Reproducible Humanities Research: Developing Extensible Databases for Recording "Messy" Categorisation, Annotation and Provenance Data

Melodee Beals, Loughborough University, UK
Albert Meroño-Peñuela, Vrije Universiteit Amsterdam, NL

Abstract

Although the Digital Humanities is fundamentally interdisciplinary in nature, all humanities research questions require a degree of interdisciplinary thinking. History, for example, draws upon most other social sciences and humanities for obtaining and analysing source materials in different contexts. The multi-modal nature of these sources, the mixing of methodologies into bespoke, project-specific frameworks and the collaboration of researchers with overlapping but distinct interpretations all require a flexible workspace. Moreover, growing calls for open research methods put pressure on humanities researchers to rethink how they document the provenance of their source materials as well as their interpretations.[1] Individual scholars often develop extensive, single-use taxonomies to categorise, encode and describe their conclusions; stored in a variety of document, spreadsheet and database systems, these are rarely disseminated and remain offline penumbra of the research process. Moreover, the prescriptive nature of out-of-the-box software may constrain the annotation process.[2] Larger collaborations may spend significant time developing extensive coding criteria resulting in over-fitted schema with little reusability or reach despite often herculean efforts of dissemination. Even when reusable, these schema may require a degree of familiarity with the bespoke systems that makes them inaccessible to those outside the project. In order to overcome these difficulties, we have developed a highly extensible database development interface, *Nisaba*.[3]

Rather than prescribe a new database structure or encoding format, *Nisaba* was developed in order to accommodate a wide variety of source materials, encoding schema and dissemination formats. To achieve this, *Nisaba* leverages World Wide Consortium (W3C) standards[4] and Linked Data publishing practices,[5] which encourage the explicit provision and reuse of vocabulary terms. Written in Python 3.6 using TKinter, a cross-platform graphical user interface (Linux, OS, Windows), *Nisaba* functions as both an input and retrieval mechanism. Users input data including text transcriptions, images and audio/visual files and apply user-created

---

[1] For more on the importance of evidence-interpretation-argumentation provenance, see D. M. Godden, "Arguing at Cross-Purposes: Discharging the Dialectical Obligations of the Coalescent Model of Argumentation", Argumentation, 17:2 (2013): 219–43.

[2] Noam Chomsky, *Language and Mind* (Cambridge: Cambridge University Press, 2006): 19.

[3] The open-source code is currently available at http://purl.org/nisaba

[4] See https://www.w3.org/TR/?tag=data

[5] Tom Heath and Christian Bizer, *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1 (Morgan & Claypool: 2011).

controlled-vocabularies, free-text annotations and an extensible selection of metadata. Once inputted, users create a segment (a selection of words, pixels or seconds of audio-visual information) and apply further metadata or annotations, allowing a single item to have multiple overlapping annotations using different schema by different users. In order to facilitate the documentation and exportation of data that is restricted or within copyright, the database encodes these segments by word number (text), or relative position (image), allowing precise locators without necessarily exporting the original materials. All data inputs are time-stamped and attached to individual user records, allowing for multiple researchers to annotate the same segments while maintaining unambiguous lines of provenance and allowing longitudinal use of the databases by multiple projects. Once inputted, the material can be retrieved through a simple browsing mechanism (controlled vocabulary) or by exporting layers of the data to non-proprietary formats, currently JSON or Turtle (RDF), allowing for deeply humanistic forms of knowledge representation in a format suitable for computational analysis.[6]

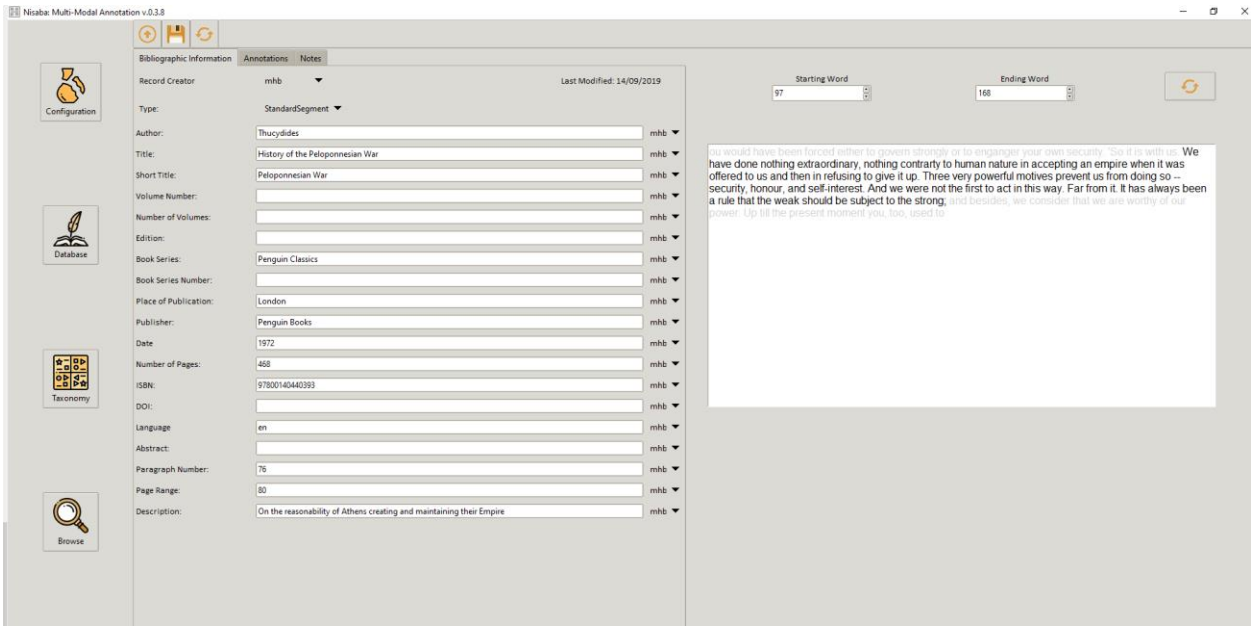

Figure 1:Text Segmentation

[6] See Dominic Oldman, Martin Doerr and Stefan Gradmann, "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge" in Schreibmann, Siemens and Unsworth, eds. *A New Companion to Digital Humanities* (Oxford: Wiley-Blackwell, 2016): 251–73; David M. Berry and Anders Fagerjord, *Digital Humanities: Knowledge and Critique in a Digital Age* (Cambridge: Polity, 2017): 77.

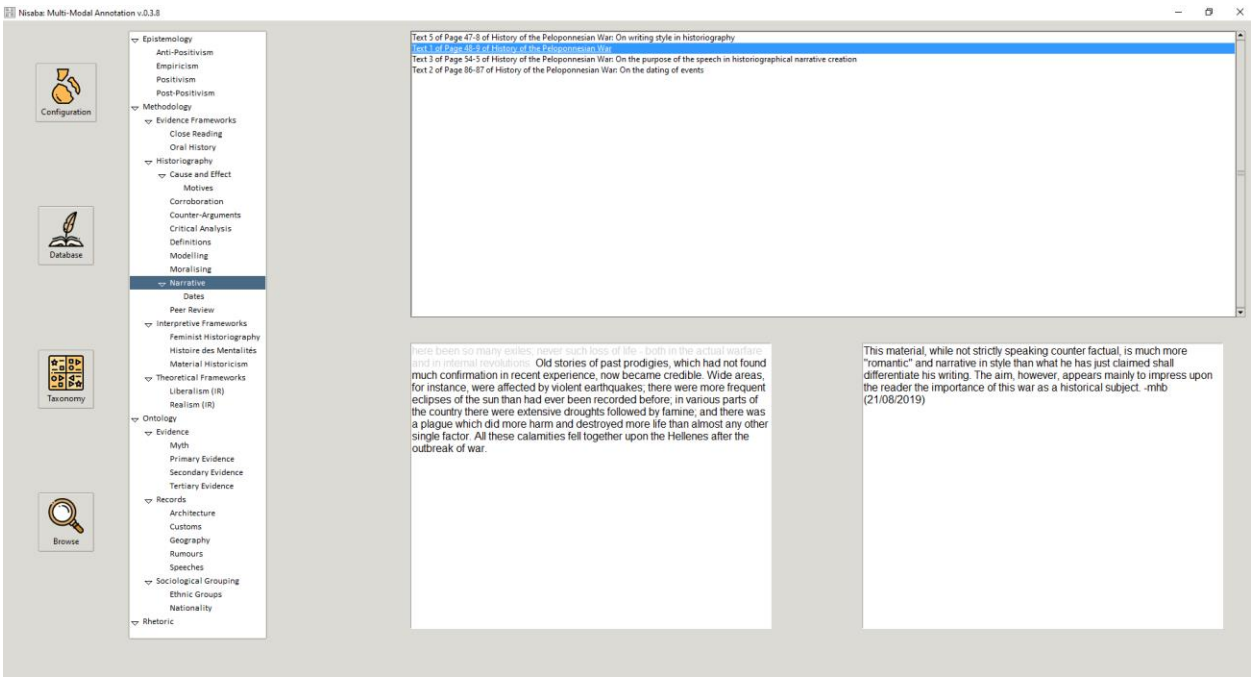

Figure 2: Browsing notes

This paper will demonstrate the use of *Nisaba* for various project types and provide guidance on how to develop an open, highly documented dataset to accompany humanities research.