TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

INTERIM REPORT
FOUNDATION PROJECTS

HERITAGE CONNECTOR
Transforming text into data to extract meaning and make connections

PI: John Stack, Science Museum Group

Science Museum Group | Victoria and Albert Museum |
School of Advanced Study, University of London

DECEMBER 2020

TABLE OF CONTENTS

Executive Summary

The aims of the *Heritage Connector* project are to make a substantial contribution to enable realisation of the ambitions within the AHRC's Towards a National Collection (TaNC) programme to make collections accessible for research and public engagement purposes.

The project is exploring three technologies that together have the potential to provide a stepchange in access and discoverability, research and public engagement by augmenting traditional catalogue data and associated keyword search through generation of a vast number of interlinked resources and content. The three technologies *Heritage Connector* explores are:

- artificial intelligence – specifically machine learning – to build links at scale from thin collection records,
- linked open data (LOD) as a scalable and flexible structuring methodology,
- knowledge graphs to store links and make them accessible.

Heritage Connector uses these technologies together to provide proof of concept for the application of artificial intelligence entity-linking approaches to making connections between online representations of heritage objects, including catalogue records, images and other assets; and to share this new and developing understanding with stakeholders and all interested parties. *Heritage Connector* seeks to demonstrate that creation of large numbers of machine-generated links is a valuable approach. Even though the accuracy of these links may fall short of those generated manually, a greater wealth of associated material may be suggested which will have practical benefits.

The project has four phases:

1. project set up and literature review;
2. initial software development with Science Museum Group (SMG) collection and Wikidata;
3. extended software development with Victoria and Albert Museum (V&A) collection and other data sources;
4. evaluation, write-up and dissemination.

Phase 1 is largely complete. The Project Coordinator is in post. An extensive literature review has been undertaken and continues; documentation of the technical approach and emerging thinking on its potential is being published on a project blog which has received over 2,800 page views. The first convening was reconfigured as a webinar in June 2020 and attracted 296 attendees.

The project is currently completing phase 2. The Research Developer is in post and there is good progress on software development. The development has been broken into three main components. Work on the first two is largely complete (processing existing links where those exist in the collection catalogue and matching entities based on the catalogue structure), and work on the third (matching entities in unstructured content such as narrative and interpretative text) continues. It is in this third area where the greatest potential lies as this approach can also be applied beyond the collection catalogue to secondary material.

Early findings have shown that the project's approach is likely to deliver the envisioned benefits. It is already highlighting areas of complexity and the technical challenge of applying artificial intelligence

techniques to cultural heritage collection catalogues. These challenges are being worked through and the options for resolving them considered; approaches taken are being documented. Publication of open-source software developed has begun.

There has been significant interest in the project within the cultural heritage sector and from HEIs, and as a result, we have added two additional outputs: a workshop at the *Linked Pasts 6* conference at the British Library (December 2020), and a paper for *Applied AI Letters* journal (submitted December 2020).

Abstract

As with almost all data, museum collection catalogues are largely unstructured, variable in consistency and overwhelmingly composed of thin records. This is largely a legacy of the development of these catalogues from handwritten paper records. The form of the catalogues means that the potential for new forms of digital research, access and scholarly enquiry remain dormant, and searching across collections is currently possible only through aggregation, which is labour-intensive to implement, or by third-party search engines where results are unreliable. In this project, we will apply artificial intelligence techniques to connect similar, identical and related items within and across collections. Our primary research question is "How can existing digital tools and methods be used to build relationships at scale between poorly and inconsistently catalogued digitised collection objects and other content sources?"

The *Heritage Connector* will be a linked data knowledge graph that will enable new forms of research and exploration. Furthermore, it will explore the opportunity for computer generated links with Wikidata to provide new levels of structure and machine-readable data that can form the foundation of new types of discovery and access. The *Heritage Connector* will use a range of technologies including machine learning; named entity recognition; open data; and persistent IDs. These methods will create a large-scale data source of links, each with a confidence ranking. Computational enquiry to generate links via an application programming interface (API) will enable the creation of a range of proof-of-concept research and discovery tools.

Aims and Objectives

The project aims are:

- To make a substantial contribution to enable realisation of the ambitions within the UKRI Research Infrastructure Roadmap to make collections accessible for research purposes.
- To provide proof of concept for the application of entity-linking approaches to making connections between online representations of heritage objects, including catalogue records, images and other assets.
- To share this new and developing understanding with stakeholders and all interested parties.

Objectives:

- To conduct a review of relevant literature and digital tools.
- To construct in software a 'Heritage Connector Engine' capable of holding a dense web of hypertext links between object records and knowledge graphs such as Wikidata for evaluation and experimentation.
- To apply a series of digital tools / computational methods to create speculative identifications between different records within the test dataset.

- To work on successively larger and varied datasets as the project proceeds in three stages, starting with SMG records, then adding those of the V&A.
- At every stage to hold focussed workshops for groups relevant to each of the project's four phases.
- To publish the resulting code as open source software.
- To write a report to AHRC on our findings congruent with the demands of the TaNC Programme Director.
- To write two articles, one for the readership of the *Science Museum Group Journal*, the other for *Digital Humanities Quarterly.*

Partnership Structure

The project is a collaboration between the Science Museum Group (SMG); Victoria and Albert Museum (V&A); and School of Advanced Study, University of London (SAS). Wikimedia UK has also joined the project's working group.

The SMG is made up of Science Museum, London; National Science and Media Museum, Bradford; Science and Industry Museum, Manchester; National Railway Museum, York; and Locomotion, Shildon. SMG is responsible for undertaking project leadership; project coordination, administration and project meetings; event hosting and organisation, project website, blog and YouTube channel hosting; and GitHub code repository management. SMG is leading on designing the technical architecture, development of the project software, and documentation of these components. In parallel with the *Heritage connector* project, SMG is undertaking a partnership with Wikimedia UK (Wikipedia's UK body) and hosting a Wikimedian in Residence and that partnership's activities are being aligned with *Heritage Connector.*

The V&A is taking part in project meetings (weekly and monthly) to input into project development and from early 2021, will provide the second collection dataset (after SMG's) to be analysed, processed and ingested into the *Heritage Connector* knowledge graph.

SAS is leading on the literature review, aligning work with the wider digital humanities field and research interests, and management of the project's Zotero literature library of publications, presentations and case studies. SAS attends the monthly project meetings.

Although not funded through the project, Wikimedia UK are joining the monthly project team meetings to provide guidance on their Wikidata project and identify opportunities to join up with other Wikidata projects in the GLAM (galleries, libraries, archives and museums) sector.

Staffing Structure

John Stack (Digital Director, Science Museum Group) is Principal investigator and responsible for overall delivery of the project and management of the Project Coordinator.

Jamie Unwin (Technical Architect: Collections Online, Science Museum Group) is Co-Investigator and is overseeing software development and approach and is manager of the Research Developer.

Professor Jane Winters (Professor of Digital Humanities & Pro-Dean for Libraries, School of Advanced Study, University of London) is Co-Investigator and is leading on literature review and will author one of the project papers.

Kalyan Dutia (Research Developer) is leading on software development and technical implementation.

Rhiannon Lewis (Project Coordinator and Doctoral research student, School of Advanced Study, University of London) is undertaking event logistics, blog authoring, project enquiries, meeting organisation and representation of the interests of digital humanities researchers.

Angela Wolff (Full Stack Developer, Digital Media, V&A) is leading on the addition of the V&A's collection catalogue data and inputting into the technical approach to enable this.

Richard Palmer (Senior Web Developer, Digital Media, V&A) is providing technical input from the V&A and overseeing V&A collection data ingestion into *Heritage Connector.*

Stuart Prior (Project Coordinator, Wikimedia UK) is providing guidance on the use of Wikidata and introductions to Wikidata specialists and related projects.

Hope Miyoba (Wikimedian in Residence, Science Museum Group) is leading a parallel project at SMG working with Wikipedia and Wikimedia Commons and is providing the link between that project and *Heritage Connector.*

Covid-19 Impacts

The Covid-19 pandemic has had only a limited impact on the project. No project staff were furloughed. The Digital Department was very stretched for a period as numerous activities moved online through 2020 but were able to ring-fence project resources.

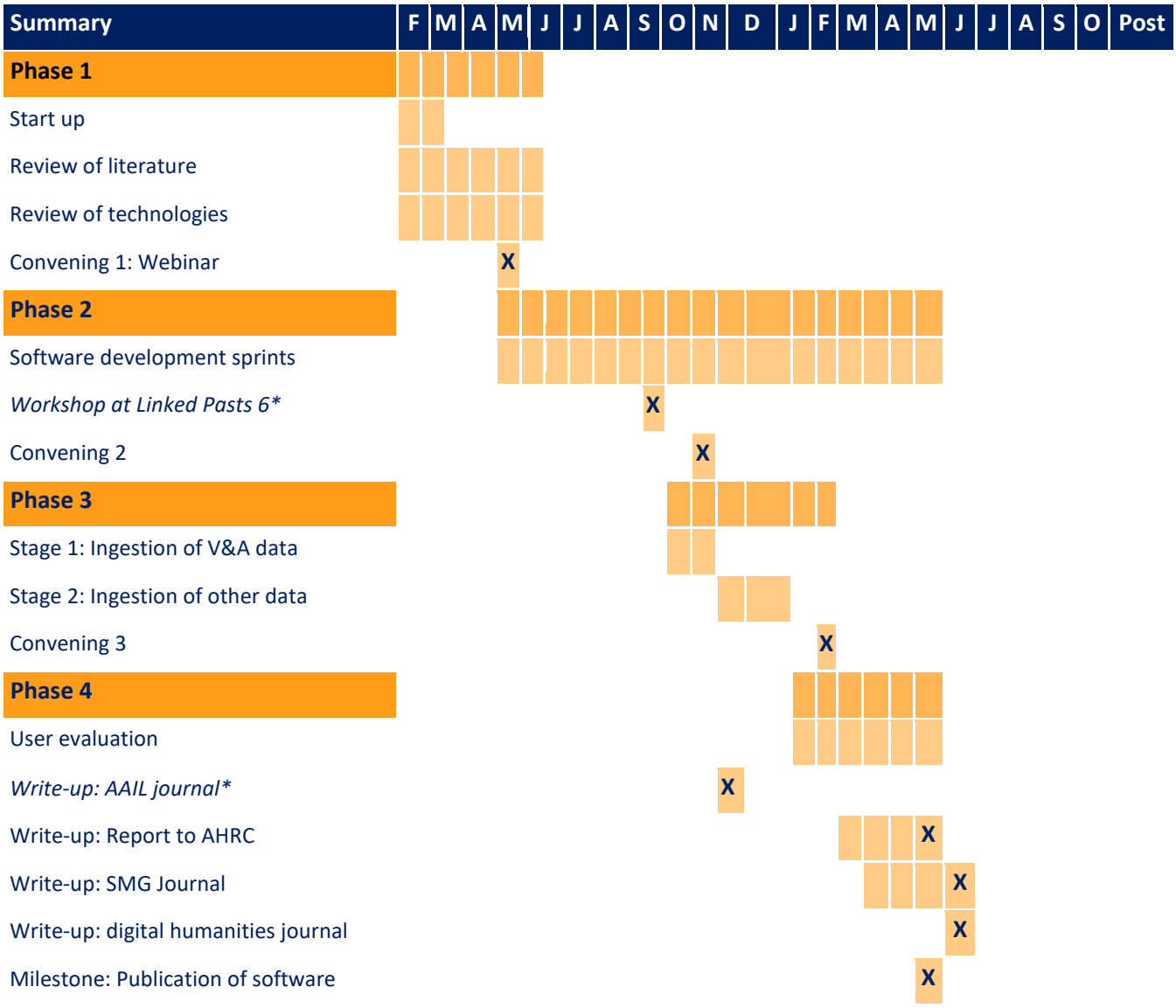
Project meetings were moved online using Microsoft Teams and a variety of other collaboration tools were implemented to facilitate ongoing working.

The project's first convening was transformed into a webinar using Zoom which enabled a larger number and greater range of speakers including one from overseas. The number of attendees of that event was also considerably larger than would have attended an in-person event at the Science Museum as had been previously envisioned.

The project's second convening which was always envisioned as a more practical workshop has been delayed into early 2021 so it can be reshaped and consideration given to an appropriate format and attendees. This will not delay the overall project timeline.

Revised Overall Programme

** New outputs added to project*



Events and consultations

Event	Date	Attendees	Outputs
Convening 1: Webinar: Wikidata and Cultural Heritage Collections	19 June 2020	296 unique webinar attendees 1,154 page views of post-webinar blog 231 responses to event survey	Blog post https://thesciencemuseum.github.io/heritageconnector/events/2020/06/22/wikidata-and-cultural-heritage-collections-webinar/ Event recordings https://www.youtube.com/playlist?list=PLkspUmkLiUBvmtIDvuGICV_yIKAjwiIZQ Survey of attendees

			https://thesciencemuseum.github.io/heritageconnector/post_files/Heritage_Connector_Webinar_1_Mentimeter.pdf
Extra event added to project: workshop at *Linked Pasts 6* conference, British Library	7–16 December 2020	39 participants	https://github.com/LinkedPasts/LaNC-workshop/tree/main/heritageconnector
Convening 2	Delayed to February–March 2021 as needs reconfiguring as virtual event		
Convening 3 and hackathon	June 2021		

Research Approach

Heritage Connector uses artificial intelligence techniques on existing collection catalogues (SMG's and V&A's) and other datasets (Wikidata and text-based content) to build an open linked data knowledge graph holding a large volume of linked open data.

Knowledge graphs were popularised by Google in their 2012 blog post *Introducing the Knowledge Graph: things, not strings*, in which they described how data structured in a graph (rather than a table which is how museum collection catalogues are stored) can help users receive better responses to their search queries, retrieve context around a specific object, and even discover new serendipitous connections. These outputs can be achieved using knowledge graphs as they can create new links (relations) between items (entities) without worrying about creating pre-defined and specific database tables to hold such information.

The value in this approach is that unlike current approaches to collection catalogue data, a knowledge graph can: hold more diverse kinds of relationships; string together relationships together; infer relationships; change and grow as required; and easily cluster related content in different ways. *Heritage Connector* also seeks to demonstrate how such approaches might sit alongside existing methods for collection digital access. Among the areas being explored by the *Heritage Connector* are ways in which this approach can:

- improve browsing interfaces where those are currently limited by the available collection catalogue data;
- enhance search by keyword where that is currently limited and where presentation of results in an ordered list is problematic;
- extract information (datapoints) from the free-text description and interpretation fields;
- create thematic and topic entry points for users of the collection(s);

- enable-cross collection linking and discovery so users can rapidly explore larger and more diverse volumes of material;
- offer onward links to related material in other collections and related sources;
- increase usage of collections through surfacing deep content more effectively;
- facilitate serendipitous discovery of material by providing a wealth of surrounding material and contextual content;
- generate links into knowledge graphs and third-party datasets that surface new data and facilitate new forms of cross-disciplinary research;
- explore playful and experimental approaches to collection access which will broaden audiences.

The *Heritage Connector* knowledge graph is being linked to Wikidata. Wikidata is the free, open, linked, multilingual and structured database which underpins Wikipedia but which is also a project in its own right. Today, Wikidata contains over 91 million items, structured as linked data and includes references to numerous external data points in cultural heritage collections and to other data sources. Because of its size and origins, Wikidata covers a vast range of subject domains and extends far beyond the areas traditionally covered by museum collection catalogues. Once built and linked to Wikidata, the knowledge graph will enable new forms of exploration, discovery and research. The full affordances of linking the collection catalogues to Wikidata remain to be explored by the project, but emerging opportunities that are shaping our approach include:

- extending the *Heritage Connector* knowledge graph to include "facts" not in the collection catalogue and presenting them to users;
- using Wikidata as a "Rosetta Stone" to provide onward links to other museum collections and data sources;
- ingesting data from Wikidata – or from other sources via Wikidata – into an index to improve discovery of collection objects, people, companies, organisations, etc.;
- using Wikidata points in the knowledge graph to infer and present new entry points into the collections such as themes, events and topics;
- Using Wikidata as a route to Wikipedia entries for articles associated with collections.

In later phases of the project, a number of prototype features and proof-of-concept applications will be built and published to evaluate the various features enabled by the *Heritage Connector* both on the SMG's public website but also as part of a final convening and hackathon event.

Early research results/outputs

Phase 1

In Phase 1 of the project a review of relevant literature was conducted, with the aim of distilling common themes, challenges and opportunities. The review encompassed journal articles, conference proceedings, working papers and reports, blog posts and conference presentations from a number of fields, including cultural heritage studies, computer science, digital humanities, and library and information science. The literature identified during the review, as well as other case studies and material relevant to the project, are collated in a public Zotero library which now includes 204 items.

Phase 1 also included setting up a project blog and YouTube channel to document the ongoing research. Six blogs have been posted which have received 2,860 page views. A set of GitHub code repositories have been set up to share the software developed. Where these have so far been made public, these are using the open source MIT Licence.

Phase 1 concluded with a webinar attended by 296 delegates of whom 231 contributed to an online survey. The recordings of this webinar and the results of the survey are published on the project blog.

Phase 2

In phase 2 of *Heritage Connector,* new links are being created between collection items and collections at scale by making use of existing metadata and mining structured data from text, as well as using Wikidata's linked open data knowledge graph.

The *Heritage Connector* knowledge graph is a unified place where all connections (from 13 data sources and counting) between the internal tables that make up our digital collection records, and where we can add new entities and relations as we discover them. The specific flavour of knowledge graph we're using is called a triple store, because every relationship is expressed in three parts: a *subject*, a *predicate* and an *object*. This representation has been chosen because any fact can be represented as a triple at its most granular level.

Our system consists of software modules for ingestion as well as two components that enable the creation of new connections in this knowledge graph and to Wikidata (See Fig. 1). Since the project began, we've built and are evaluating three main methods to create links internally and to Wikidata:

1. converting existing URLs and IDs in the collection to Wikidata IDs;
2. machine learning for creating new links between the SMG collection and Wikidata;
3. named entity recognition (NER) for creating internal links: adding new entities to the graph from free text fields.

It's important to note at this point that we're not aiming to link each and every collection record to Wikidata; that would be impossible, as many SMG people, organisations and objects will not have Wikidata referents. Instead, our aim is to use information from Wikidata through the creation of links where possible and focus on creating structure in *the Heritage Connector* that we can contribute back to Wikidata at a later date.

1. Converting existing URLs and IDs in the collection to Wikidata IDs

Our analysis showed that around 10,000 records in the SMG collection have IDs or URLs from databases like *Oxford Dictionary of National Biography*, *Grace's Guide* or Wikipedia added to them in the museum's collection management system (Mimsy). Wikidata holds references to external databases using External IDs, making it reasonably straightforward to convert these external references to Wikidata IDs.

We've found that there's one more step to ensure these connections are accurate. As well as an ID that relates to the record, curators regularly add IDs or URLs to *related* entities in the Notes field of a collection item, e.g. someone's father or company. To ensure that each link is accurate, we've added an extra step to the process which checks that the labels and significant dates (e.g. birth or death date) are similar for both the SMG and Wikidata entity if they are both present.

2. Machine learning for creating new links between the SMG collection and Wikidata

We have created a disambiguator tool which learns to distinguish whether an internal (SMG) record and a Wikidata record refer to the same real-world entity. Using machine learning this enables the creation of links between internal SMG records and Wikidata records faster than any human could, with a measurable accuracy.

There are several deep-learning-based methods which use tools including knowledge graphs and word embeddings to perform this disambiguation. But they tend to require large numbers of data to perform well (in the order of thousands), and knowledge graph embeddings require a graph representation which captures some 'structure' of the information (which we didn't have, due to record thinness). For this reason, we chose to use a classical machine learning method, which requires much less data and computational power.

3. Named entity recognition (NER) for creating internal links: adding new entities to the graph from free text fields

Named Entity Recognition (NER) is a natural language processing (NLP) technique which aims to find words (entities) in a piece of text of a number of predetermined types. It does this by learning both what these entities of a type tend to look like and where they tend to appear in a sentence.

By training an NER model on types of entities that we'd want to add to the *Heritage Connector* – for example people, places, historical events or movements – we have created a method to add new typed entities to the knowledge graph from unstructured text.
Robustly extracting relations between entities is harder: existing methods for relation extraction haven't proved reliable in tests on our data. This is mainly due to their wide-domain application. Later in the project we plan to research methods for reliably extracting relations from text and aligning them with predicates for application in the heritage domain.

We've just finished building mechanisms to load in tabular data to a knowledge graph, load links from external IDs, and the disambiguator. We've tested the disambiguator on people, organisations and objects.

At this stage of the project, our data has the properties described in the table below, where 'average no. predicates' is the number of categorical fields on average that each record type has, and 'average no. triples' is the number of values for categorical fields that each record type has (some fields, such as a person's occupation, can have multiple values). We expect to see the values on the last two rows increase in the NER phase of the project as we look to create more links between records in the collection.

	People	Objects	Organisations
Total number	10,352	282,259	7,743
Number with Wikidata link	5,343	551	1,692
Percentage with Wikidata link	51.61%	0.2%	21.85%
Average no. predicates	9.44	4.76	5.62
Average no. triples	11.15	13.47	7.70

Early findings

Headline survey results of the project's June 2020 webinar:

- 26.7% of respondents were using Wikidata IDs and other IDs with their collections, 4.8% were using Wikidata IDs only, 24.0% were using other IDs but not Wikidata, and 44.5% were using no external IDs.
- 59% of respondents from cultural heritage institutions said that a major hurdle to them adopting Wikidata IDs in their collection was time, resources, or the large amount of work required.

Findings from literature review:

- Motivations for GLAM institutions working with Linked Open Data include: a concert to make cultural heritage more visible; an interest in exposing 'hidden' collections, or 'hidden' aspects of relatively well-known collections; the enrichment of catalogues and metadata; the encouragement of data reuse in new contexts; the desire to create a better user experience; the challenges of dealing with large volumes of data when resources are scarce.
- Many projects involve only one or at most two institutions, and international collaboration is relatively rare.
- Cultural heritage databases are rich, large and complex, and there is limited standardisation.
- Institutional histories and cultures can make standardisation challenging.
- Barriers to Linked Open Data (LOD) in the cultural heritage sector fall under four broad headings: technical, conceptual, legal and financial.
- Working with LOD at any kind of scale is both time consuming and resource intensive.
- A great deal of LOD work to date has focused on people rather than objects.
- It is not a question of *if* human intervention and curation is needed, but at what point in the pipeline it should be introduced and how it may be most usefully focused.
- Many LOD projects envisage personalisation as an important outcome, but this remains a mid- to long-term goal.
- Quality, authority and trust are crucial for cultural heritage organisations, but these can hold back experimentation and present a challenge for scalability.
- It is rare for promising experimental projects to move beyond the prototype stage.

Findings from software development:

- Aligning specific free-text fields to entities (in our case collection item types and locations) is important but can take a significant time using existing tools such as OpenRefine. Faster and more robust methods therefore exist in the *Heritage Connector.*
- You can expect varying success disambiguating records with Wikidata depending on their type, due to the nature of the records in Wikidata. We've had most success with people and organisations and expect that we'll be able to find Wikidata links for a much smaller proportion of objects as they are less likely to exist on Wikidata.
- The separately described steps of creating external and internal links work better when used iteratively rather than when they're treated as two separate 'run-once' processes. As you use NER to create more entities and relations in the graph, the effectiveness of the disambiguator will increase.
- Where possible, it's best not to bulk query Wikidata, especially through SPARQL. We've circumvented this by creating an Elasticsearch index we can use to perform text searches on Wikidata in a faster and more stable way. We've open-sourced the tool we used to do this as *elastic-wikidata*.

Next steps

Areas for consideration in forthcoming blog posts are:

- What are the end user contexts in which *Heritage Connector*'s data might be used? Are there interfaces and applications emerging as possibilities from it?
- How can "fuzziness" and uncertainty of machine-generated links/content be presented?
- How is it best to define the edges of who and what is not included in the catalogue, in other datasets and in the various software toolkits being used?
- Are already marginalised/under-represented groups and histories in collections moved further to the margins through existing or historical data usage?
- What are the relevant ethical issues in machine learning generally, and how do these manifest when programming to learn specifically from museum collection catalogue data?
- How is it best to be transparent to users about how links were created and consider how this transparency might speak to different audiences?
- How might users feedback on the usefulness of the outputs and how might we potentially incorporate this into the output interfaces?

Remaining software development:

- Addition of V&A collections data.
- Extend data ingest beyond collections and Wikidata to include text content such as articles and blog posts.
- Exploring how knowledge graphs enable new forms of interaction and discovery in practice.
- Build and test a robust internal link creation method for heritage collections data. We'll open source any heritage-specific models that we create using the Spacy NLP library so they can be used for any NLP applications in the heritage sector.

In addition to the planned two post-project journal articles, we are submitting an article to the *Applied AI Letters* journal in December 2020.

Contacts

- **John Stack (PI)**
  Digital Director, Science Museum Group
  John.Stack@ScienceMuseum.ac.uk

- **Jamie Unwin (Co-I)**
  Technical Architect: Collections Online, Science Museum Group
  Jamie.Unwin@sciencemuseum.ac.uk

- **Professor Jane Winters (Co-I)**
  Professor of Digital Humanities & Pro-Dean for Libraries,
  School of Advanced Study, University of London
  jane.winters@sas.ac.uk

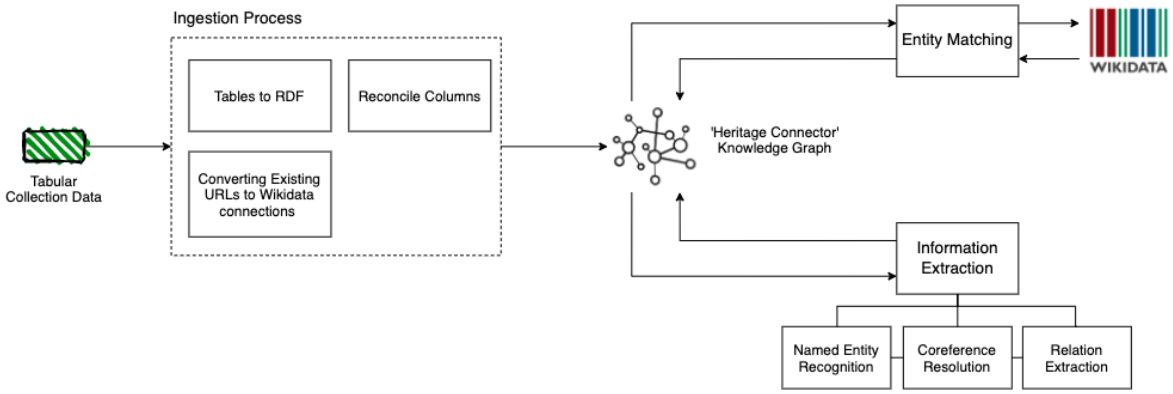
- **Kalyan Dutia**
  Research Developer, Science Museum Group
  Kalyan.Dutia@ScienceMuseum.ac.uk

- **Rhiannon Lewis**
  Project Coordinator, Science Museum Group, Doctoral research student,
  School of Advanced Study, University of London
  Rhiannon.Lewis@ScienceMuseum.ac.uk

Annexes and links

Diagrams

Fig.1: An overview of the components in the Heritage Connector



Links

Project webpage: https://www.sciencemuseumgroup.org.uk/project/heritage-connector/
Project blog: https://thesciencemuseum.github.io/heritageconnector/
Project Zotero library: https://www.zotero.org/groups/2439363/heritage_connector/library
Project YouTube channel: https://www.youtube.com/channel/UCzO6jroIvj-JbFuiQ9BpZdQ
Project Github software repositories:

- https://github.com/TheScienceMuseum/heritage-connector
- https://github.com/TheScienceMuseum/heritage-connector-nlp
- https://github.com/TheScienceMuseum/fuseki-docker
- https://github.com/TheScienceMuseum/heritage-connector-data
- https://github.com/TheScienceMuseum/elastic-wikidata
- https://github.com/LinkedPasts/LaNC-workshop/tree/main/heritageconnector