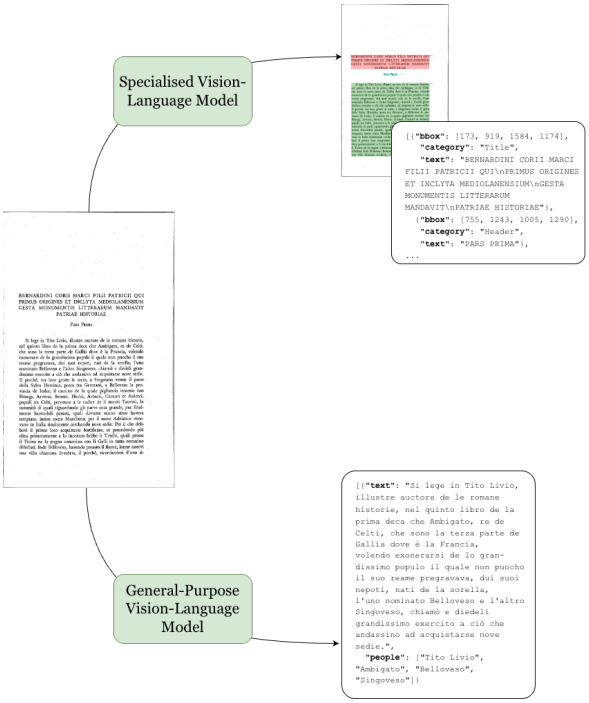

插图 3：专用视觉语言模型与通用视觉语言模型处理路径的输出差异。

3.3. 精炼

精炼阶段通过一系列清理、整合和汇总操作，将页面级抽取结果集合转化为一个连贯统一的文档表示。

页内校正。此操作用于处理单页输出中的伪影和不一致之处。典型的校正包括重建跨行拆分的连字符断词

、规范排版惯例（例如，统一引号样式、规范空白字符），以及根据定义的模式验证提取的数据，以确保结构完整性。

跨页解析。此操作用于解析跨越页面边界的依赖关系和连续性。在历史文献中，内容单元经常横跨多个页面，例如一篇报纸文章在一页开始、在另一页结束，或一场议会演讲跨越数个对开页。跨页解析可确保正确识别并关联此类碎片化内容，并在分页处恰当地传递元数据。

文档聚合。将经过精炼的页面级输出合并为统一的数据结构，以表示完整的源文档。聚合后会生成一个连贯的整体，作为后续增强阶段的输入。

3.4. 增强

增强阶段通过将结构化数据与外部知识源关联、获取更多语义信息，并为存储和后续分析做好准备，提升其价值。

实体链接。对转录文本中识别出的命名实体（如人物、组织和地点）进行消歧，并将其链接到外部知识库中的规范标识符（例如维基数据、虚拟国际规范文档或特定领域的规范档）。此过程不仅能消除歧义（例如区分同名人物），还能将文档置于更广泛的知识图谱中，从而支持交叉引用和关系查询。

语义推理。此操作利用大语言模型执行高级分析任务，并推断源文本中未明确呈现的信息。根据研究目标，此类任务可能包括主题分类、情感分析、事件抽取、时间推理、针对特定领域实体类型的命名实体识别，或抽象式摘要。该处理流程由模式驱动的架构使这些推断出的标注能够系统地纳入结构化输出中。在抽取阶段，也可以使用通用视觉语言模型执行这些操作。

索引。索引操作会依据目标研究社群适用的既定标准，对丰富后的数据进行格式化，并将其载入适当的存储系统。对于人文学科的学术研究，XML-TEI（文本编码倡议）等标准序列化格式有助于与现有的数字人文基础设施实现互操作性。对于定量分析，表格格式（例如逗号分隔值）或与数据库系统（例如关系数据库、全文搜索引擎、向量数据库或图数据库）的集成，可实现高效查询和统计处理。

插图 4展示了增强操作的不同示例：通过语义推理，我们可以抽取命名实体，例如人物和地点；编码器可以计算文本元素的嵌入，以便将其编入索引进行语义搜索；我们还可以将文本中的实体与外部知识库进行匹配。

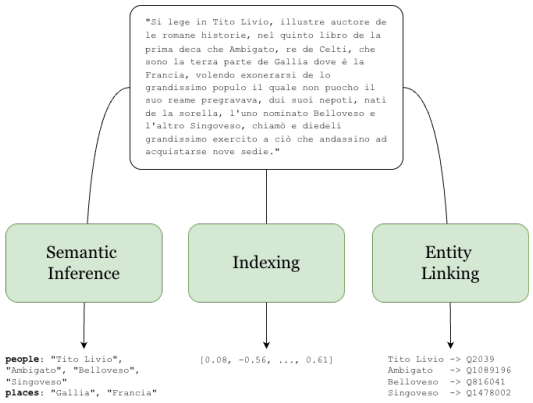

插图 4：增强操作示例。

4. 评估

为评估 VERITAS 流水线的有效性，我们在一个具有代表性的历史文献集上开展了实证评估。本次评估重点考察两个主要维度：(i) 转录准确率，通过标准 OCR 质量指标进行测量；(ii) 计算效率，通过处理时间分析进行评估。此外，我们还评估了该处理流程进行准确版面分析和元素抽取的能力。

4.1. 案例研究

本次评估以贝尔纳迪诺·科里奥（1459年至约1519年）撰写的文艺复兴编年史《米兰史》（《米兰史》）为对象。这部作品的数字化是我们大学一项更广泛的跨学科项目的一部分，该项目汇集了历史学家、计算语言学家和计算机科学家，旨在让这份一手史料可供大规模计算分析。Corio 的著作是研究米兰公国历史最重要的史料之一，涵盖从古代到15世纪晚期的历史。本研究采用了安娜·莫里西·格拉编订的校勘本，该版本由 UTET 于1978年在都灵出版，作为《史学经典》

丛书（Corio，1978年）。该版本共两卷，总计1,688页。

尽管这部校勘本采用现代排版，因此避免了历史字体或手稿笔迹带来的挑战，但它仍然包含多种自动化处理难题，这也是将其选作评测平台的原因。文本在白话语和现代意大利语之间交替，因此抽取系统必须在同一文档流中处理两种不同的语言。页面布局多种多样，包括密集的正文、大量脚注、插图、表格、索引和前置内容；版面分析组件必须对这些内容进行正确识别和分类。

4.2. 评估方法

真实标注构建。从语料库中采用分层抽样选取了100页，以确保充分涵盖本版本中多种多样的页面布局（例如，包含密集正文、大量脚注、标题页以及插图或表格的页面）。三位领域专家依据一套共同的转录指南，手动转录这些页面，以建立参考真实标注。此外，同一批标注人员还对视觉语言模型识别出的结构元素（文本、页眉、脚注等）进行了人工核验和修正，为版面分析评估提供真实标注。

基线。为评估 VERITAS 流水线的性能，我们将其与 AB- BYY FineReader 进行了比较。后者是一款广泛采用的商业光学字符识别解决方案，也是目前我们大学图书馆使用的标准数字化工具。我们采用软件的默认配置设置运行程序；鉴于校勘本使用的是无需专门历史文档配置文件的现代字体，这一做法是合适的。

流水线配置。在本次评估中，VERITAS 流水线配置如下。在预处理阶段，我们采用了多种图像归一化操作，以处理数字化过程中引入的瑕疵。原始扫描图像以横向拍摄，画面中部分可见相邻页面；此外，扫描图像为彩色。为准备抽取图像，我们进行了旋转校正以恢复纵向方向，使用目标检测模型识别并分离出与目标页面相对应的感兴趣区域（Boillet 等人，2021），并将图像转换为灰度。此外，我们还采用了自适应阈值处理技术，以增强文本与背景之间的对比度并提高内容的可读性。在抽取阶段，处理流程配置为使用路径 A（专用视觉语言模型），采用文档专用视觉语言模型，具体为`dots.ocr` （Li 等人，2025年），随后仅进行少量后处理操作，包括行末连字符校正和输出格式归一化（移除 Markdown 痕迹）。

评测指标。我们使用两项标准指标评估转录质量：词错误率（WER），即将预测转录转换为真实文本所需的最少词级插入、删除和替换次数，再除以参考文本的总词数；字符错误率（CER）则以类似方式在字符层面定义，能够更细致地评估转录保真度。两项指标均在语料库层面计算，即基于所有评估页面的拼接文本进行计算，以避免页面长度差异可能造成的偏差。为将实质性转录错误与表面格式差异区分开来，我们报告了两种条件下的结果：原始（不进行归一化）和归一化（将文本转换为小写并删除标点）。我们还使用 F1 分数评估了元素抽取：若抽取出的元素在空间位置（边界框重叠度）和语义标签两方面均与真实元素匹配，则视为真阳性；假阳性包括错误检测或标签分配错误；假阴性则指漏检的元素。

4.3. 结果

表格1 展示了两个系统的语料库级转录错误率。结果表明，在所有评测指标和归一化条件下，VERITAS 采用的基于视觉语言模型的方法始终优于商业基线。在归一化评估中，VERITAS 的词错误率为 1.1%，字符错误率为 0.7%；与 ABBYY FineReader 相比，分别相对提升了 67.6% 和 50.0%。在原始评估条件下，性能差距则更加显著。

	ABBYY		VERITAS	
Text	WER	CER	WER	CER
Raw	0.142	0.031	**0.055**	**0.013**
归一化	0.034	0.014	**0.011**	**0.007**

表格1：基于全部 100 个评估页面的拼接文本计算得出的语料库级转录错误率。数值越低，性能越好。

关于处理效率，虽然基于视觉语言模型的方法在逐页单独处理时会产生高得多的计算成本——

（每页 46 秒，而 ABBYY 仅需 4 秒），但现代推理框架可通过自动请求批处理实现高效的并发处理。在我们的实验配置中，我们采用了 vLLM（Kwon 等人，2023年），这是一种高吞吐量服务框架，可动态批处理传入请求并优化图形处理器内存利用率。vLLM 的连续批处理机制仅使用单块 Nvidia H100 NVL 94Gb 图形处理器三分之一的内存分配，就将每页的实际处理时间缩短至 0.89 秒，比商业解决方案快 4。5×倍。然而，除了自动化处理时间外，还必须将校正输出所需的人工投入纳入处理流程的实际影响考量。领域专家记录了校正 ABBYY FineReader 转录文本所需的时间，平均每页为 2 分 15 秒。按比例推算至全部 1,688 页的语料库，人工后校正工作量约为 63 小时。由于 VERITAS 将归一化词错误率相较于 ABBYY 降低了 67.6%，可以合理假设校正工作量也会按比例减少，使估计的每页校正时间降至约 44 秒，整个语料库的预计总耗时约为 20 小时，节省 40 多小时的专家工作时间。综合自动化处理与人工校正，为整部版本制作经核验转录文本的预计总耗时，从使用 ABBYY 时的约 65 小时降至使用 VERITAS 时的约 21 小时，端到端工作量减少至三分之一。不过，应结合所选抽取模型的计算特性来解读这些收益。尽管并发推理显著降低了实际的每页延迟，但在图形处理器内存、服务基础设施和能耗方面，基于视觉语言模型的处理仍比传统光学字符识别要求更高。

最后，基于视觉语言模型的方法在元素抽取方面的表现证实了其在版面分析中的可靠性，模型的 F1 分数达到 0.966。这些结果表明，专用视觉语言模型能够可靠地识别并正确分类文档页面中的结构元素。这项能力对于处理流程后续的优化与增强阶段至关重要，因为准确的元素抽取有助于正确进行内容聚合和语义标注。

5. 下游应用

前述评估表明，VERITAS 流水线能够实现高保真转录和可靠的版面分析。然而，数字化框架对人文研究的最终价值——

——不仅在于其输出的准确性，更在于这些输出在多大程度上能够支持实质性的学术研究。为展示这一潜力，我们开展了一项探索性研究，将 VERITAS 流水线生成的《米兰史》完整转录文本导入检索增强生成（RAG）系统，并向模型提出了一系列立足历史背景的研究问题。

5.1. 设置

基于 VERITAS 增强阶段的索引结果，我们开发了专门的检索增强生成流程，以适应 Corio 编年史的结构。我们通过定制的数据摄取模块处理了 1,688 个数字化页面，为每页添加时间和结构元数据，并将其划分为符合编年史年份标记和章节划分、时间顺序连贯的片段。最终语料库包含约 1,331 个内容片段和 200 条脚注，使用 BAAI/bge-m3[1] (Chen 等人，2024年) 进行嵌入，并以余弦距离索引至 ChromaDB 集合中。

在查询时，基于嵌入的路由器会将每个问题分类为具体（针对事件、人物或日期）或一般（涉及主题、风格或阐释性问题）。具体查询会触发按年份筛选的语义搜索，并进行脚注增强；一般查询则采用最大边际相关性（MMR）重新排序，以确保检索到的文本块在时间和主题上具有多样性。随后，将检索到的段落输入生成模型 GLM-4.7-Flash[2] (等人，2025年），并使用提示模板确保模型以源材料为依据，并按照查询所用的语言作答。

一组历史学家制定了一系列研究问题，涵盖多个分析维度：事实检索、阐释性分析、主题综合、群体传记重建，以及对事件和行动者的交叉参照。这些问题旨在反映学者研读这份一手史料时自然会开展的探究类型。我们从中选出三个具有代表性的问题进行详细讨论，每个问题都体现了一种不同的历史研究方式。

5.2. 典型查询

事实实体提取。 问题“斯福尔扎时代的公爵秘书和大臣是谁？

[1] https://huggingface.co/BAAI/bge-m3 [2] https://huggingface.co/zai-org/GLM-4. 7-Flash

”要求模型从编年史中分散出现的提及项识别出具体人物及其机构职务。系统根据检索到的段落，生成了一份条理清晰的回答。它正确地指出 Cicco Simonetta 担任总秘书，并引用编年史的记载说明，他在 1477 年 Galeazzo Maria Sforza 去世后正式获任，还描述了他对内政与外交事务广泛的行政权力。回答还检索到 Giovanni Francesco Marliano：1499 年 Ludovico Sforza 离开意大利期间，他被任命为法学家和总督；此外也提及了一些次要人物，例如 Bernardino Curtio 及其兄弟 Iacopo，他们在 1489 年公爵患病期间分别被任命为米兰堡垒的长官和指挥官。值得注意的是，系统还重现了编年史所述的机构重组：设立两个元老院，一个在 Corte dell’Arenga 处理民政事务，另一个在城堡内负责国家事务的审议；Simonetta 及其同僚在其中发挥了决定性影响。该回答展现了系统将散布于数百页内容中的事实信息整合为结构化结果的能力。若仅通过传统的精读方式完成这项任务，将需要投入大量人工。

立场解读检测。问题“作者是否流露出自己的政治倾向？”要求采用一种性质迥异的分析操作：模型必须综合整部作品中体现作者偏见的证据，并阐明一种解读性判断。系统的回答指出了一种多层次的政治立场，这种立场随着编年史的推进而变化。它识别出 Corio 对斯福尔扎王朝的明确忠诚。回答检测到一种早期的辩护姿态，例如作者赞扬 Francesco Sforza 修复米兰防御工事，并将公爵权力辩护为抵御民众动乱的屏障。然而，系统也发现作品后半部分逐渐流露出幻灭感：Corio 将 Giovanni Galeazzo Maria 描绘成一位被大臣贪婪腐化的统治者，并将傲慢刻画为斯福尔扎王朝覆灭的根本原因，这表明他能够对自己的庇护者作出自我批判性的判断。这一回答展现了系统超越字面抽取、进行历史解读的能力，并识别出源文中的意识形态张力。

主题综合与因果推理。问题“作者记录了哪些疫情？采取了哪些卫生措施来限制传染？”要求模型从跨越数世纪的叙述中识别反复出现的主题线索，并针对每个实例，将事件与任何相关的政策应对措施联系起来。该系统识别出了多起疫情事件，其中最突出的是1485年的猛烈瘟疫。Corio称，这场瘟疫迫使他退居乡间，并直接促使他创作这部编年史；此外还有1450年的瘟疫，据报造成约三万人死亡，并严重影响了禧年庆典。模型针对每起事件都提供了源文本中的背景细节，包括Corio关于逃往乡间的自述。关于卫生措施，回答正确指出，编年史反映的主要是应急且非正式的应对方式：退居乡间作为一种隔离手段；要求统治者（如亨利七世皇帝）自费维护城市基础设施（桥梁、道路）；以及公民集会（*Credentia*）在协调危机管理方面发挥的作用。这一回答展现了系统进行多跳推理的能力：从编年史中提取分散各处、主题相关的段落，并将其综合为连贯的分析叙述。

5.3. 讨论

以上探索性结果表明，基于检索增强生成的大语言模型系统若采用 VERITAS 处理流程生成的高质量转录文本，便可成为开展初步历史分析的有效工具。不过，尽管这些结果令人鼓舞，仍应谨慎解读。相关回答尚未经过系统验证，而基于 RAG 的系统仍容易产生幻觉，尤其是在查询需要推断源材料中未明确陈述的内容时。与领域历史学家合作，对事实准确性和诠释有效性开展严格评估，是未来研究的重要方向。尽管存在这些局限，这项演示仍凸显了将高保真转录流水线与基于大语言模型的查询系统相结合的潜力：此举能够降低大规模历史文档分析的门槛，使学者得以跨越大量语料库，更高效地提出并探索研究假设，而传统人工方法难以做到这一点。

6. 结论

我们介绍了 VERITAS，这是一种模块化、与模型无关的框架，将历史文献数字化重新构想为一项整合流程，在统一流水线中涵盖转录、结构分析和语义增强。该框架采用模式驱动架构，使研究人员能够以声明式方式指定提取目标，确保流水线的输出

能够满足不同学术群体的分析需求。

我们对 Corio 的《米兰史》校勘本进行的评估表明，基于 VLM 的抽取处理流程显著优于商业 OCR 基线：在归一化条件下，词错误率相对降低了 67.6%；与此同时，并发推理使每页的有效处理时间缩短至原来的 1/4.5。考虑到人工校正仍是任何数字化工作流程中不可或缺的环节，这些改进意味着，处理完整的 1,688 页语料库所需的端到端时间预计可减少至原来的三分之一。通过基于 RAG 的系统将处理流程的输出用于下游应用，进一步表明高保真、结构化的转录文本能够直接支持实质性的历史研究，涵盖从事实实体提取到立场解读检测等任务。

未来仍有若干研究方向。首先，需要对增强阶段进行系统评估，包括实体链接准确率，以及基于大语言模型的语义推理在历史文本上的可靠性，从而验证完整处理流程超越转录能力的表现。其次，本文所呈现的基于 RAG 的下游分析仍属探索性研究；必须与领域历史学家合作，严格评估事实准确性和解读有效性。第三，我们计划在古文字学挑战更大的文献集上评估 VERITAS，例如手稿资料和采用非标准字体的早期印刷书籍，以考察该方法的普适性。最后，我们计划将这一框架作为开源工具包发布，以推动其在数字人文学科中的应用和由社区驱动的扩展。

参考文献

Mélodie Boillet、Solène Tarride、Yoann Schneider、BastienAbadie、LionelKesztenbaum 和 Christopher Kermorvant。2024年。socface 项目：大规模收集、处理和分析一个世纪以来的法国人口普查资料。载于文档分析与识别——*ICDAR 2024*，第57–73页，Cham。Springer Nature Switzerland。

Mélodie Boillet、Christopher Kermorvant 和 Thierry Paquet。2021年。多文档数据集预训练通过深度神经网络提升文本行检测效果。载于*2020*年第*25*届国际模式识别会议（*ICPR*），第2134–2141页。

Emanuela Boros、Maud Ehrmann、Matteo Romanello、Sven Najem-Meyer 和 Frédéric Ka- plan。2024年。使用大语言模型对历史文本转录稿进行后校正：一项探索性研究。载于第八届文化遗产、社会科学、人文与文学计算语言学联合研讨会论文集（*LaTeCH-CLfL 2024*），第133–159页，马耳他圣朱利安斯。计算语言学协会。

Jianlv Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2024年。BGE M3-Embedding：通过自知识蒸馏实现多语言、多功能、多粒度文本嵌入。

Floriane Chiffoleau。2024年。保持开放：一种基于文本编码倡议的历史文献出版处理流程。《文本编码倡议期刊》，第15期。

Thomas Constum、Lucas Preel、ThéoLarcher、Pierrick Tranouez、Thierry Paquet 和 Sandra Brée。2024年。手写文献中的端到端信息抽取：理解1880年至1940年的巴黎婚姻记录。

贝尔纳迪诺·科里奥。1978年。《米兰史》。《史学经典》。都灵印刷出版联合会，都灵。《祖国史》的校勘本。

GLM团队等。2025年。GLM-4.5：智能体、推理与编程（ARC）基础模型。

Mark Humphries、Lianne C. Leddy、Quinn Downton、Meredith Legace、John McConnell、Isabella Murray 和 Elizabeth Spence。2024年。解锁档案：使用大语言模型转录手写历史文献。

Philip Kahle、Sebastian Colutto、Günter Hackl 和 Günter Mühlberger。2017年。Transkribus——用于历史文献转录、识别和检索的服务平台。载于*2017*年国际模式识别协会第*14*届国际文档分析与识别会议（*ICDAR*），第04卷，第19–24页。

Jenna Kanerva、Cassandra Ledins、Siiri Käpyaho 和 Filip Ginter。2025年。历史文献中的 OCR 错误后校正与大语言模型：没有免费的午餐。载于第三届资源匮乏语言资源与表征研讨会论文集（*RESOURCEFUL-2025*），第38–47页，爱沙尼亚塔林。爱沙尼亚塔尔图大学图书馆。

Benjamin Kiessling、Robin Tissot、Peter Stokes 和 Daniel Stökl Ben Ezra。2019年。eScriptorium：用于历史文档分析的开源平台。载于*2019*年国际

文档分析与识别研讨会（*ICDARW*），第2卷，第19–19页。

Seorin Kim、Julien Baudru、Wouter Ryckbosch、Hugues Bersini 和 VincentGinis。2025年。关于大语言模型在历史记录 OCR/手写文本识别任务中优于传统系统的早期证据。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 和 Ion Stoica。2023年。采用 PagedAttention 实现大语言模型服务的高效内存管理。载于*ACM SIGOPS* 第*29*届操作系统原理研讨会论文集。

Maria Levchenko。2025年。面向历史文献 OCR 的大语言模型评估：数字人文的方法论框架。

Yumeng Li、Guang Yang、Hao Liu、Bowen Wang 和 Colin Zhang。2025年。dots.ocr：在单一视觉语言模型中进行多语言文档版面解析。

Guenter Muehlberger、Louise Seaward、Melissa Terras、Sofia Ares Oliveira、Vicente Bosch、Max- imilian Bryan、Sebastian Colutto、Hervé Dé- jean、Markus Diem、Stefan Fiel、Basilis Gatos、Albert Greinoecker、Tobias Grüning、Guenter Hackl、Vili Haukkovaara、Gerhard Heyer、Lauri Hirvonen、Tobias Hodel、Matti Jokinen、Philip Kahle、Mario Kallio、Frederic Kaplan、Florian Kleber、Roger Labahn、Eva Maria Lang、Sören Laube、Gundram Leifert、Georgios Louloudis、Rory McNicholl、Jean-Luc Meunier、Johannes Michael、Elena Mühlbauer、Nathanael Philipp、Ioannis Pratikakis、Joan Puigcerver Pérez、Han- nelore Putz、George Retsinas、Verónica Romero、Robert Sablatnig、J oan Andreu Sánchez、Philip Schofield、Giorgos Sfikas、Christian Sieber、Niko- laos Stamatopoulos、Tobias Strauß、Tamara Terbul、Alejandro Héctor Toselli、Berthold Ulre- ich、Mauricio Villegas、Enrique Vidal、Johanna Walcher、Max Weidemann、Herbert Wurster 和 Konstantinos Zagoris。2019年。通过手写文本识别革新档案学术研究：以 Transkribus 为例。《文献学杂志》，75(5):954–976。

Clemens Neudecker、Konstantin Baierer、Maria Federbusch、Matthias Boenig、Kay-Michael Würzner、Volker Hartmann 和 Elisa Herrmann。2019年。OCR-D：用于历史印刷文档的端到端开源光学字符识别框架。载于第三届文本文化遗产数字访问国际会议论文集，DATeCH2019，页码 53–58，美国纽约州纽约。计算机协会。

Sina J. Semnani、Han Zhang、Xinyan He、Merve Tekg"urler 和 Monica S. Lam。2025年。CHURRO：利用开放权重大型视觉语言模型，让历史变得易于阅读，实现高精度、低成本的历史文本识别。载于*2025*年自然语言处理经验方法会议论文集（*EMNLP 2025*）。

R. Smith。2007年。Tesseract 光学字符识别引擎概述。载于第九届国际文档分析与识别会议（*ICDAR 2007*），第 2 卷，第 629–633 页。

Alan Thomas、Robert Gaizauskas 和 Haiping Lu。2024年。利用大语言模型对历史报纸进行 OCR 后校正。载于第三届历史与古代语言技术研讨会（*LT4HALA*）@ *LREC-COLING-2024* 会议论文集，第 116–121 页，都灵，意大利。ELRA 和 I CCL。