Integrating Community-Generated Digital Content into the UK National Collection

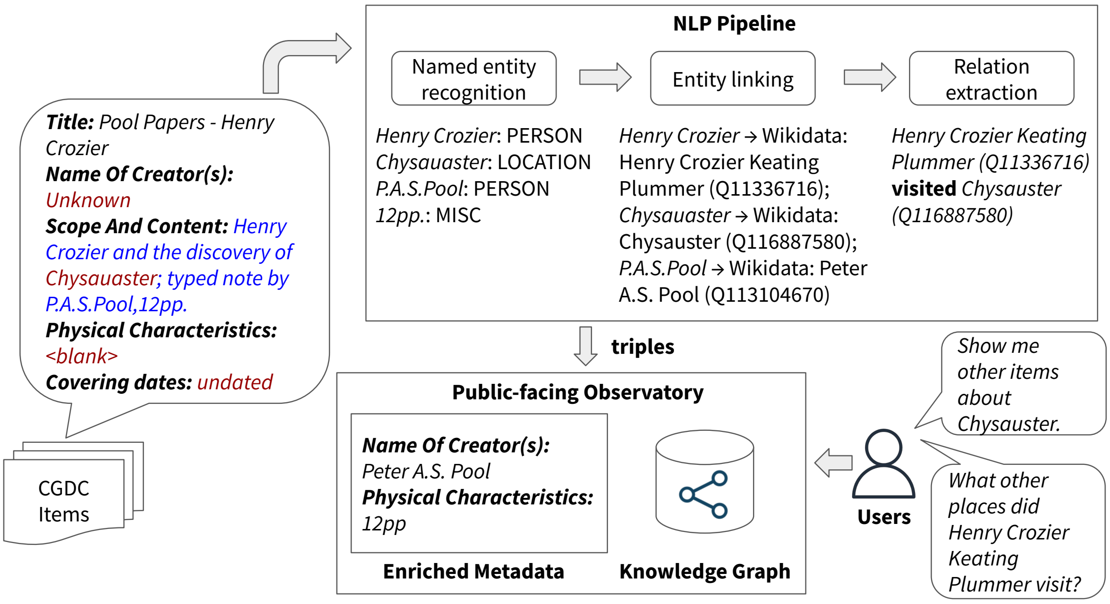
Youcef Benkhedda, Viktor Schlegel , Goran Nenadic & Riza Batista-Navarro

Department of Computer Science, University of Manchester, Manchester M13 9PL, UK
*Corresponding author E-mail: youcef.benkhedda@manchester.ac.uk

Abstract for Contributed Talk

Community-generated digital content (CGDC) pertains to data collections that have been developed by communities to create a democratic reservoir of knowledge about cultural heritage. Despite its important role in enriching people's understanding of their heritage, CGDC collections remain hard to find and severely under-represented in national collections. The Our Heritage, Our Stories (OHOS) project aims to rectify this by employing cutting-edge natural language processing (NLP) techniques to enrich CGDC metadata, making CGDC collection items discoverable in the UK national collection. To this end, we have developed an NLP pipeline for analysing raw CGDC metadata and extracting fine-grained information that helps produce enriched semantic metadata. The pipeline begins with named entity recognition (NER) to identify names of persons, locations, organisations and miscellaneous entities. The pipeline then performs entity linking to identify these entities within Wikidata, thus enabling the assignment of authority identifiers (where applicable), as well as connectivity with linked open data (LOD), including items in mainstream collections. Furthermore, relation extraction identifies meaningful relationships between the recognised entities, leading to the generation of subject-predicate-object triples. Results generated by the NLP pipeline provide information that can be used not only to complete missing CGDC metadata, but also to populate an RDF-compliant knowledge graph with entities and any relationships between them (which are otherwise left obscured in textual descriptions). Notably, each entity in the knowledge graph retains a reference to its original collection and item record, in order to provide context and facilitate traceability. The knowledge graph forms the backbone of a public-facing observatory of the UK national collection that enables users to perform searches over CGDC. By employing advanced NLP techniques and creating a knowledge graph, the project facilitates seamless integration of CGDC into the UK national collection, while providing a powerful tool for researchers and historians to uncover new perspectives on modern British cultural heritage.

Keywords: Cultural Heritage, Community-Generated Digital Content, Natural Language Processing, Named Entity Recognition, Entity Linking, Relation Extraction, Knowledge Graphs


NLP Pipeline
Named entity recognition
Entity linking
Relation extraction
Henry Crozier: PERSON
Chysauaster: LOCATION
P.A.S.Pool: PERSON
12pp.: MISC
Henry Crozier → Wikidata: Henry Crozier Keating Plummer (Q11336716);
Chysauaster → Wikidata: Chysauster (Q116887580);
P.A.S.Pool → Wikidata: Peter A.S. Pool (Q113104670)
Henry Crozier Keating Plummer (Q11336716) visited Chysauster (Q116887580)
Title: Pool Papers - Henry Crozier
Name Of Creator(s): Unknown
Scope And Content: Henry Crozier and the discovery of Chysauaster; typed note by P.A.S.Pool,12pp.
Physical Characteristics: <blank>
Covering dates: undated
CGDC Items
triples
Public-facing Observatory
Name Of Creator(s): Peter A.S. Pool
Physical Characteristics: 12pp
Enriched Metadata
Knowledge Graph
Users
Show me other items about Chysauster.
What other places did Henry Crozier Keating Plummer visit?

A natural language processing (NLP) pipeline that analyses raw metadata in community-generated digital content (CGDC) to provide enriched metadata and populate a knowledge graph.