Beyond Notability Knowledge Base Documentation

December 2024

Intro

This documentation describes the production of the data in the Beyond Notability Knowledge Base, the 'completeness' of the archival sources used to produce that data, and our reflections on anticipated uses.

Package Description

The Beyond Notability Knowledge Base contains extensive linked data on women's work in archaeology, history, and heritage circa 1870-1960. The University of London manage a related data deposit that contains the following project outputs.

1. 'archival-digitisation': a directory containing two files:
    - 'beyond-notability_archival-digitisation_raw.zip': a zip containing raw archival photography with preservation metadata.
    - 'beyond-notability_archival-digitisation_packages': a zip containing packaged archival photography with item description and preservation metadata.
2. 'blogs': a directory containing all posts on the Beyond Notability Blog.
3. 'derived-data': a directory containing data derived from the Beyond Notability Knowledge Base, including a summary spreadsheet of all women in the wikibase, an alphabetical directory of all women in the wikibase, and related bibliographic data from the Archaeological Data Service and Library of Congress.
4. 'github': a directory containing all GitHub repositories (containing data, code, and documentation) published during the project.
5. 'queries': a directory containing all SPARQL queries and results from the Beyond Notability Knowledge Base.
6. 'storyteller': a directory containing briefs for Beyond Notability storytelling outputs.blogs
7. 'wikibase-backup': a directory containing one Browsertrix Crawler capture and one Wikiteam dumpgenerator capture of the Beyond Notability Knowledge Base dated December 2024.
8. 'Beyond-Notabilty_Knowledge-Base-Documentation.docx': this file which describes the production of data in the Beyond Notability Knowledge Base.

Background

Production of the Beyond Notability Knowledge Base (hereafter BNKB) began in October 2021 and ceased in December 2024. Production of the BNKB was part of the Arts and Humanities Research Council (UK) funded project *Beyond 'Notability': Re-evaluating Women's Work in Archaeology, History and Heritage in Britain, 1870 – 1950* (Project Reference AH/V01384X/1). Ontology development, archival research, and BNKB editing were co-ordinated by Katherine Harloe, Amara Thornton, and James Baker. The majority of contributions to BNKB were made by Thornton, Baker, and Ammandeep K. Mahal. Eva Alexander, Holly Russell, Madelaine Watson, and Taylor Thompson joined as student interns in Autumn 2023 to make additional contributions. Baker led on quality assurance and Sharon Howard used data led approaches to identity areas of inconsistency within BNKB and to propose

solutions. The BNKB ontology is loosely based on biographical approaches to linked data used by the Wikidata community, but with significant divergence to accommodate subject specific detail and temporal specificity. From a basic framework, the ontology developed as we encountered archival material and data was gradually entered into BNKB, which in turn shaped the ontology. Given the focus on our research, the development of our ontology focused on creating biographical profiles of individuals rather than, say, rich data on publications, institutions, and places, and where possible BNKB links outwards – using unique identifiers – to other knowledge bases and sources of structured data to enable linked queries. Although most of the data subjects in the BNKB are no longer living it was possible during the research that some would be, and it was also possible that some personal data uncovered during *Beyond Notability* may have been sensitive, or potentially sensitive, in character. The project team was aware throughout that the publication of personal information was not to be handled lightly. We therefore consulted with our partners the Society of Antiquaries of London (SAL) and the Royal Archaeological Institute about both GDPR and copyright issues in relation to publication of the data their archives contain. In particular, we agreed with the SAL and RAI that we would consult with them before publication if any issues of potential sensitivity arose, and that in any cases of doubt there would be a presumption in favour of anonymity. After this due diligence was undertaken, all content in the BNKB was and continued to be licensed under a Creative Commons Zero 1.0 Universal (CC0 1.0) Public Domain Dedication.

For a detailed elaboration of how and why the database construction took place, see Chapter 1 of Thornton and Harloe, *Beyond Notability* (forthcoming University of London Press, 2025).

Source Categories

The source material for the BNKB fell into 4 main categories:

1. Main sources: the archives of the Society of Antiquaries of London, the Royal Archaeological Institute, and the Congress of Archaeological Societies. For these sources we entered all data relevant to our work.
2. Additional sources: the archives of the Royal Historical Society and Victoria County History, as well as information pertaining to births, deaths, and marriages. For these sources we entered all data available for women identified in SAL, RAI, and CAS archives only, where those women appeared (or were identifiable) in these sources.
3. Ad hoc sources: information in first two categories augmented by ad hoc sources such as newspapers, university calendars and reports, fellows lists, and personal papers. 62 'sources' are listed in the BNKB (linked to item Q2319).
4. Deep dives: for a small number of women in the BNKB, linked data biographies were augmented by extensive archival research.

In all sources, reference statements are used to identify our sources. As a result, any users taking a query led approach to our data and therefore interested in 'complete' runs of records will be able to include in their queries only data attributed to the main sources listed above.

Notes on our approach to sources

Main sources

To extract information from these records members of the project team consulted the records in the reading room at the Society of Antiquaries of London. There we manually browsed the records, typically in chronological order, looking for names of people we might reasonable assume were women (e.g. by the use of a gendered honorific, a given name coded feminine, or a names we recognised as representing a woman working in archaeology, history, and heritage in our period). We then recorded in a spreadsheet some brief notes on why these women appeared in the records and photographed each relevant page for later reference. After visiting the archive, members of the project team would then create items, properties, and triple statements that rendered the information as linked data in the BNKB.

We consider the data extracted from these records to be **complete**: that is, all in-scope information – information relating to women's work in archaeology, history, and heritage circa 1870 to 1950 – has been recorded. This is different to all information about women being recorded. For example, we did not record in the BNKB evidence of a Society of Antiquaries of London committee agreeing to send a letter of condolence to a spouse of an archaeologist when that spouse undertook no discernible work in archaeology, history, or heritage.

In all three cases (Society of Antiquaries of London, the Royal Archaeological Institute, and the Congress of Archaeological Societies), archival resources need to be read alongside printed records (e.g. *SAL Proceedings* (and after 1920 *Antiquaries Journal*), *RAI Journal*, *CAS Reports*). In addition, one major dataset which overlaps all three record sets is the Congress of Archaeological Societies Annual Indexes of Archaeological Papers (1890-1907) and the volume that takes the list of papers back to 1665.

<u>Incompleteness of archival material</u>

The records used to produce the BNKB are incomplete for a number of reasons:

- Records contain temporal runs and gaps. For example, Congress of Archaeological Society Research Committee Records exist for the 1930s only, whilst Congress of Archaeological Society Earthworks Committee Records exist from the 1900s to the 1930s.
- The level of detail across records series can vary due to changes in how procedures were recorded. For example, in records of the Royal Archaeological Institute, for most of our period the individuals who nominated prospective fellows are not recorded.

<u>Record improvement to enable historical research</u>

Towards the end of the *Beyond Notability* project we embarked on improvements to data in the BNKB focused on facilitating questions emerging from our historical research.

One group of work involved inputting data from new sources into the BNKB. Much of this work was undertaken by Alexander, Russell, Watson, and Thompson, with Q&A by Baker and Thornton. This included adding to the BNKB information relating to: education at Cambridge and Oxford colleges; extension lectures outside of London; and archaeological excavations listed in sources including Historic England, Hampshire Field Club, and the Journal for Roman Studies;

A second group of work involved resolving – or attempting to resolve – gaps in the data that precluded longitudinal study. This work was undertaken by Thompson, Baker, and Thornton. For example, we identified roughly 200 women whose entries BNKB had 10 or more statements (e.g. places of residence) but were missing either dates of death, dates of birth, or both. Record improvement work more than halved this number, and fewer than 30 entries for women in this category contain no information about birth and death dates.

A third group of work involved ad hoc additions based on emerging interests: for example, connecting items about historic/country houses, which some individuals in the BNKB resided at, to other structured information sources about those places, including their current ownership. Most of this work was undertaken by Thornton.

Deep dives

In some cases, individuals in the BNKB and/or information about individuals has been included on an in ad-hoc via 'deep dives'' that pull together a variety of sources. The entries for these women have been added in part because of prior knowledge of their work, but in some cases because their experiences complement or overlap with work being done by other women – indeed some of the ad-hoc entries were collaborating with (or related to) women who are referenced in the three record sets listed above. In other cases, women have been included because they represent backgrounds, fields of work, or geographical areas, not otherwise recorded on the database. Adding these individual to the database enabled us to test ways to model new statements while enriching our the breadth of the BNKB. These decisions were made based on the commitment to build routes through which women's work could be represented that may otherwise of fallen through the cracks (meaning they are not represented in the archives and sources focused on by the *Beyond Notability*, but were active in the fields we are looking into during our period of focus).

Reuse of BNKB

Following principles developed by 'Datasheets for Datasets' (Gebru et al 2020), the present documentation focused on responding to the question:

> Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

The short answer is yes, for the reasons given in this document. In particular, we advise users of the BNKB to cautious if they choose to use the data in an attempt to analyse trends and patterns across the period. Whilst effort has been made to consistency check the data in the BNKB, our archival sources are incomplete meaning that our data is incomplete. For example, where possible we have recorded instances of individuals having children using 'had child in' property statements, however this statement is used for fewer than 10% of the women in the BNKB. This cannot be all the possible data, is likely a sample skewed towards those women who were more notable and whose activities are recorded across multiple sources, and is limited by a lack of publicly available information. Statistical inferences made using the BNKB will there be rarely appropriate. Our data essays (Sharon Howard and James Baker, *Beyond Notability Data Essays* (https://beyond-notability.github.io/beyond-notability-observable-essays, 2024)) explore our reflections of this kind of analysis in more detail.

As part of this data publication we have published a range of SPARQL queries (and their outputs) that facilitate navigation of the BNKB data and enable users to grasp the character of the data. However, as our ontology was developed iteratively by specialists in the field of study, we have inevitably internalised its logics. Property descriptions, History and Discussion pages for properties, and entries in the *Beyond Notability* blog (https://beyondnotability.org/blog/) capture some of these logics, but users of the BNKB may wish to contact the project team before using the data for advanced analysis.

Contact

The core project team were:

- The School of Advanced Studies, University of London
    - Katherine Harloe https://orcid.org/0000-0002-0207-5212
    - Amara Thornton https://orcid.org/0000-0001-6227-6481
- The University of Southampton
    - James Baker https://orcid.org/0000-0002-2682-6922
    - Sharon Howard https://orcid.org/0000-0002-6051-6274
    - Ammandeep K. Mahal

We have included ORCIDs in this document in an attempt to support the medium to long term ability of users to contact the project team.