

面向文化遗产数据的结构化处理与交互式可视分析平台设计与实现

——以中国出土青铜器数据为例

钟 瑞, 郭艳芬

成都信息工程大学计算机学院, 四川 成都

收稿日期: 2026年8月12日; 录用日期: 2026年9月30日; 发布日期: 2026年10月14日

摘 要

大量文化遗产资料至今仍以图像、扫描文档等非结构化形式保存, 难以直接用于统计分析和关系探索。本文以中国出土青铜器数据为对象, 将这类资料的处理分为三个环节: 先借助OCR和文本抽取把原始文献整理成可计算的二维数据, 再按时间、空间、类别组织成面向分析的字段结构, 最后通过交互式可视分析平台完成多维展示与关联发现。该流程在2310件青铜器记录上得到验证, 年代分布、地域分布、器型构成等分析都能落到具体数据上, 为文化遗产数据的深度利用提供了一条可行的路径。

关键词

文化遗产数据, 青铜器, 结构化处理, 可视分析, 交互式平台

Design and Implementation of a Structured Processing and Interactive Visual Analytics Platform for Cultural Heritage Data

—A Case Study of Chinese Excavated Bronze Data

Rui Zhong, Yanfen Guo

School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan

Received: August 12, 2026; accepted: September 30, 2026; published: October 14, 2026

Abstract

Large volumes of cultural heritage material remain preserved in unstructured forms such as images

文章引用: 钟瑞, 郭艳芬. 面向文化遗产数据的结构化处理与交互式可视分析平台设计与实现[J]. 软件工程与应用, 2026, 15(5): 664-677. DOI: 10.12677/sea.2026.155062

and scanned documents, which cannot be directly used for statistical analysis and relational exploration. Taking bronze artifacts excavated in China as its research subject, this paper decomposes the handling of such materials into three stages. First, OCR and text extraction convert the raw literature into computable tabular data. Second, the data are organized into analysis-oriented fields of time, space, and category. Third, an interactive visual analytics platform supports multi-dimensional display and relational discovery. The workflow was validated on 2310 bronze records, where analyses of temporal distribution, regional distribution, and vessel-type composition could all be grounded in concrete records, offering a feasible path for the in-depth utilization of cultural heritage data.

Keywords

Cultural Heritage Data, Bronze Artifacts, Structured Processing, Visual Analytics, Interactive Platform

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着数字技术的发展, 文化遗产保护与研究逐渐由传统资料保存向数字化管理方向发展, 数字资源建设逐渐由简单保存向数据组织和知识利用方向演进。数字博物馆、文物数据库以及数字档案系统的建设[1], 使大量文化遗产资料以文本、图像、文献等形式得到保存和传播, 为文化资源管理和公众访问提供了基础条件。

然而, 针对文化遗产数字资源进一步分析利用的需求, 现有研究仍存在数据转换、面向分析的数据组织以及关系探索方式等方面的问题, 主要表现在: (1) 文化遗产资料多以非结构化形式存在[2], 难以直接用于数据分析, 需要将原始资料转化为具有明确字段结构的数据形式, 为后续分析提供基础。(2) 已有结构化数据仍缺少面向分析任务的组织方式, 仅完成数据字段提取并不能满足分析需求, 需要合理组织数据, 对不同属性进行关联分析。(3) 文化遗产数据包含复杂的多维关系, 传统固定的信息展示方式难以支持开放式探索, 需要结合可视化表达和交互技术, 提高文化遗产信息的理解效率。

青铜器本身具有复杂的多维属性, 其历史价值不仅体现在单个文物的信息描述中, 还包含年代演变、地域分布、器物类型以及文化特征之间的关联关系。因此, 本文以中国出土青铜器数据为研究对象, 重点研究面向分析需求的文化遗产数据结构化处理方法、数据组织方式以及交互式可视分析实现方法, 提出一种的文化遗产数字分析利用的方案。

2. 相关研究

2.1. 文化遗产数据构建研究

文化遗产数字资源的利用效率要提升, 数据结构化处理是基础。近年来, 相关研究多从这一环节入手, 借助文本识别、信息抽取和知识组织等技术, 将非结构化资料转化为具有明确语义的数据形式。光学字符识别(Optical Character Recognition, OCR)是其中基础但有效的一环。它直接从图像或扫描文档中提取文字[3], 很好应对海量纸质或者未整理电子材料。文本信息抽取方法则更进一步, 从文本中识别时间、地点、类别等关键实体[4], 使原始资料具备可分析的基础。另有一部分研究从知识组织切入, 构建关联

关系来描述文化遗产信息之间的联系, 以求改善数据的管理与检索能力。

在文化遗产信息提取和资源组织方面, 相关研究已有一定积累, 但多数工作围绕提高信息获取效率、搭建资源管理体系展开, 数据如何真正服务于具体分析任务, 考虑得并不充分。不少研究的样本量也有限。文化遗产分析的要求不止于抽取基本信息。研究目标不同, 需要组织和关联的数据维度也随之变化, 这一步不到位, 后续分析便无从谈起。就覆盖面而言, 相关探索从通用博物馆文物图谱[5]、非遗档案资源整合[6]逐步拓宽到大规模古籍典籍关联[7], 为跨媒介文化遗产数据的语义互联提供了理论基础。针对可移动文物, 高劲松等给出了基于 RDF 与 Neo4j 的关联数据存储方案[8], 为本文的结构化存储设计提供了参考。

2.2. 文化遗产可视分析研究

可视化技术能够通过视觉编码降低复杂数据理解难度[9], 并支持用户在数据探索过程中的模式发现, 因此被广泛应用于复杂信息分析领域。但传统可视化大多围绕某个既定的展示需求展开, 把功夫下在信息的呈现效果上, 比如采用时间线、地图、统计图表以及关系网络等方式[10], 至于用户如何在复杂数据里主动发现关系, 着墨不多。文化遗产分析带有明显的探索性, 研究者会根据分析结果不断调整关注范围, 仅靠静态图表, 难以满足这种开放式需求。

交互式可视分析把计算能力和人的认知结合起来, 通过筛选、缩放、联动等操作[11]-[13], 用户可以动态探索数据关系。对文化遗产这类多维数据而言, 交互并不只是展示功能的补充, 更是理解数据、发现关系的重要途径。

总的来看, 已有工作分别落在数字保存、数据结构化和可视化展示这三块, 为文化遗产数字资源的保存与访问打下了基础。但要从原始资料走到分析应用, 中间还缺一条完整的流程: 资料先要结构化, 变成可分析的数据, 再按研究任务重新组织, 最后结合可视分析, 让研究者有能力去探索其中的复杂关系。

本文围绕文化遗产数据的分析需求, 设计了一套包含资料处理、数据组织到交互式可视分析应用的处理流程, 着眼提升数字资源的分析利用能力; 并以中国出土青铜器数据为例, 搭建交互式可视分析平台, 验证这套流程在文化遗产数据探索中的实际效果。

3. 数据处理与组织方法

3.1. 数据来源与特点

本文所用数据取自《中国出土青铜器全集》[14]。这套资料系统收录了全国各地区的出土青铜器信息, 包括文物名称、年代、出土地、器物描述和图片等, 针对文化遗产数据分析是一个典型的例子, 也是数据选择的原因之一。

数据集按图文结合的方式组织。单件文物通常列有名称、所属年代、出土地、尺寸、器物描述和图片, 一切以方便阅读、方便展示为准, 字段之间没有统一的结构, 要直接用于统计和多维探索并不容易。年代、地域、器型这类信息散落在页面文字里, 研究者想比较, 只能先人工翻查、再手工整理。要改变这种局面, 就得对原始资料做结构化处理, 把有分析价值的信息抽出来, 整理成规范的数据字段, 后续的数据组织和可视分析才有依托。

3.2. 数据结构化处理流程

文化遗产资料的非结构化是首先要解决的问题。本文为此设计了一条数据处理流程, 依次包含资料获取、文本识别、信息整理和结构化存储四个环节, 目标是提取出能支撑分析任务的关键字段, 把文本

资源转换为可分析的数据。流程如图 1 所示。

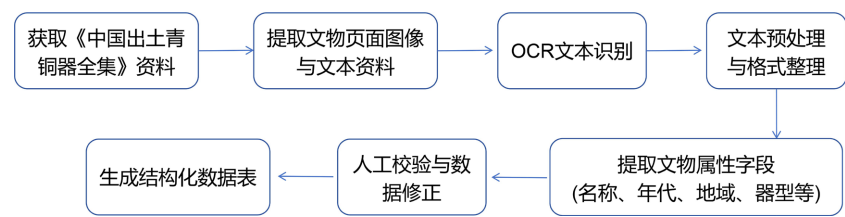


Figure 1. Data structuring processing flow
图 1. 数据结构化处理流程图

《中国出土青铜器全集》的电子资料图文混杂，原始形态是扫描文档，文字以图像形式存在，必须先经过 OCR 才能变成可编辑文本。这一环节本文交给了阿里云大模型 OCR 服务。它建立在多模态大模型底座之上，叠加了 OCR 技术模型，能够理解复杂版式的版面结构，并按要求抽取其中的内容。

使用前先定义一份 JSON Schema，把文物名称、所属年代、出土地点、器物类型等目标字段固定下来，作为提取约束传入模型接口，具体组织结构参考 3.3 节。之后模型逐页读取文物资料，按这套字段结构定位并返回内容。拿到返回的 JSON 后，脚本再做解析和拆分，把嵌套字段展平成关系型表格，得到的结果是可以直接用于统计的二维数据表。

各文物对象按统一的字段格式入库，形成面向后续分析的数据基础。与过去人工录入、逐条查找相比，这套流程省去了大量前期准备工作，原本散落在文献里的信息也具备了可计算、可统计的能力，这是后续可视分析的基础。

3.3. 面向分析的数据组织

传统数据存储通常关注信息完整性，而文化遗产分析更关注不同属性之间的关系。考虑到后续分析任务，数据组织以单件文物为基本实体，围绕它建立三类属性模型：时间属性、空间属性和类别属性。时间属性对应文物所属年代，历史演变分析的基础就在这里。空间属性记录出土地，地域分布分析据此展开。类别属性涵盖器型、纹样等文化特征，为属性关系探索提供素材。经过这样的整理，原本零散的文物记录变成了带有多维属性的数据对象，同一份数据由此能服务于不同的分析任务。例如，用户可以根据年代字段观察不同历史时期文物数量变化，也可以结合地域字段分析不同区域的文物分布情况。本文提取的核心青铜器数据字段组织如表 1 所示。

Table 1. Examples of core bronze artifact data field organization
表 1. 核心青铜器数据字段组织示例

字段类别	字段名称	说明
基本信息	id、name	文物唯一编号和名称
时间信息	dynasty	所属年代
分类信息	type、pattern	器型与纹样信息
尺寸信息	length_cm、width_cm、height_cm	文物尺寸
空间信息	unearthed_site、region	出土地和区域
来源信息	museum、source_册	收藏单位及资料来源
描述信息	description	文物文字描述
图像信息	image_1、image_2	文物图片路径

其中, 尺寸和描述原文字段用于保留原始资料, 长度、宽度、高度、地区、器型和纹样等字段属于面向当前分析任务生成的结构化属性。后者受到字段定义、词表范围和材料表述方式影响, 不应视为对原始文献全部语义的完整替代。实际使用中可以根据文化遗产分析需求, 对已有结构化数据进行二次信息抽取和合理组织。

关键提示词如下:

你是青铜器图录数据提取专家。从文本提取每件文物, 输出 JSON 数组, 每件一对象。同页多件按编号逐条提取不合并。只输出 JSON, 无解释, 无法提取填 null。

字段: id (数字)、name (器名)、dynasty (夏/商/西周/春秋/战国/秦/汉/东汉/三国/晋/南北朝/隋/唐/宋/元/明/清/其他)、type (兵器/马具/礼器/生活用器/乐器/其他)、length_cm/width_cm/height_cm/weight_g (浮点)、size_raw (尺寸原文)、unearthed_site、unearthed_year (整数)、region (省级)、tomb_id、museum、pattern (从描述提取全部纹样名, 多个纹样以顿号分隔; 仅有铭文时写“有铭文”; 无纹时写“素面”); has_inscription (true/false)、description (形制描述, 不含尺寸、出土地和馆藏)。

规则: 同段文字拆分尺寸至 size_raw、地点至 unearthed_site、馆藏至 museum, 其余内容至 description; pattern 须完整提取原文纹样, 不可遗漏。

该提示词主要包含对象边界、字段范围和输出形式三类约束: 以编号区分同页多件文物, 以 null 保留无法确认的信息, 以 JSON 数组约束机器可读的输出形式。后续本地脚本只负责记录拆分、字段规范化和 CSV 转换, 不对原始描述进行独立大模型改写。

4. 交互式可视分析设计

文化遗产数据往往牵涉多个维度: 年代、地域、器型, 每个属性都对应不同的分析需求。实际分析中, 用户可能在观察数据的过程中不断从新的角度观察文物, 如先看某个时期的文物数量变化, 接着又想知道这一时期的器物出在哪些地方、属于什么器型。这样的使用方式意味着, 系统仅展示信息是不够的, 还得提供灵活的筛选和关联探索, 让用户能随分析需要调整观察范围。

本文据此把可视分析任务分成两块: 多维信息的可视化展示与交互式探索。前者用不同的视觉形式呈现数据特征, 后者靠交互操作帮用户发现属性之间的关系。

4.1. 多维数据可视表达设计

本文确定的青铜器的属性大致归为时间、空间和类别三类, 各类数据的特征不同, 适合的视觉形式也不一样。接下来按照这三个维度分别设计可视化方式, 让数据特征能直观地呈现出来。

4.1.1. 时间维度表达

文物所属年代天然带有先后顺序, 按时间统计各时期的数量, 能直接看出文物在不同历史阶段的变化。折线图最为适合: 横轴按历史时期排列, 纵轴对应该时期的文物数量, 不同时期之间的数量差异、样本在时间维度上的集中区间, 一眼就能比较出来。比起逐行查看数据表格, 这种画法把时间变化关系摆在了明面上, 理解历史演变过程的门槛也随之降低。

4.1.2. 空间维度表达

文化遗产的形成和传播带有明显的地域特征, 出土地点反映着不同区域之间的文化分布。要把这种分布呈现出来, 地图是首选。本文以省级地图为主, 地区数量排序图为辅: 地图用颜色深浅表示文物数量, 排序图补充列出出土较多的地区及其数量差异, 两者合起来, 样本在空间上的集中程度和覆盖范围一目了然。用户既能比较不同区域的文物数量, 也能结合时间与属性信息, 进一步看地域和文化特征之

间的关系。

4.1.3. 属性维度表达

器型和纹样都是离散类别数据, 不适合按数值大小去画, 只能做分类统计。器型方面, 柱状图直接比较各类别的数量差异; 纹样方面, 频次统计看哪些纹样出现得多。问题在于, 一件文物上往往同时出现多种纹样, 单独的频次统计看不出来它们之间的搭配关系。为此, 系统额外引入共现矩阵, 记录纹样两两同时出现的次数, 把这一层组合信息补上。

4.2. 交互设计

交互设计遵循简洁、功能多元、可持续探索的原则, 以增强用户认知。本文根据需求选择的交互方式是筛选、联动和详情查看, 帮助用户逐步发现数据之间的关联关系。

条件筛选帮助用户根据年代、地域、器型等属性选择目标数据范围, 从整体数据中快速定位感兴趣的对象, 减少无关信息对分析过程的干扰。

多视图联动用于增强不同数据维度之间的关联表达。当用户调整某一分析条件时, 相关时间、空间和属性视图同步更新, 帮助用户观察同一数据对象在不同维度中的表现, 从而从单一属性观察转向多维关系分析。

详情查看支持从统计结果进一步查看具体文物信息, 通过由整体到局部的分析过程, 使宏观数据规律与具体文物对象相互衔接, 实现统计分析 with 个体信息的结合。

4.3. 异常考虑

考虑到“未知”“其他”以及无法进一步细分的记录在数据中占有一定比例, 系统不将其直接删除, 也不与明确类别合并。在统计结果中, 这部分数据保留独立计数, 并在图表说明中同步给出其数量及占比。为了达到辩证看待的效果, 特地对此类数据呈现模式采取特殊标记, 采用叠加浅色斜线纹理, 作为不确定或待细分数据的辅助标记; 用户由此既能看到总体分布, 也能识别某一分析维度中数据缺失或分类不足的程度, 可以快速定位问题, 避免把“未知”和“其他”误读为具有明确文化含义的类别, 促进辩证思考。

5. 交互式可视分析平台实现

5.1. 可视化平台实现

整个平台按组件化的思路搭建, 前端框架用 React, 构建和开发环境由 Vite 负责, 折线图、柱状图、地图、饼图和矩阵等可视化内容则交给 ECharts 渲染。界面代码分成四块: 页面模块管整体布局, 图表组件接收处理好的数据、生成对应视图, 数据处理模块做字段筛选、分类统计和格式转换, 全局状态模块统一维护共享的筛选状态。这样切分之后, 数据计算和界面呈现解耦, 不同分析页面之间重复的数据处理逻辑和冲突减少, 性能也能得到相应的释放。

结构化结果以 JSON 文件保存, 每条记录包含文物编号、名称、年代、器型、出土地区、纹样、尺寸、馆藏信息、文字描述和图像路径等字段。应用启动时把这些数据一次性读入公共状态, 之后各页面不再重复加载原始文件, 只按当前页面和筛选条件从内存里取相应的子集。其中字段为空、或无法归入明确类别的记录, 系统在展示阶段做一次缺失值判断, 统一归入“未知”或“其他”, 避免空值直接进入图表计算。各字段在前端怎么处理, 架构里也已约定好: 年代、器型用规范化字段, 地区对应地图上的省级名称, 纹样则用预先拆分的标准化数组。

实际加载渲染管线要经过五步: 数据读取、条件过滤、字段聚合、图表格式转换和组件渲染。数据

处理模块先按当前筛选状态, 从全量数据里捞出符合年代、器型、地域和纹样条件的记录; 再按视图需要的字段分组、计数, 把原始记录变成“类别 - 数量”“时间 - 数量”这类统计结果; 最后按 ECharts 的要求组装成横纵坐标、系列数据、地图数值或分类标签, 交给对应图表组件渲染。

四类视图的处理方式分别为: 时间视图按年代计数, 再按预设的历史顺序排列, 生成折线图需要的横坐标和数值序列——顺序由程序指定, 避免出现直接按字符串排序带来的年代错乱; 器型视图统计各器型的数量画柱状图, 数量太少或归不了类的, 统一放进“其他”; 空间视图先汇总各省的记录数, 再把地区名称和地图上的行政区对应起来, 形成着色的数据序列, 同时把结果按数量降序排列, 生成主要出土地区的排名; 纹样视图读的是预先拆好的纹样数组, 统计每种纹样出现的频次, 遇到一件文物上有多多个纹样的, 还能据此生成共现关系数据。

每个分析页面都按“主图表 + 辅助图表”来布局, 所以核心点是主图表, 也是当下页面的聚焦研究点。年代页把全局年代趋势放在中间, 器型、地域、纹样等统计图作为辅助; 地域页的主角是省级地图, 旁边配地区排名、年代分布和器型构成; 器型页以类别数量统计为主, 辅以年代分布、主要出土地和文物列表; 纹样页则同时呈现纹样频次、年代变化、器型构成和共现矩阵。这些图表外观各异, 但都取自同一数据集, 所以同一页面上各项统计结果是完全一致的。页面内容与图表区域各自做成独立组件, 由页面负责统一组合与布局。

样式方面, 颜色、文字、边框和间距统一由一套配置控制。图表在普通、悬停、选中三种状态下用不同的视觉编码: 当前选中的年代、地区、器型或纹样以高亮边框或强调色突出, 其余类别压暗一些。图表容器随页面宽度自适应, 宽屏下主要区域和辅助区域按一定比例排布, 保证地图、趋势图和分类统计图同屏时可读。这套样式配置集中管理, 各页面不再各自定颜色和尺寸, 界面表现才不至于各处不一。平台主界面如图 2 所示。



Figure 2. Main interface of the interactive visual analytics platform
图 2. 交互式可视分析平台主界面

5.2. 交互功能实现

交互功能有四个: 组合筛选、图表选择与多视图联动、文物列表更新、详情查看。

筛选条件和当前选中的对象都放进一个全局状态里, 年代、地域、器型、纹样四个页面共用这份状

态。用户做一次筛选或点选图表, 状态先更新, 各图表组件随即读取新的数据子集、重新计算统计结果, 页面内容便同步刷新。

条件筛选做在右侧边抽屉里, 年代、器型、出土地区、铭文情况都可以作为条件。多个条件可以在同一个面板里勾选, 点下应用, 系统把它们并进一个统一的筛选对象, 再按字段取交集过滤全量数据, 只留下同时满足所有条件的记录。过滤之后, 统计图表、结果数量和文物列表都基于新的数据子集重算。这个筛选状态在页面切换时不会丢, 用户点清空才会重置筛选对象、恢复全量数据。

多视图联动由主视图的选中状态驱动。四个页面各自记住自己当前选中的年代、地区、器型或纹样, 点击对象也按数据特点来定。点中年代折线图的某个节点, 系统记下这个年代, 从当前筛选结果里再提取该时期的数据, 器型、出土地、纹样等从属视图随之更新; 点中地图中的某个省份后, 系统按该地域重新生成数据子集, 同步刷新相应的年代、器型和文物信息。器型页、纹样页的联动采用相同逻辑。

一次交互之后, 当前选择会被保留, 对应图形高亮, 用户据此知道现在限定在什么条件下。再点选别的对象, 原来的选择被替换, 从属视图跟着更新; 取消选择, 页面回到当前筛选条件下的总体分布。各图表之间并不直接改来改去, 而是各自读取更新后的选择状态和数据子集, 跨组件联动由此成立。

文物列表直接由当前数据子集生成。列表组件按文物唯一编号取出名称、年代、器型、地区和图像路径, 以卡片形式排出来。用户在统计图表里把范围收窄, 列表会同步缩小到相关对象; 点任意一张卡片, 系统把编号当作路由参数跳进详情页, 再按编号查出这条记录, 图像、尺寸、馆藏、纹样、描述和铭文都在这里。从详情页返回时, 之前的筛选和选择状态原样保留。交互功能的部分效果见图 3。

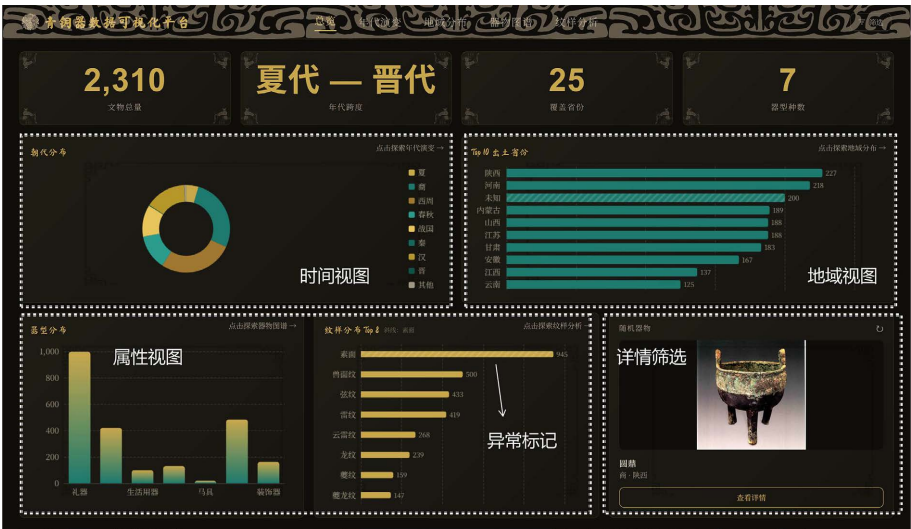


Figure 3. Examples of interactive function implementations
图 3. 交互功能实现示例

多次交互产生的结果必须一致, 其做法是把全局筛选和页面内主视图选择分开对待。全局筛选划定当前页面能用的基础数据范围, 页面内的选择再在这个范围内切出一个局部子集。比如用户先在筛选面板里限定“西周”和“礼器”, 再到地域视图中点内蒙古, 系统先按全局条件拿到西周礼器这批数据, 再依内蒙古做二次过滤。执行顺序固定为全局筛选、主视图选择、从属视图更新, 图表、结果数量和文物列表因此始终面对同一份数据范围。

图表重绘时并不会回头再取一次原始文件, 过滤和聚合都在内存里的全量数据上完成。当前大约两千多条文物记录, 这个规模足以对筛选和点击即时响应, 也省掉了反复读数据的开销。统一状态、数据

过滤函数和图表组件彼此配合, 用户从做出选择、完成条件计算, 到多视图更新、查看具体对象, 这就是整条交互链路的逻辑。

6. 结果与分析

为了验证本套方法的有效性, 在每个环节的关键方面进行以下的分析检验和说明。

6.1. 数据结构化验证

数据作为可视化的原始材料, 也是方法链条的第一步, 在此处设计定量实验以及评价指标来评估信息抽取结果质量。

本文从 9 册资料中分别随机抽取 5 个包含文物记录的页面, 每页选择 1 件文物, 共得到 45 条校验记录。人工对照原始 PDF, 对名称、年代、器型、地区、纹样和描述六个字段进行校验。其中, 名称、年代、器型和地区采用字段级评价, 纹样采用纹样词项级评价, 描述采用内容字符级评价, 均报告 Precision、Recall 和 F1; 描述字段另报告字符错误率(表 2)。

Table 2. Evaluation results of key fields for information extraction

表 2. 信息抽取关键字段评测结果

字段	评价粒度	Precision	Recall	F1
名称	字段级	80.00%	80.00%	80.00%
年代	字段级	97.78%	100.00%	98.88%
器型	字段级	100.00%	100.00%	100.00%
地区	字段级	100.00%	88.89%	94.12%
纹样	纹样词项级	97.78%	94.62%	96.17%
描述	内容字符级	83.99%	99.57%	91.12%

注: 字段级以单条记录的单个字段为单位, 纹样词项级以拆分后的纹样名称为单位, 内容字符级以去除空格和标点后的字符为单位。Precision 表示识别结果中正确内容所占的比例, 主要衡量是否多提、错提。Recall 表示人工校验内容中被正确提取的比例, 主要衡量是否漏提。F1 为二者的调和平均, 只有“错误较少”和“遗漏较少”同时满足, F1 才会较高。

评测结果显示, 年代字段 F1 为 98.88%, 地区字段 F1 为 94.12%, 纹样字段 F1 为 96.17%, 能够为本文采用的时间、地域和纹样分布分析提相对准确的数据支撑。名称字段 F1 为 80.00%, 差异主要集中在青铜器名称中的低频生僻字和形近字。器型字段在本次人工复核中未发现明显差异, 但该字段主要由识别服务依据名称和描述进行归纳, 没有经过文博专家独立标注, 因此这里只将其视为当前样本下的字段一致性结果, 不进一步推断其专业分类效度。

描述字段的整段完全一致率为 64.44%, 即本条记录若有一处与原文不符, 比如多写, 漏写, 即标记失误, 仅作为严格补充。字符级评测结果显示, 其 Precision 为 83.99%, Recall 为 99.57%。45 条记录中有 40 条的字符错误率不超过 5%, 仅有 4 条存在较明显的跨记录多提取。参考文本共包含 3451 个有效内容字符, 其中替换字符 13 个、删除字符 2 个、插入字符 642 个, 说明精确率下降主要来自相邻记录被一并带入, 而不是正文内容普遍遗漏。将记录边界问题单独区分后, 其余 41 条描述的字符 Precision 为 99.53%, Recall 为 99.59%, 字符错误率为 0.53%。这说明当前流程在原文内容保留方面表现较好, 主要不足集中于少数专业生僻字和复杂排版页面中的记录边界判断。

差异记录进一步归纳为五类: 一是生僻字和形近字造成的字符误识, 主要出现在名称与少量描述文

字中,如“𠩺、𠩺、𠩺、𠩺、𠩺、𠩺、𠩺、𠩺”。值得一提的是,有相当比例生僻字被正确识别,如“簠、盞、𠩺、𠩺、𠩺、𠩺、𠩺”,甚至缺字占位符“口”。后续可尝试引入文物专业词表和上下文纠错机制以提高生僻字和特殊表达的识别稳定性;二是同页多件或连续排版造成的记录边界偏移,表现为相邻文物文字被一并带入;三是口径、腹径、通高、残长和数值范围等复杂尺寸难以直接映射到统一长宽高字段;四是地点、馆藏信息与正文处于同一文本块时产生字段切分不足;五是预设纹样词表不能完全覆盖目雷纹、联珠纹等细粒度或复合纹样。按记录计数,上述五类分别涉及 9、4、30、7 和 5 条,同一记录可能同时涉及多类问题;上述问题主要影响派生字段的精细使用,完整信息被保存在了原文字段之中,此类失误可做二次提取处理,不等同于原始文字整体丢失。

综合来看,当前结构化结果能够支持年代、地域、器型和纹样等总体分布的探索性分析,但不同字段的可靠程度并不完全相同。本文采用的是围绕当前分析目标设计的轻量化字段结构,对于未参与主体分析的复杂尺寸、考古编号和细粒度文化属性,仍以保留原文和提供回溯依据为主,不将其视为已经完成专业化抽取。结构化处理的价值由此更多体现在降低资料整理成本、形成可计算的核心字段,并为后续针对具体问题的再处理提供基础。

6.2. 多维数据分析

数据结构化完成后,本节借助可视分析平台从时间、空间和属性三个角度观察这批青铜器数据,检验前面的数据组织方式能否支撑不同方向的分析。

6.2.1. 时间维度表达

在时间维度方面,图 4 给出了各时期的数量分布,下方显示总体器型、主要出土地和纹样构成。2310 件记录里,西周 634 件、商代 629 件,各占 27.4%和 27.2%,合起来超过一半,构成样本最主要的时间范围。春秋 298 件、战国 271 件、汉代 349 件,仍有一定规模;夏代 105 件,秦、晋及更晚的时期则少得多。

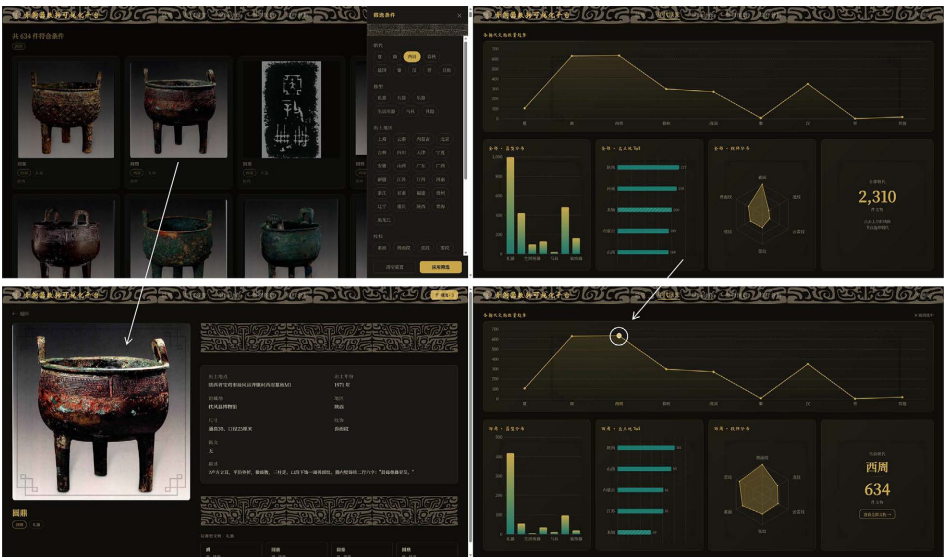


Figure 4. Temporal analysis results
图 4. 时间分析结果

数量从夏代进入商代后明显走高,商、西周两代维持在高位,春秋、战国回落,汉代又有所回升。这种分布说明样本在时间上很不均衡,各时期数量差距悬殊,不能只凭绝对数量判断青铜器发展的强弱;

与此同时, 商、西周样本充足, 比较这两个时期的器型、地域和纹样特征有较扎实的依据, 而数量稀少的时期更适合翻具体文物记录做个案观察。

选定某一时期, 器型、出土地和纹样视图会同步更新, 时间维度由此成为进入其他属性分析的入口, 可以比较不同时期内部的结构差异。

6.2.2. 空间维度分析

在空间维度方面, 地图用于呈现样本的空间覆盖范围和集中程度, 地区排序用于补充各省份之间的数量差异。用户在地图中选择某一省份后, 系统以该省记录作为当前数据范围, 相关统计图表随之更新, 从而进一步观察该地域内部的年代与属性构成。图 5 为联动数据表。



Figure 5. Spatial analysis results
图 5. 空间分析结果

从总体数量看, 陕西有 227 件、河南有 218 件, 位居明确地域记录的前两位; 内蒙古 189 件, 山西和江苏均为 188 件, 甘肃 183 件、安徽 167 件、江西 137 件、云南 125 件。此外, 还有 200 件记录的出土地无法归入具体省份, 在统计中被单独标记为“未知”。

上述结果表明, 当前样本在空间上并非均匀分布, 而是主要集中于陕西、河南、山西和内蒙古等地区, 江苏、甘肃、安徽、江西和云南等地也具有一定数量。通过地图可以直观辨识样本的主要集中区域, 结合地区数量排序则可以进一步比较不同省份之间的规模差异。

需要注意的是, “未知”地区记录达到 200 件, 其数量接近部分主要省份, 说明原始文献中的地域信息仍存在缺失或标准化不足。开展省际比较时, 这部分记录既不能分摊至其他省份, 也不应直接忽略。因此, 空间分析呈现的是当前数据集的地域分布特征, 而不能直接等同于青铜器在历史时期中的完整地理分布。

在具体地域分析中, 用户可以通过地图选择目标省份, 年代分布、器型构成和纹样统计等视图将基于该省数据同步更新。由此, 地域不仅是空间分布的展示维度, 也成为连接年代、器型和纹样等属性的分析条件, 使用户能够从总体地域格局继续进入特定地区的内部结构分析。

6.2.3. 器型属性分析

在器型方面, 图 6 是各类别的数量对比。礼器 997 件、占 43.2%, 居首; “其他” 483 件、占 20.9%; 兵器 420 件、占 18.2%; 装饰器 162 件、乐器 130 件、生活用器 99 件, 马具最少, 仅 19 件。类别之间

差距悬殊，器型构成的不均衡很明显。



Figure 6. Attribute analysis results
图 6. 属性分析结果

礼器接近样本总量的一半，在当前资料中占据主导；兵器与“其他”也有相当规模，其余类别构成尾部。这一分布既让用户能快速把握数据的主体器型，也提醒后续比较要留意类别间的样本量差异：礼器、兵器适合做年代和地域层面的整体统计，马具这类低频类别则更适合结合具体文物逐件考察，不宜由少数几件推出普遍结论。

选中某类器型，下方的年代和出土地视图会显示它在各时期的分布和主要出土地，“这类文物数量多”的笼统判断便具体成“它集中在哪些时期、哪些地区”。文物列表则保留了从统计结果回到具体记录的通道，宏观分布和个体信息由此接上。

“其他”类占到 20.9%，说明器型分类里还有相当数量没有细分下去。这会影响器型结构的精细比较，后续可以借助文物名称和描述字段进一步细化类别标准，减小宽泛分类带来的偏差。

6.3. 交互式可视分析

多维分析回答了数据本身长什么样，交互探索则要回答用户能不能按自己的需要去查：既能逐步收窄分析范围，也能在不同属性之间来回查看关联关系。

6.3.1. 基于条件筛选的信息探索

分析目标明确时，用户的工作是从整批数据里一步步筛出符合条件的文物：先限定年代，再限定器型，再限定出土地。系统支持多条件组合筛选，筛选过程中当前已选条件和结果数量始终可见，每一步收窄了多少，基于上步筛选，有递进感性认知。这里以“西周时期、内蒙古地区出土的礼器”为例，说明渐进式筛选是怎么进行的。

先选器型“礼器”，得到 997 件；再加年代“西周”，范围收窄到 417 件；最后补上地域“内蒙古”，剩 16 件同时满足三个条件的文物。筛选完成后，页面顶部显示当前条件与结果总数，下方列表呈现对应的文物卡片，点开可看单件文物的详细资料。整个过程的阶段与结果变化见表 3。

筛选过程的条件每加一重，结果就缩一截，从 997 件一路收到 16 件。礼器在全部样本里数量偏大，单靠器型一条筛不出来，必须把年代、地域叠加上去才落得到具体对象。这说明统一组织后的字段可以自由组合成查询条件，分析能从“一大类”逐步走到“特定集合”。

一次性把条件填满也能得到同样的结果，但过程中看不到每一步的代价；渐进式筛选把过程摊开，每个条件让范围缩小多少都清清楚楚。筛出的数据也不只是一张列表，继续做器型细分、纹样构成、具体文物特征分析，都以它作入口。说到底，筛选的价值不在“最终剩几条”，而在把结构化字段变成可组

合、可调整的条件, 研究者能根据看到的结果随时修正问题。

Table 3. Progressive conditional filtering process and changes in results
表 3. 渐进式条件筛选过程及结果变化

筛选阶段	新增条件	当前组合条件	结果数量	分析作用
1	器型: 礼器	礼器	997 件	从整体数据中提取礼器类别
2	年代: 西周	西周 + 礼器	417 件	限定目标历史时期
3	地域: 内蒙古	西周 + 礼器 + 内蒙古	16 件	定位特定时期、地区和器型的文物集合

6.3.2. 基于多视图联动的关系探索

在多维数据分析过程中, 年代、地域、器型和纹样既可以作为独立的分析入口, 也可以在其他条件被选中后作为从属视图呈现关联结果。系统采用主视图驱动的联动机制: 用户选择某一维度中的目标属性后, 其余视图基于相同的数据子集同步更新。**图 7** 分别展示了年代、器型和地域作为主要分析条件时的联动结果。

四个维度没有固定的主从之分: 任何一方被点中, 都能作为主条件, 其余视图随之变成它的从属。



Figure 7. Multi-view linkage results under different primary analysis conditions
图 7. 不同主分析条件下的多视图联动结果

以年代为主条件时, 点中“西周”, 器型、出土地、纹样等视图便换成这一时期的构成, 时间分布从“总体统计”变成了进入其他维度的入口。

以器型为主条件时, 礼器的从属视图透露出更有用的信息: 它的年代集中在商、西周两代(西周约 417 件), 出土地偏于河南、陕西、山西一带(144 件、139 件、132 件)。礼器数量多, 并不是各个年代、各个地区均匀堆出来的, 而是时间上集中、空间上有明显倾向。单看总量只能知道“礼器多”, 联动之后才看清它多在哪里。

以地域为主条件时, 点中“陕西”, 该区域青铜器的年代、器型构成同步呈现。地图此时不只是在标位置, 也成了限定数据范围的入口, 可以继续比较这个地区内部的结构。

四个维度彼此没有固定的单向依赖, 根据需求选择主条件, 其余视图跟着当前子集反馈。分析可以挑任何属性切入, 同一界面里就能看到这组数据在别的维度上的分布, 省去在多个图表之间来回比对的功夫。需要说明, 以上结果只反映当前这批结构化样本的分布, 不能直接当作整个历史时期青铜器生产、使用和传播的规律。

7. 结论与展望

文化遗产数字资源的积累到了一定规模之后, 分析利用能力就成了决定其价值的关键。本文围绕资料结构化程度不足、多维信息关系难以表达、数据探索方式有限三个问题, 以中国出土青铜器数据为对象, 完成了从资料结构化、数据组织到交互式可视分析应用的完整处理流程, 为青铜器数据的数字化分析提供了一种可行的实现方式。

这套方案也有几处明显的边界。其一, 数据只来自《中国出土青铜器全集》一种来源, 规模与覆盖都有限, 据此得出的结论不宜外推到整个历史范围, 后续需要纳入更多类型的文化遗产资料来扩充。其二, 结构化结果在原文内容保留方面总体较好, 但复杂文本和文化属性的自动抽取仍有提升空间。后续可引入文物专业词表、上下文纠错或领域知识模型; 当研究目标转向尺寸比较、铭文释读或细粒度纹样分析时, 还需重新设计字段结构, 并开展针对性的二次提取和专业复核。其三, 目前的分析只围绕时间、空间和属性三个维度展开, 文物之间的关联、历史事件、文化传播路径等更丰富的关系还没有被利用起来, 这些是下一步可以做深的方向。未来可根据分析对象的特性针对性建立模型分析数据结构, 或许可进一步引入文物关联、历史事件和文化传播路径等更丰富的数据关系, 以支持更加深入和便捷的文化遗产知识探索。

基金项目

成都信息工程大学 2024 年本科教学工程项目“数智项目”(JYJG2024210)。

参考文献

- [1] 位通, 张婷, 邓秋怡, 等. 大规模文化遗产图像数据化研究[J]. 中国图书馆学报, 2025, 51(2): 58-73.
- [2] 高劲松, 韩牧哲. 考古发掘资料图数据库的语义关联构建研究[J]. 图书情报工作, 2021, 65(9): 105-116.
- [3] 田鹏辉, 张君, 张正全, 等. 基于 OCR 技术的复杂背景图像文字识别方法分析[J]. 电子技术, 2025, 54(12): 120-121.
- [4] 李小松. 自然语言处理技术下文本信息语义抽取方法[J]. 现代电子技术, 2025, 48(23): 176-180.
- [5] 丁利民. 基于数字化下博物馆文物知识图谱的构建[J]. 文物鉴定与鉴赏, 2022(17): 40-43.
- [6] 赵雪芹, 路鑫雯, 李天娥, 等. 领域知识图谱在非遗档案资源知识组织中的应用探索[J]. 档案学通讯, 2021(3): 55-62.
- [7] 欧阳剑, 梁珠芳, 任树怀. 大规模中国历代存世典籍知识图谱构建研究[J]. 图书情报工作, 2021, 65(5): 126-135.
- [8] 高劲松, 张强, 李帅珂. 可移动文物的知识图谱构建及关联数据存储——以湖北省博物馆为例[J]. 现代情报, 2022, 42(4): 88-98.
- [9] 陈为, 沈则潜, 陶煜波. 数据可视化[M]. 北京: 电子工业出版社, 2023.
- [10] 刘杰, 石战结, 王征宇, 等. 基于 ArcGIS 的遥感与地球物理考古探测数据综合管理系统研究[J]. CT 理论与应用研究, 2020, 29(3): 319-327.
- [11] 胡汗林, 邓三鸿. 知识图谱在青铜器数字馆藏建设中的应用[J]. 数字图书馆论坛, 2023, 19(4): 1-8.
- [12] 张雨辰. 从修复到解读: 交互式可视化在文物修复展示中的应用——以《蒋懋德画山水图贴落》为例[J]. 文物天地, 2026(4): 147-151.
- [13] 沈恩亚. 交互式可视化关键技术研究[D]: [博士学位论文]. 北京: 国防科技大学, 2017.
- [14] 李伯谦. 中国出土青铜器全集[M]. 北京: 科学出版社, 2018.