PaleOrdia: Semantically Describing (Cuneiform) Paleography using Paleographic Linked Open Data

Timo Homburg[1,*]

[1]*Mainz University Of Applied Sciences, Lucy-Hillebrand Straße 2, Mainz, Germany*

Abstract

This publication describes PaleOrdia, a web application developed to visualize (cuneiform) paleographic sign variants in Wikidata and the data model developed in Wikidata to represent paleography. Modeling paleographic sign variants of (ancient) scripts in linked open data is a relatively new development. It will enable better descriptions of digital scholarly editions with paleographic annotations supported by established web annotation data model vocabularies. As a use case for showcasing the capabilities of PaleOrdia, the cuneiform annotation tool Cuneur is presented as one way to harness the paleographic-linked open data for digital scholarly editions.

Keywords

PaleOrdia, Paleography, Cuneiform, Annotation, Paleographic Linked Open Data

1. Introduction

Describing the paleography of inscriptions on cultural heritage objects is a common task in many digital scholarly editions [1] projects. In a digital scholarly edition, a set of texts is commonly transcribed, annotated, and finally translated or interpreted so that the respective scholar can address the targeted research question. While the contents of the respective textual materials are likely the main focus of the scholar's work, a closer inspection of the stylistic choices made in writing the text is necessary in many disciplines. Such analysis might hint at the detection of authors of texts by writing style, the identification of particularities of writing in a specific time and space, and finally, may give hints about the circumstances in which a text has been written. To make an accurate assessment of authorship, a detailed knowledge of not only the preferred choice of words but also the shape of the characters the author uses for writing, i.e., paleographic features, is important. An automatic analysis of paleographic data needs, at best, accurately described training data, which may assist in automated analysis of the given work of text at hand. This advocates for knowledge graphs of paleographic features, which may be reused in different research contexts. This publication wants to highlight the need for paleographic-linked open data by discussing the cases of cuneiform digital scholarly editions, which will serve as the primary, but not exclusive, application case made possible by this work. Section 2 will give some background on cuneiform signs and the terminologies

First International Workshop of Semantic Digital Humanities, co-located with ESWC 2024, Hersonissos, Greece

*Corresponding author.

timo.homburg@hs-mainz.de (T. Homburg)

https://situx.github.io/ (T. Homburg)

0000-0002-9499-5840 (T. Homburg)

used in paleography, Section 3 explains how paleographic linked open data can be represented in Wikidata. Finally, the tool PaleOrdia, a tool to manage paleographic linked open data in Wikidata, is introduced in Section 4, and the usefulness of paleographic linked open data is shown in an application example in Section 5.

2. Foundations and related work on the cuneiform LOD cloud

The cuneiform script is one of the earliest writing systems used in Ancient Mesopotamia as the script for many languages such as Sumerian, Akkadian, and Hittite. Throughout its existence, for 3000 years, cuneiform signs have considerably evolved and been simplified. They usually depict a transformation from a pictograph to increasingly simplified cuneiform wedge configurations—the basis of all known cuneiform signs. Figure 1 shows the evolution of

Time Period Designation Point in time	Late Uruk ca. 3100	Djemdet Nasr ca. 3000	Early Dynastic III ca. 2400	Ur III ca. 2000	Old Assyrian ca. 1900	Old Babylonian ca. 1700	Middle Assyrian ca. 1200	Neo-Babylonian ca. 600	meaning
Glyph (U+12295)									SAG "head"
Gottstein	N/A	N/A	a3b2c1d1	a3b2c1d1		a3b4	a3b3c1	a3b4	
Glyph (U+120FB)									NINDA "ration"
PaleoCode Gottstein	N/A N/A	N/A N/A	:sa~a-:c~;<F a2c1d1	a-:sa-sa~~~;f a3d1	sa-sa;w-sa a3c1	sa-:sa;w-sa a3c1	sa-sa;w-sa a3c1	f-;>c-:sb-a a1c2d1	
Glyph (U+12165)									GU_7 "to eat"
Gottstein		N/A	a5b2c6d2	a5b2c4d2		a6b7c2d1	a5b5c1	a5b4c2	

Figure 1: Evolution of cuneiform signs over different centuries: Cuneiform signs SAG, NINDA and GU7 along with character description codes, Unicode references and time period annotations

cuneiform signs throughout space and time. For example, the cuneiform sign SAG for head starts as a pictograph of a head in the Late Uruk period and is subsequently simplified. Starting from the Early Dynastic period, cuneiform signs are comprised of cuneiform wedges to represent cuneiform signs. Cuneiform wedges as atomic components of every cuneiform sign allow for a more simple and formalized drawing of the grapheme and change in composition and positioning of the cuneiform wedges in the subsequent centuries.

2.1. Sign variant terminology

This section briefly introduces terminology that may be used to describe paleographic sign variants.

Definition 1. ***Character*** *A unit of information that often corresponds to a grapheme or a symbol.*

The definition of a character can be seen as equivalent to something that can be represented by a Unicode codepoint. In contrast, a Unicode code point stands for a variety of shapes of

the same grapheme, e.g., may be depicted by different fonts, which may visualize a Unicode codepoint to the user.

Definition 2. ***Sign*** *A cuneiform sign is a set of shapes/graphemes of cuneiform wedge configurations across time and space that have been classified under a common identifier.*

A sign identifier for a cuneiform sign may be represented in different ways. The cuneiform research community defines so-called sign names for cuneiform signs, often derived from the most common reading of the sign in the Sumerian language. Many, but not all, sign names are equivalent to a Unicode codepoint. Figure 1 shows the three sign names SAG, NINDA and GU7, which each have a corresponding Unicode codepoint in the Unicode standard. However, there are several sign names that have no Unicode codepoint attested at the time of writing. Most cuneiform signs without a Unicode codepoint are combinations of already existing cuneiform signs.

Definition 3. ***Sign Variant*** *A sign variant is a class of representations of a cuneiform sign that has been attested in time and space.*

A sign variant is a set of distinct configurations of cuneiform wedges that have been classified by a sign name and have been assigned a time period and/or location. Figure 2 shows a set

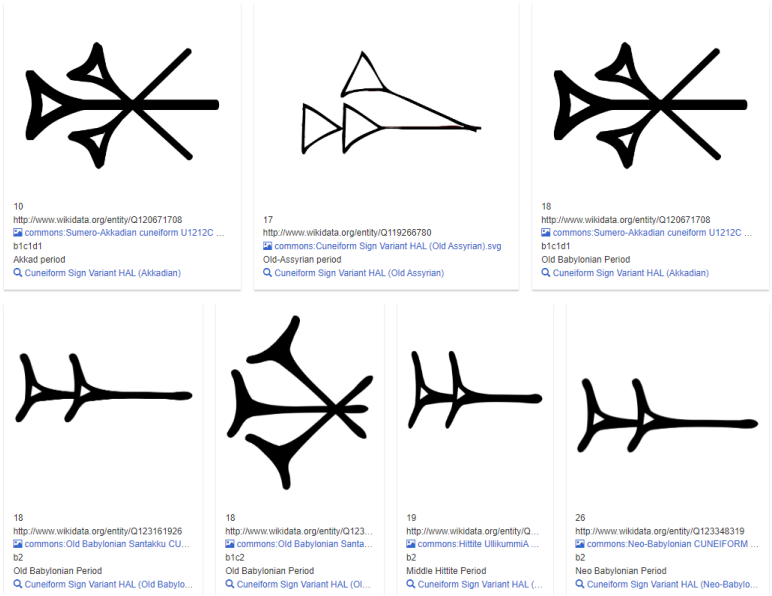

Figure 2: Sign variants of the cuneiform sign HAL (U+1212C) occurring in time and space, ordered from the oldest to the most recent time period

of cuneiform signs that differ in the number of wedges, their positioning, and the shapes of the cuneiform wedges used to build the signs. The shapes of the individual cuneiform wedges shown in Figure 2 differ slightly due to font variations and due to semantic expressions font creators might want to convey. For example, a filled wedge head of a cuneiform sign might hint at the sign variant being created on a stone surface vs. on a clay surface.

Definition 4. ***Stylistic Sign Variant*** *A stylistic sign variant is a stylistic change of a sign variant that does not constitute a change of its characteristic elements.*

However, in reality, subtle changes in writing on the cuneiform clay tablet, expressed, for example, in the length of individual strokes, the angle of individual strokes in certain cuneiform sign variants, or the pointiness of wedge heads, might reveal the style of a particular writer or even writing school. The aforementioned changes would typically be part of a stylistic sign variant that hints at a specific author of cuneiform texts.

2.2. Capturing sign variants using character encodings

Various approaches have been researched in the past to capture the essence of cuneiform sign variants. Different encodings such as [2] and [3] try to capture the number of cuneiform wedges per wedge type, and in the case of PaleoCodage, the positioning of wedges towards each other in a String encoding to create a unique identifier for each of the cuneiform sign variants in existence. These codes purposefully do not capture the elements that describe a stylistic sign variant but can be used as elements in knowledge graphs [4] capturing paleographic information. Character encodings such as this form the building blocks for identifying and classifying cuneiform signs in knowledge graphs.

2.3. Related Work

The first ideas of modeling paleography with linked open data technologies emerged in the digital humanities community in 2020. [5] proposed modeling paleographic features with linked open data vocabularies and creating a formalized vocabulary. [4] described approaches to generalize a paleographic vocabulary which can be used in conjunction with the Ontolex-Lemon [6] model to express the paleographic variants in which different words can be written. In particular, this model also introduces the connection between lexemes and paleographic descriptions and the concept of a paleographic sign variant occurrence for annotation purposes. Further approaches to capture features of inscriptions include the CIDOC CRMtex extension [7], which allows the description of characters on inscriptions of surfaces of cultural heritage objects. In general, though, digital humanities and computational linguistics approaches rarely use paleographic information for classifications for the time being.

3. Modeling paleography in Wikidata

In the previous sections, foundations for modeling paleographic-linked open data were defined. This section explains how Wikidata can be used to represent paleographic features such as cuneiform signs. Many cuneiform signs have been described in the Unicode standard, which Wikidata adopts as QIDs. Hence, a starting point for a paleographic description are the Unicode signs themselves. However, coverage of the Unicode codepoints for cuneiform signs would not be sufficient. Cuneiform signs may appear as ligatures, represented as more than one Unicode codepoint, or cuneiform signs may not occur sufficiently often to be considered for addition to the Unicode standard. These signs will be added as new items to Wikidata and referenced in

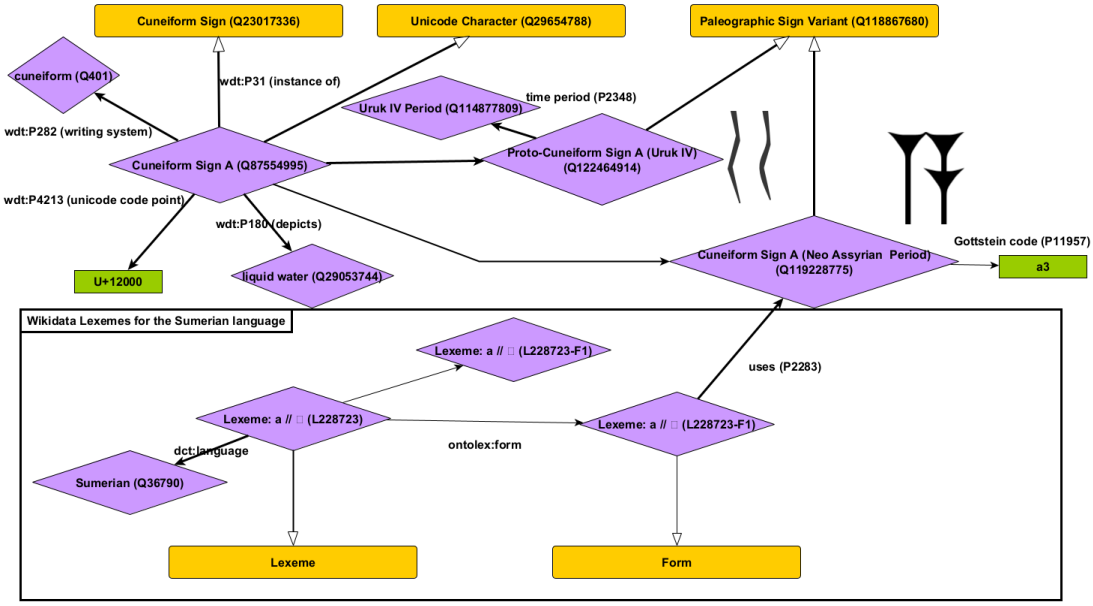

Figure 3: Paleographic data model in Wikidata: Cuneiform sign instances are linked to paleographic sign variants. Paleographic sign variants may be associated with Forms of Wikidata Lexemes to state that these forms have been attested in these particular paleographic representations

respective literature. Figure 3 shows the data model for paleographic data adopted in Wikidata. The data model relates cuneiform signs to their paleographic sign variants, which are classified by time periods and - if applicable - encodings for their description. Lexicographical data within Wikidata, such as the Sumerian word "a" (water) (L228723) may link to paleographic sign variants used within their respective forms. In this way, scholars can express not only that a Lexeme has been occurring in a specific time period, at a specific place, and in a specific text source but also in which paleographic shape the Lexeme form has been attested. As shown in fig:cuneiformsignhalcompoenents, many cuneiform signs consist of other cuneiform signs. These can be modeled using Wikidata has parts (wdt:P527) relations. Finally, the semantics of the original pictographs can be captured as their meanings in Wikidata using the depicts (wdt:P180) relation. This leads to the capture of cuneiform sign meanings not only on the level of the grapheme but also on a semantic level.

4. PaleOrdia: A tool to visualize paleographic linked open data

PaleOrdia[1] is a fork of the tool Ordia [8], which has been developed as a view on Lexeme data in Wikidata. PaleOrdia differs from the original Ordia tool in two fundamental ways:

- PaleOrdia is a static web application which runs on a Github Page, as opposed to the original Ordia, which needed to be run on a web server

[1] https://situx.github.io/paleordia/script?q=Q401&qLabel=cuneiform

- PaleOrdia combines the view of Wikidata Lexemes with Wikidata Paleography data

PaleOrdia offers the following functionalities to highlight cuneiform paleographic data:

- Listing of cuneiform sign variants by time period and reference work
- Identification of cuneiform signs which have not been included into Unicode[2]
- Listing of cuneiform signs by type (compound[3] and allograph signs[4])
- Representation of cuneiform sign etymology and compounds

4.1. Character Data / Cuneiform Signs

PaleoOrdia allows users to view information about a cuneiform sign, including the reference works and reference databases in which it is attested, its readings (phonetic values), its classification as a compound sign, allograph, or cuneiform sign, and whether a single Unicode codepoint represents it. Besides metadata such as the attestations of a cuneiform sign in sign lists and on actual cuneiform tablet texts, each PaleOrdia page for a cuneiform sign also lists the different sign variants[5] the sign has been attested with, as shown in Figure 4 Compound

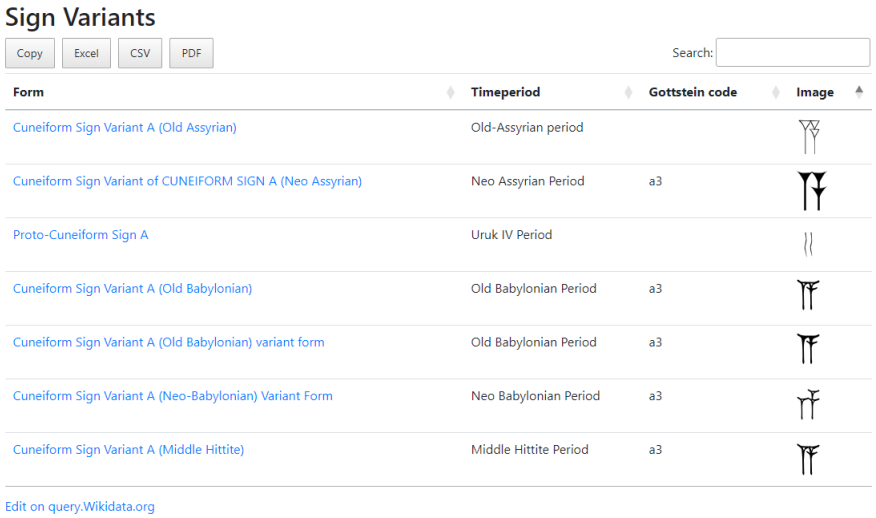

Figure 4: Cuneiform Sign Variants in PaleOrdia along with images from freely available fonts uploaded to Wikimedia Commons to be used as a reference. Sign variants in this view come with an associated time period and a description code.

signs may be shown per cuneiform sign and cuneiform sign variant. Figure 5[6] shows compound signs containing the cuneiform sign HAL in a sign variant common in the Old Babylonian and

[2]https://situx.github.io/paleordia/no_unicode/?q=Q401&qLabel=cuneiform&qb=Q401
[3]https://situx.github.io/paleordia/compoundsigns/?q=Q401&qLabel=cuneiform&qb=Q401
[4]https://situx.github.io/paleordia/allographs/?q=Q401&qLabel=cuneiform&qb=Q401
[5]https://situx.github.io/paleordia/c/?q=Q87554995&qLabel=%F0%92%80%80
[6]https://situx.github.io/paleordia/cf/?q=Q120671708&qLabel=Cuneiform%20Sign%20Variant%20HAL%20(Akkadian)

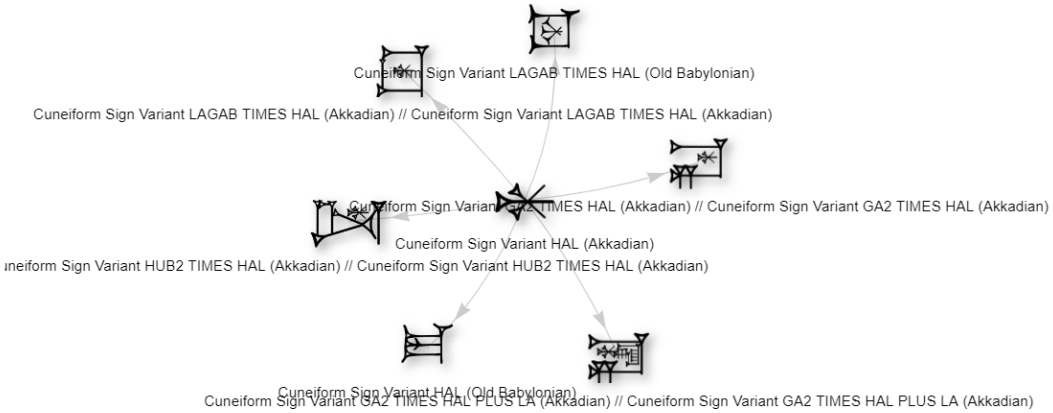

Figure 5: Cuneiform signs containing the cuneiform sign HAL as one of its components. This image displays the sign variants for the Old Babylonian and Akkadian time periods

Akkadian periods. Similar visualizations exist for other time periods, and a generic visualization based on Unicode codepoints is generated for the Unicode sign itself. Finally, users can list the Lexemes in which the Unicode codepoint appears and the attested readings for a cuneiform sign, which may be used to search for the sign and the sign variants in a research context. In essence, PaleOrdia hereby acts as the view on a linked data-based paleographic sign registry that can be extended collaboratively and provides a basis for discussion by scholars.

5. Application Case: Digital Editions of Cuneiform Tablets

A digital edition of cuneiform tablets encompasses a variety of steps by a scholar but usually requires the following components:

1. A transliteration of the written contents of the cuneiform tablet's sides into the Latin alphabet
2. The annotation of interesting text passages
3. The annotation of interesting features on image media depicting the cuneiform tablet (e.g., broken parts, cuneiform signs, or seal impressions)

Figure 6 shows common elements of a digital edition of cuneiform tablets. Figure 7 shows an example of an image annotation in the image annotation application Cuneur - Cuneiform Annotator[7]. The annotation, saved as an annotation in the W3C Web Annotation Data Model [9], currently includes the annotation of the tablet surface, line, and character index to locate the annotated sign. In addition, it includes a character encoding - here, a PaleoCode - to identify the cuneiform sign variant. With the availability of a paleographic sign registry in Wikidata, annotation tools such as Cuneur can be extended with search functionality for sign variants

[7] https://fcgl.gitlab.io/annotator-showcase/

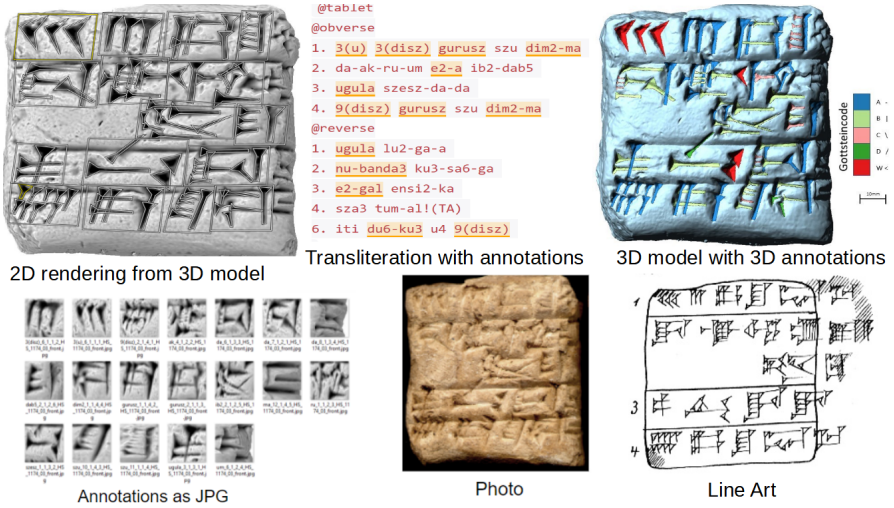

Figure 6: Typical elements of a digital edition of cuneiform tablets. Each element may be seen as a part of a digital edition knowledge graph and is connected to other knowledge graphs using web annotations

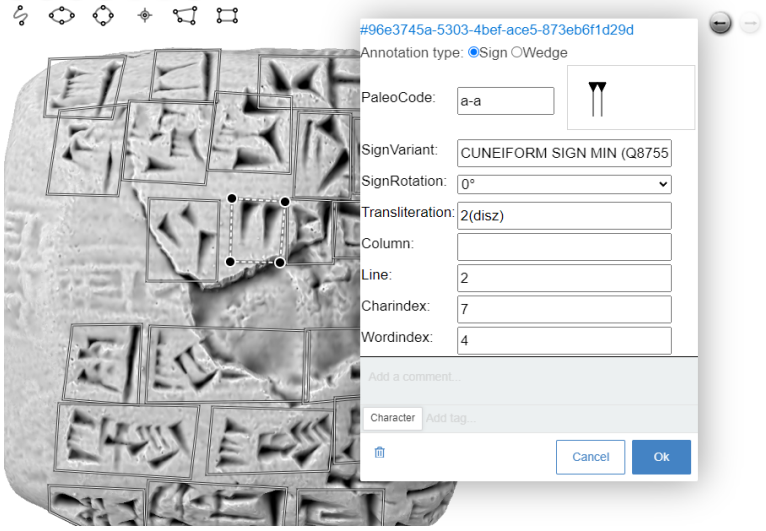

Figure 7: Annotations of cuneiform signs in the Cuneiform Annotator application with an added field for sign variant identification in Wikidata

and with URIs, which allow for the reusage of paleographic sign variants across the boundaries of a single cuneiform digital edition.

6. Conclusions

This publication introduced the application of a paleographic-linked open data model in Wikidata. The model was tested using the cuneiform script as an example use case and has been used to describe actual cuneiform sign variants found on images and renderings of cuneiform

tablet surfaces. The tool PaleOrdia gives an overview and can manage the entered cuneiform sign variants using only a static homepage on Github. This allows cuneiform scholars to get an overview of available cuneiform signs, allows them to compare these sign variants to their findings on the cuneiform clay tablets, and create a linked open data graph of image annotations that are linked to this sign variant registry, which has therefore emerged within Wikidata. The results of such annotations can not only prove valuable for the cuneiform scholar community but may also provide the basis and training data for a variety of machine-assisted classification methods. Finally, this case study based on the cuneiform script might inspire the modeling of sign variants of other scripts in Wikidata, contributing to an interconnected linked open data cloud for paleography.

6.1. Future Work

Future work should enhance the paleographic linked open data cloud with metrics that allow the calculation of the similarity of different graphemes in the linked open data cloud either by semantic similarity, by image similarity metrics, or by metrics built from encodings which allow the expression of a characters characteristics, such as a PaleoCode [3] or a Gottstein code [2] for cuneiform. Expressing these metrics will allow for the implementation of a better semantic search for paleographic sign variants and may prove valuable for approaches for automatic annotation of cuneiform signs.

References

[1] P. Sahle, What is a scholarly digital edition?, Digital scholarly editing: Theories and practices 1 (2016) 19–39.

[2] S. V. Panayotov, The gottstein system implemented on a digital middle and neo-assyrian palaeography, CDLN, London (2015).

[3] T. Homburg, Paleocodage—enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding, Digital Scholarship in the Humanities 36 (2021) ii127–ii154.

[4] T. Homburg, T. Declerck, Towards the integration of cuneiform in the ontolex-lemon framework, 2022.

[5] T. Homburg, Towards paleographic linked open data (plod): A general vocabulary to describe paleographic features., in: DH, 2020.

[6] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The ontolex-lemon model: development and applications, in: Proceedings of eLex 2017 conference, 2017, pp. 19–21.

[7] M. Doerr, F. Murano, A. Felicetti, Definition of the crmtex, 2020.

[8] F. Å. Nielsen, Ordia: A web application for wikidata lexemes, in: The Semantic Web: ESWC 2019 Satellite Events: ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers 16, Springer, 2019, pp. 141–146.

[9] R. Sanderson, P. Ciccarese, B. Young, Web annotation data model, W3C recommendation 23 (2017).