# Collecting Pieces of Historical Knowledge from Documents: Introduction of HIMIKO (Historical Micro Knowledge and Ontology)

## Ogawa, Jun

htjk6513khbk@yahoo.co.jp
ROIS-DS Center for Open Data in the Humanities, Japan

## Ohmukai, Ikki

i2k@l.u-tokyo.ac.jp
Graduate School of Humanities and Sociology, The University of Tokyo

## Nakamura, Satoru

nakamura@hi.u-tokyo.ac.jp
Historiographical Institute, The University of Tokyo

## Kitamoto, Asanobu

kitamoto@nii.ac.jp
ROIS-DS Center for Open Data in the Humanities, Japan

### Historical knowledge representation and HIMOKO

Historical knowledge is made from various information extracted from various sources. This fact makes it difficult to develop a versatile data model for historical knowledge representation and must be the reason why models proposed in the field of digital history tend to be limited to specific areas, such as Factoid for prosopography [1], Knowledge Claims for Chinese literature [2], DEPCHA for accounting [3], etc. In this context, we propose HIMIKO ( **Hi**storical **Mi**cro **K**nowledge and **O**ntology) as a potential versatile model for Historical Linked Data. The ‘micro-knowledge’ is a basic data unit representing a specific statement in a document mentioning historical situations, actions, or states of affairs.

### HIMIKO Model

HIMIKO should effectively collect pieces of knowledge represented in various documents, semantically organize, and link them to universal and unique entities existing outside of the documents. The last part of the process, which is the linking between the inside and outside of the documents, has been already realized by many studies. They annotate entities in documents and link them to external counterparts by using semantic web technologies [4]. On the other hand, they have hardly organized internal semantic relations within a document, which we call ‘intra-knowledge’. It is this problem that HIMIKO works through: how specific historical phenomena were represented in a particular document, who or what is mentioned, which word or image is used to represent certain objects or concepts, or with what kind of causal or temporal relations statements are connected? Fig. 1 is the data created with HIMIKO.

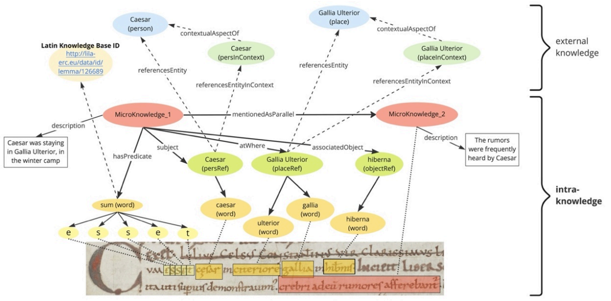
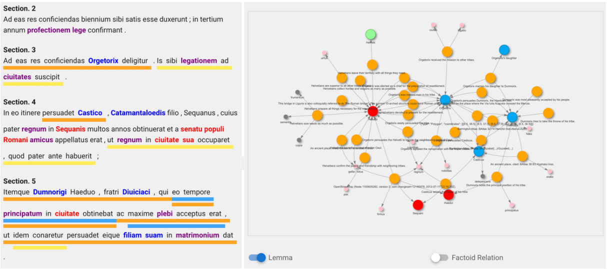


Fig. 1: Data constructed with HIMIKO model (Image of the manuscript is taken from British Library MS Viewer: http://www.bl.uk/manuscripts/Viewer.aspx?ref=add_ms_10084_f011r

The basic structure resembles Factoid, which structures the biographical data based on primary sources. While referring to it, HIMIKO achieves a more elaborate description of intra-knowledge in two points: firstly, word and character level structuring guarantees high referentiality to the primary source descriptions, and secondly, as we semantically connect and nest micro-knowledge resources, the whole semantic structure of the document can be preserved as data.

Such a document-based model enables us to deal with, not only the biographical data concerning mainly persons and places, but also any pieces of information written in sources, such as a mention of small objects or concepts, descriptions about the natural condition or spatial arrangement, or verbs used to represent historical events, and their semantic relationships. One example demonstrating the advantage of this model is the visualization in Fig. 2.



**Section. 2**
Ad eas res conficiendas biennium sibi satis esse duxerunt ; in tertium annum **profectionem lege** confirmant .

**Section. 3**
Ad eas res conficiendas **Orgetorix** deligitur . Is sibi **legationem** ad **ciuitates** suscipit .

**Section. 4**
In eo itinere persuadet **Castico** , **Catamantaloedis** filio , Sequanus , cuius pater **regnum** in **Sequanis** multos annos obtinuerat et a **senatu populi Romani amicus** appellatus erat , ut **regnum** in **ciuitate sua** occuparet , quod pater ante habuerit ;

**Section. 5**
Itemque **Dumnorigi** Haeduo , fratri **Diuiciaci** , qui eo tempore **principatum** in **ciuitate** obtinebat ac maxime **plebi** acceptus erat , ut idem conaretur persuadet eique **filiam suam** in **matrimonium** dat .

Fig. 2: Visualization of text and network based on HIMIKO data

The network contains various entities mentioned in the document, such as statements, persons, organizations, places, concepts, and objects. These elements are semantically connected, and the network as a whole shows the outline of the semantic structure of the document. Since each node on the network is linked to a specific word or sentence in the text, it is possible to search the text from the network and vice versa. Obviously, it is possible to explore the data quite in detail with SPARQL queries, for example, “What is the verb frequently used in the statements mentioning any specific person?”, or “Which concept or object is mentioned together with some specific place in same statements?” Being able to ask these questions by using linked data technologies must bring new possibilities for digital historical analysis.

### Data construction workflow of HIMIKO

To effectively construct a large amount of data, we are now developing HIMIKO Editor and setting a workflow (Fig. 3) for the entire process. HIMIKO Editor consists of two separate sub-editors: Entity  Editor is for annotating entities and accordingly editing the XML text file, while Micro Knowledge Editor is used to describe intra-knowledge in the form of RDF, linking it to external knowledge [5].

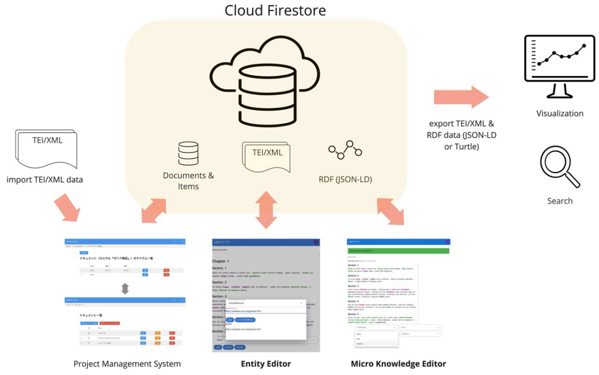


Fig. 3: Image of HIMIKO data construction workflow

We have put two sets of documents to the editor and tried to construct micro-knowledge data. One is Caesar's *De Bello Gallico*, and the other is the diary documents provided by the Shibusawa Eiichi Diary website [6]. As a result, even though these two documents greatly differ in terms of their language, genre, and style of writing, we found that the model and the editor could be equally applied to both. This proves that HIMIKO has the potential to be a versatile data model for collecting pieces of historical knowledge from various documents and connecting them to broader knowledge both within and without the documents. Furthermore, as HIMIKO provides a package containing both the model and the editing system to make the data construction process much closer to normal reading activities, it would help historians without digital skills to produce highly sophisticated, structured data without much difficulty and encourage them to participate in the research of digital history.

# Bibliography

**‘The Factoid Prosopography Ontology’**, Factoid Prosopography, https://www.kcl.ac.uk/factoid-prosopography/ontology, (accessed: November 4, 2022).

**‘Semantic annotation’**, Chinese Text Project, https://ctext.org/instructions/annotation, (accessed: November 4, 2022).

**‘Bookkeeping Ontology for Historical Accounts’**, DEPCHA: Digital Edition Publishing Cooperative for Historical Accounts, http://gams.uni-graz.at/archive/objects/o:depcha.bookkeeping/methods/sdef:Ontology/get, (accessed: November 4, 2022).

Even the editors for this purpose have been developed. The most famous example of those must be **Recogito** ( https://recogito.pelagios.org/).

We present this editor at DH Budapest 2022. https://elte-dh.hu/en/dh_budapest_2022-dariah-days-conference-program/.

**[6] Shibusawa Eiichi Diary**, https://shibusawa-dlab.github.io/app1/en, (accessed: November 4, 2022).