Heritage Connector: A Machine Learning Framework for Building Linked Open Data from Museum Collections

Kalyan Dutia[1] and John Stack[1]

[1]Science Museum Group

January 6, 2021

Abstract

As with almost all data, museum collection catalogues are largely unstructured, variable in consistency and overwhelmingly composed of thin records. The form of these catalogues means that the potential for new forms of research, access and scholarly enquiry that range across multiple collections and related datasets remains dormant. In the project Heritage Connector: Transforming text into data to extract meaning and make connections, we are applying a battery of digital techniques to connect similar, identical and related items within and across collections and other publications. In this paper we describe a framework to create a Linked Open Data knowledge graph (KG) from digital museum catalogues, connect entities within this graph to Wikidata, and create new connections in this graph from text. We focus on the use of machine learning to create these links at scale with a small amount of labelled data, on a mid-range laptop or a small cloud virtual machine. We publish open-source software providing tools to perform the tasks of KG creation, entity matching and named entity recognition under these constraints.
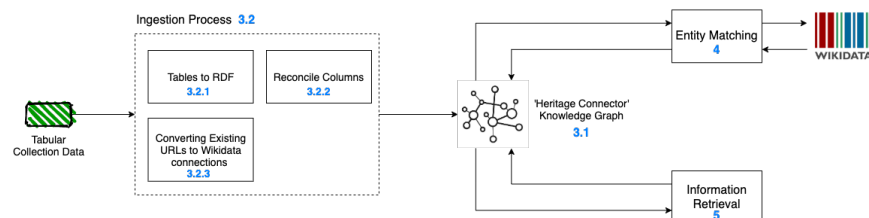
Hosted file

`Heritage_Connector_draft_3.pdf` available at https://authorea.com/users/387788/articles/502720-heritage-connector-a-machine-learning-framework-for-building-linked-open-data-from-museum-collections

Hosted file

`Heritage_Connector_draft_3.tex` available at https://authorea.com/users/387788/articles/502720-heritage-connector-a-machine-learning-framework-for-building-linked-open-data-from-museum-collections
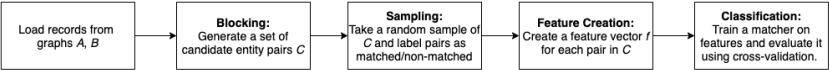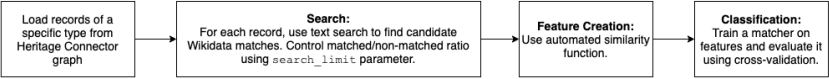
Hosted file

`wileyNJD-AMA.bib` available at https://authorea.com/users/387788/articles/502720-heritage-connector-a-machine-learning-framework-for-building-linked-open-data-from-museum-collections

Standard approach

```
┌─────────────────┐   ┌─────────────────────┐   ┌──────────────────────┐   ┌────────────────────┐   ┌────────────────────┐
│ Load records from│──▶│ Blocking:           │──▶│ Sampling:            │──▶│ Feature Creation:  │──▶│ Classification:    │
│ graphs A, B      │   │ Generate a set of   │   │ Take a random sample │   │ Create a feature   │   │ Train a matcher on │
│                  │   │ candidate entity    │   │ of C and label pairs │   │ vector f for each  │   │ features and       │
│                  │   │ pairs C             │   │ as matched/non-matched│  │ pair in C          │   │ evaluate it using  │
│                  │   │                     │   │                      │   │                    │   │ cross-validation.  │
└─────────────────┘   └─────────────────────┘   └──────────────────────┘   └────────────────────┘   └────────────────────┘
```

Heritage Connector approach

```
┌─────────────────┐   ┌──────────────────────────────────┐   ┌────────────────────┐   ┌────────────────────┐
│ Load records of a│──▶│ Search:                          │──▶│ Feature Creation:  │──▶│ Classification:    │
│ specific type    │   │ For each record, use text search │   │ Use automated      │   │ Train a matcher on │
│ from Heritage    │   │ to find candidate Wikidata       │   │ similarity         │   │ features and       │
│ Connector graph  │   │ matches. Control matched/non-    │   │ function.          │   │ evaluate it using  │
│                  │   │ matched ratio using search_limit │   │                    │   │ cross-validation.  │
│                  │   │ parameter.                       │   │                    │   │                    │
└─────────────────┘   └──────────────────────────────────┘   └────────────────────┘   └────────────────────┘
```

ThesaurusMatcher without EntityFilter

[Charles Babbage PERSON] 's calculating engines are among the most celebrated icons in the prehistory of computing. His Difference Engine No. 1 was the first successful automatic calculator and remains [one ORG] of the finest examples of precision [engineering ORG] of the time. The portion shown was assembled in 1832 by Babbage's engineer, [Joseph Clement PERSON] .

ThesaurusMatcher with EntityFilter

[Charles Babbage PERSON] 's calculating engines are among the most celebrated icons in the prehistory of computing. His Difference Engine No. 1 was the first successful automatic calculator and remains one of the finest examples of precision engineering of the time. The portion shown was assembled in 1832 by Babbage's engineer, [Joseph Clement PERSON] .