Entity Linking for Untangling Provenance of Colonial Heritage Object

Anonymous authors

No Institute Given

Abstract. The adoption of Semantic Web and Linked Open Data (LOD) in the cultural heritage domain, specifically focusing on museums leveraging these technologies to enrich repositories is growing through the past decade. The current study underscores the challenges museums face in managing collections, particularly in the absence of essential provenance information, with a spotlight on the complexities associated with colonial-era acquisitions. It proposes a solution through automated entity linking across datasets, offering a scalable approach to information enrichment. The research centers on supporting provenance researchers in colonial contexts by integrating collector biographies into a comprehensive knowledge base. The central challenge involves linking data across Knowledge Graphs (KGs) from multiple heritage institutions, addressing characteristics such as naming variations, missing attribute values, and errors in historical datasets. The comparative analysis of entity linking approaches, ranging from basic string matching to sophisticated methods, informs the selection of the most suitable method. The chosen method is then applied to integrate data from diverse sources, resulting in a unified knowledge graph. Evaluation using competency questions, derived from prior research, assesses the integrated knowledge graph's capacity to meet the specific needs of domain-specific researchers, contributing valuable insights to the discourse on data integration challenges and entity linking methodologies in the cultural heritage domain.

1 Introduction

The adoption of Semantic Web and Linked Open Data (LOD) has witnessed a notable upsurge within the domain of cultural heritage. By providing a standardized method for publishing and interlinking diverse datasets, Linked Open Data further contributes to the enrichment of cultural heritage repositories. Museum data publishers are increasingly adopting Linked Open Data practices to enhance the accessibility and discoverability of their exhibits. Utilizing semantic technologies, museums can link information about individual artifacts, including artist details, creation dates, and thematic categories. The vision of interconnected web of data is to enable researchers and the public to explore relationships between diverse artifacts, uncovering hidden narratives and contributing to a more nuanced understanding of cultural heritage.

Museums today grapple with a significant challenge in managing their collections, particularly concerning the ethical and legal complexities of determining

rightful ownership when essential provenance information is lacking. The global discourse on colonial-era acquisitions highlights the difficulties museums face due to limited documentation accompanying artifacts taken from their original locations. Addressing this challenge necessitates transparent and collaborative efforts across institutions, recognizing the importance of provenance research in shedding light on the history of colonialism and its implications for cultural heritage.

include definition of museum provenance research

The comprehensive investigation of the provenance of individual museum objects holds the potential to unravel intricate historical narratives, although it is acknowledged that this process can be both time-consuming and expensive. Conversely, the avenue for augmenting information enrichment on a large-scale presents itself through the application of automated entity linking across diverse datasets. This method provides a distinctive opportunity to improve the accessibility of information across institution by establishing connections to relevant information and context. Through the utilization of entity linking techniques, the potential arises to create associations among individuals, locations, and events referenced in various data sources. This, consequently, facilitates a more thorough understanding of the contextual significance or validity of certain information, thereby streamlining the research process for scholars in the cultural heritage domain.

This research aims to support provenance researchers in colonial contexts by enriching museum object databases with collectors' background information. Given the often-lost historical context of heritage objects, the study focuses on integrating collector biographies into a comprehensive knowledge base. The central challenge involves linking data across Knowledge Graphs (KGs) from multiple heritage institutions, specifically targeting lesser-known individuals. The goal is to facilitate disambiguation of the same person entity from different data sources. In this context, successful entity linking approaches will have to deal with the following characteristics: 1) Naming variations across and within institution (e.g. military databases may record designations and others do not); 2) Missing attribute values, and 3) Errors and historical variations, common in historical data-sets.

In this study, we delve into the intricacies of the entity linking and data integration challenge keeping in mind the specific data property mentioned above. Our investigation involves a comparative analysis of various entity linking approaches, ranging from basic and ad-hoc string matching techniques to more complex off-the-shelf methods previously used in other digital humanities literature. Through performance assessments on ground truth data, we identify the most suitable method among these approaches for this task. Subsequently, we employ the chosen method to integrate data from different sources, culminating in the creation of a unified knowledge graph. To gauge the efficacy of this integrated knowledge graph, we leverage competency questions derived from prior research, thereby evaluating its capacity to address the specific interests and requirements of domain-specific researchers.