arXiv:2304.13606v1 [cs.CV] 26 Apr 2023

# SIMARA：用于从整页手写文档中进行键值信息抽取的数据库

Sol`ene Tarride[1][0000−0001−6174−9865]、M´elodie Boillet[1]、2[0000−0002−0618−7852]、Jean-Fran.cois Moufflet[3],和 Christopher Kermorvant[1]、2[0000−0002−7508−4080]

[1] TEKLIA，法国巴黎[2] LITIS，诺曼底大学，
法国鲁昂[3] 法国国家档案馆，法国巴黎

**摘要**。 我们提出了一个用于从历史手写文档中进行信息提取的新数据库。该语料库包含来自六个不同系列的5,393份档案检索工具，时间跨度为18至20世纪。档案检索工具是包含描述较早档案的元数据的手写文档。这些文档存放于法国国家档案馆，供档案管理员用于识别和查找档案文献。每份文档均经过页面级标注，并包含七个待提取字段。每个字段的位置信息并未提供，因此该数据集能够促进无分割信息提取系统的研究。我们提出了一个基于 Transformer 架构的模型，该模型经过训练，可执行端到端信息抽取；同时提供训练集、验证集和测试集三组数据，以确保与未来研究进行公平比较。该数据库可免费访问[4]。

**关键词：** 开放数据集 · 键值抽取 · 历史文档 · 无分割方法。

## 1 引言

继网络上的信息检索发生变革之后，机器学习和深度学习技术如今已广泛应用于档案领域[3]。与网络上的原生数字数据或传统信息系统相比，最大的区别在于，档案中保存的大多数文件都是纸质文档。近年来，这些文档已大规模实现数字化，但馆藏总量中只有很少一部分完成了数字化；由于数量极其庞大，档案的全面数字化并未被列入计划。因此，有必要根据文档的实用价值，选择需要进行数字化的文档。在档案保存的文档中，档案检索工具发挥着特殊作用。数十年来，档案服务部门一直在制作这些文档，以便于获取文档中包含的信息。它们是通往大量馆藏文档的主要入口，因此在任何研究中都会被优先查阅。

[4] https://zenodo.org/record/7868059

因此，对这些档案检索工具进行数字化，并将其中包含的索引信息转换为数字格式，对于改善向用户提供的服务至关重要。

虽然对原始历史文献的自动处理通常旨在生成其内容的文本转录[11], 但档案检索工具则并非如此：对于后者，自动处理必须能够以打字形式和标准化形式提取其中包含的索引信息，以便将其插入数据库。因此，这种处理更侧重于提取文本信息，而非纯粹的转录。因此，这些文件的自动处理流程除传统的文档分析和手写识别阶段外，还必须包括信息或实体（命名实体）抽取 [7] 以及规范化阶段。

具有悖论意味的是，尽管档案检索工具是获取档案信息的核心，但它们在文档分析领域几乎未得到研究：目前没有公开可用的此类数据库，文档识别系统也未在这类文件上进行评估。为了填补这一空白，我们向该领域公开本文所述的 SIMARA数据库。此外，我们还希望向该领域介绍一种供自动处理模型使用的新型任务：通过处理整页内容，从历史文献中提取键值信息的任务。该任务在以下方面不同于传统的文本转录或命名实体提取任务：

– 必须考虑整份文档才能执行信息提取（整页）– 在文档包含的全部文本中，只需提取其中一部分– 每条提取的信息都必须具有类型，即构成一个键（类型）-值关联– 无需记录信息在文档中的位置– 提取的信息（值）必须经过规范化，以满足其将要填充的数据库的约束。

本文提出两项主要贡献。首先，我们介绍 SIMARA，这是一个新的档案检索工具图像数据集，其中标注了转写文本及其在页面级别对应的字段。我们还提出了一个在 SIMARA 数据集上训练的端到端基线模型，该模型通过单个前向步骤执行目标键值信息的抽取。

本文的组织结构如下。第 2 节回顾与 SIMARA 相关的不同数据集，以及近期提出的、用于从整页图像中提取文本和实体的模型。第 3 节详细介绍 SIMARA 数据集及其不同系列。最后，第 4 节介绍从 SIMARA 图像中进行键值信息抽取的自动模型训练初步实验及结果。

## 2 相关工作

与当前文献中的研究相比，SIMARA 数据集针对的是一项截然不同的任务。其挑战不仅在于识别整页中的手写文本，还在于识别与每段文本相对应的字段（键）

（值）。这些字段处于命名实体与基于版面的标记之间。

许多公共数据集已被用于历史文档图像识别。Nikolaidou 等人通过其系统综述所开展的大量工作，已将所有这些数据集列出，并按子任务进行了分组。[12]

在本节中，我们介绍一些与 SIMARA 最为相似的数据集。我们还详细介绍了针对完整页面端到端手写识别所提出的各种方法，这些方法既可以包含实体抽取，也可以不包含实体抽取。

### 2.1 数据集

许多数据集已被用于自动手写识别，但据我们所知，其中没有一个包含数字化档案检索工具。一般而言，根据标注程度，数据集可以分为两种主要类型：

**为文本行识别设计的数据集：**一些数据集是为基于行图像的手写识别而设计的。因此，其主要目标是从预先分割的文本行中识别文本。例如，NorHand 数据库[2] 包含书写于 19 至 20 世纪、使用挪威语的文件，但仅包含文本行图像及其对应的转写文本。同样，IAM [9] 数据库也是为基于行图像的文本识别而设计的。在该数据库上评估的大多数算法都是基于文本行进行训练的，尽管其中也提供完整页面，并且可以轻松获取阅读顺序。

**为整页识别设计的数据集：**其他数据集包括整页图像和标注，其中包含转写文本和详细的版面信息。这些信息通常编码在 XML PAGE 真实标注文件中。例如，以下数据集提供了页面级的完整真实标注：READ 2016（德语，15至19世纪）[16], Digital Peter（俄语，18世纪）[13], ESPOSALLES（加泰罗尼亚语，18世纪）[8], HOME-ALCAR（拉丁语，12至14世纪）[15], 以及 POPP（法语，19至20世纪）[4]。

其中一些数据集还包括命名实体标注。在 ESPOSALLES [8],中，每个词都关联有两个标签：类别和人物。因此，该数据库支持在完整页面上训练用于手写识别和命名实体识别的端到端模型。在 POPP [4], 中没有明确的命名实体，但可以利用转写文本中包含的列分隔标记检索这些实体。最后，HOME-ALCAR [15], 提供了两个命名实体（地点和人物）。

尽管这些数据集提供了整页级标注，但大多数研究人员关注的是文本行检测或文本行识别。

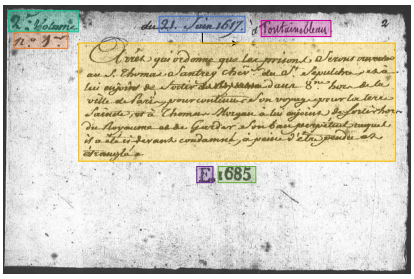
### 2.2 端到端页面识别模型

本节介绍用于端到端手写页面识别的自动化方法，无论是否包含实体抽取。我们只关注能够处理文件完整页面的系统。

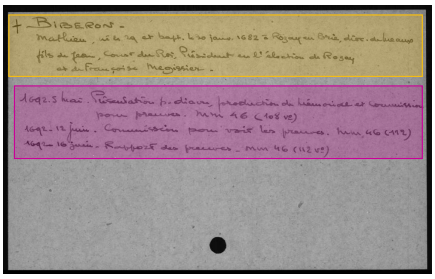
迄今为止，已有两种主要方法被提出。第一种是基于分割的方法，将多个模型结合起来，每个模型都针对特定任务进行设计：文本行检测、文本识别，以及可选的命名实体识别。这类方法非常流行，因为其实现更为简单。另一种方法则是直接从无分割的完整页面执行手写识别，并可选择执行命名实体识别。

**基于分割的模型**Constum 等人 [4] 在 POPP 数据集的法国人口普查图像上执行表格检测、页面分类、文本行检测和手写识别。类似地，Tarride 等人 [17] 提出了一个用于魁北克教区登记册端到端识别的完整工作流。该工作流包括页面分类、文本行检测、文本识别和命名实体识别。

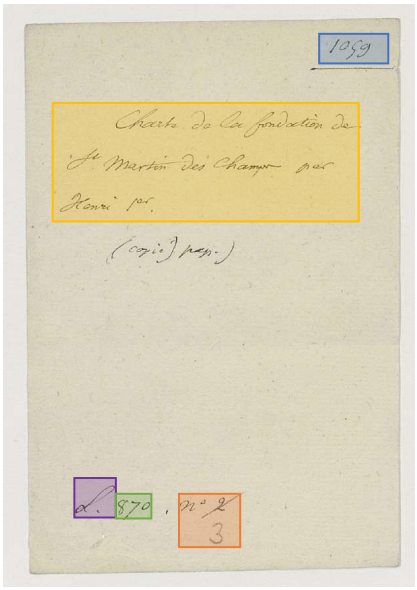
**无分割模型**Start Attend Read（SAR）是首个为段落识别提出的手写文本识别模型。该模型是一种基于注意力的 CRNN-CTC 网络，能够关注字符并读取多行文本。Start Follow Read（SFR） [18] 将区域提议网络（RPN）与 CRNN-CTC 网络相结合。因此，它能够联合学习行检测和识别。Origa miNet [19] 是一种 CRNN-CTC 网络，可将多行特征展开为单行特征。垂直注意力网络（VAN） [6] 是一种基于注意力的模型，由 FCN 编码器、混合注意力模块和循环解码器组成。该模型可在行或段落级别运行，并利用转录中的换行来关注行特征。所有这些模型都能够识别单页或段落中的文本，但无法处理多栏文本块。文档注意力网络（DAN） [5] 是首个能够有效处理完整文档的模型，包括具有多栏和复杂阅读顺序的文件。该模型基于 Transformer 架构，并联合学习表示布局信息的字符和特殊标记。该模型在 RIMES [10] 和 READ [16] 上取得了最先进的结果。最后，Rouhou 等人 [14] 提出了一种用于 ESPOSALLES 数据库中记录的手写文本识别和命名实体识别的 Transfor mer 模型，其中命名实体也由特殊标记表示。他们表明，与在文本行上训练的模型相比，直接从记录中提取信息能够使模型受益于更多的上下文信息。该模型仍然需要记录分割，但有可能应用于完整页面。



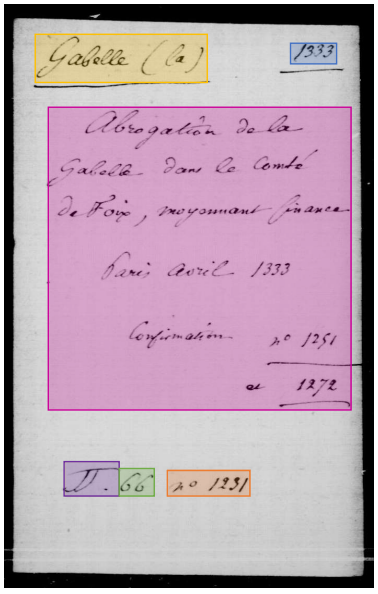
(a) E系列



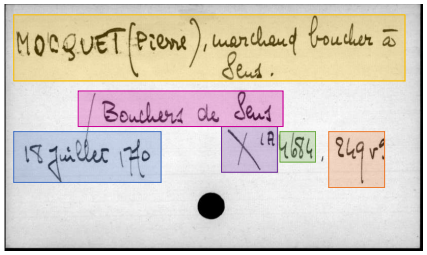
(b) M 系列



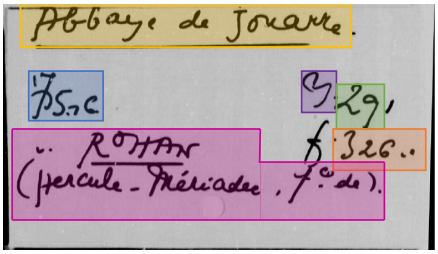
(c) L系列



(d) Dou¨et d'Arcq（JJ 系列文档的记录ument)



(e) X1a 系列



(f) Y系列

图1：SIMARA 数据集不同系列及其包含信息的示意图。**图例**：日期；标题；分析；编排；系列；文章编号；卷号。

## 3 SIMARA 数据集

本节介绍 SIMARA 数据集，如图 1 所示。我们将描述文件的来源、用途、结构及其内容。包含图像和标注的数据库可通过 https://zenodo.org/record/7868059 免费获取。

### 3.1 文件的来源

SIMARA 数据集中的文件由法国国家档案馆制作。法国国家档案馆是负责保存国家档案遗产的机构[5]。其主要职责包括收集兼具法律和历史价值的文件，无论其载体为何都对其进行保存，并向公民、学者和行政机构提供这些文件。

国家档案馆从中央现行行政机构（总统府、各部委、公共机构）收集文件，时间是在这些文件的行政用途结束之后。作为法国大革命的产物，国家档案馆还继承了 1789 年被废除的机构所遗留下来的档案：主要包括自 $6^{\text{th}}$ 世纪以来由王室机构和地方巴黎行政机构制作的所有文件、巴黎教堂和修道院的特许状，以及巴黎公证人档案。

如果不使用元数据集对这些文件（380 公里的纸质文件和 18 To 的数字档案）进行整理和描述，用户便无法找到可能感兴趣的文件，因此也就无法传递这些文件。档案管理员通常将这些元数据集称为“档案检索工具”。用于识别档案的最低限度元数据包括：标题、日期和标识符（通常是箱、文件甚至单份文件的唯一参考编号）。还可以补充其他信息，例如尺寸、物理描述和额外的描述性内容。所有这些信息都按照全球通用的特定文档类型定义，以 XML 格式编制：EAD[6]。这些信息发布在网站上，供用户通过搜索引擎检索。

### 3.2 档案检索工具

档案在档案检索工具中的描述方式可能有很大差异。较新的文件的元数据会直接以 XML 格式编制，而中世纪和近代最早期的系列则配有手写检索工具，因为它们最初是在 19世纪整理的。目标是将这些检索工具转换为 XML EAD 文件，但这会带来若干挑战：

– 异质性：这些文件从 $18^{\text{th}}$ 世纪末（有时甚至早于法国大革命）一直到 $20^{\text{th}}$ 世纪中叶陆续产生。因此，它们在版面、内容、语言和手写方面存在巨大差异。

[5] https://www.archives-nationales.culture.gouv.fr/en_GB/web/guest/ 我们是谁 [6] 编码档案描述。官方网站：https://www.loc.gov/ead/

– 缺乏标准化：这些文件创建于数字时代之前，当时档案著录不像今天这样规范，因此，不同检索工具之间的内容呈现方式差异很大。– 无法访问：这些档案检索工具直到2017—2019年才被数字化为图像格式（灰度、300 dpi，有时来自缩微胶片或缩微平片），目的是便于进行数字转换，但公众无法访问它们，无论是在阅览室还是互联网上。– 数量：据估计，索引文件约由80万张纸质卡片组成。采用传统方法对这些文件进行数字转换需要数年时间：手动输入内容，然后再手动将其编码为XML，是一个耗时漫长且可能令人望而生畏的过程。

这些考量的结果是，需要使用自动化流程将此材料转换为数字数据。

表1：SIMARA 数据集中不同系列所含信息的说明。

| **系列** | **待提取信息** |
|---|---|
| E | - 日期（日期）<br>- 决定的分析（标题）<br>- 唯一标识符的第 1 部分（系列、文章编号）<br>- 唯一标识符的第 2 部分（卷号）<br>- 原始参照（编排）<br>- 卡片编号（分析） |
| L | - 宪章年份（日期）<br>- 宪章主要分析（标题）<br>- 进一步详情（分析）<br>- 宪章索引（系列、文章编号、卷号） |
| M | - 个人姓名及信息（标题）<br>- 贵族身份描述及证明（分析） |
| X1a | - 第一方姓名（标题）<br>- 第二方姓名（分析）<br>- 审判日期（日期）<br>- 判决索引（系列、文章编号、卷号） |
| Y | - 人物姓名（标题）<br>- 人物信息（分析）<br>- 日期（日期）<br>- 参考文献（系列、文章编号、卷号） |
| Douet d’Arcq | - 记录标题（标题）<br>- 补充信息（分析）<br>- 日期（日期）<br>- 参考文献（系列、文章编号、卷号） |

### 3.3 手写检索工具的描述

在本节中，我们介绍 SIMARA 数据集所包含的不同档案检索工具系列及其面临的困难。表1总结了每个系列中需要提取的字段。我们还在图1中标出了各系列中的这些字段。

**E系列（“国王会议”）**

档案类型：记录国王会议决议的登记册，涵盖 $17^{\text{th}}$ 世纪和 $18^{\text{th}}$ 世纪。

检索工具：40 480张来自 $18^{\text{th}}$ 世纪的索引卡片，逐项分析了国王会议的决议。由于采用古法语书写体，且语言古旧，这一系列难以辨读。

**L系列（“教堂与修道院的宗教生活”）**

档案类型： 巴黎圣马丁教堂的中世纪特许状。

检索工具： 623张索引卡片，每张都包含对一份特许状的分析。它们写于 $19^{\text{th}}$ 世纪中叶。主要困难在于信息的异构版式，以及标识符的近期更正（旧的19世纪标识符被铅笔划掉，并由当代书手予以更正）。

**M 系列（“贵族身份证明”）**

档案类型： 与马耳他骑士团有关且包含其贵族身份证明的文件。

检索工具： 4,847张按字母顺序排列的卡片。它们写于20世纪50年代左右。这些文件没有任何特殊困难。

**X1a 系列（“议会”）**

档案类型： 记录议会判决的登记册；议会是最高司法法院。

检索工具： 101036张卡片，包含对涉及两方的 $18^{\text{th}}$ 世纪审判的分析。这些卡片按当事人姓名的字母顺序排列。卡片写于 $20^{\text{th}}$ 世纪，不存在任何特别困难。

**Y系列（“沙特莱”）**

档案类型： 记录于沙特莱的 $18^{\text{th}}$ 世纪公证文书。沙特莱是一个巴黎机构，公证人必须在那里报告其记录的所有交易。

检索工具： 61,878张卡片，包含对公证文书的分析。这些

卡片按取得契据的人物姓名的字母顺序排列。它们写于20世纪50年代。主要难点包括数字化质量差、手写风格、异构版面以及文本密度高。

**Dou¨et d'Arcq档案**

档案类型： 与前述档案不同，该文件对多个系列进行了索引，主要是"历史系列"（J、K、L、M），这些系列主要对应最古老且最具声望的文件（中世纪王室档案、来自巴黎修道院和教堂的特许状、巴黎大学、圣殿骑士团和医院骑士团等）。其中涵盖了不同类型的文件（特许状、登记册、账目）。

检索工具： 在19$^{th}$ 世纪中叶制作的118 093张卡片，分为三个子集：一套人名卡片、一套地名卡片，以及最后一套按主题排列的卡片。我们只关注后一套（6 284张卡片）。

### 3.4 档案检索工具和原始文献

为了让读者更清楚地了解档案检索工具和原始记录，图2展示了档案检索工具以及其所描述的原始文件复制件示例。

我们突出显示了记录中的主要信息类别（如图1所示），并标明了原始文献中部分信息的所在位置。

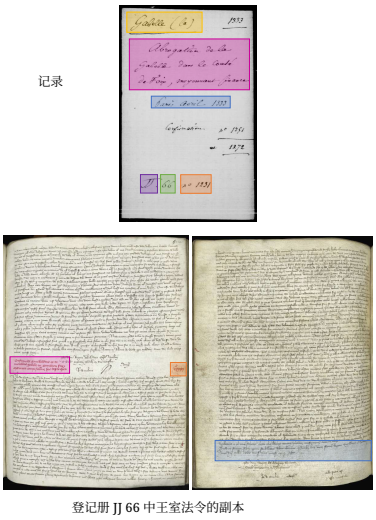
### 3.5 标注数据的生产

SIMARA 数据集标注数据的生成采用了不同于传统方法的方案。通常，由于处理流程由行检测、手写识别和命名实体抽取等不同模型组成，因此需要为每个模型生成特定的标注：图像上的行位置、转写文本行，以及文本中标出位置的实体。每种标注都需要专门的界面，而用于验证生产阶段模型预测的界面也与之不同。多种界面和多种训练数据带来的结果是，生产过程中验证的数据无法用于改进模型。要重新训练模型，就必须在验证任务进行的同时重复开展标注活动，这会导致任务重复。本研究采用了截然不同的方案：用于生成模型训练所需首批数据的标注界面，与用于验证生产数据的界面相同。这样，所有生产数据都可以用于迭代改进模型。

SIMARA 数据集的独特之处在于，真实标注可以通过填写档案管理员设计的表单轻松创建。每位标注者负责



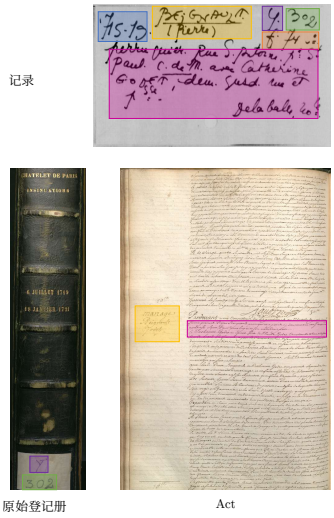

（a）L系列




（b）Dou¨et d'Arcq 文件




（c）X系列




（d）Y系列。

图2：四个系列中包含原始文件的记录示例。

通过在检索工具的图像上查找信息，将真实标注输入表单。标注界面如图 3 所示。

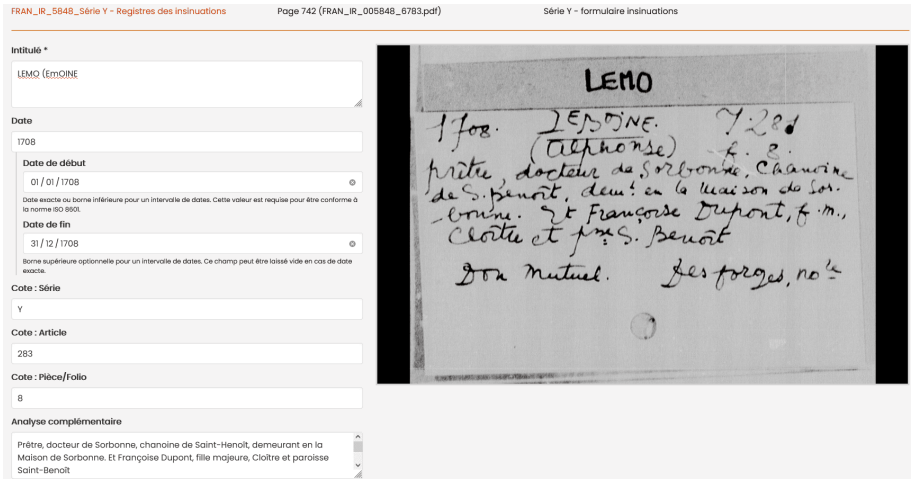



图 3：标注员界面：同一界面用于生成真实标注数据，以及在生产阶段验证模型的建议。

模型训练完成后，将在生产阶段用于根据建议预填表单。在此阶段，标注者的任务有所不同，因为他们必须阅读、确认或更正每个字段中建议的信息。这一步无疑是整个流程中最耗时的环节，但至关重要，因为标注者所做的更正将用于改进信息提取模型。验证界面与生成真实标注时使用的界面相同。

转录质量通过在有疑问时进行双重验证的流程得到保障：当标注者无法确定某条信息时，他们可以将其标记出来，由档案管理员进行核查和验证。

## 4 实验结果

本节报告了使用我们提出的整页键值信息抽取系统在 SIMARA 数据集上开展实验所取得的结果。

### 4.1 数据集

随着项目推进，越来越多的文件完成标注并可用于训练，我们的训练集也随之不断发展。在训练

标注流程不同阶段的数据时，我们在不同的数据划分上训练了三个模型：

– *Split-v1*：780 张图像用于训练，50 张用于验证，50 张用于测试；– *Split-v2*：3659 张图像用于训练，783 张用于验证，784 张用于测试；– *Split-v3*：3778 张图像用于训练，811 张用于验证，804 张用于测试。

表2： *Split-v3* 划分（a）每个系列的图像数量

| **系列** | **训练集** | **验证集** | **测试集** | **总计 (%)** |
|---|---|---|---|---|
| E系列 | 322 | 64 | 79 | 8.6 |
| L系列 | 38 | 8 | 4 | 0.9 |
| M 系列 | 128 | 21 | 27 | 3.3 |
| X1a 系列 | 2209 | 491 | 469 | 58.8 |
| Y系列 | 940 | 205 | 196 | 24.9 |
| Dou¨et s'Arcq 系列 | 141 | 22 | 29 | 3.5 |
| 总计 | 3778 | 811 | 804 | 100.0 |

（b）实体数量（词数）

| **实体** | **训练集** | **验证集** | **测试集** | **总计 (%)** |
|---|---|---|---|---|
| `date` | 8406 | 1814 | 1799 | 10.4 |
| 标题 | 35531 | 7495 | 8173 | 44.4 |
| 系列 | 3168 | 664 | 676 | 3.9 |
| 分析 | 25988 | 5130 | 5602 | 31.8 |
| 卷号 | 3913 | 808 | 813 | 4.8 |
| 文章编号 3181 | | 665 | 678 | 3.9 |
| 编排 | 644 | 122 | 153 | 0.8 |
| 总计 | 80831 | 16698 | 17894 | 100.0 |

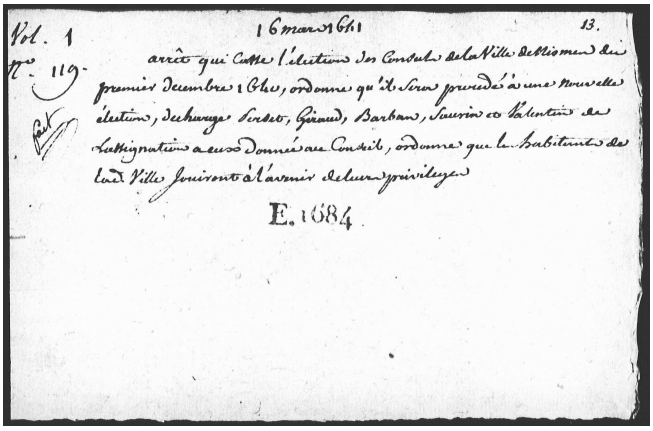
### 4.2 键值信息提取模型

我们基于DAN架构 [5] 训练了一个模型，用于从SIMARA的整页中进行键值信息抽取。DAN[7] 是一种开源的基于注意力机制的Transformer模型，用于手写文本识别，并且可以直接处理页面。该模型采用交叉熵损失函数进行训练，其最后一层是一个线性层，并通过Softmax激活函数计算词汇表中每个字符的概率。我们通过以下方式解决键值信息抽取任务：

[7] https://github.com/FactoDeepLearning/DAN

使用位于每个部分开头的特殊标记对每个字段进行编码，如表3所示。该模型的完整描述和评估见 [1]。

表3：用于训练基线模型的转写文本示例。特殊标记位于每个文本片段的开头，用于标识相应字段。请注意，即使字段顺序与图像中的阅读顺序不一致，转写文本中的字段也始终按相同顺序出现。




| 转录 |
| --- |
| <参考编号>编号 119<日期>16 三月 1641 <分析>13<标题>法令，撤销尼姆市领事的选举，并于1640年12月1日下令重新进行选举，免除Perset、Giraud、Barban、Saurin和Valentin应向议会履行的传唤义务，并下令该市居民今后享有其特权。 <系列>E<文章编号>1684 <编排>第1卷 |

### 4.3 指标

手写文本识别指标：手写识别质量采用标准的字符错误率（CER）和词错误率（WER）进行评估。评估对象为全文，但此类评估不考虑命名实体。

*NER* 指标： 我们使用 Nerval[8] 评估工具包来评估命名实体识别结果。Nerval 能够通过在字符级别将自动转录与真实标注对齐，处理含噪文本——

[8] https://gitlab.com/teklia/ner/nerval

表 4：三个模型在测试集上的结果。

| 模型名称 | 划分 | CER (%) | WER (%) | 图像数量 |
|---|---|---|---|---|
| *Model-v1* | *Split-v1* | 12.08 | 22.41 | 50 |
| *Model-v2* | *Split-v2* | 7.39 | 15.07 | 784 |
| *Model-v3* | *Split-v3* | **6.46** | **14.79** | 804 |

预测值实体与真实标注实体在标签相似且编辑距离低于阈值时，被视为匹配。在我们的实验中，该阈值设为 30%。基于这一对齐结果，我们计算精确率、召回率和 F1分数，以评估命名实体识别。

### 4.4 结果

本节介绍我们基线模型在 SIMARA 上进行键值信息提取所达到的性能。首先，我们研究训练数据集规模的影响，然后详细说明使用最佳模型所获得的结果。

**训练数据集规模的影响**我们使用规模不断增大的训练集，在三个不同的数据划分上训练了三个模型。表4比较了使用不同版本训练集所获得的结果。正如预期，随着训练集规模的增大，结果也会有所改善。值得注意的是，即使训练数据量仅小幅增加，例如*Split-v2*和*Split-v3*之间增加了一百张图像，在我们的案例中，识别率也提高了一个CER点。

**最佳模型的详细手写文本识别结果**我们在表5a中给出了最佳模型（*Model-v3*）的详细结果。在每个文档系列上，分别针对测试集计算字符错误率和词错误率。*L*系列的识别效果最佳，尽管它在训练集中的代表性并不高。需要注意的是，该系列在测试集中仅包含4份文件，因此这一观察结果的统计解释受到限制。最难识别的系列是Dou¨et d'Arcq，其WER接近25%。这可以解释为，该系列具有独特的版面，并且训练集中的示例很少。其他最难识别的系列是Y系列（CER较高）和E系列（WER较高）；尽管它们在训练集中具有较高的代表性，但分别在版面和语言方面存在一些困难。总体而言，文本识别结果非常令人满意，WER低于15%。

**最佳模型的详细命名实体识别结果**从转录的角度来看，这些文件难以评估，因为机器的阅读顺序不一定与标注者的阅读顺序一致。因此，

表5：*Model-v3*的详细评估结果。

(a) 在测试集上分别计算每个系列的字符错误率和词错误率。在此评估阶段，会移除用于编码各字段的特殊标记。

| 系列 | CER (%) | WER (%) | 图像数量 |
|---|---|---|---|
| E系列 | 6.89 | 18.73 | 79 |
| L系列 | **3.33** | **9.35** | 4 |
| M 系列 | 7.70 | 15.87 | 27 |
| X1a 系列 | 4.97 | 11.16 | 469 |
| Y系列 | 8.23 | 16.19 | 196 |
| Dou¨et d'Arcq | 10.24 | 24.97 | 29 |
| 总计 | 6.46 | 14.79 | 804 |

（b）在测试集上计算每个字段的命名实体识别精确率、召回率和 F1分数。

| Tag | 预测值 | 匹配值 | 精确率（%） | 召回率（%） | F1 值（%） | 实体数量 |
|---|---|---|---|---|---|---|
| date | 1808 | 1784 | 98.67 | **99.17** | **98.92** | 1799 |
| 标题 | 8059 | 7650 | 94.92 | 93.60 | 94.26 | 8173 |
| 系列 | 677 | 664 | 98.08 | 98.22 | 98.15 | 676 |
| 分析 | 5449 | 5248 | 96.31 | 93.68 | 94.98 | 5602 |
| 文章编号 | 672 | 664 | **98.81** | 97.94 | 98.37 | 678 |
| 卷号 _ | 805 | 776 | 96.40 | 95.45 | 95.92 | 813 |
| 编排 | 206 | 119 | 57.77 | 77.78 | 66.30 | 153 |
| 总计 | 17676 | 16905 | 95.64 | 94.47 | 95.05 | 17894 |

字符或词语层面的高错误率不一定意味着转录质量差，反而可能是阅读顺序不同所致。

对键值信息（命名实体）的检测进行评估，有助于克服阅读顺序方面的问题。表5b展示了使用 *Model-v3* 进行自动提取的结果，并按信息类型进行了详细分类。我们注意到，日期、序列号和项目编号等数值信息的提取效果非常好。标题和分析等文本信息也以很高的准确率被提取出来。只有训练集中较少出现的信息（如编排字段）的表现较差。总体而言，超过 95% 的信息都被正确提取。

## 5 结论

本文介绍了 SIMARA，这是一个新的公共数据库，包含数字化检索工具的图像及其中所含信息的转录。该数据库使我们能够为自动历史文档处理系统定义一种新型任务。这项任务是从文档页面中提取键值信息。我们提出了一种基线系统

该系统基于 Transformer神经网络，可通过单次处理提取这些信息。该模型的性能构成了参考结果，可用于评估后续为解决这一任务而提出的模型在 SIMARA 数据库上的进展。

## 参考文献

1. 匿名：从整页手写文档中进行键值信息提取。已投稿至国际文档分析与识别会议（2023年） 2. Beyer, Y., K ̈asen, A.: NorHand / 挪威语手写文本识别数据集（2022年5月）。https://doi.org/10.5281/zenodo.6542056, https://doi.org/10.5281/zenodo.6542056 3. Colavizza, G., Blanke, T., Jeurgens, C., Noordegraaf, J.: 档案与人工智能：当前讨论与未来展望概述。《计算与文化遗产期刊》**15(1)**（2021年） 4. Constum, T., Kempf, N., Paquet, T., Traounez, P., Chatelain, C., Bree, S., Merveille, F.: 历史手写表格的识别与信息提取：探索20世纪初的巴黎人口普查。载于：第15届文档分析系统国际研讨会（DAS）。第143–157页（2022年5月）。https://doi.org/10 2 10 5 1007 978 3 031 06555 . / - - - - . Coquenet, D., Chatelain, C., Paquet, T.: DAN：用于手写文档识别的无分割文档注意力网络。《IEEE模式分析与机器智能汇刊》（2023年）。https://doi.org/10 3235826 6 1109 2023 . /TPAMI. . . Coquenet, D., Chatelain, C., Paquet, T.: 使用垂直注意力网络进行端到端手写段落文本识别。载于：《IEEE模式分析与机器智能汇刊》。第45卷，第508–524页（2023年）。https://doi.org/10 3144899 7 1109 2022 1, . /TPAMI. . . Cunha L.F., Ramalho, J.C.: 微调BERT模型，从档案检索工具中提取命名实体。载于：第26届数字图书馆理论与实践国际会议（2022年） 8. Forn ́es, A., Romero, V., Baro, A., Toledo, J., S ́anchez, J.A., Vidal, E., Llad ́os, J.: ICDAR2017历史手写记录信息提取竞赛。载于：国际文档分析与识别会议。第1389–1394页 227 9 11 2017 10 1109 2017 ( ). https://doi.org/ . /ICDAR. . . Marti, U.V., Bunke, H.: IAM数据库：用于离线手写识别的英语句子数据库。载于：《国际文档分析与识别期刊》。第5卷，第39–46页（2002年11月）。https://doi.org/10 100320200071 10 1007. /s . Menasri, F., Louradour, J., Bianne-Bernard, A.L., Kermorvant, C.: a2ia法语手写识别系统参加RIMES-ICDAR2011竞赛。国际光学工程学会会议录 **8297**，51–（2012年1月）。https://doi.org/10 911981 11 1117 12. / . . Muehlberger, G., Seaward, L., Terras, M., Oliveira, S., Bosch, V., Bryan, M., Colutto, S., D ́ejean, H., Diem, M., Fiel, S., Gatos, B., Gr ̈uning, T., Greinoecker, A., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Zagoris, K.: 通过手写文本识别变革档案学术研究：以Transkribus为例。《文献学期刊》（2019年7月）。https://doi.org/10.1108/JD-07-2018-0114

12. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: 历史文档图像数据集综述。《国际文档分析与识别期刊》（IJDAR）**25**(4)，305–338（2022年12月）。https://doi.org/10.1007/s10032-022-00405-8, https://doi.org/10 8 13 1007 10032 022 00405 . /s - - - . Potanin, M., Dimitrov, D., Shonenkov, A., Bataev, V., Karachev, D., Novopolt- sev, M., Chertok, A.: Digital Peter：新数据集、竞赛与手写识别方法。载于：第6届历史文档图像与处理国际研讨会。第43–48页。HIP '21，美国纽约：计算机协会（2021年）。https://doi.org/10.1145/3476887.3476892, https://doi.org/10 3476892 14 1145 3476887 . / . . Rouhou, A.C., Dhiaf, M., Kessentini, Y., Salem, S.B.: 用于历史文档中手写识别与命名实体识别的联合Transformer方法。《模式识别快报》**155**，128–134（2022年）。https://doi.org/https://doi.org/10.1016/j.patrec.2021.11.010, https://www.sciencedirect.com/science/article/pii/S0167865521004013 15. Stutzmann, D., Torres Aguilar, S., Chaffenet, P.: HOME-ALCAR：对齐并标注的文书集（2021年10月）。https://doi.org/10.5281/zenodo.5600884, https: //doi.org/ 10 5281 5600884 . /zenodo. . Sanchez, J.A., Romero, V., 16 Toselli, A.H., Vidal, E.: ICFHR2016 READ数据集手写文本识别竞赛。载于：2016年第15届手写识别前沿国际会议（ICFHR）。第630–635页（2016年）。https://doi.org/10 0120 17 1109 2016 . /ICFHR. . . Tarridea, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., V´ezina, H., Kermor- vant, C.: 从魁北克教区手写记录中大规模提取家谱信息。《国际文档分析与识别期刊》（2023年） 18. Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B., Cohen, S.: 开始、跟随、阅读：端到端整页手写识别。载于：Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y.（编）《计算机视觉——ECCV 2018》。第372–388页。施普林格国际出版公司，尚（2018年） 19. Yousef, M., Bishop, T.: OrigamiNet：通过学习展开实现弱监督、无分割、一步式整页文本识别。载于：IEEE计算机视觉与模式识别会议（CVPR）。第14698–14707页（2020年6月）