Empowering Data Linking in Cultural Heritage: A Modular Approach for Usability in the Digital Humanities Domain

Anonymous authors

Institute
emails

Abstract. The incorporation of Linked Open Data practices among museum data publishers is increasingly prevalent, utilizing semantic technologies to improve the accessibility and discoverability of exhibits. Within the realm of Digital Humanities, the challenge of data linking is pivotal, especially when researchers navigate decentralized cultural heritage data sources. Existing solutions often cater to specific data properties, potentially limiting usability for a broader audience. This study addresses the data linking challenge in Digital Humanities, emphasizing characteristics such as non-standard naming variations, missing attribute values, imprecision, and errors in data values. In response, a modular command-line tool is proposed, empowering users to select algorithms for entity and data linking based on dataset-specific characteristics. The tool offers user control, enabling exploration, method selection, and validity investigation. Through comparative analysis, strengths and weaknesses of existing algorithms are evaluated, highlighting the tool's user-friendliness and adaptability for supporting data linking tasks in the digital humanities domain.

Keywords: Linked Open Data · cultural heritage · Digital Humanities · data linking · entity linking · user empowerment

1 Introduction

The proliferation of Semantic Web and Linked Open Data (LOD) has significantly impacted the cultural heritage domain, particularly in museum data publishing. This adoption facilitates a standardized approach for publishing and interlinking diverse datasets, thereby enriching cultural heritage repositories. The incorporation of Linked Open Data practices by museum data publishers is increasingly prevalent, aiming to enhance the accessibility and discoverability of exhibits through the utilization of semantic technologies. This interconnected web of data enables the exploration of relationships between individual artifacts, fostering a nuanced understanding of cultural heritage.

Within Digital Humanities, data linking emerges as a central challenge, particularly when researchers seek insights from decentralized cultural heritage data

sources. The amalgamation of datasets is crucial for addressing research inquiries, yet existing solutions often focus on specific data properties, potentially hindering usability for broader audience/users. This work investigates the data linking challenge in a Digital Humanities use case, emphasizing specific data properties, including non-standard naming variations, missing attribute values, imprecision, and errors in data values.

Reflecting on the complexities of data linking in our scenario, we contend that there is no one-size-fits-all solution in the digital humanities data linking. As a response, we propose a modular command-line tool that empowers users to choose algorithms for entity linking and data linking based on their dataset's specific characteristics. The tool aims to provide users with control, allowing them to explore data, select appropriate methods, and investigate the validity of chosen approaches. Through a comparative analysis, we evaluate the strengths and weaknesses of existing algorithms in the context of our specified data properties. The proposed tool is designed to be user-friendly and customizable, enhancing its usability in supporting data linking tasks within the digital humanities domain.

2 Related Work Not relavent for now

Explain how most of the paper only covers majority of the scenario, but in digital humanities we are interested for entity that do not have much for us.

Establishing identity links among RDF datasets stands as a pivotal and formidable task to achieving the goals of the Data Web initiative. It is a commonly recognized fact that information originating from diverse sources can exhibit significant dissimilarity. In many instances, two entities that refer to the same real-world entity are frequently described, structure, and mentioned in distinct manners, or in a manner that mutually complements each other [1]. The activity of aligning instances within Knowledge Bases (KBs) which correspond to identical real-world entities (e.g., an individual appearing in two distinct KBs) is termed as Instance Matching (IM) [2].

Commonly IM task is challenged by data heterogeneity, namely multilinguality, data format variation and poor data quality (inconsistency, incompleteness and incorrectness) address by [6], as we recognise it in our case-study. There are wide range of the IM techniques, applied for the RDF KBs, available in literature. However, no algorithm seems to target all these properties of data, rather made-available solutions that are suit for the purpose at hand. Moreover, the solutions available for IM makes assumptions about data properties that may not hold true in many situation. For example, schema alignment or the knowledge of ontology is required for Legato [1], LIMES [10], SILK [9], and Lenticular Lens [8]. On the other hands, the success of IM algorithm provided by [11] and [3] depends on high accuracy of data value.

Common features of historical data include large volume and diverse form of key attribute, adding complexity to the processing [4]. Most of Digital Humanities literature handles instance matching based on only literal value of the key attribute [11,5,7]. While others uses domain knowledge to construct rules for