Linked open data for languages written in cuneiform script

Dissertation

zur Erlangung des

Doktorgrades der Naturwissenschaften

(Dr. rer. nat.)

der Naturwissenschaftlichen Fakultät III

Agrar- und Ernährungswissenschaften,

Geowissenschaften und Informatik

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Timo Homburg

Gutachter:innen

Juniorprofessor Dr. Hubert Mara

Professor Dr. Kai-Christian Bruhn

Juniorprofessorin Dr. Eliese-Sophia Lincke

Tag der Verteidigung: 24.10.2024

Zusammenfassung

Keilschrifttafeln haben eine mehr als 3000jährige Geschichte, die es uns erlaubt Rückschlüsse auf das Leben und die Kulturen der ersten Zivilisationen der Erde zu ziehen. Die Entzifferung und Übersetzung von Texten in Keilschrift kann in verschiedene Probleme in der Informatik heruntergebrochen werden, welche alle von einer gut definierten und reichen Datenbasis für die Ausführung verschiedener Algorithmen profitieren.

Diese Doktorarbeit widmet sich der Frage wie ein Datenmodell für die Integration von Keilschriftartefakten, dargestellt in unterschiedlichen Medien wie Fotos und 3D Modellen interoperabel mittels Linked Open Data Technologien definiert werden kann und welche neuen Erkenntnisse aus der Verbindung der verschiedenen Daten gewonnen werden können. Besondere Schwerpunkte liegen hierbei auf der Integration von 3D Modellen von Keilschrifttafeln, auf der Modellierung und Integration von sprachübergreifender Keilschriftpaläographie und schließlich auf der Darstellung aller Inhalte auf Keilschrifttafeln vom Artefakt ausgehend bis zum Einzelkeil.

Um die Arbeitsweise von Assyriologen ebenfalls in das Linked Data Datenmodell zu übernehmen, müssen zudem Beobachtungen der Assyriologen im Editionsprozess der Keilschrifttafeln mit integriert werden. Dies bedingt die Entwicklung von Annotationsmodellen auf verschiedenen Medien und die Abbildung des Interpretationsprozesses in einer digitalen Edition. In diesem Prozess soll die bisherige Arbeitsweise von Assyriologen von einer traditionellen Edition von Keilschrifttexten in einer linked data basierten digitalen Edition abgebildet und durch digitale Technologien ergänzt werden. Hierbei soll die Grundlage für eine Verknüpfung unterschiedlicher Datenquellen auf der Bedeutungsebene (semantischen Ebene) erschlossen werden um für weitere darauf aufbauende Verfahren eine Grundlage wohldefinierter und kurartierter Daten bereitstellen zu können.

Die Linked Open Data Modelle werden anschließend mittels einer Toolchain für Assyriologen erschließbar und an praktischen Beispielen auf ihre Eignung geprüft. Abschließend werden eine Reihe von Anwendungsszenarien für das neu erstellte Datenmodell erläutert und Ausblicke auf die Adaption des Datenmodells in bereits bestehenden Datenrepositorien gegeben.

Keywords— Keilschrift, 3D Modelle, Paläographie, Linked Open Data, Ontologie, Annotationen

D

Abstract

Cuneiform artifacts have a 3000-year-old history and are testimonies of life in some of the oldest civilizations on Earth. The decipherment and translation of cuneiform texts may be broken down into various computer science tasks from OCR to machine translation, all of which depend on and benefit from a well-defined and rich availability of data for the algorithms to work with.

This thesis deals with the question of how a data model that describes cuneiform artifacts in different media (e.g., photos, 3D models, transliterations) could be created based on linked open data technologies and which new insights can be gained by exploiting such a data model. Particular emphasis is given to integrating 3D models of cuneiform artifacts and integrating language-independent paleographic descriptions of cuneiform signs. Furthermore, the model should include every aspect of a cuneiform artifact, from the physical artifact down to a single cuneiform wedge, and allow for the exploitation of data useful to many research communities.

Another important aspect of this thesis is the digital modeling of a typical workflow of an Assyriologist, i.e., the creation of a linked-data-based digital scholarly edition, following the aforementioned ontology model, as they interpret cuneiform texts and will document their interpretations in a targeted manner. This warrants the definition of new types of annotations on respective media and the enrichment and replacement of media commonly used in a non-digital way. Moreover, the thesis discusses the tools that enable the input of so-formatted data by Assyriology researchers so that linked open data, along with other standard data formats for research communities, might become the norm in the future.

Finally, a variety of application cases for the newly defined data model are explored, and the adaptation of the newly defined data model for implementation in already existing data repositories is discussed.

Keywords— Cuneiform, Paleography, 3D Models, Annotations, Linked Open Data, Ontology

F

Danksagung

An dieser Stelle möchte ich allen Personen danken, die mich bei der Anfertigung meiner Dissertation unterstützt haben.

Allen voran gilt mein Dank meinem Doktorvater Hubert Mara für die fachliche Unterstützung, hilfreichen Diskussionen und die Begleitung der Dissertation, der erschienenen Publikationen und der Möglichkeit des Networkings innerhalb der Community der digitalen Assyriologie. Desweiteren danke ich Kai-Christian Bruhn für die Ermöglichung der Dissertation im Rahmen des Forschungsprojekts "Digitale Edition der Keilschrifttexte aus Haft Tappeh", für den guten Rat und für ein besseres Verständnis der Perspektive der Geisteswissenschaften auf die Problemstellung, sowie der Gewährung von genügend Freiraum für die Anwendung der Ideen dieser Disseration im Forschungsprojektkontext.

Im Kreis des i3mainz Instituts der Hochschule Mainz und des mainzed Netzwerks gilt mein Dank den folgenden Personen:

- Doris Prechel, Eva-Maria Huber und Tim Brandes (Uni Mainz) für die konstruktiven Diskussionen, der Einbindung von digitalen Methoden in die Lehre und der Evaluierung von Ansätzen aus dieser Arbeit im Forschungsprojektkontext
- Anja Cramer und Laura Raddatz (RGZM und i3mainz) für die gemeinsame Entwicklung des Metadatenmodells für die Beschreibung von 3D Meshes
- Robert Zwick und Marc Häuser für die Durchführung von zwei Praxisprojekten welche praktische Anwendungen in der Paleographie und der 3D Annotation entwickelt haben

Im Weiteren gilt mein Dank verschiedenen Personen aus der Community der digitalen Assyriolgie, namentlich:

- Adam Anderson (UC Berkeley) für einen regen Austausch über die Adäquatheit der Linked Data Modellierungen und ihres praktischen Nutzens
- Emilie Page-Perron (Oxford) für die Einbindung in die tägliche Arbeit der CDLI und einen Einblick in eine der größten Forschungsdateninfrastrukturen für die Assyriologie
- Katrien De Graef (UGent) und Hendrik Hameeuw (KU Leuven) für die Bereitstellung von Keilschrifttafeln für Showcases dieser Dissertation, für Feedback von Tools die im Laufe dieser Disseration entwickelt wurden und für die fachlichen Austausch, sowie Einblicke in die Arbeitsweise im CUNE-IIIF-ORM Projekt

- Stefan Jakob (Uni Heidelberg) für die Bereitstellung und Korrektur von Annotationen auf einer in dieser Dissertation gezeigten Keilschrifttafeln

Zuletzt möchte ich noch Personen aus dem Kreis der Computerlinguistik, der Geodatenverarbeitung und der Anwendung von Semantic Web Technologien danken:

- Christian Chiarcos und Thierry Declerck, sowie dem Nexus Linguarum Netzwerk für die Einbindung in die Ontolex-Lemon Arbeitsgruppen zur Standardisierung von Ontologiemodellen für linguistische Knowledge Graphen
- Verschiedenen Kollegen aus der OGC GeoSemantics Domain Working Group und der W3C Spatial Data On the Web Working Group für den Austausch und die Diskussionen bezüglich der Standardisierung und Darstellung von Geodaten und CRS im Allgemeinen und komplexeren Geodaten wie 3D Meshes im Speziellen

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The cuneiform script is one of the oldest written scripts in the world [RR11, DB96]. Scribes have used it as the script of more than five different languages, usually on clay tablets, and it has a history of more than 3000 years of practice [VdM15]. Since the decipherment of the cuneiform script in the 19th-century [Cat83], Assyriologists have been transliterating and translating cuneiform texts manually, creating dictionary resources [BGP00, Klo07], word glossaries [Fox22, Stu08], grammars [Jag10, Hue18, Lau14] with the prospect of further understanding textual contents, and have been working on gazetteers [Rat19] to map the locations in which texts were found and to which texts were referring to. This work is invaluable for understanding the culture and living conditions and for socioeconomic research questions of interest to Middle Eastern Studies.

However, this work of word glossary curation, processing text corpora of cuneiform tablets, and interpreting said digital texts is still often done without much digital assistance. While nowadays, many cuneiform text edition projects result in digital text corpora which are accessible over the internet, we can observe the creation of so-called data silos [SH10], i.e., databases using customized data formats and transliteration styles, that have been created often for a particular project, which are accessible for the individual researcher (human-accessible), but which are not necessarily compatible on a machine-readable level.

On top of that, digital scholarly editions [Gab10, Pie16] of cuneiform texts change how researchers work with cuneiform texts. Modern digital scholarly editions produce many digital artifacts such as 3D models, renderings of 3D objects, Transliteration text files, photos, and their metadata and annotations on these mediums, which need to be stored securely to be of value for researchers not only in the field of Assyriology but also for neighboring fields such as computational linguistics, Optical Character Recognition [MNY99] (OCR) and more. This fact is also recognized by interdisciplinary demands for more interoperable digital scholarly edition concepts [Sch14].

Figure 1.1 shows the bandwidth of digital media used to describe a cuneiform tablet. The representation of these media results in data of heterogeneous formats and contents, which are often challenging to consult for the average researcher (especially from a non-Assyriology field) and much less for machine-processable algorithms. Experiments using machine learning data on specific corpora usually need considerable time for data collection, preprocessing, filtering, and conversion from multiple different formats until the

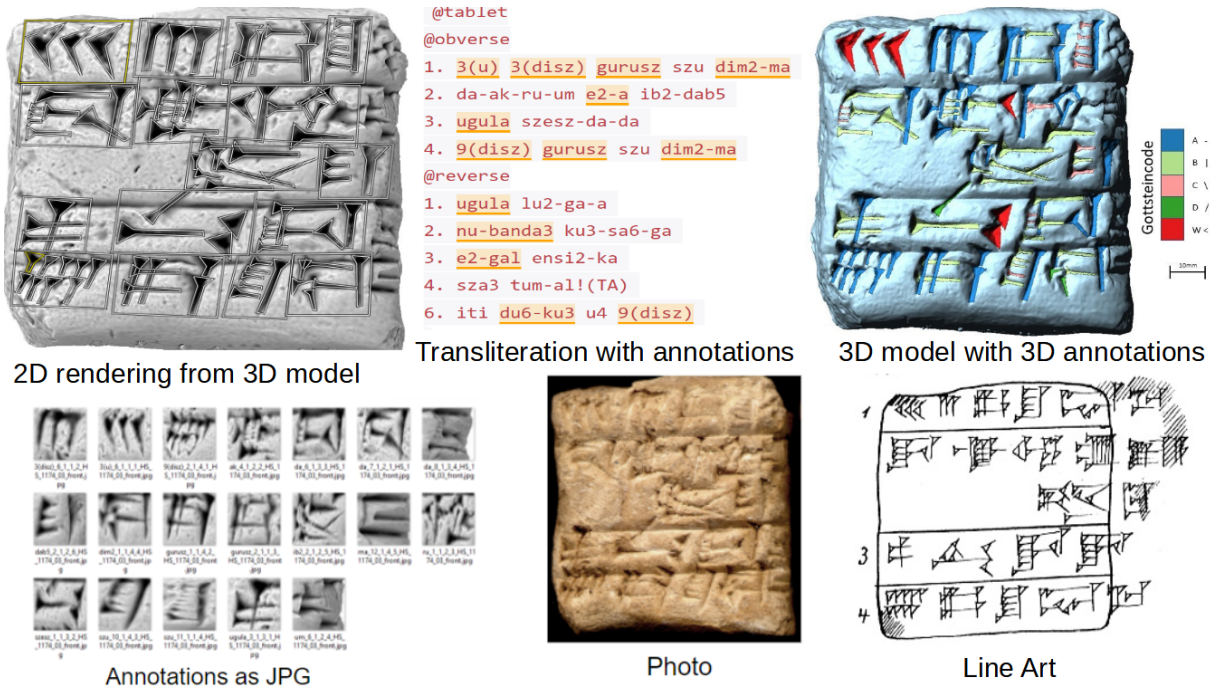


Figure 1.1: Digital representations of a single cuneiform tablet on the example of cuneiform tablet HS1174: Transliterations, Line Arts, photos, 3D renderings, 3D meshes, and annotations on each of these mediums

appropriate data for a specific experiment has been extracted. Similarly, Assyriologists need an overview of the specificities of cuneiform text corpora and image representations to conclude the contents and the historical context of the cuneiform tablets they investigate.

A digital scholarly edition and a digital publication of cuneiform data ask of many, if not all, components of a traditional edition of cuneiform tablet corpora and, at the same time, request the resources which are created in such a digital scholarly edition to be machine-readable, interoperable, annotatable, and reusable in different usage contexts. To this day, however, these resources are often fragmented and non-standardized and mandate the definition of further vocabularies for these classifications. A simple dictionary lookup in a machine-readable way over many corpora or a lookup of cuneiform sign variants present on many images of the current cuneiform artifacts shared across different databases is, despite data that may be available, a manual process that must be done by the Assyriologist without many technical aids.

The data, which might help in creating, e.g., suggestion systems for part of speech tagging or semantic annotations, are part of various data repositories which are currently not interlinked. For instance, the contents in the cuneiform data repositories of the Cuneiform Digital Library Initiative (CDLI)[1] and the Open Richly Annotated Cuneiform Corpus (ORACC)[2] differ in their attested transliterations, their transliteration styles, and the number of features they provide to preprocess their given data. At the time of writing,

[1] https://cdli.earth
[2] http://oracc.museum.upenn.edu

ORACC corpora provide part of speech tags and lemmatized forms in their data, while CDLI does not. In contrast, CDLI is likelier to include image data and line art about the cuneiform artifacts they display. Currently, because of the lack of a standardized data format, but also due to a lack of Application Programming Interface (API) support and a definition of semantics, exchanges between these data repositories depend on the repositories being willing to import the other's data or on the single researcher manually combining cuneiform tablet data to answer their particular research question - a task which is possible for a researcher in data science, but not likely to be accomplished by an Assyriologist.

Linked open data [BK11] promises and has the potential of providing a unifying data model, which helps integrate different cuneiform data repositories and potentially allows live querying of distributed cuneiform-related data on-the-fly, as they develop in their original curated repositories. Considering that Assyriologists often specialize in one of the cuneiform languages and may neglect information that is uncovered by researchers on a related location, language, time period, or in terms of paleographic features simply because it is not easily accessible, a machine-readable representation of the facts that researchers have gathered over time can significantly improve joint information exchange between research communities, but also between digital media.

1.1 Motivation

The discipline of Assyriology has strived in recent years to digitize their scholarly text editions and documentations of cuneiform tablets as evidenced by several online repositories to archive transliterations [Eng16], repositories for a linguistic assessment of cuneiform text corpora such as ORACC [TR14] and further research project websites like The Munich Open-access Cuneiform Corpus Initiative [RW18] (MOCCI) [SR18] which try to highlight research results about a respective lemmatized and to an emerging extent open access corpus.

These digital representations of cuneiform data are plentiful and cover cuneiform texts from a variety of languages and time epochs, but are not necessarily easy to consolidate for Assyriologists, as data has to be researched from different places (data silos) in possibly slightly different representations (image vs. 3D mesh vs. different Transliteration styles), different annotations, and with a possibly different focus on corpus creation in mind. Navigation through the different corpora, especially across language and epoch boundaries, is not an easy task. A cross-lingual, cross-corpora, cross-epoch, and cuneiform-feature-specific search engine for cuneiform artifacts is lacking for enabling a more rapid, efficient, and substantiated scholarly discourse for teaching, a more accessible publication of their research, also for non-specialist communities, and stricter digitally-linked, evidence-based, and statistically grounded research in Assyriology. In particular, digital representations in Assyriology repositories lack cross-media linkages, connections between areas on, e.g., images, textual interpretations, and further representations such as 3D models or Line Arts.

Computational linguistics benefit from research data created in such a way that they are machine-readable and annotated with linguistic metadata so that they may be used for language classification tasks such as word segmentation [HC16], named entity recognition [LBHL15], or machine translation [PSCPP20]. Current data portals only begin to expose cuneiform research data in machine-readable ways[3] but, more often than not, expose corpora data only in human-readable form[4], in Hypertext Markup Language [BLC95] (HTML) or Portable Document Format [BCASMV93] (PDF). Given a variety of transliteration styles, transliteration formats, dictionary formats, and a variety of formats to represent other media and a lack of data conversion tools, data integration is, at best, a challenging endeavor.

Machine learning classification tasks usually rely on a representation of the individual language in a Unicode representation or a transliterated version of the given script. For cuneiform, however, Assyriologists are often reluctant to work with Unicode cuneiform representations. For them, it is at least as important to consider the cuneiform paleography of the individual signs when interpreting a cuneiform text as the meaning of a word might change with a varying paleographic representation - an aspect that is hard to fulfill by a given font that can only represent the shape of the specific epoch which was encoded into it. However, what is true for an Assyriologist who wants to interpret a cuneiform text must also be valid for an automated approach to analyzing the language. Hence, to automatically interpret cuneiform texts, the inclusion of paleographic features should become necessary, which has to be captured into data supporting such classifications. This necessity, exemplified for cuneiform, is not necessarily applicable to cuneiform only, as there are further languages for which paleography plays an essential role in classification tasks [VP22].

In addition, paleographic variants are essential for correctly recognizing cuneiform scripts using machine-learning approaches, as they can often, apart from the context of the given word, be a decisive factor when classifying in computational linguistics. To date, computational linguistic approaches taking advantage of formalized paleographic representations of their analyzing script are a minority. The formalization of scripts has often only been considered when generating fonts as parts of sign description languages. Still, it has not been a part of, e.g., machine learning approaches in computational linguistics. One reason is that scripts of specific languages are sufficiently normalized, e.g., in the Latin alphabet. In these cases, paleographic considerations are not necessary for a machine-learning setting. For languages such as cuneiform, in which an analysis of paleographic aspects might be crucial, the research community faces a lack of formalized paleographic descriptions, standards, and availability of annotated data. Hence, researchers have yet to explore the synergies of combining paleographic features with computational linguistics features.

In Assyriology, the insufficient digitalization of paleographic features even leads researchers not to use specific digital technologies such as cuneiform fonts and Unicode representations of cuneiform signs since they cannot convey the individual paleographic particularities that can be found on the cuneiform artifact itself. While these can only be found on media depicting the original artifacts, the images the fonts display on the screen

[3] http://oracc.museum.upenn.edu/epsd2/json/index.html
[4] e.g., https://www.assyrianlanguages.org/akkadian/list.php

are seen as misleading, often even when using a time period specific font. The preferred method of working is to create a transliteration in the Latin script and use other digitally not linked image media for contextualization. In doing so, researchers in Assyriology rarely capture and even define paleographic sign variants in their usual scientific work, only in specialized studies, and seldom in a digitized way, allowing the comparison and embedding in a digital setting.

Cuneiform studies are often separated into different sub-research communities, which distinguish themselves by the cuneiform languages they research, time period, and various other factors. This has led to various data silos of varying quality and interoperability, hosted on different servers and often unsustainable financing, eventually leading to their deactivation. Moreover, there is an increased interest in working with cuneiform-related data and in the digitization and analysis of so-called low-resource languages such as cuneiform in computational linguistics, but also in OCR research and related more technology-oriented fields. These related areas do not necessarily view cuneiform data resources from a philological standpoint, and researchers in these disciplines typically have no means of reading the cuneiform script itself. Hence, they rely on data structures that allow the processing and researching aspects of the cuneiform script in their respective area. To do proper research in any of the aforementioned related fields, one needs to work with data models that allow an accurate description of the contents given.

To that end, better integration of cross-disciplinary data concerning cuneiform artifacts will benefit various research communities and constitute one of this thesis's main motivations.

1.2 Research questions

In light of the motivations given in the previous section, this thesis focuses on solving the following four research questions:

RQ1 How can paleographic features of cuneiform signs be represented in linked data to maximize their impact on different research communities?

RQ2 How should an annotation model for cuneiform data on different digital mediums be organized?

RQ3 How can an interconnected linked open data graph on different digital mediums of cuneiform contribute to cuneiform sign recognition approaches performed by machine learning algorithms?

RQ4 How can Assyriologists be enabled to provide cuneiform data which is interoperable also for digital application cases?

The solution to these research questions is attempted with linked data technologies. At the time of writing this thesis, linked open data provides one, if not the most advanced, methodology of sharing semantically interpreted data of various kinds and making them accessible, and should be investigated as the medium of interconnection between the different heterogeneous data sources.

1.3 Contributions

This thesis contributes:

CON1 A linked data-based digital representation of cuneiform signs and sign variants to capture paleographic data

CON2 Representation of 3D models and annotations in 3D in linked open data, including annotation and interconnection to different other mediums

CON3 A holistic linked open data model to represent cuneiform artifacts and their interpretations with a case study of machine learning applications

CON4 Prototypical application and case study of the applicability of the aforementioned concepts in practice

Together, these contributions allow for creating a cuneiform linked open data cloud, enabling researchers of different research disciplines to access the kinds of information about cuneiform-related media they need for their particular research task.

1.4 Contributions in context

The contributions of this thesis need to be seen in the context of the work of Assyriologists and computer scientists alike and consist of a data model that interlinks the different media used in the daily work of Assyriologists. A digital and machine-readable representation should hinder neither the work of Assyriologists nor the work of computer scientists who want to work with cuneiform data. The former need easy access to cuneiform tablet data, the preservation of cultural heritage objects, and better possibilities to highlight the remains of cuneiform signs for transliteration purposes. The latter has an interest in conducting natural language processing analysis in cuneiform languages, cuneiform character recognition, and classification on 3D scans, photos, or renderings, and, in the long run, the creation of tools for automated translations from cuneiform artifact images to target languages such as English or German.

One prerequisite for computer science approaches to be successful and help Assyriology researchers by delivering valuable tools and classifications is data repositories that provide standardized, documented, and open data that are easily consumable by computers and humans alike. In other words, this advocates for implementing the Findable Accessible Interoperable Reusable [WDA+16] (FAIR) principles for data and research software [LGK+20] so that research data is fully documented, reproducible, accessible, and easily categorizable. For example, suppose computer scientists prepare data for a machine-learning experiment. In that case, it should be easy for them to query for exactly the kind of data they need, irrespective of their understanding of the cuneiform script, and to customize their training data set compared to downloading and manually filtering and merging data in heterogeneous formats, such is the case today.

Assyriologists need digital tools to automate certain possibly repetitive tasks (e.g. part of speech tagging, annotations), which a knowledge base may support. In this way, they can fully focus on the most valuable tasks that are hardly automatable, i.e., the interpretation of the text corpus they are currently working on. Digital tools should exploit the knowledge stored in these repositories to give researchers an optimal choice to e.g. interpret sign variants and assign semantic meanings to words and part of speech tags to existing texts. Similarly, Assyriologists could benefit significantly from databases of sign variants which could be compared to signs on the cuneiform tablet they currently observe.

To conclude, a linked open data model of cuneiform artifact data would likely elevate the exchange of data between cuneiform repositories and research disciplines, give leverage to tools that provide assisting tasks to Assyriologists working on digital scholarly editions, and would provide a simple, interpreted, and easily accessible data model computer scientists may exploit for any experiment that they see fit.

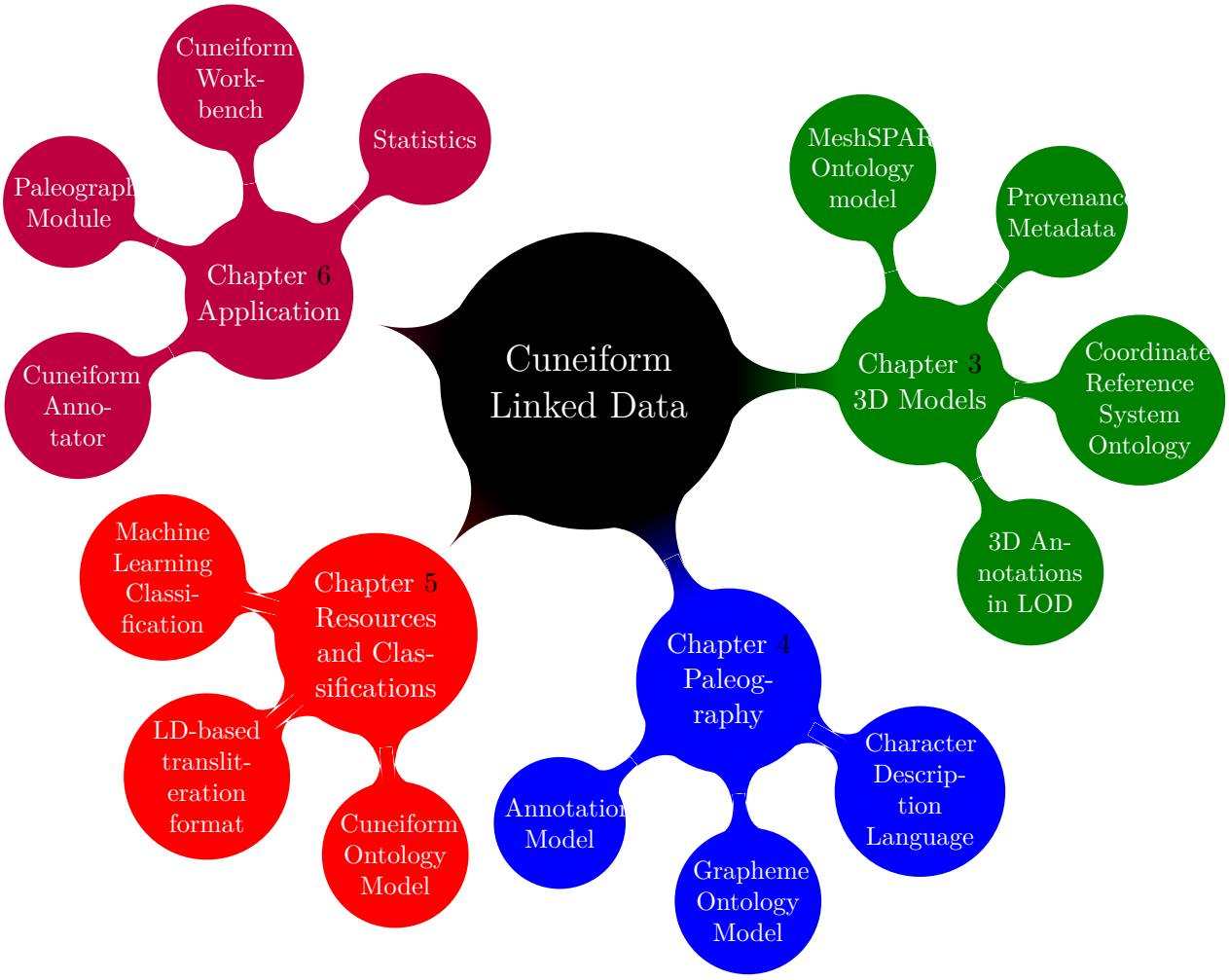


Figure 1.2: Contributions in of this thesis in clockwise direction: Description of 3D models of cuneiform tablets in linked data, followed by a machine-readable description of cuneiform paleography, an overarching ontology model for cuneiform resources and applications as well as tools applying the aforementioned contributions

Figure 1.2 shows the contributions of this thesis and breaks them down in the following chapters. After discussing the foundations of cuneiform languages and linked open data with a particular focus on linked open data in computational linguistics and spatial data, Chapter 3 deals with the integration of 3D meshes and annotations on these media in linked open data in general and 3D meshes of cuneiform tablets in particular. Next, Chapter 4 defines a cuneiform sign description encoding (cf. Section 4.2) and an ontology model to capture cuneiform paleography about existing vocabularies in computational linguistics. The ontology model is extended by a more general approach to describing paleographic items in any written script (cf. Section 4.3) and by a case study on the applicability of the integration of paleographic information in already established transliteration formats.

Chapter 5 combines the aforementioned developed ontology models to create an overarching ontology model which describes cuneiform artifacts, their texts, and interpretations and can link to paleographic, linguistic, and semantic classifications of their contents. A new linked-data-based transliteration format linked to this ontology model is proposed in the following (cf. Section 5.2). Next, the ontology model is exploited to create MaiCuBeDa, a new machine-learning data set, and to execute two machine-learning classification tasks. Chapter 6 combines the work of the previous chapters in a case study of creating a set of tools called the CuneiformWorkbench, a digital edition environment for Assyriologists based on the previously developed technologies. This chapter will also highlight the anticipated difference in the workflow that a digital scholarly edition imposes on digitally-trained Assyriologists. Finally, the results of the thesis are discussed and set in relation to future work in Chapter 7.

Chapter 2

Foundations

This chapter introduces the reader to foundations in semantic web technologies, the components and representation of spatial data, specifics of cuneiform languages, and the representation of linguistic data in knowledge graphs. To create an overarching data model for the connection of cuneiform digital artifacts, it is essential to be aware of current practices in the respective Assyriology communities and to relate these to already existing related work in computer science so that the reusability of already existing technologies is maximized.

2.1 Semantic Web Technologies

Since the emergence of the World Wide Web [BLCL+94] (WWW) as the largest repository of human knowledge globally, computer scientists have worked on making this document-centered and human-optimized data source better accessible for machine processing. To that end, the semantic web [BLHL+01] acts as an extension of the world wide web through a series of standards defined by the World Wide Web Consortium [Bro15] (W3C). The semantic web aims to create a machine-readable web of data that is structured utilizing knowledge graphs [FŞA+20]. The literature typically distinguishes:

Definition 1. ***Linked Data***
Linked data is structured data that is interlinked to other datasets.

Definition 2. ***Linked Open Data***
Linked Open Data [BK11] (LOD) is structured open data interlinked to other openly accessible datasets.

A more precise technical definition of linked open data can be found in the 5-star model for linked open data[1] following the linked data principles[2] and thereby implementing the FAIR principles [WDA+16] for publishing open data. Linked data and linked open data are described using the following standards:

- Resource Description Framework [Pan09] (RDF) [CWL14]

[1] https://5stardata.info/en/
[2] https://www.w3.org/DesignIssues/LinkedData.html

- Resource Description Framework Schema [BG14] (RDFS)
- Web Ontology Language [Gro12] (OWL) [MPSP12]
- Shape Constraint Language [KK17] (SHACL)
- SPARQL Protocol And RDF Query Language [SH13] (SPARQL) [HS13]

RDF defines resources on the web, which may be described using Uniform Resource Identifier [BLFM05] (URI) or Internationalized Resource Identifier [DS05] (IRI), the latter allowing Unicode characters as part of its identifier definition. RDFS defines a schema to describe, structure, and classify web resources using RDF, OWL defines constraints and class hierarchies, and SHACL sets constraints on graph structures and facilitates the validation of graph structures. SPARQL is the semantic web query language and allows querying RDF graphs and their contents. The global graph described by all linked open data resources is called the linked open data cloud (LOD cloud)[3] [MAB+19] and is subdivided into different sub-clouds representing different domains of knowledge. In particular, this thesis will discuss extensions to the spatial LOD cloud and the linguistic LOD cloud. The former includes geospatial and spatial data representations and the latter deals with linguistic classifications of textual artifacts. All linked data resources developed in this thesis follow the best practices for implementing fair vocabularies and ontologies on the web [GP20] and the (spatial) data on the web best practices [LBC17, vdBBT+19].

2.1.1 Spatial Data Representations and the spatial linked data cloud

The term spatial data describes all data objects representing one or many locations. While different ways to represent spatial data exist (e.g., 3D meshes, raster images, vector graphics), all spatial data consists of or can be mapped to a set of (possibly interconnected) points in a given spatial reference or spatial coordinate system. Spatial data may be associated with a reference; e.g., in the case of geospatial data, a vector geometry might be referenced to a position of an approximated version of planet Earth. Still, mobile objects such as 3D scans may also be provided without a (geo)reference, with isolated coordinates in a coordinate space or as a sequence of geometries over time. The spatial linked data cloud [vdBBT+19] consists of all linked data elements connected to a spatial location.

Currently, only vector data is sufficiently represented in linked data through vocabularies such as the Geographic SPARQL Protocol And RDF Query Language ([BK12]) (GeoSPARQL) vocabulary [PH12, NTM+22, CH22], NeoGeo [HG14] or the W3CGeo vocabulary[4]. Query languages such as the GeoSPARQL query language [PH12, NTM+22] provide operators to compare, access, and convert spatial Data and to access their semantics. GeoSPARQL support in triple stores is still fractured, [JHS21a, JHS21b], and support for raster data or 3D meshes in semantic web graphs is despite initial work in this direction [HSJ20, AVL22] still not well-established.

[3] https://lod-cloud.net
[4] https://www.w3.org/2003/01/geo/

2.1.2 Linked data representations of cultural heritage objects

Cultural heritage objects have been represented using linked data technologies because of the flexibility of its data model and the fact that cultural heritage objects are always to be seen in various knowledge contexts (e.g., historical, material science, sociocultural, etc.). The arguably most advanced ontology model for cultural heritage representation is the CIDOC-CRM [Doe05] model covering the representation of cultural heritage objects, their storage, archaeological context, and remarkable features on cultural artifacts (e.g., drawings and areas of inscriptions). CIDOC-CRM is used in a variety of cultural heritage research contexts [BCG17, Nic17] and among others provides extensions for modeling spatio-temporal contexts [HDE17], textual surfaces [DMF20]. Current usages of the CIDOC-CRM model for cuneiform texts can be found in the implementation of the new CDLI framework[5] providing certain metadata of cuneiform artifacts as linked data using the CIDOC-CRM model.

2.2 Digital representation of cuneiform texts

Cuneiform texts can be digitally represented in various media. Some media require an interpretation of the cuneiform tablet's contents by a human being or algorithm. Other media accurately represent the cuneiform artifact itself without prior interpretation.

2.2.1 Captured media

Photos and 3D scans are media that an Assyriologist has not yet interpreted. Instead, the quality of these captured media depends on the capturing devices and the responsible person's skill and knowledge. In the absence of the physical cuneiform tablet, however, these captured media may be the only source of the interpretation of the cuneiform tablet in creating a digital scholarly edition. Apart from that, each media may include important information not represented in another media representation describing the same cultural heritage object. For example, photos may include color information not necessarily present in a 3D scan. In contrast, the 3D scan contains information about the depth of the cuneiform wedges and better accessibility for the Assyriologist, as the 3D model can be rotated. Captured data may be subject to already established metadata schema mandated by the format they use for saving their respective data. Examples are the Exchangeable image file format [Tac01] (EXIF) standard and the Extensible Metadata Platform [Ado04] (XMP) standard, which are established standards for saving image metadata. Similar standards for 3D meshes have not been established as of the time of writing.

[5] https://cdli.earth

2.2.2 Interpreted media

Line arts and transliterations represent interpretations and formalizations of Assyriologists examining a cuneiform tablet. A Line Art is a hand-drawn or computer-drawn interpretation of the contents of the cuneiform tablet, typically of its cuneiform signs, lines, damaged areas, and further areas of interest. Typically, Line Arts deviate from the contents of photos or 3D scans in the following ways:

- They include human errors when copying contents from the cuneiform tablet, e.g., missing cuneiform wedges
- Simplification and normalization of cuneiform signs in the process of copying for the purposes of readability
- Errors because the readability of the source material is not clear (e.g., photos are blurred, or the amount of wedges of a cuneiform sign is not necessarily distinguishable)

A transliteration constitutes an interpretation of a researcher, which is formalized in a transliteration format, usually in the Latin alphabet. The researcher will, in this process, interpret the language, lines, words, and individual signs, including their reading on the cuneiform tablet's surfaces.

An interpretation of line arts and transliterations alike always consists of the following steps:

1. Sign recognition: The cuneiform signs present on the cuneiform tablet are recognized and are used as the basis for the transfer of this information to the target medium
2. Content interpretation: The recognized sign contents are interpreted in their context and adapted before introducing them in their respective representation

During sign recognition, a researcher may make assumptions about the nature of cuneiform signs that are not readable or missing on the tablet (e.g., broken, non-readable signs). The researcher may interpret a seemingly unknown sign as an already known sign based on its context, or a researcher may interpret the shape of a sign as a new sign variant of an already known sign. The insights the researcher gained in this process will be reflected in the medium of representation, i.e., the line art might represent abstracted or reconstructed versions of the signs visible on the cuneiform tablet. The transliteration will include the interpretations of signs recognized on the cuneiform tablet in the form of a set of readings for each individual sign or sign combination. The Assyriologist has to choose the most plausible reading from a set of possible readings, possibly even from different cuneiform languages (e.g., choose between an Akkadian and Sumerian reading for the same sign combinations or choose a reading depending on the semantic context of the sign derived from the current sentence). The same sources of error valid for line arts also apply to transliterations. That is, signs may have been misrecognized, and therefore, the transliteration may contain errors. However, contrary to line arts, transliterations are the primary discourse medium in the Assyriologists community. Opinions on sign readings are not necessarily unanimous and improve in the scientific discourse.

2.2.3 Transliteration data formats

The community of Assyriologists creates transliterations in various traditions. Over the years, many transliteration dialects have emerged, and in the digital age, many data formats compete to represent cuneiform transliterations. This section briefly aims to overview the available data formats, contents, and transliteration styles.

All transliteration data formats are text-based and can be classified by different families. The ASCII Transliteration Format (ATF) family[6] consists of a textual representation of the cuneiform text content and provides a limited amount of operators to describe broken signs and sign reconstructions, the amount of which varies depending on the often repository-dependent dialect of ATF. ATF formats generally contain only a transliteration text composed of sign readings connected by **-** or **.**, with logograms typically written in uppercase. In its simplest form, ATF texts provide no room for textual annotations or comments inline. Annotations are only possible by defining new syntax elements, thereby creating yet another ATF dialect or format [DFMMO18].

The TEI/XML [IV95] family of transliteration formats consists of cuneiform-specific extensions of the TEI/XML formats. TEI/XML can capture additional annotation information and metadata compared to ATF. Also, because it is an XML-based format, it can be easily extended for further applications. However, TEI/XML is only used in a few Assyriology projects such as Electronic Text Corpus of Sumerian Literature [Rob98] (ETCSL)[7] and is not the primary medium of exchange for digital cuneiform text data.

Lastly, JSON-based [Bra17] transliteration formats such as the JSON Transliteration Format (JTF)[8] or ORACC-JSON[9] are currently emerging file exchange formats. Like TEI/XML, JSON formats allow for better extensibility and the inclusion of metadata and annotations at the cost of a lack of user readability. For the latter two format families, viewers are required to show transliteration contents. No transliteration format at the time of writing supports encoding different cuneiform media, e.g., connections to images or 3D scan components.

2.2.4 Transliteration contents

Transliterations contain the interpretation of the contents of the cuneiform tablet writings in the Latin alphabet. In doing so, the transliteration captures the interpretation of the following elements that the researcher recognized on the cuneiform tablet:

- The reading of the sign at a defined position on the cuneiform tablet, i.e., the recognition of the sign itself and its reading value depending on the context
- The boundaries of recognized words
- The boundaries of lines
- Certain functions of a sign, when realized in the spelling of a word form, for example:

[6] http://oracc.museum.upenn.edu/doc/help/editinginatf/
[7] https://etcsl.orinst.ox.ac.uk/edition2/etcslmanual.php
[8] https://github.com/cdli-gh/jtf-lib
[9] http://oracc.ub.uni-muenchen.de/doc/opendata/json/index.html

  - Marking of determinative signs (classifiers) using curly brackets or as superscript (e.g. {dingir} to mark a deity name)
  - Logogram signs written in uppercase

- Damages which are visible on the tablet, e.g., dingir# for a damaged sign
- Reconstructions of only partly visible words on the cuneiform tablet
- Comments about specific parts of the transliteration by the scholars

All of the elements, as mentioned earlier, constitute important parts of understanding the contents of the cuneiform tablet and are often the only form of publication available for a scholar. Other media are not always legally allowed to be published, or simply unavailable, and are not digitally connected to the transliteration contents.

2.2.5 Transliteration styles

Apart from differing data formats to store transliterations, the Assyriology community also has differing opinions on how readings should be represented. While researchers broadly agree on how to name cuneiform signs, sign readings in transliterations may be disputed on a semantic and syntactic level. The first level determines which reading makes sense in the context of the given sign. The second level concerns the representation of said reading in a transliteration format and with its formatting. The conventions of the latter are not necessarily unanimously accepted in the Assyriology community but can be broken down into the following distinctions:

- Use of ASCII vs. Unicode diacritics (e.g. *e2* vs. *é*)
- Representation of special characters in cuneiform studies such as *š*, *sz*, *sh*
- Usage of sign indices vs. diacritic signs
- Text formatting decisions, such as superscript and subscript for specific cuneiform signs
- Definition of cuneiform word features, i.e., how to format broken signs or reconstructed signs in a transliteration
- Addition of data typically found in annotations to the text format encoding

This diversity in transliteration syntax illustrates that transliteration syntaxes are often region-specific, not necessarily developed for interoperability with other transliteration formats, and often contain customized extensions that are only present in specific data formats. In these efforts, semantic annotation, part of speech tagging, and textual contents are often intertwined so that they become part of the markup of the given format and are hence hard to translate to other formats. The idea behind such an entangling of

transliteration and annotations is that a modified transliteration format is better suited to be integrated into the workflow of an Assyriologist. While this may be true in some cases, the question of interoperability is not always tackled in the aftermath [DFMMO18] and often requires project-specific conversions if interoperability should be a priority.

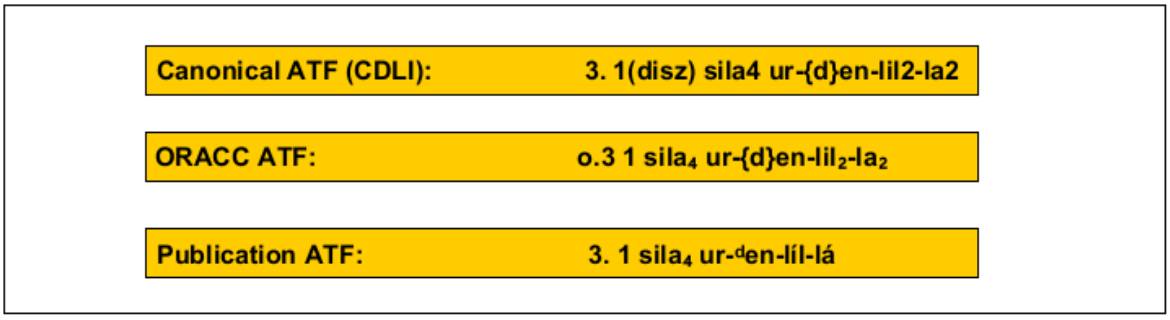


Figure 2.1: Transliteration variations of obverse line 3 of cuneiform tablet P123456 visualized in three common transliteration styles

Figure 2.1 shows three common transliteration styles. The differences between these styles are not only of stylistic relevance. While the first two transliteration styles may be expressed as text files, the publication variant needs a markup language to express some superscript characters unavailable in Unicode. This makes the publication ATF even harder to process, as it has to be encoded in a markup format that must be parsed. Because of this, it is rarely considered by digital databases but usually formatted by hand in text processing software that prepares a traditional publication. Software libraries to convert between the different transliteration styles are, at the time of writing, only existent for the ORACC and CDLI ATF variants[10]. Other transliteration styles have rarely been formalized, and consolidation efforts towards a commonly accepted data and transliteration format for different Assyriology communities are, if at all existent, currently only a concern of data repositories and not of individual researchers.

2.3 Languages written in the cuneiform script

The cuneiform script has been used as a writing medium from the 31st century BC to the 2nd century AD [RR11]. Cuneiform signs are comprised of groups of particularly aligned cuneiform wedges, which represent their atomic parts. Cuneiform signs may be reused as part of other cuneiform signs, thereby, as a component of semantic meaning or a stylistic choice.

[10] https://github.com/oracc/pyoracc

2.3.1 Language classifications

Several languages from different language families, such as Sumerian (Isolate), Akkadian (Semitic), and Hittite (Indo-European), were written in cuneiforms. While the Sumerian language mainly used the cuneiform signs as pictographic and ideographic descriptions similar to Chinese, Akkadian, and Hittite, as later developments until Old Persian, tend to use cuneiform signs more in a syllabic way, i.e., cuneiform signs represent distinct syllables rather than representing the meaning of the individual sign. Still, remnants of cuneiform signs as ideographs can be observed even in Akkadian and Hittite, which contain certain ideographic signs from either Sumerian or Akkadian, so-called Sumerograms, and Akkadograms.

2.3.2 Cuneiform digital scholarly edition process

In a digital scholarly edition, cuneiform texts go through two different interpretation phases. The first phase comprises the detection of the individual cuneiform signs by the Assyriologist or an image recognition algorithm. In most cases, each sign is described using a character name corresponding to a Unicode code point. Optionally, a drawing of the exact or interpreted sign variant on the cuneiform tablet is created as a Line Art. This practice is usually obsolete if other media, e.g., 3D scans, provide a similar or better substitute.

Next, the readings included in the transliteration are inferred from already identified cuneiform signs for the respective identified cuneiform language, again using either a human to interpret the cuneiform text or an algorithm to create a (probably non-perfect) transliteration. Therefore, a transliteration is an interpretation of a human or a machine. Texts prepared this way are published in repositories such as the CDLI and as text corpora for analysis in, e.g., computational linguistics settings. Meanwhile, other mediums might be present that describe the given cuneiform text. Photos, 3D scans, renderings of 3D scans, and manual re-drawings of the shape and characters of the cuneiform tablet are common elements of a digital scholarly edition and are related to the transliteration representation.

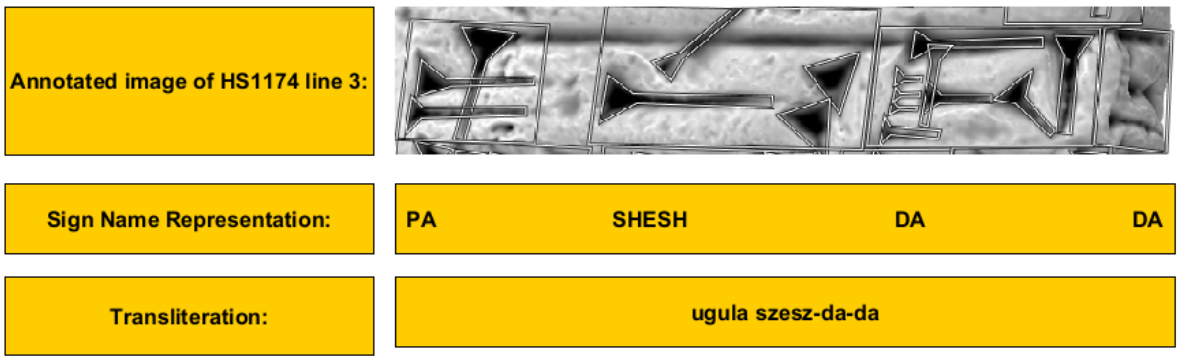


Figure 2.2: Transliteration process of one line on cuneiform tablet HS1174: In the first step, the cuneiform signs are recognized (represented here with sign names from the Unicode proposal), in the second step, the readings and word boundaries are assigned

Figure 2.2 visualizes the transliteration process using one example of the transliteration of one line of a cuneiform tablet. While two steps must be done to arrive at the transliteration result, the intermediate step of writing down the respective sign names is usually not documented by Assyriologists. Sign names are implicitly conveyed using the readings in the transliteration.

2.4 Linguistic resources in linked open data

The Linguistic Linked Open Data [CCMG20] (LLOD) cloud consists of various vocabularies and standards representing linguistic classifications. Computational linguistics usually targets the dissection of texts in Lexemes and their lemmatization, if necessary, the classification of Lexemes into grammatical categories, and finally, the identification of sentences and text sections. In particular, to achieve the classification of lexemes, several tagging standards compete for the application in different languages [PDM12, Atw08]. To assist with these goals, various semantic web vocabularies have been created to aid in the data storage of the required information.

2.4.1 Part Of Speech Tagging

To formalize complete part of speech tagging approaches, the Ontologies of Linguistic Annotation [CS15] (OliA) provide a language-specific and overarching framework for creating vocabularies of part of speech tags and the grouping of different tags to POSTag sets. These POSTags capture necessary linguistic features such as the lexical category of the given word, its word case, grammatical gender, tense, and person and are language-specific. A set of parts of speech tags needs to be created for Sumerian, Akkadian, and further cuneiform languages if these are to be useful. POSTags are necessary for more advanced machine learning classifications such as machine transliteration. POSTags described in linked open data may be linked to tagsets described by treebanks[11] for interoperability between semantic representations and already created treebank resources.

2.4.2 Dictionary representation

Linked data dictionaries combine linked data vocabularies and semantic meanings with morphological representations of words (word forms), their classification, and, if available, their (many) normalized forms. In doing so, they reference vocabularies for classifying word forms, i.e., part of speech vocabularies with vocabularies to represent dictionaries. The most common standard to represent linked data dictionaries is the OntoLex-Lemon model [MBGG+17], which has been recently released as a W3C community report recommendation[12]. It provides the backbone of several Lexeme resources on the web, such as Wikidata[13] [VK14] or BabelNet [NP10]. In short, the following definitions describe the elements of a linked data dictionary:

[11] https://universaldependencies.org/u/pos/
[12] https://www.w3.org/2019/09/lexicog/
[13] https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

Definition 3. ***Lexeme***=*lemon:LexicalEntry*
A Lexeme ($l \in \mathbb{L}$) is a unit of lexical meaning related to one or many syntactical and semantic representations.

A Lexeme (cf. Definition 3) may have many syntactic representations comprised out of one or many words in many different inflections, which are called Lexical Forms (cf. Definition 4).

Definition 4. ***Lexical Form***=*lemon:LexicalForm*
A Lexical form ($lf \in \mathbb{FO}$) is a syntactical representation of a Lexeme in one or many grammatical cases.

Typical deviations of syntactic representations include inflections of verbs or nouns in Sumerian cuneiform and different readings of the same cuneiform word. For example, the Lexeme lugal (wd:L643713), has among others, a plural representation lugal-ene (wd:L643713-F7) and a representation in the genitive case lugal-ak (wd:L643713-F3). Apart from different syntactical representations, the semantic representations of a Lexeme are known as Lexical Senses (cf. Definition 5).

Definition 5. ***Lexical Sense***=*lemon:Sense*
A lexical sense ($ls \in \mathbb{S}$) is a semantic representation of a lexeme's meaning.

For example, in Sumerian cuneiform, a Lexeme comprised of the cuneiform sign AN (wd:L220924), may have the Lexical Senses of sky (wd:L220924-S1) or heaven (wd:L220924-S2). Lexemes make use of, e.g., cuneiform signs to represent their meanings but do not describe the signs themselves in terms of meaning or in terms of their composition, even though the meanings of Lexemes are likely to be related or derived from the meaning or composition of the individual cuneiform sign. The composition and representation of individual cuneiform signs are not a part of the Ontolex-Lemon model.

Definition 6. ***Dictionary***=*lemon:Lexicon*
Dictionary $dict = (l_0, l_1 ... l_n), l \in \mathbb{L}$
$l = (lf_0, lf_1 ... lf_m), (ls_0, ls_1 .... ls_n), lf \in \mathbb{FO}, ls \in \mathbb{S}$

A dictionary is, according to Definition 6 comprised of a set of lexemes, each represented by a set of forms and senses. Dictionaries are created by language so that for cuneiform, there might be relations between words of different cuneiform languages written with the same cuneiform signs but differing in meaning or etymological relations. Linked data dictionaries for cuneiform have not gained traction for the time being and have only been developed as a byproduct of this thesis for Sumerian[14] and Akkadian[15] cuneiform.

2.5 Related Work on digital scholarly editions

Digital scholarly editions are an essential element of the academic discourse in many Humanities disciplines. [Sah16] defines a digital scholarly edition as

[14] https://ordia.toolforge.org/language/Q36790
[15] https://ordia.toolforge.org/language/Q35518

> A scholarly edition is the critical representation of historic documents.

and highlights four main requirements that a digital scholarly edition needs to fulfill:

1. **Representation:** Recoding of a document (...) in the same or another kind of media
2. **Critical:** Incorporation of information that goes beyond the contents of the text (e.g., includes its historical context, features of its visual appearance, etc.) and the criticism and commentary of this information
3. **Documents:** Information about the document that is discussed
4. **Historic:** Editions are created for historical documents only, as current documents are in no need of a critical examination since their context is clear

Digital scholarly editions, i.e.,

> "digital scholarly editions are scholarly editions that are guided by a digital paradigm in their theory, method and practice"

[Sah16], are no new inventions. However, digital scholarly editions are not merely transformations of traditional editorial work. Instead, digital scholarly editions should take advantage of features available in the digital representation but not in a traditional publication. Digital scholarly editions matching these definitions for cuneiform can mainly be found in projects hosted by ORACC, e.g., for the The Royal Inscriptions of the Neo-Assyrian Period [Fra11] (RINAP) corpus[16]. ORACC projects like RINAP usually provide a translation and a linguistic analysis but mostly neglect image media, which might be necessary for cuneiform studies.

The fact that image media are considered in digital scholarly editions can be shown in the example of the Vincent van Gogh letters[17], published by the van Gogh museum, which allows for a display of an image, transcription and a transliteration side by side. In contrast to, e.g., letters provided by the van Gogh museum, which are in specific handwriting, but on a unified medium (paper) and with distinguishable authors, cuneiform artifacts provide more varied paleography and need a sign-by-sign interpretation by scholars. Digital scholarly editions like RINAP allow retracing the second part of the digital scholarly edition, i.e., which readings have been used in the transliteration but are oblivious to the sign recognition part, as image media and especially annotations on image media are usually not included.

Finally, when digital scholarly editions are present in standardized and community-accepted formats such as TEI/XML they are rarely thought of in terms of data sources that can be exploited but rather as scholarly works of individual researchers on textual material that should stand on their own. However, interest in converting digital scholarly edition formats to RDF have been increasing in recent years [BHL+21, CCMG20]. For cuneiform languages, the advantages of a linked data representation of digital scholarly editions have been acknowledged by publications such as [NF18, p. 348],

[16] http://oracc.museum.upenn.edu/rinap/corpus/
[17] https://vangoghletters.org/vg/letters/let003/letter.html

> the Assyriological community has yet to extensively embrace semantic web technologies

but have not been explored further in the respective research communities.

2.6 Summary

The last sections showed the related work concerning different aspects of cuneiform objects, their metadata, and previous work on semantic web vocabularies connected to cuneiform digital objects.

It becomes apparent that there are many ways to represent cuneiform artifacts in different image media, and considerable heterogeneity is present not only in the data representation but also in data formats representing these different aspects. Media representations are not necessarily digitally connected and need special knowledge for interpretation even by human beings, e.g., to find a particular sign on a cuneiform artifact if only its transliteration is given. Digital scholarly editions for cuneiform texts are, despite wanting to live up to their claim of providing enhanced access using digital means, not always fulfilling those ambitions in daily practice. In this sense, this thesis would like to contribute to creating digital scholarly editions that are useful for scholars and provide data that can be useful to researchers of many disciplines.

Transliteration formats and syntax of transliterations vary considerably between different schools of research and language, and the research communities do not always accept cuneiform representations in Unicode. Yet, natural language processing and linked open data provide interesting technologies that can overcome this heterogeneity and enrich the transliteration information without needing to (re-)invent or alter existing syntax specifications. However, as will be detailed later on, not only varying transliteration syntaxes but also lacking specifications in semantic web technologies currently hinder the adoption of linked open data in cuneiform studies. This thesis wants to contribute to overcoming these obstacles in interlinking these different sets of data.

The next chapter will find out about the requirements for such a linked data-driven digital scholarly edition and aims to semantically describe and represent 3D models in linked open data, one medium of representation for cuneiform tablets that have become increasingly interesting and typical for digital scholarly edition projects and which are currently often absent from cuneiform digital scholarly edition environments.

Chapter 3

3D Mesh representation in linked data

This chapter examines the integration of 3D models as one central component of the analysis of cuneiform texts into the linked open data cloud. In most scenarios, integrating 3D model representations directly into a knowledge graph is not advisable, as the size of the 3D model hinders an efficient execution of queries to the database. Instead, bigger files, such as 3D models, are usually linked from the knowledge graph, described as best as possible using linked data models, and only loaded on demand.

In the following, the essential parts of this linked data description, metadata representation, coordinate (reference) systems, and annotations are developed and discussed before they are integrated into the workflow of an Assyriologist in the later chapters. To achieve this, a linked data representation of the 3D model and its metadata is explained in Sections 3.2 and 3.3, after establishing the foundations on 3D meshes in Section 3.1. One notable part of metadata, the coordinate (reference) system, is exemplified in Section 3.4. Finally, an annotation model for 3D meshes in linked data is proposed in Section 3.5 so that interpretations of Assyriologists may be captured as annotations on 3D meshes and may be set in relation to annotations on other mediums.

Three publications complement the ideas presented in this chapter. The first publication [HZBM22] describes best practices for publishing 3D meshes in Assyriology so that they could develop a maximum impact in both the Assyriologist and computer science communities. Its accompanying data publication [HZMB22] shows the examples highlighted in [HZBM22]. The third publication [HCRM21] describes a metadata model for provenance metadata of 3D models.

3.1 Background on 3D meshes

3D meshes as a medium for analysis and publication in Assyriology have been around for many years. In 2001 [WFA+01] conducted the first experiments of capturing cuneiform clay tablets in 3D. Further work has been done with respect to cultural heritage preservation [KCD+03]. 3D meshes have been used by Assyriologists to better examine certain signs on cuneiform tablets and better understand cuneiform artifacts that are not physically available for analysis. This also facilitated sharing of 3D models online [Mar19] so that the same analysis of one Assyriologist could be reproduced by other scientists of the same field or even by computers.

Another usage of 3D models for Assyriologists is the replacement of hand-drawn Line Arts of cuneiform tablet contents with better resolved 3D models or renderings of 3D model surfaces. 3D models usually resolve a detailed picture of the inscriptions in different customizable lighting conditions, avoid errors when creating the Line Art, and provide the bassis for renderings or screenshots that are of sufficient, if not superior, quality for a traditional publication. Information lost compared to Line Arts are interpretations added to the Line Art itself. For example, a cuneiform sign that is only partially visible on the cuneiform tablet may be reconstructed in the Line Art for better comprehensibility for the cuneiform scholar. This loss of information is usually negligible for Assyriologists, as the transliteration of the cuneiform artifact should contain this information in an abstract but equivalent way.

For computer scientists, 3D meshes and their renderings provide interesting but often non-annotated datasets which may be used in machine learning applications, for example, for period classifications of cuneiform tablets [BM20] or as the basis of machine learning challenges [MB19]. Currently, data corpora with cuneiform 3D models are, despite first publications, still in their infancy, and more publications are needed to get sufficient coverage of cuneiform tablets across the most relevant time epochs and relevant find spots. Sufficiently documented and interconnected 3D model representations, which would allow for a more granular creation of training data, are currently nonexistent to the author's knowledge. Hence, this chapter would like to provide the means for an interoperable linked data model to capture the specificities of 3D models.

3.2 Description of a mesh with linked data

To better access 3D meshes in the context of a cuneiform linked open data cloud, 3D meshes first need to be described as linked data entities so that they can be made queryable and accessible for machine-readable and human-readable use. Hence, this section discusses the representation of a 3D mesh in linked data and provides a way to represent 3D meshes in the linked open data cloud. At first rudimentary definitions for the description of a 3D mesh are introduced and set into relation to classes in the upcoming ontology model:

Definition 7. ***Vertex and 3DVertex***
$Vertex\ v = (x_0, ..x_i..., x_n), v \in \mathbb{R}^n, n = |v|,\ x, y, z \in \mathbb{R}$
$3DVertex\ v3d = (x, y, z), v \in \mathbb{R}^3,\ x, y, z \in \mathbb{R}$

Definition 7 introduces points in an n-dimensional space called vertices which are the common building blocks of a mesh. While n-dimensional meshes are the basis of super concepts modeled in the linked data model, for practical purposes, only vertices in 3D are considered in the following.

Definition 8. ***Edge***
$Edge\ e = (v_1, v_2), e \in \mathbb{E},\ v_1, v_2 \in \mathbb{R}^3,\ v_1 \neq v_2$
$Edgelist\ el = (e_0, ...e_i, ...e_n), e_i \in E, n = |el|$

Definition 8 describes a connection between two unequal vertices, which are the basis for the structure of the mesh. This definition is equivalent to a LineString definition of two elements in [PH12]. An edge list is an ordered collection of edge definitions.

Definition 9. ***Point Cloud***=*msp:PointCloud*
$PointCloud\ pc = \{v_0, v_1, ..v_i, .., v_n\},\ pc \in \mathbb{PC},\ v_i \in \mathbb{R}^n, i \in \mathbb{N}_0, n = |pc|$
$PointCloud3D\ pc3d = \{v_0, v_1, ..v_i, .., v_n\},\ pc3d \in \mathbb{PC},\ v_i \in \mathbb{R}^3, i \in \mathbb{N}_0, n = |pc3d|$

Definition 9 defines a point cloud as a set of points in an n-dimensional space. Again, for practical purposes, only point clouds in a 3D space are considered in the following. In this definition, each vertex has been assigned a unique identifier, but in general, points in a point cloud follow no particular order. A point cloud at this level could form the first representation of a cuneiform tablet. Hence it is represented as a class msp:PointCloud in the ontology model.

Definition 10. ***Face and TriangularFace***
$Face\ f = (e_1, e_2..e_i..e_n), e_i \in \mathbb{E}, f \in \mathbb{F}$
$TriangularFace\ trif = (e_1, e_2, e_3),\ e_1 \neq e_2 \neq e_3 \in \mathbb{E},\ e_1.v_1 = e_3.v_2, trif \in \mathbb{TF}, \mathbb{TF} \subseteq \mathbb{F}$
$FaceList\ fl = (f_1, f_2..f_i..f_n), f_i \in \mathbb{F}$
$TriangularFaceList\ trifl = (f_1, f_2..f_i..f_n), f_i \in \mathbb{TF}$

Definition 10 introduces faces of a 3D mesh. Faces are polygons defined by an ordered list of edges that describe the grid structure and, thereby, the composition of the surface geometry of a 3D mesh. The surface geometry could be composed of an arbitrary polygonal base geometry but usually consists of triangles or quadrilaterals. Only triangular meshes are discussed in the following, but the ontology model will consider meshes of all grid types. A mesh is defined in Definition 11:

Definition 11. ***Mesh and 3D Mesh***=*msp:Mesh and msp:3DMesh*
$Mesh\ m = (pc, el, fl), m \in \mathbb{M}$
$3DMesh\ 3dm = (pc3d, el, trifl),\ 3dm \in 3DM \subseteq \mathbb{M}$

A mesh defines a structural 3D build consisting of a set of vertices, edges, and faces determining the mesh's grid structure. For practical uses, the sets of vertices, edges, and faces should not be empty. The vocabulary to represent 3D meshes distinguishes further mesh subtypes categorized by the shape of their grid, i.e., the base geometry of faces. For brevity, these will not be discussed in this section but can be found in the documentation of the ontology model. However, the description of meshes in linked data includes further means to categorize meshes independent of their structure. To enable the categorization of meshes, an ontology model for meshes should include the following most distinct attributes:

- One or many coordinate (reference) system definitions to describe the object Coordinate System (CS)
- A set of metrics describing the properties of a mesh
- The possibly many semantic contexts in which the mesh is used
- A classification of the object type that the mesh represents
- A classification of the mesh grid type

- A variety of links to related mesh types and mesh instances
- A link to the actual mesh content
- A vocabulary of relations between different meshes

To model the aspects as mentioned above, this thesis introduces MeshSPARQL, an ontology model which models these relations and is inspired by the GeoSPARQL vocabulary and its extension GeoSPARQL+ [HSJ20].

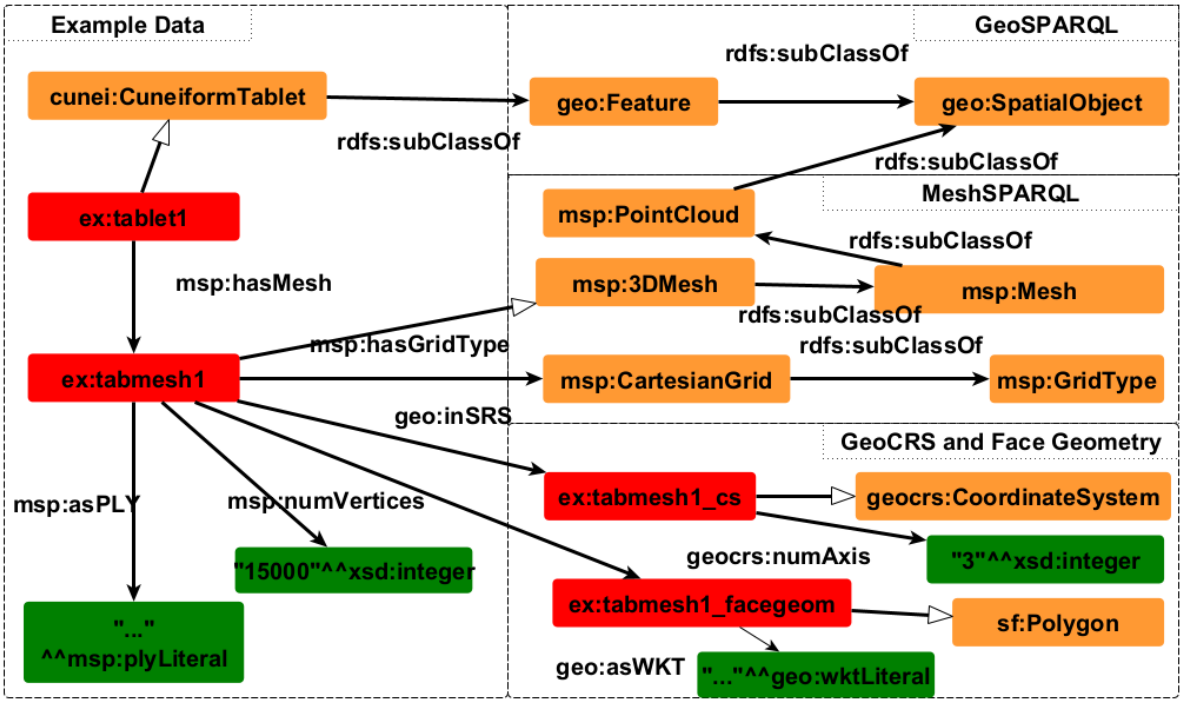


Figure 3.1: Ontology model for mesh descriptions inspired by the GeoSPARQL vocabulary: Mesh representations as instances of subclasses of spatial objects. These are connected to a coordinate system definition (to be defined in Section 3.4), at least one serialization, and a set of properties describing them. The mesh is further described as a spatial feature, that is, by the nature of the object it represents. In this and all following ontology representations, classes are modeled in orange, instances in red, and literals in green, as described in Table A.1

Figure 3.1 shows an overview of the proposed ontology model. It extends the GeoSPARQL class geo:SpatialObject defined in the GeoSPARQL vocabulary with a new geometry type msp:PointCloud, of which a general class of a Mesh (msp:Mesh) is derived. It further defines a class hierarchy of mesh types which describes the main characteristics of meshes, e.g., if the mesh is irregular or which geometric shape the mesh uses as its base form (e.g., polygon, triangular, etc.). In addition, MeshSPARQL can be used to describe metrics of 3D meshes as property relations. For example, the number of vertices, faces, or simply the bounding box of the mesh count as such. A set of these metrics is currently available as an export function in the current development version of the software Gigamesh [MKJB10] and may be used to create a MeshSPARQL-compatible subgraph of mesh metrics.

Listing 3.1: Minimum example of a mesh modeled with MeshSPARQL including two serializations, a CRS reference and some metrics

```
ex:ctablet1mesh rdf:type msp:3DMesh ;
                msp:asPLY "http://www.....ply"^^xsd:anyURI ;
                msp:numberOfVertices "9362"^^xsd:integer ;
                msp:numberOfFaces "7642"^^xsd:integer ;
                geo:hasBoundingBox ex:ctablet1mesh_bbox ;
                geo:inSRS ex:ctablet1mesh_cs .
```

Listing 3.1 shows a minimal example of a represented mesh in the knowledge graph. Three metrics describe a mesh's number of vertices and faces and refer to one instance in the PLY format by defining a URI. The mesh instance's bounding box and its coordinate system's definition are referenced as another URI. While the bounding box may be described using the GeoSPARQL vocabulary, the representation of the SRS system warrants a more detailed discussion in a later part of this chapter.

To relate instances of meshes, MeshSPARQL is extended with a vocabulary for relations. A vocabulary of relations may be understood as describing the relation of two meshes in the same 3D space, i.e., if the meshes touch, intersect, or are disjoint, as is shown in the example in Figure 3.2.

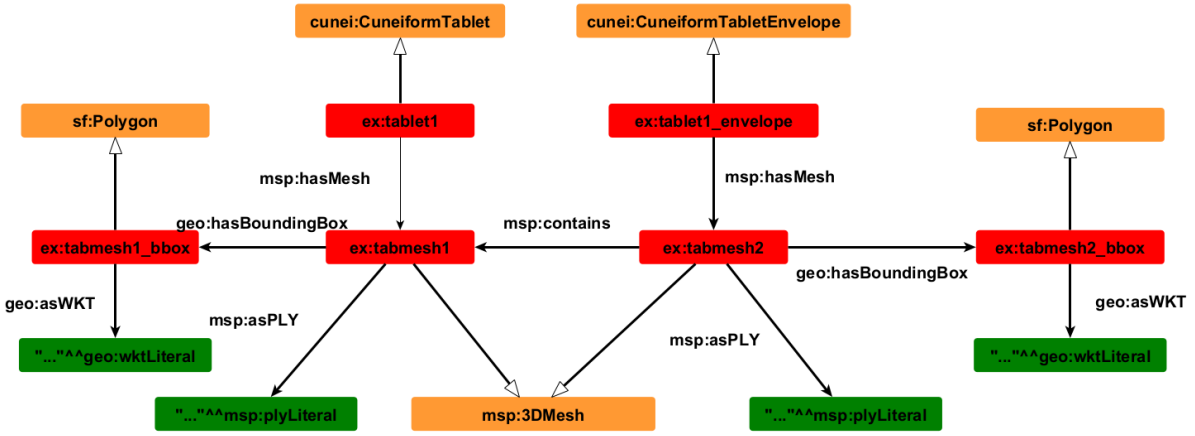


Figure 3.2: Example of the representation of two related meshes of cuneiform artifacts: A cuneiform tablet and its clay envelope that usually encompasses the tablet are represented using two 3D meshes. The clay envelope contains the mesh of the cuneiform tablet. Styles in this graphic follow Table A.1

Vocabularies of relations between geospatial objects have already been defined in related work, for example, in the DE-9IM model [Str08], which describes relations between geospatial objects in a 2D space. Out of these geospatial relations, 3D equivalents msp:equals3D, msp:disjoint3D, msp:intersects3D, msp:touches3D, msp:contains3D, and msp:isContainedBy3D are used in the MeshSPARQL vocabulary.

Another possible interpretation is the position of meshes towards each other if a set of meshes are parts of a bigger scanned object. This direction vocabulary could follow cardinal directions and map positions in the mesh coordinate systems to merge mesh contents. Cardinal directions are modeled with the following terms: msp:above, msp:below, msp:rightOf, msp:leftOf, msp:behind, msp:front. Clearly, this is an insufficiently precise model to relate exactly two 3D meshes in the same coordinate system. Still, it is sufficient to allow for an approximate semantic description of how, e.g., 3D meshes of cuneiform tablet fragments relate to each other. A more precise relation of 3D meshes, e.g., with a boundary surface, is left to future work.

Throughout this chapter, MeshSPARQL provides the basis of how meshes are modeled in linked open data. Further sections of this chapter define extensions to the MeshSPARQL vocabulary, including necessary components to define meshes that are not part of a mesh classification.

3.3 Metadata of 3D meshes

Describing a mesh and its properties is a necessary first step to identifying potentially interesting meshes for a particular application task. Meshes can now be selected and filtered by e.g., their bounding box or the number of vertices. However, an equally important aspect for researchers to evaluate a mesh's suitability is its accompanying metadata. To make research reproducible and document its contents, metadata must be added to a given data publication. The literature distinguishes, among others, the following kinds of metadata applicable to a published 3D mesh [Jef98]:

- Descriptive metadata: Metadata identifying the data entity that it targets
- Structural metadata: Metadata about data formats and their composition
- Reference metadata: Metadata about the contents of the given dataset
- Legal metadata: Information about the license, copyright, and ownership of the dataset

All aforementioned metadata may be applied to a 3D mesh, and related work has already established vocabularies that can be used to model these using semantic web vocabularies [WK00]. However, one essential type of metadata is usually not considered when publishing 3D metadata online: Provenance metadata.

3.3.1 Provenance metadata

Provenance metadata describes the creation process of a 3D mesh from scanning the original object to the eventual publication in a data repository. As one contribution of this thesis, a metadata model to capture precisely this provenance information is introduced. The work has been published [HCRM21][1] and relies on a processing pipeline divided into several stages, shown in Figure 3.3.

[1] http://objects.mainzed.org/

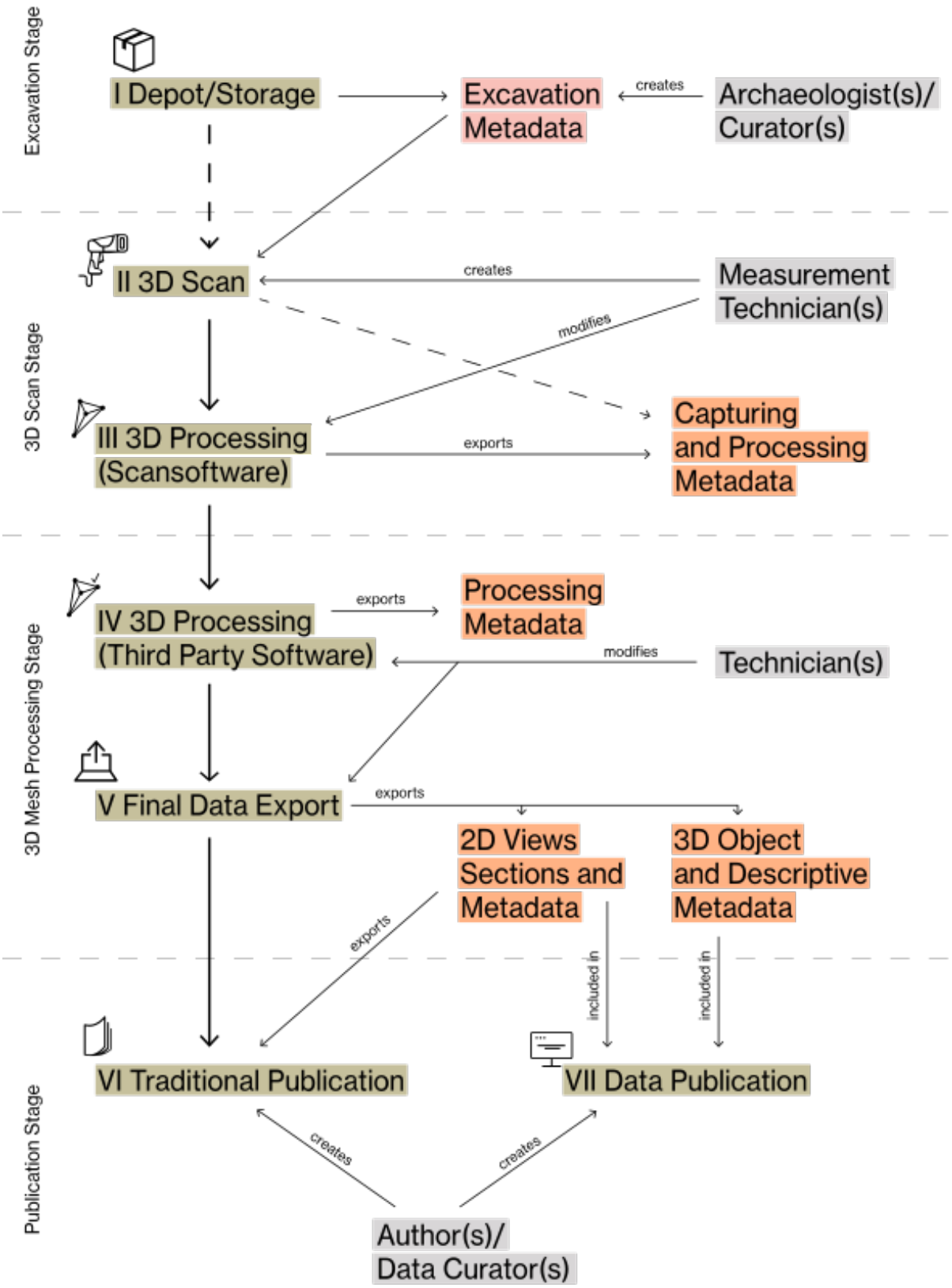


Figure 3.3: The metadata schema for capturing metadata of 3D meshes [HCRM21]

In the first excavation stage, an artifact from a depot/storage possibly derived from an excavation (e.g., a cuneiform clay tablet), modeled as a subclass of cidoc:E22_ManMadeObject, is taken as the basis for further metadata to be enriched. This artifact is expected to be digitally described using excavation metadata, possibly using the CIDOC-CRM vocabulary [Doe05]. Therefore, the result of the excavation stage is a linked data description of the artifact to be scanned, identifying the artifact uniquely. In the second stage, the 3D Scan Stage, this artifact is 3D-scanned using scanning software by a measurement technician in a specific measurement setup environment.

The scanning software will yield capturing and processing metadata that must be appended to the 3D scan creation process metadata, including the technicians' identity, expertise, and the scanning setup's conditions. The result of the 3D Scan stage is e.g., a msp:Mesh or a msp:PointCloud that could be published as is but is likely to be further processed. Hence, until the creation of the final scan result, an arbitrary number of mm:IntermediateMeshResult might be created and documented by the scanning software. In the case study conducted in [HMB21] and afterward, three scanning software could be accessed for metadata of this stage. Results rendered in HTML of three examples are available on Github [CRH22].

The third stage, the 3D Mesh Processing Stage, includes the post-processing steps applied to the given 3D Mesh, e.g., Mesh cleaning or simplification steps. Often, third-party software such as Gigamesh [MKJB10] is used for post-processing. The result of the 3D Mesh Processing Stage is a new 3D mesh derived from the resulting mesh of the 3D Scan Stage. This derivation of meshes, a provenance relationship, is modeled using the provenance vocabulary [LSM13]. After post-processing, the mesh is prepared for publication, i.e., the mesh is assigned a 3D object and descriptive metadata, which will describe the finally created mesh. In this stage, derivations of the to-be-published mesh might need to be created. These might be 2D renderings or renderings of the mesh in different shades or colors. 2D renderings also need to be documented with metadata of the 3D mesh and metadata of their generation process. The accumulation of this provenance information and the addition of it to the linked open data graph is essential for various cuneiform-specific tasks, as it allows researchers to define quality metrics for their respective use cases.

3.3.2 Deducing Data Quality for 3D meshes

A critical aspect of capturing provenance metadata besides mesh metadata is to define data quality indicators for mesh data in general and cuneiform 3D meshes in particular. Data quality can, in this context, be defined as shown in Definition 12:

Definition 12. ***Data Quality:*** *Data that are fit for use by data consumers*

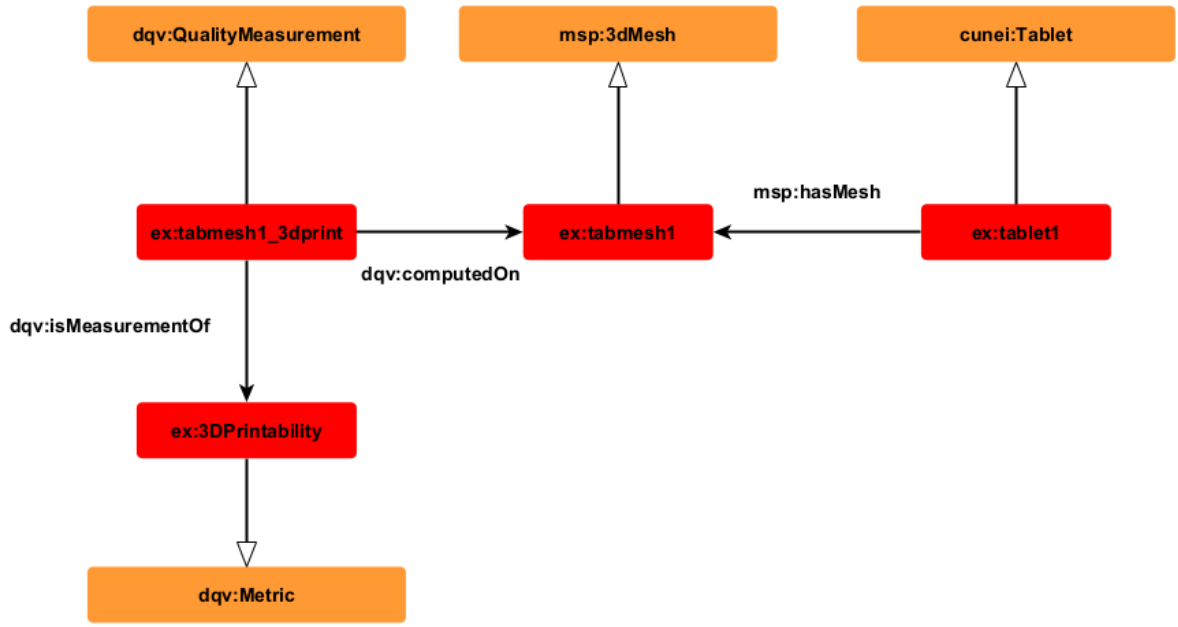


Figure 3.4: Data quality statements in knowledge graphs using the dqv vocabulary [HAI21]. Styles in this graphic follow Table A.1

If Definition 12 is to be applied to the work of a cuneiform scholar, expectations on what constitutes a 3D mesh of good quality may vary depending on the usage context. Most scholars would use the 3D mesh to identify the written text on the cuneiform tablet. For this, the resolution of the 3D scan must be so that cuneiform signs are recognizable. Another application case for a cuneiform scholar might be to print the 3D model using a 3D printer. For this application, the 3D mesh must be printable and fulfill the parameters for printability.

Finally, one can think about more advanced applications such as comparing 3D models of cuneiform tablets, either of the same cuneiform tablets scanned multiple times or comparing annotations on their own. For these comparison operators, knowledge graphs would be expected to include data quality statements rather than only provenance statements, created either from information in the knowledge graph itself (via reasoning) or through information added to the knowledge graph by a third-party application. Figure 3.4 shows how data quality aspects may be modeled using the data quality vocabulary (DQV) [HAI21]. The knowledge graph can define data quality metrics and data quality metric results and semantically accessible preferable statements of suitability. This allows for creating and defining a comprehensive set of data quality metrics for 3D meshes, which should be explored in future work or may be project-specific. Given a sufficient amount of standardized data quality statements, researchers and algorithms alike can judge the suitability of 3D model representations for tasks such as sign recognition or simply to find cuneiform tablet representations that are not scanned with a suitable 3D scanner.

3.4 Ontology model for spatial reference systems

One part of modeling a 3D mesh in linked data is the representation of at least its coordinate system and, if available, its spatial reference system. This information is crucial to understand distances in meshes and distances of meshes to related objects and to enable automated classifications of mesh properties, such as subparts of mesh contents. Apart from one related work[2] [TAA14], which initially modeled coordinate reference systems for name conversion purposes, we lack a representation of spatial reference systems and coordinate systems in the linked open data cloud - especially of previously not standardized coordinate systems such as the one's used in 3D meshes. To that end, the following section will introduce the geocrs vocabulary, which aims to model spatial (reference) systems of spatial data and, in that context, also of 3D meshes.

The complete ontology model[3] has been tested on the EPSG database, one of the largest repositories of (geo-)spatial reference systems available and is subject of an in-progress OGC discussion paper, the first step to an eventual standardization [HKAA22]. The test case contains information on the following components browsable in HTML[4]:

- Definition of local coordinate systems in linked open data
- Definition of spatial references, in general, to relate spatial data to geolocations or to other spatial references describing different spatial data contents
- Description of conversion algorithms between coordinate (reference) systems, in particular, projections
- Definition of reference types and reference bodies (e.g., planetary spheroids)
- Definition of application types for spatial reference systems

While the vocabulary is designed to fit the needs of geospatial data and borrows terms from [Lot15] to be consistent with existing non-semantic standards, this thesis focuses on the necessary components of the ontology model to describe 3D meshes in the following.

3.4.1 Defining spatial reference systems

At first, spatial reference systems need to be defined to be used in the linked open data model.

Definition 13. ***Spatial Reference System=geocrs:SpatialReferenceSystem:*** *System for identifying the position in a spatial context*

A spatial reference system (cf. Definition 13) consists of two parts: Its coordinate system (Definition 14) and, if available, its spatial reference. At first, coordinate systems are defined here, spatial references are described in Section 3.4.2.

[2] http://data.ign.fr/def/ignf/20190213.en.htm
[3] https://situx.github.io/proj4rdf/proj4rdfextracted.html
[4] https://situx.github.io/proj4rdf/data/def/crs/EPSG/0/4326/

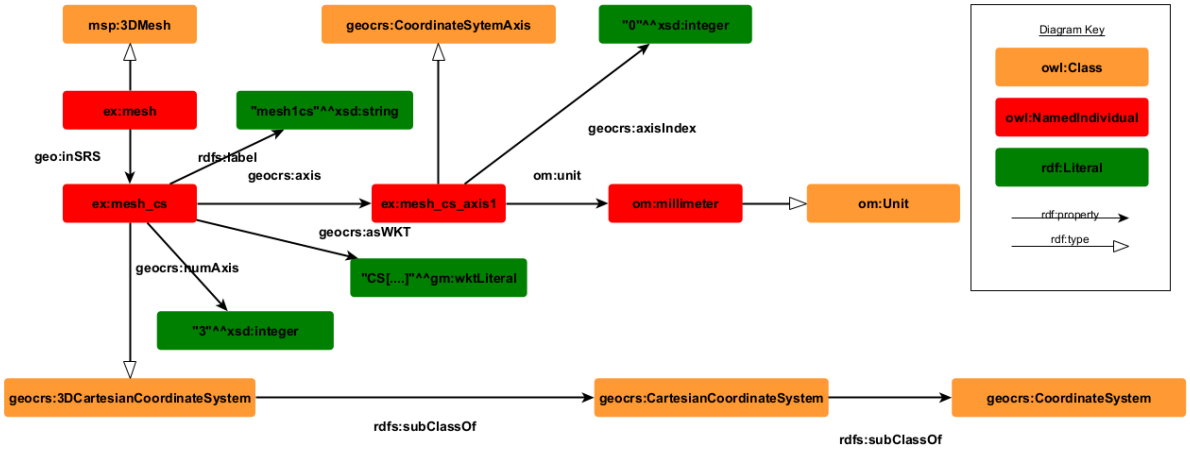


Figure 3.5: Representation of a coordinate system for a given mesh using a newly defined vocabulary for coordinate systems and the om Units of Measurement ontology [RVAT13] for representing units.

Definition 14. ***Coordinate System=geocrs:CoordinateSystem:***
CoordinateSystem $cs = (ax_0, ax_1, ...ax_i, ...ax_n), i \in N_0, n = |cs|, n > 0$

A coordinate system (cf. Definition 14) is a non-repeating sequence of coordinate system axes (cf. Definition 15) that span a given coordinate space.

Definition 15. ***Coordinate System Axis=geocrs:CoordinateSystemAxis:***
A reference coordinate line that originates in the coordinate system origin, with a defined direction and a subdivision of distances with a given unit

The fact that most mesh data formats usually do not capture the particularities of their given coordinate system makes the definition of a coordinate system with or without a spatial reference necessary in the provided mesh metadata. Otherwise, opening the same 3D mesh in different software might require manual user input, such as the unit of the coordinate system. In automated processes, this information is very hard to provide, primarily when the metadata of the meshes does not reveal coordinate system information and when meshes need to be compared.

Related work on modeling coordinate reference systems in the Geosciences provides data formats and specifications on how to represent coordinate reference systems, i.e., coordinate systems with a reference to a spheroid approximating a planet. The Well-known text [Lot15] (WKT) format for describing coordinate reference systems and the proj4 and PROJ libraries [War06] are the de-facto standards for defining coordinate reference systems. The WKT representation is also common in 3D processing software such as Agisoft Metashape[5] for only defining coordinate systems as sidecar file definitions, but without a semantic web vocabulary. To prepare WKT for semantic access, its semantics need to be modeled explicitly to enable, e.g., filtering of mesh contents by parameters of the coordinate system, such as its type, axis definition, or axis unit. Figure 3.5 shows this case's application of the GeoCRS ontology model. A 3D cartesian coordinate system

[5] https://www.agisoft.com

is expressed as an RDF instance connected to a mesh instance In certain instances, a 3D scan of a cuneiform artifact might need to be georeferenced. This happens when a 3D scan is conducted of a non-mobile object, such as a part of a temple wall, or when a mobile cuneiform artifact is arranged at a fixed position along with other artifacts over a more extended period of time. In these cases, the ontology model shown in Figure 3.5 needs to be extended by a reference component, which references coordinates in the mesh coordinate system to coordinates in a world coordinate system.

3.4.2 Modelling general spatial references

In this case, the ontology model must be extended to accommodate a georeference. While not a core topic of this thesis, a draft to create such an ontology model with a conversion of the most critical database of coordinate reference system definitions, the European Petroleum Survey Group Geodesy [NS08] (EPSG) database, has been converted as a proof of concept and has been published as proj4rdf[6] on Github.

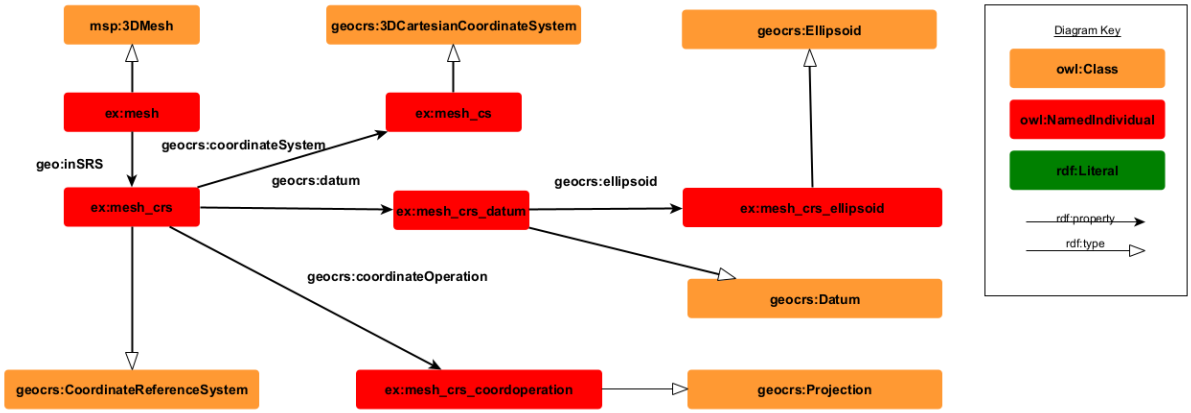


Figure 3.6: The coordinate system of a 3D mesh with RDF instances needed to add a georeference, including a geodetic datum instance with an ellipsoid and the definition of a transformation function from the local coordinate system to the world coordinate system (projection).

Figure 3.6 shows how to extend the coordinate system description of a 3D mesh with an added georeference in principle. This will allow relating representations of cuneiform texts on non-mobile objects, such as inscriptions on temple walls, to their real location on planet Earth. For practical examples of coordinate reference systems modeled in this way, the EPSG use case of the GeoCRS model can be consulted[7].

[6] https://github.com/situx/proj4rdf
[7] https://situx.github.io/proj4rdf/data/def/crs/EPSG/0/4326/

3.5 Annotation model for 3D cuneiform

Especially in cuneiform studies, a 3D mesh is a valuable asset for the scientific discourse. Properly published on the internet, it allows all interested researchers to access a precise digital representation of the cuneiform tablet and to draw their own conclusions about their respective research areas. To enable such a scientific and formalized discourse in the digital age, researchers need to be able to create specific statements about sections of the respective object, called annotations, which are commonly further discussed in the relevant specialist communities. An annotation in the context of this thesis is defined in Definition 16.

Definition 16. ***Annotation:***
Additional information which is related to parts of a digital object

Annotations may be added by the original creator of the digital object before an initial publication or may be created and related by anyone after the digital object has been published. Annotations may be stored alongside a 3D mesh, linked to a 3D mesh, or even be present inside a 3D mesh file. Besides the provision of the 3D mesh itself, annotations provide added value, which helps interpret specific parts of the cuneiform tablet depicted in the 3D mesh. While creating an annotation in a 2D space may be considered relatively straightforward (a subarea of the image in question delineated by a 2D bounding box in pixel coordinates), annotations in the third dimension may take shapes depending on their purpose. This thesis presents two different types of annotations practiced on 3D models and explores how these annotations can be created in principle and can finally be represented in linked data.

3.5.1 Annotations on 3D models

There are different ways to annotate 3D models, which depend on the shape of the annotation and the purpose of the annotation in question.

3.5.1.1 Annotation by a bounding cuboid

This type of annotation (cf. Figure 3.7) creates a minimum bounding cuboid around the area on the 3D mesh to be annotated. The shape of the annotation is very simple, and so are its representation and storage. They can be stored as GeoSPARQL Geometry String literals relative to the mesh coordinate system and can be created externally.

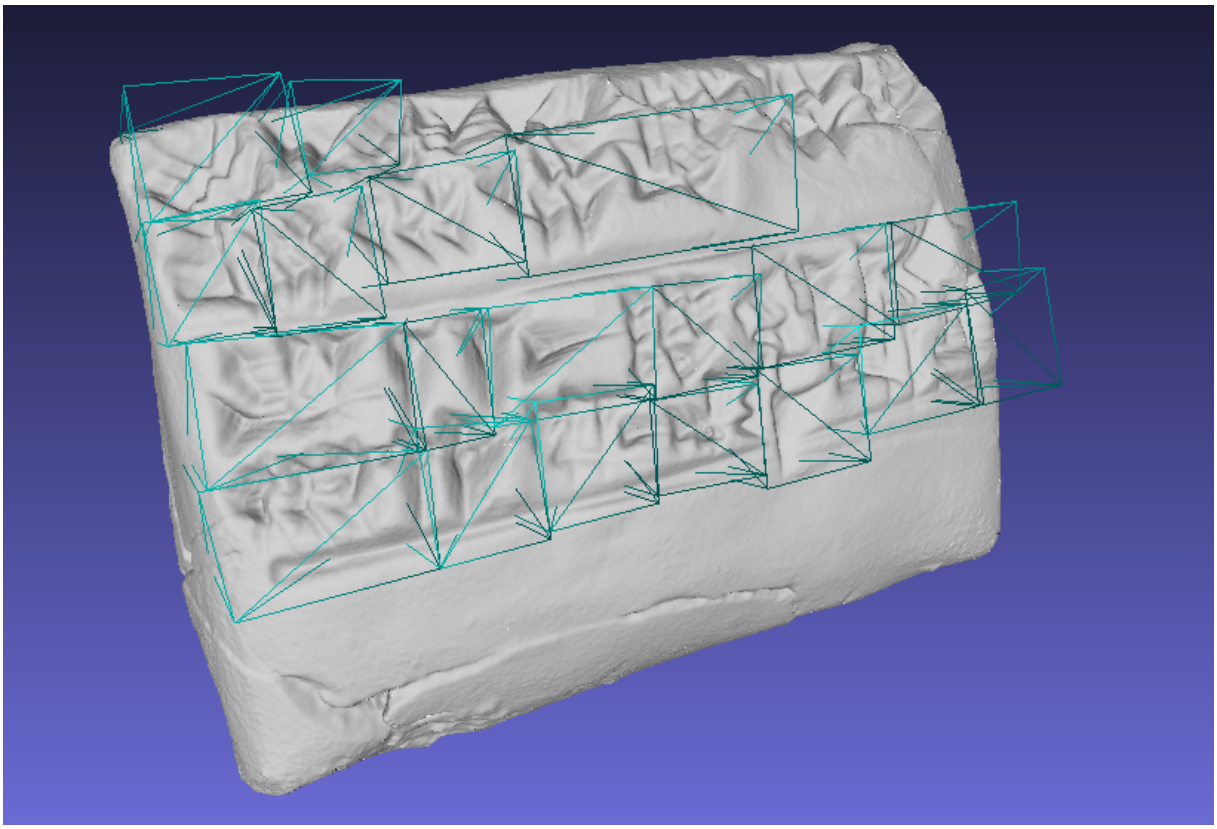
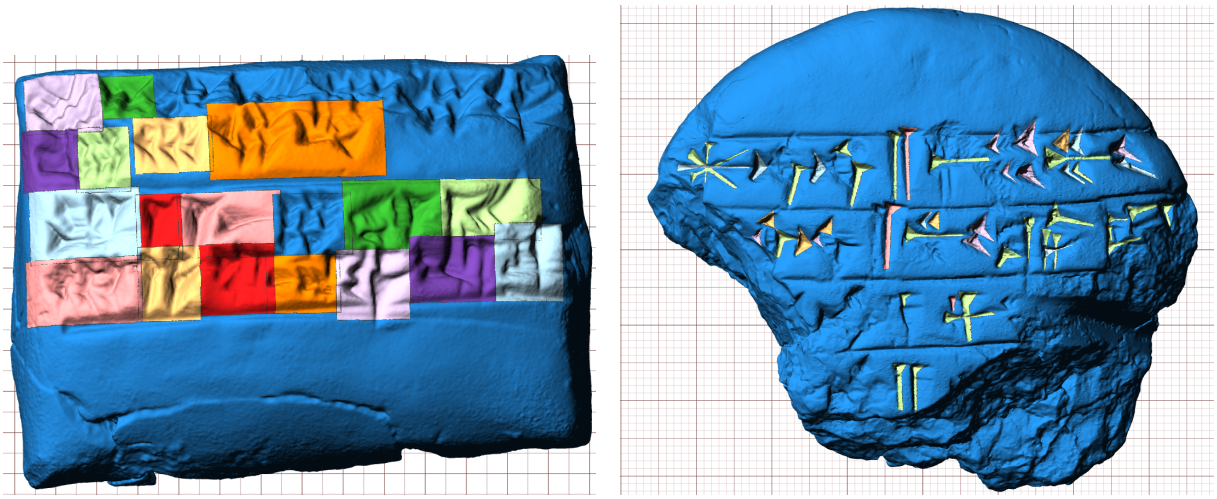


Figure 3.7: Annotation by bounding cuboid example: Bounding cuboids are created around cuneiform signs with a fixed height. This allows a precise location of inscriptions on the surface of the cuneiform tablet.

A drawback of bounding cuboid annotations is that they may not be precise enough for, e.g., sign classification tasks. In addition, cuboid annotations do not encompass the actual mesh contents which they are targeting, but rather only deliniate an area of the 3D mesh in which the cuneiform sign is present. If the contents of bounding cuboids should represent cuneiform signs, a better method would be to capture the point set and face set, which captures the shape of the cuneiform sign in question. The method proposed in Section 3.5.1.2 next introduces one kind of annotation which allows for a more precise expression of areas of a cuneiform mesh.

3.5.1.2 Annotation by labeling

This annotation method labels components of a given mesh (cf. Figures 3.8a and 3.8b) and associates the labeled components with annotation content. Each point of a given mesh associated with a given annotation is assigned a unique ID. The sum of all faces, which includes labeled points, comprises the given annotation.



(a) Annotation by labeling example 1: Cuneiform signs are labeled in the 3D mesh and highlighted using a particular color code in the Gigamesh application

(b) Annotation by labeling example 2: Single cuneiform wedges are labeled in the 3D mesh and highlighted using color codes per recognized wedge type in the Gigamesh application

Figure 3.8: Annotations by labeling: Examples of single wedge and sign annotations on 3D meshes by coloring vertice areas

This method of annotation in 3D meshes allows the selection of exact volumes of parts of the 3D mesh (e.g., the volume comprising a cuneiform wedge) that are relevant for classification tasks. Even the annotation of single wedges is possible and can be used for classification tasks if deemed applicable. However, this annotation depends on the 3D mesh being labeled in the first place. Hence, this annotation type is likely to be used by the original mesh creator as a form of pre-publication annotation. Further additions or modifications of labeled mesh components require users to re-publish the given mesh.

3.5.2 Defining annotations in linked data

One standard way to represent any kind of annotations in linked data is the usage of the W3C Web Annotation Data Model [CYS17], which was published with the idea in mind to make annotations interoperable, compatible, comparable and shareable across the web. The W3C Web Annotation Data Model has been adopted by various tools, such as Annotorious[8] for image annotation, or Recogito[9] for text annotation, therefore enabling its widespread usage as a data exchange format.

[8] https://annotorious.github.io/
[9] https://github.com/recogito/recogito-js

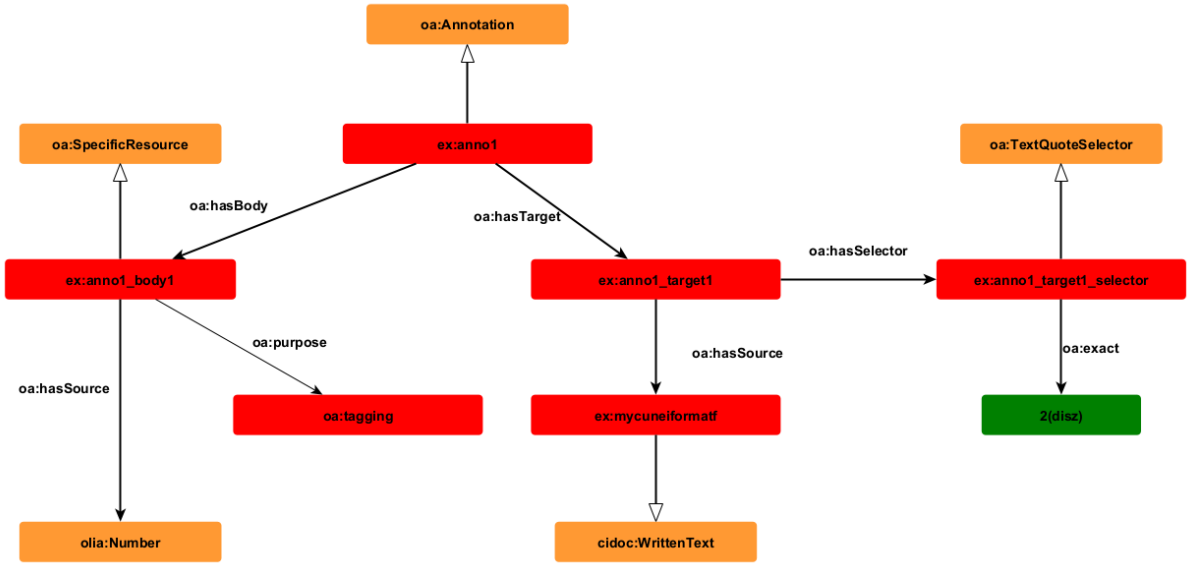


Figure 3.9: W3C Web Annotation Data model annotating a Part of Speech Tag for a number in a given text: The oa:Annotation is comprised of one or more many annotation bodies with annotation contents such as URIs or values and an optional annotation purpose (here oa:tagging). An annotation might have one or many annotation targets. Here, a cidoc:WrittenText is defined as the target source. The exact annotation selection is defined by a oa:TextQuoteSelector. Styles in this graphic follow Table A.1

An annotation in the W3C Web Annotation data model (cf. Figure 3.9) consists of an annotation definition identified by a URI connected to a set of annotation bodies and a set of annotation targets. Annotation bodies contain the contents of the annotation. Annotation targets specify the resource(s) to be annotated. If a subpart of a given resource is annotated (e.g., a text passage on a homepage), the annotation target needs to identify this subpart using an appropriate oa:Selector definition. One example of an annotation selector is the SVG [DFS+11] selector (cf. Listing 3.2). This selector may, for example, be used to select areas on a given JPEG image [Ham04].

Listing 3.2: SVG Selector as defined in the W3C Web Annotation Data Model, used to annotate a road on a given JPEG image representation on a map

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/anno27",
  "type": "Annotation",
  "body": "http://example.org/road1",
  "target": {
    "source": "http://example.org/map1.jpg",
    "selector": {
      "type": "SvgSelector",
      "value": "<svg:svg> ... </svg:svg>"
    }
  }}
```

An SVG selector defines an area on a 2D image. Hence, its definition is limited to a 2D space in the SVG coordinate system definition, and only 2D coordinates can be used to describe an annotation space. In the following, this definition will be expanded to encompass 3D annotation targets and selectors.

3.5.3 Defining selectors for 3D mesh annotations

This section proposes an extension to the W3C Web Annotation Data Model that will define new selector types to express the 3D annotation types mentioned previously.

3.5.3.1 WKT 3D Selector

At first glance, the definition of a new selector type for the W3C Web Annotation data model might look straightforward. The selector would simply need to include the coordinates in the mesh coordinate system to uniquely identify a position in the mesh. To achieve the representation of the mesh coordinates, the SVG format is unsuitable, as it cannot represent 3D coordinates. Therefore, the msp:WKTSelector represents 3D coordinates in a Well-Known Text string, including one of the following Well-Known Text geometry types: Polygon Z, MultiPolygon Z, Point Z, or LineString Z [H+11]. Next, a reference to a representation of the mesh coordinate system needs to be given. This reference might be given with the definition of the mesh itself, as defined in Section 3.4.2 or might be defined in the annotation if the annotation coordinate system differs from the mesh coordinate system.

Listing 3.3: Annotation of the bounding cuboid of the first cuneiform sign in line 1 on the obverse side of a cuneiform tablet on a given 3D mesh using a WKT selector referencing the mesh coordinate system

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/cuneianno1",
  "type": "Annotation",
  "body": "http://example.org/tablet1_front_line1_char1",
  "target": {
    "source": "http://example.org/mymesh.ply",
    "selector": {
      "type": "3DSelector",
      "srs": "http://example.org/mymesh_cs",
      "value": "POLYGON Z (...)"
    }
  }
}
```

Listing 3.3 shows an example of a msp:WKTSelector defining a bounding cuboid in WKT. This annotation is external; it can be hosted independently of the given mesh.

3.5.3.2 Labeling Selector

The following selector type allows referencing labeled areas in a 3D mesh. As established previously, most mesh formats allow the representation of unique IDs to be added to vertices of the given mesh format. However, even though labeled areas are defined, they lack semantic descriptions, which are in most mesh formats not supported [HZBM22] and may be defined by a third-party online.

Listing 3.4: Annotation of the bounding cuboid of the first cuneiform sign in line 1 on the obverse side of a cuneiform tablet on a given 3D mesh

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/cuneianno1",
  "type": "Annotation",
  "body": "http://example.org/tablet1_front_line1_char1",
  "target": {
    "source": "http://example.org/mymesh.ply",
    "selector": {
      "type": "MeshLabelSelector",
      "value": "v[labelid=10]"
    }
  }
}
```

Listing 3.4 shows an example of a msp:MeshLabelSelector targeting the same 3D mesh as in Listing 3.3. The mesh label selector contains a specific String literal, the msp:meshselectorLiteral expression. The syntax of the msp:meshselectorLiteral allows for the selection of mesh components which may be addressed by their names as variables and defines the following keywords:

- **v**: Selector for vertices
- **f**: Selector for faces

Keywords and variables may be combined with the following operators:

- Selection operator []: Allows to select a specific subset of faces or vertices
- Comparison operators $>, <, <=, >=, =$: To filter vertices and faces based on other attributes assigned to them
- Algebraic operators: $+, -, *, /$ to manipulate numeric expressions
- Concatenation operator $+$: To concatenate two expressions inside a msp:meshselectorLiteral

In Listing 3.4, the expression **v[labelid=10]** is used to select precisely the number of vertices that are associated with a property ”labelid” which has a value of 10. For brevity, the formal syntax specification of the msp:MeshLabelSelector will be left to future work, which would also need to define a parser grammar and eventual implementations.

3.5.3.3 Mesh reference selector

The third selector type introduced here is the msp:MeshReferenceSelector. This selector targets the last common way of sharing mesh annotations: Mesh annotations exported as new mesh instances.

Listing 3.5: Annotation of the bounding cube of the first cuneiform sign in line 1 on the obverse side of a cuneiform tablet on a given 3D mesh using a mesh reference

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/cuneianno1",
  "type": "Annotation",
  "body": "http://example.org/tablet1_front_line1_char1",
  "target": {
    "source": "http://example.org/mymesh.ply",
    "selector": {
      "type": "MeshReferenceSelector",
      "srs": "http://example.org/mymesh_cs",
      "value": "http://example.org/mymeshannos/anno1mesh.ply"
    }
  }
}
```

Listing 3.5 shows one example of the msp:MeshReferenceSelector. Its value is a URI which refers to the mesh annotation in a common mesh format. As the referenced annotation mesh only acts as the target of the annotation, the mesh format does not need to feature annotation-specific components. The selector can be extended by a coordinate (reference) system and coordinate operation definition, as shown in Section 3.4, if the coordinate system used in the annotation differs from the mesh coordinate system and requires parameters for their conversion.

3.5.4 Ensuring reusability of annotations

A common problem when working with 3D mesh data is that new 3D meshes might be created from the same artifact at some point. There may be a variety of reasons for this. Meshes might be recreated to monitor the decay of the original artifact [HMB21] because new scanning technologies allow the creation of better mesh representations or simply because other scanning software has been used to modify already existing mesh data. However, mesh annotations created on a previous version of a 3D scan are not necessarily compatible with a newly scanned version, as the mesh coordinate system parameters and the coordinates representing corresponding points in the two versions might differ drastically. To overcome this obstacle and ensure the reusability of the aforementioned 3D annotation types, which depend on coordinates, computed references on the actual 3D mesh artifact data can help approximate the position of annotations on a new scan of the same cuneiform artifact.

3.5.4.1 Computing References: Principal Component Analysis (PCA)

A Principal Component Analysis (PCA) [MR93] of a collection of points (e.g., in a 3D mesh) yields a set of vectors, called the principal components, which are defined in Definition 17:

Definition 17. ***Principal Components:*** *The principal components of a set of points in a given coordinate space are a sequence of unit vectors p, whereas the ith vectors follow the direction of a line that describes the given set of points, and each $i-1$ vector is orthogonal to the previously defined ones.*

Suppose a principal component analysis is performed on a 3D mesh. In that case, it can be configured to yield exactly three vectors, which can be seen as akin to the coordinate system axis of a coordinate system based on these PCA vectors. These coordinate system axes will likely deviate from the coordinate system assigned by the scanning software. Still, they can be expected to be reproducible, as they are based on the coordinate system generated from the data in the actual 3D mesh. Hence, a PCA-based coordinate system can act as a computed reference. In linked data, this reference is expressed as a msp:ComputingReference, such as is shown in Listing 3.6.

Listing 3.6: Computing Reference of a 3D mesh representing a cuneiform clay tablet

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/cuneianno1",
  "type": "Annotation",
  "body": "http://example.org/tablet1_front_line1_char1",
  "target": {
    "source": "http://example.org/mymesh.ply",
    "selector": {
      "type": "MeshReferenceSelector",
      "srs": "http://example.org/mymesh_cs",
      "value": "http://example.org/mymeshannos/anno1mesh.ply"
    },
    "computingReference": {
        "type":"PCAReference",
        "value":"[[....]]",
        "stable":true
    }
  }
}
```

While a PCA calculation in this way will always yield three vectors that comprise the vectors of a PCA coordinate system, a coordinate system definition requires an ordering and naming of its axis. A viable way to achieve an ordering of PCA vectors to coordinate the system axis is to assign the longest vector the designation of the X axis, the second longest vector the designation of the Y axis, and the remaining vector the designation of the Z axis. This method will work unambiguously for a large set of objects, including typically formed cuneiform tablets, given a sufficient length difference between the different vectors.

A problem with this assignment is posed by, for example, round objects, which will yield PCA vectors of almost equal or equal length. In that case, the computing reference is to be considered as not stable and is designated as such in the knowledge graph by the property msp:stable with values true or false. A stable indicator with the value "true" means that the computing reference is sufficiently stable for the transformation of annotations to a new scan, and a stable indicator with the value "false" means that other computing reference methods, for example, based on mesh saliency calculation [NALM20] could be explored. Suppose no automated or semi-automated way to create a computing reference is available, but a rescan of a given mesh has been conducted. In that case, the meshes might need to be manually compared and adjusted. The result of such a comparison may be a transformation function, which might be modeled in the GeoCRS vocabulary.

3.5.5 Annotation vocabulary

This section presents the beginnings of an annotation vocabulary that can be used to annotate contents on cuneiform tablets. This annotation vocabulary is to be seen as a starting point for further development and covers essential elements found on most cuneiform tablets and is published under the namespace http://www.purl.org/cuneiform/annotation.

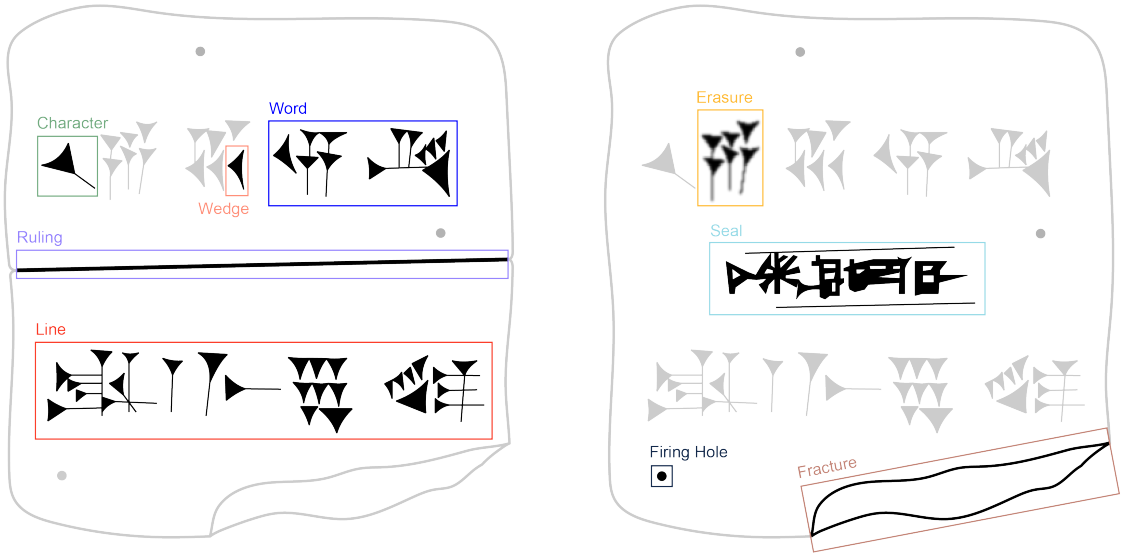


Figure 3.10: Areas of interest concerning annotations on cuneiform tablets, as published in [HZBM22]. These areas are of interest regardless of a 2D or 3D representation and constitute commonly remarked parts of the cuneiform by Assyriologists

Figure 3.10 gives an overview of these basic elements, which are also formalized as classes in the ontology model. These contain columns, lines, characters, and words on a cuneiform tablet's surface, seals, broken areas, rulings, and firing holes. The annotation of cuneiform

signs and single wedges will be most important for the contents of the upcoming chapters. To describe these sufficiently in an annotation, their position on the cuneiform tablet, relative to a transliteration, their orientation (e.g., is rotated by, e.g., 90°), and their physical properties (e.g., their color and depth) need to be captured. The annotation vocabulary allows capturing these elements as annotation bodies present in the Web Annotation Data Model. (cf. Listing 3.7)

Listing 3.7: Annotation body that classifies an annotation af as a cuneiform character, referencing to the annotation vocabulary. Further classifications yield more annotation bodies.

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/cuneianno1",
  "type": "Annotation",
  "body": { "id":"http://example.org/tablet1_front_line1_char1",
            "purpose":"tagging",
            "value":"cunei:Character"
  },
  "target": {
    ...
  }
}
```

A more fine granular development of this vocabulary would need to be consulted by experts in the field, especially concerning research on cuneiform seals [Rofrm-e1] and possibly with colleagues of the material sciences. However, annotating the examples using the web annotation data model would remain applicable for all available annotation targets.

3.6 Annotation transformation

The previous sections have discussed how to represent 3D meshes, their metadata, and which kinds of 3D annotations can be represented in linked open data. This section will discuss the creation process of 3D annotations, especially concerning minimizing the workload for Assyriologists. One idea for achieving this is to take advantage of 2D annotations on renderings generated from 3D models [HZMB22].

The rationale is: If 3D models are available for assessing cuneiform artifacts, 3D renderings of all sides of the cuneiform artifact may be a technically easily accessible way for annotation, as software for image annotations in 2D images is readily available. Creating annotations on 3D renderings also arguably removes the burden of creating 3D annotations, as 2D annotations can be projected to cuboid or volume representations in the 3D model. This section introduces the tooling and two possible ways to convert 2D annotations created on renderings back to the original 3D model.

3.6.1 Annotation of 2D renderings generated from 3D models

At first, 2D annotations need to be created from 2D renderings, at best in consolidation with the transliteration workflow. Figure 3.11 shows an example of an annotation on a 2D

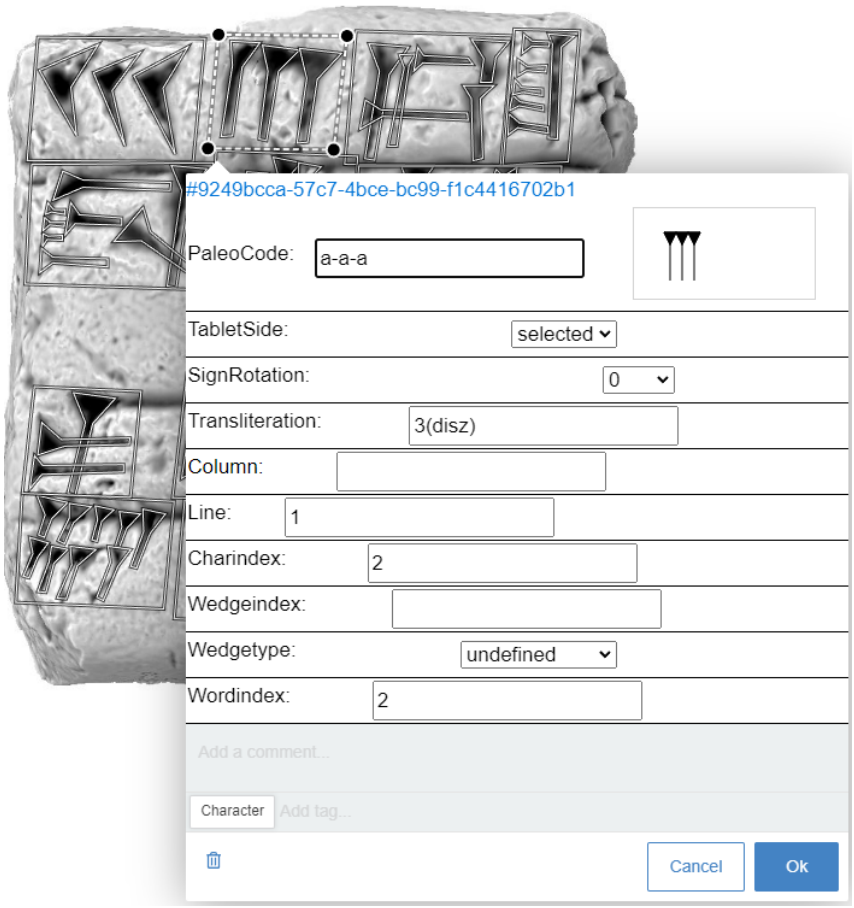


Figure 3.11: One example of a cuneiform tablet rendering that has been annotated on a sign level. The content of the annotation bodies reveals referencing information and description information of the cuneiform sign in question

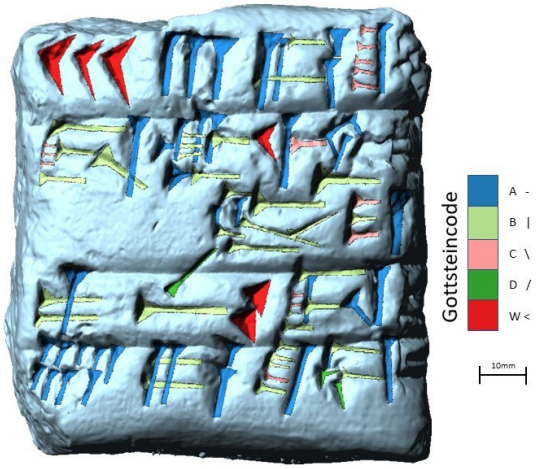
rendering. The annotation includes referencing information to relate an image annotation on the 2D image to a place in the corresponding transliteration. It may use the annotation vocabulary introduced in Section 3.5.5 to classify annotations on these renderings. In this way, the Assyriologist can create a transliteration beforehand and reference the image annotations in a later step once it has been established how many lines, words, and cuneiform signs have been recognized on the cuneiform tablet surface. Image annotations of cuneiform signs are indexed by their annotation content, i.e., the annotation bodies need to define a line index and a cuneiform sign index relative to the selected cuneiform line as defined in the transliteration. In addition, a sign name or its transliteration value is added in the image annotation so that it may be used independently of a linked transliteration. This interconnection allows applications to highlight cuneiform signs by

hovering over each respective medium (e.g., text or transliteration) and in linked data to collect these different representations for the tasks at hand. Once annotations on 2D renderings sourced from 3D models have been created, these annotations can be converted to 3D representations.

3.6.2 Postprocessing of annotations: Deriving 3D annotations

Image annotations are saved in the JSON-LD format using the W3C Web Annotation Data Model, as described in this chapter. They may be picked up by postprocessing scripts that perform image cropping and 3D conversion tasks.

(a) Cuneiform tablet HS1174 annotated in 2D. Annotations include cuneiform sign annotations and cuneiform wedge annotations. Only the front surface is visible in this depiction.



(b) 2D annotations of HS1174 transformed to 3D using the algorithms described in this section. Here, labeling in the 3D mesh has been used for representation. Colors have been chosen according to a detected wedge type

Figure 3.12: Input and output visualizations of the conversion of 2D annotations to 3D on a sample cuneiform tablet

To derive 3D annotations from 2D rendering annotations (cf. Figures 3.12a and 3.12b), two algorithms have been proposed and used in the accompanying publications [HZMB22, HMB21].

The first algorithm converts the X/Y coordinates of the 2D bounding box of the 3D rendering annotation to X and Y coordinates in the mesh coordinate system. This transformation is a transformation of scales per coordinate and requires the following inputs:

- The minimum and maximum X or Y values P_{min} and P_{max} of the 2D rendering image in pixel coordinates
- The minimum and maximum X or Y values M_{min} and M_{max} of the 3D model in mesh coordinates
- The X or Y coordinate C to be converted

The scale conversion value C_{conv} is then calculated using the following formula:
$C_{conv} = ((C - P_{min})/(P_{max} - P_{min})) * (M_{max} - M_{min})) + M_{min}$

Subsequently, a second algorithm extracts the Vertex Set of mesh coordinates within the X/Y coordinate bounding box. The coordinate set is then filtered to return the maximum and minimum Z values around the targeted mesh surface used to determine the annotations' bounding cuboid. The return value is a set of coordinates delineating the bounding cuboid as required by Section 3.5.3.1.

The returned point set may then be indexed using a triangulation algorithm [BE95] to achieve a mesh representation of the annotation volume. The result of this algorithm might be used to label the mesh, i.e., to act as a basis for msp:MeshLabelSelector, or to create a new mesh instance which could act as the target of a msp:MeshReferenceSelector. Either way, the annotation data may be added to the knowledge graph using one of the previously described 3D annotation types.

If cuneiform signs span more than one cuneiform surface, for example, if they have been written around the edges of cuneiform tablets, one annotation per surface will be present in 2D. The algorithms must be executed per annotation, and their results can be merged later or labeled with the same identifier.

3.7 Summary and Discussion

This chapter has investigated the requirements to represent 3D meshes in linked data vocabularies and how to properly annotate images and 3D models of cuneiform tablets. The solution provides the MeshSPARQL ontology model (cf. Section 3.2), which for the first time can capture essential properties of a 3D mesh in a linked open data graph. Applications of the ontology model in [HCRM21] and as an implementation in the Gigamesh Software Framework[10] showed the feasibility of using the ontology model in practice. However, adopting the ontology model on a larger scale will depend on the availability of a crucial number of implementations, repositories, and a perceived need for interconnectivity of mesh data by the respective research communities. Also, as the name MeshSPARQL suggests, the increased exposure of 3D meshes in linked open data would warrant an investigation into extending the SPARQL query language for operators capable of processing mesh data (parts). This question is left to future work in this thesis. Still, it is likely to be discussed in the standardization efforts of GeoSPARQL 1.3 query language, which targets the inclusion of 3D data of any kind.

The provenance data model for 3D meshes (cf. Section 3.3.1) shows, using two examples, that capturing information on 3D mesh data is a necessity when sharing 3D data becomes more common in web repositories, even independent of the application case of cuneiform tablet representation. The provision of provenance metadata is even more complex, as exports of scanning metadata need to be accessible in the scanning software of the respective 3D scanner. Companies may choose not to disclose this information and

[10] https://gitlab.com/fcgl/GigaMesh/-/tree/develop

hinder the metadata collection using the ontology model. To adopt the ontology models, a standardization process of an organization such as W3C or OGC would greatly help accelerate this process and strengthen the possibilities for adoption in products in the scanning industry.

The same applies to describing coordinate reference systems definitions in linked data (cf. Section 3.4). The ontology model serves the need to describe coordinate systems and even coordinate reference systems for cuneiform tablets and spatial objects in general. Still, software is needed to take advantage of these representations. This need, however, is not necessarily new. Many scanning software, such as Agisoft Metashape[11], can already process Well-Known Text information, so an extension to include further encoded coordinate reference system data might not be too far off. Thinking of coordinate reference system representations in general, these would also need to be supported by triple store implementations that process linked open data. Current solutions to represent (geo)spatial data in linked data usually rely on additional relational databases of predefined coordinate reference systems in Well-known text [Lot15] (WKT). A standardization effort, as envisioned in the OGC GeoSemanticsDWG, should normalize the usage of coordinate reference system definitions in spatial knowledge graphs and, therefore, also motivate developers of triple stores to read these definitions from within a knowledge graph and to provide additional SPARQL extension functions to convert and compare spatial objects in even previously unknown and, especially common for meshes, local coordinate reference systems.

Equally crucial for the representation of 3D meshes in linked open data is the definition of its annotation model in Section 3.5. With the extension of the W3C Web Annotation Data model with three different annotation selector types for 3D mesh contents, users have three possibilities to create annotations for 3D meshes which are compatible with all currently known 3D mesh formats. This lays the foundations to create a semantic web of 3D annotations in general and for the cuneiform scholar community in particular, as it enables the comparison of annotated contents, the creation of very targeted statements about cuneiform artifacts in the third dimension, and finally, a scientific discourse about the annotation contents. Again, the challenges here lie in finding appropriate hosting services for 3D models and the annotation contents and insufficient support to create annotations in yet-to-be-developed applications.

Section 3.6 proposed an annotation process that can convert 2D representations of mesh surfaces to 3D representations. While this process works for many cuneiform tablets, as the signs are usually depicted on one of the surfaces, annotations of cuneiform signs around the edges need merging algorithms of duplicated 2D annotations, which annotated signs written around the corner of cuneiform tablets. Depending on the needs of scholars, the reverse process, that is, annotating cuneiform 3D meshes directly in a 3D viewer, is thinkable. This would allow researchers the freedom to annotate areas on cuneiform tablet surfaces irrespective of signs being written across them. In this thesis and in practical work, however, it was perceived that the access to tools that allow the annotation on 2D images is usually easier to understand, whereas 3D annotations, depending on the tooling, might need a more precise and possibly more error-prone annotation method.

[11] https://www.agisoft.com

As a final aspect, Section 3.5.4 showed in the example of PCA that the transformation of mesh annotations between 3D mesh rescans using additional references derived from the 3D model itself is possible for various 3D meshes, but in particular for meshes of regularly shaped cuneiform tablets. As mesh analysis algorithms evolve, further references, for example, from machine learning algorithms, are possible. They would only add to a growing pool of metrics that can be used to transform existing annotations into new scans of cuneiform tablets correctly. Thus minimizing the effort of annotations to be recreated for a limited amount of experts who are the only people able to do so.

Sign annotations alone and as 3D volumes are a necessary and significant contribution to Digital Assyriology but are of limited use without a way to make the paleography of cuneiform signs that are annotated machine-readable and accessible. The next chapter will introduce a way to achieve this for cuneiform signs independently of the given language.

Chapter 4

LOD representation of cuneiform paleography

The analysis of Paleography in cuneiform studies is one of the main decisive factors in identifying texts, classifying a cuneiform text's historical and linguistic context, and is invaluable for documenting its transliteration. Paleographic particularities are often remarked in the comments of transliterations and are the subject of a variety of works on the paleographic analysis of respective text corpora [Ell02, Win70, Wee18, Pop16, SH07]. A missing gap in digital cuneiform studies is how to represent these paleographic features digitally, describe variants of cuneiform sign writings in data, and systematically document them over time and space. This chapter introduces an ontology model and linked data representation of cuneiform signs and their paleographic features in a linked data environment. After introducing the foundations of cuneiform paleography in Section 4.1, a character encoding of cuneiform signs, PaleoCodage, which can be used to create abstractions of cuneiform signs (graphemes) for machine-readable comparison, is introduced in Section 4.2. With a digital representation of sign variants, it is possible to integrate sign variant representations in already established ontology models for dictionaries. Section 4.3 discusses how this can be achieved. Finally, Section 4.4 outlines how to integrate paleographic information in already established transliteration formats while preserving the linked data representation and connecting it to existing approaches of annotations for texts.

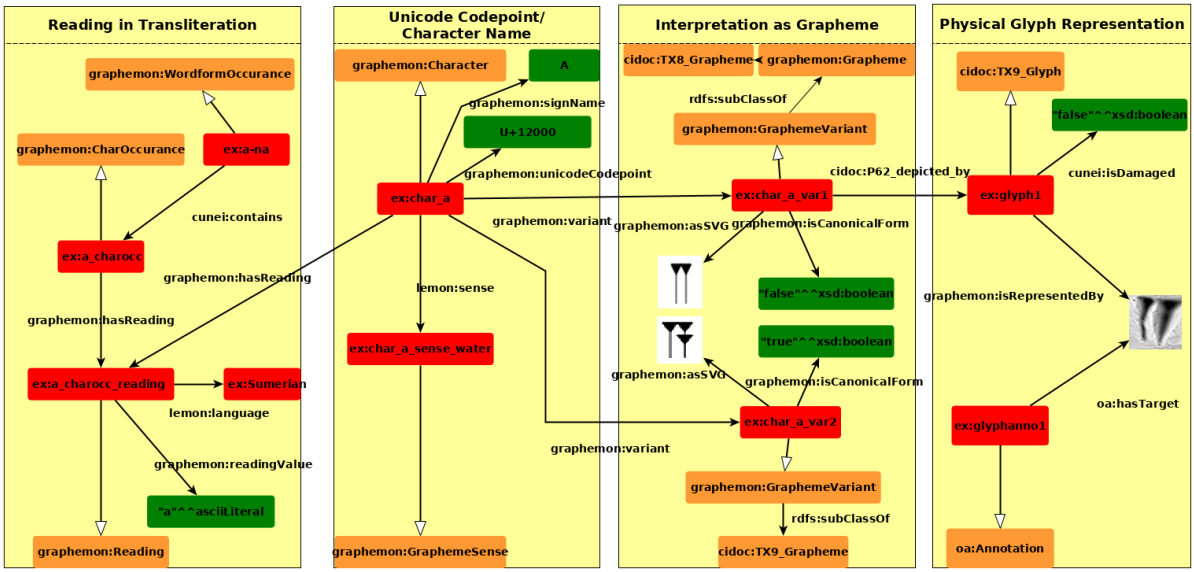


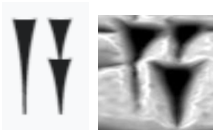
Figure 4.1: Relation between glyphs, graphemes, characters and Unicode Codepoints, character names, and transliteration representations. A reading in a transliteration can be 1:1 related to a Unicode code point, which may describe senses, dictionary references, and further information about the cuneiform sign. The Unicode Codepoint is related to a set of graphemes depicting the Unicode representation. A grapheme is related to an arbitrary number of Glyph representations on cuneiform tablets.

Figure 4.1 shows a general relation between important elements discussed in this chapter. In a digital representation, we distinguish a glyph, i.e., a physical representation of a cuneiform sign on a given medium, typically a clay cuneiform tablet. This definition does not include a digital representation. However, glyphs may be visualized using images, renderings, or 3D models and, in these forms, may also be targets of annotations. Glyphs are grouped and abstracted in the form of graphemes. Graphemes are idealized representations of the forms in actual glyphs on cuneiform tablets, retaining their most important semantic components. If a grapheme is common, it is likely related to a Unicode code point, its machine-readable representation, and rendered in many epoch-specific cuneiform fonts. However, any cuneiform font can only incorporate one grapheme per character or Unicode Codepoint, limiting the expression of grapheme variants mapped to the same Unicode code point. Per the Unicode cuneiform proposal [Con22], each Unicode code point is assigned a unique Sign Name, often derived from one of its Sumerian readings, which is also used in Assyriology to describe a cuneiform sign. From this Sign Name, the Assyriologist derives readings of signs in the assumed target language of the cuneiform tablet. More specifically, the Assyriologist will assign the reading of a Lexeme in the transliteration, which fits the perceived meaning of the sign in the context of the given line/text on the surface of the cuneiform tablet. In certain instances, even this reading in the transliteration is not the final iteration, as passages on cuneiform tablet surfaces may be subject to interpretations in different languages written in cuneiform.

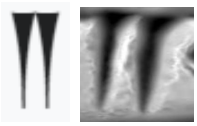
After this general overview, it is important to understand the approximate amount, variety, and nature of grapheme variants that exist for a single Unicode Codepoint. Because of this, already existing approaches to collect and describe grapheme variants are summarized in the following.

4.1 Background: Paleography in cuneiform languages

Many researchers in the past have conducted paleographic studies on cuneiform texts, and paleographic particularities may help Assyriologists to date cuneiform texts to identify scribes of cuneiform tablets by their writing style. Due to a large variety of cuneiform sign variants, it is often not entirely possible to identify a cuneiform sign only by its shape. Often, the context of the cuneiform sign needs to be considered by the researcher to make the claim of classification of the cuneiform sign. Depending on the context, a cuneiform sign with the same amount and positioning of wedges might be classified as a different sign.



(a) The cuneiform sign A with its standard form once as grapheme and once as an actual occurrence in the cuneiform text HS 367, front side, column 1, line 3, sign 4



(b) The cuneiform sign A with an alternative form is more common in older cuneiform texts once as grapheme and as an actual representation in HS 1163, back side column 1, line 14, sign 4. This form also resembles the cuneiform sign for the number two 2(disz).

Figure 4.2: Example of sign variants of cuneiform sign A on specific cuneiform tablets

Figures 4.2a and 4.2b show an example of the cuneiform sign A, which consists of three vertical cuneiform wedges in a common version and of two vertical cuneiform wedges in a variant form. The variant form, however, has the same shape as the cuneiform sign for the number 2 (MIN).

While Assyriologists can infer cuneiform signs from their context, a computer might not take this deduction as easily. The first step in enabling a computer to make suggestions for sign classification has to be machine-readable documentation of cuneiform sign variants so that the computer can learn how to classify sign variants in the future. In Assyriology, mainly two works would depict the state of the art in paleographic research. [Lab02, Fos26] are two books that described paleographic sign variants in detail, [Fos26] even using an indexing system. As the basis of the Unicode proposal, the Borger sign list [Bor04] gives a universally accepted list of cuneiform signs, but not necessarily of all sign variants. These works are often consulted when interpreting texts to clarify sign variants and their occurrences in different time periods.

In the following, the requirements to digitize these cuneiform sign representations, to make these representations machine-accessible and comparable, as well as their representation in knowledge graphs, are discussed as a means to facilitate not only searching for cuneiform sign variants for Assyriologists but also to enable a computer to make its deductions about the materials provided by a cuneiform digital scholarly edition.

4.1.1 Image resources vs. abstracted representations

Cuneiform signs can be represented digitally in two main different ways. The first way is the representation of cuneiform signs with image resources. For example, cuneiform signs could be annotated on a JPEG photograph and then cropped as one JPEG image per annotation. This representation provides a very accurate representation of the relevant cuneiform sign on the cuneiform tablet and can be used as a resource, e.g., for machine learning classifications. A second way is choosing an abstract representation of a cuneiform sign in a digital medium. Abstracted representations of cuneiform signs have always been created in Line Art and recently created as SVG drawings. These may arguably already be used as a resource for abstracted cuneiform signs but may, depending on the drawing style of the individual Assyriologist, still be quite heterogeneous in shape. In addition, SVG drawings, if not created sign by sign, lack a consistent machine-readable character description language, as it is common practice in, e.g., Chinese [BC03] for font generation.

4.1.2 Character description languages

A character description language is an encoding that can capture a structured, regular script and allows its reproduction using an algorithm. For many non-alphabetic languages composed out of strokes, such as Japanese or Chinese, encodings have been proposed to describe their character composition. The Chinese character description language [BC03] can compose Chinese characters for font generation. Similar character description languages like Kanji Vector Graphics (`https://kanjivg.tagaini.net`) (KanjiVG)[1] exist for Japanese. To the author's knowledge, fonts for cuneiform languages [Píš05, ML17, Pí12] have been based on either SVG drawings or JPEG images of cuneiform signs. Hence, unlike the Chinese character description languages, they have not relied on character description languages to describe their respective cuneiform signs. Images will accurately represent the character in question but do not encode semantic information about the context of the character and its composition, which would be essential for a proper digital representation of structured scripts.

To date, despite the structured nature of cuneiform signs, no sign description language that could capture the essential parameters of the shape of a cuneiform wedge has been proposed. The next section will introduce such a sign description language, which will help encode cuneiform signs and integrate them into a knowledge graph.

[1] `https://kanjivg.tagaini.net`

4.2 PaleoCodage: Digital Representation of Cuneiform Paleography

In Section 2.3.2, several methods to represent cuneiform in Unicode, in font representations, and in images have been described. The representation of paleography for cuneiform characters is of utmost importance for the Assyriologist to accept a digital representation of cuneiform script, apart from image resources. Cuneiform signs vary in shape and form over their thousands-year-old history. It is not uncommon to find different cuneiform sign variants on the same cuneiform tablet in the same place dated to the same time period. Even being represented by the same Unicode code point, the same character described might not even remotely resemble an image given in a font, not even if it is specified for the same time period. The reasons for this variety of sign forms are rooted in the extreme diversity of cuneiform writing materials across space and time, as described in Chapter 2.

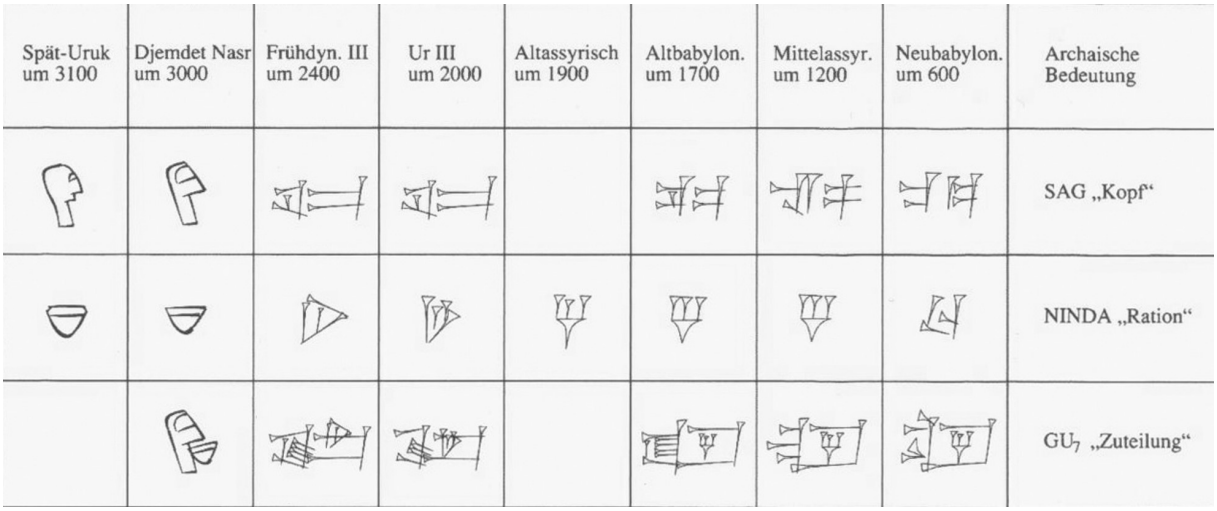


Figure 4.3: The etymology of selected cuneiform characters as described in [Lab02] from earlier pictographic representation to more recent depictions using cuneiform wedges

Figure 4.3 shows the development of some cuneiform signs from their first attestation to the changes they were subjected to over the centuries. One can observe a trend of abstraction of a pictographic or ideographic representation of a cuneiform sign to a more simplified representation of the latter, as in many other similar scripts, such as Chinese or Japanese [Fun19]. While the pictographic representations of the earlier centuries are hard to formalize, a machine-readable representation of later cuneiform sign variants seems like a feasible prospect, as they are built from one single atomic component: The cuneiform wedge.

4.2.1 Developing the PaleoCodage encoding

Related work [Pan15] already showed that using cuneiform wedges as atomic components makes cuneiform sign variants searchable. To achieve indexing of cuneiform signs for search purposes, they utilized the Gottstein system [Got13]. The Gottstein system describes cuneiform signs according to four different wedge types, as shown in Figure 4.4.

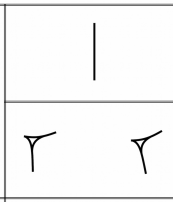
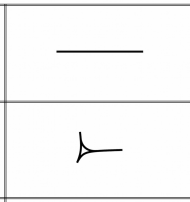
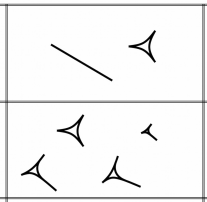
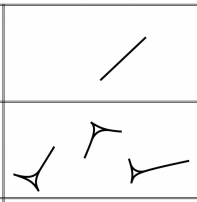
the elements					Sign EME 
designation	a	b	c	d	abc
parameters	sum of the designations and the indices				a3 b5 c1
category	number of elements				9 = 3+5+1

Figure 4.4: The Gottstein System [Got13] defining four different wedge types and describing the cuneiform sign EME

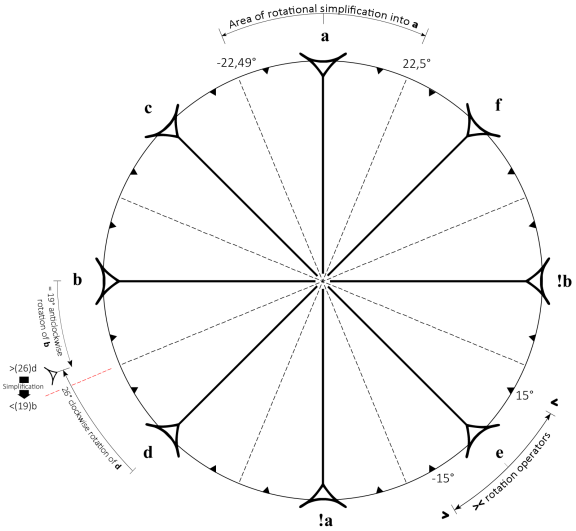
Gottstein defines the wedge type **a** to represent a vertical wedge, the wedge type **b** to represent a horizontal wedge, wedge-type **c** to represent either the Winkelhaken wedge or a diagonal wedge from the upper left to the lower right and the wedge type **d** to represent a wedge going from lower left to upper right. The number of wedges per type of a given cuneiform sign becomes the sign variant's identifier. While the Gottstein system works sufficiently for implementing a search functionality to find cuneiform sign variants, it has several shortcomings:

- The Gottstein System does not encode the positions of the cuneiform wedges towards each other
- The sizes of the wedge head and the length of the wedge stroke cannot be encoded
- The encoding cannot model broken individual wedges
- The system cannot represent parts of signs which are repeated in other signs

In addition, it does not uniquely model cuneiform sign variants, as seen in Figure 4.5a.

Sign	Sign Name	Gottstein Code
	MASZ	a1b1
	BAR	a1b1
	LAL	a1b1
	ME	a1b1

(a) Cuneiform signs described with the same Gottstein Code of one vertical wedge (a1) and one horizontal wedge (b1): The Gottstein system is usable as a search system for cuneiform signs, but not to describe a cuneiform sign variant unambiguously



(b) PaleoCodage wedge description model [Hom21]

Figure 4.5: Comparison between the Gottstein system and the PaleoCodage encoding

To accommodate these shortcomings, this thesis presents the PaleoCodage encoding, [Hom21], which extends the Gottstein encoding by various expressions. The core of the PaleoCodage encoding is its wedge description model, shown in Figure 4.5b. It proposes different identifiers for wedges depending on their position on the unit circle. Certain frequently occurring wedges are assigned a character akin to the Gottstein encoding. These adopted wedge types, which are still present in PaleoCodage, are shown in Table 4.1:

Wedge Type	Description	Image
a	Vertical wedge	
b	Horizontal wedge	
c	Diagonal wedge type 1	
d	Diagonal wedge type 2	

Table 4.1: Wedge types adopted and adjusted from the Gottstein encoding. Wedge type **c** has been redefined not to include the Winkelhaken wedge.

In addition, PaleoCodage introduces the wedge types described in Table 4.2:

Wedge Type	Description	Image
e	Diagonal wedge type 3	
f	Diagonal wedge type 4	
w	Winkelhaken wedge	
x	half-rounded stylus imprint	
y	full stylus imprint	

Table 4.2: New wedge types introduced to the PaleoCodage encoding

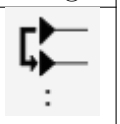
The addition of the aforementioned wedge types allows capturing the diagonal direction of every cuneiform wedge as a vector description. In addition, it distinguishes the Winkelhaken wedge exclusively as a new type compared to the Gottstein model. The wedge types **x** and **y** are less frequently occurring wedges that describe half-rounded and rounded stylus impressions, mainly occurring in earlier cuneiform texts.

These defined wedge types represent the most frequently occurring wedge positions on the unit circle [EMW+32]. Still, they leave room for the definition of less frequently occurring variants, which can be modeled with the rotation operators <> on the unit circle, as shown in Figure 4.5b. For example, the expression >**a** describes a horizontal wedge rotated by 15° in the clockwise direction. In comparison, the expression <**a** describes the same wedge rotated by 15° in the counterclockwise direction. With the option to represent a cuneiform wedge in every position on the unit circle, the next section discusses a set of necessary operators to describe the relations between individual wedges.

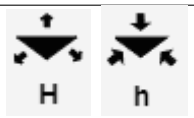
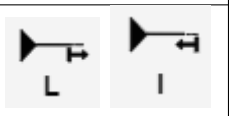
4.2.2 PaleoCodage operators

The PaleoCodage model extends the Gottstein model with a few additional operators partially inspired by the Manuel de Codage Encoding [Goz13, BGH+88] for Egyptian Hieroglyphics. The main motivations of this character description model are as follows:

The description of wedge sizes and individual wedge components (wedge head and wedge stroke), the relation of wedges towards each other, a description and the option of reusage of reoccurring wedge parts, and a better classification of wedge types based on their function. In the following, the main operators of PaleoCodage, including some examples, are presented. The full range of operators and application cases is described in detail in [Hom21]. At first, a set of positioning operators is introduced in Figure 4.6a:

Operator	Description	Image
:	above operator	
-	right of operator	
.	diagonal operator	
[]	reusage operator	[A]

(a) PaleoCodage positioning operators:
Allow the positioning of cuneiform wedges next to each other (**-**), above each other (**:**) and diagonally (**.**).
So-modelled signs can be reused in other sign components with the [] operator, therefore simplifying the construction of compound signs

Operator	Description	Image
H/h	Head size	
L/l	Line length	
G/s	Wedge size	
{}	Aggregation	s{a-a}

(b) Wedge modification operators: Wedge heads (**H/h**), wedge lines (**L/l**) and complete wedges (**s/g** or wedge type in capital letter) may be modified in steps of sizes, as wedges are represented bigger and smaller in relation to other wedges in the same cuneiform sign. The aggregation operator (**{}**) allows the modification of wedge groups

Figure 4.6: PaleoCodage operators for positioning and modification

These operators allow for the positioning of wedges relative to each other, as referenced in standard literature [EMW+32]. That is, PaleoCodage demands the representation of each individual wedge with a specific wedge-type character. The sequence of wedges determines how the cuneiform sign is written. For example, consider the cuneiform sign A introduced in Figure 4.2a. It consists of three cuneiform wedges of type **a**. To represent this sign in PaleoCodage, the sign is constructed from left to right and up to down. That is, at first, a cuneiform wedge of type **a** is represented, and right of (**-**) this wedge, a second cuneiform wedge of type **a** can be observed. The second cuneiform wedge is located above (**:**) a third cuneiform wedge of type **a**. The resulting PaleoCode is, therefore, **a-a:a**.

For many cuneiform signs, it is known that they occur as parts of other cuneiform signs. For example, the cuneiform sign A occurs as part of the cuneiform sign A TIMES A.

Figure 4.7: Cuneiform sign A TIMES A. The cuneiform sign A is reused in the middle of another cuneiform sign

Figure 4.7 shows this occurrence of the sign A in the middle of A TIMES A. To model this sign as a PaleoCode and to explicitly describe that another sign definition occurs within the sign variant described, the PaleoCode could look as follows: **a-[A]-a:a**. This PaleoCode uses the reusage operator [] to reproduce an already-known PaleoCode in its place. The reusage operator takes a sign identifier assigned by a cuneiform sign registry and allows it to be used instead of the PaleoCode for the respective sign. In case the reused sign is smaller or bigger in size or rotated, operators from the PaleoCodage wedge description model can be applied even to signs inserted with the reusage operator. In the case of Figure 4.7, the PaleoCode could be modified to **a-s[A]-a:a**, to acknowledge the fact that the included sign A is smaller than comparable wedges in the same cuneiform sign. Further examples showcasing more capabilities of the PaleoCodage encoding can be found in Appendix B.

4.2.3 PaleoCodage normalization

The PaleoCodage encoding aims to solve two crucial goals. The first goal is the human readability and accessibility of a PaleoCode. After some initial training, a human should be able to read a PaleoCode and understand its meaning. The second goal is that the PaleoCode is machine-processable and convertible to other common representations such as SVG. A machine-readable code should be unique and reproducible. In contrast, a human-readable code may not necessarily need to be unique when the code is constructed but should be convertible to a unique and minimal representation that a machine may understand. Therefore, specific normalization steps must be taken before a PaleoCode can be saved in a paleographic database.

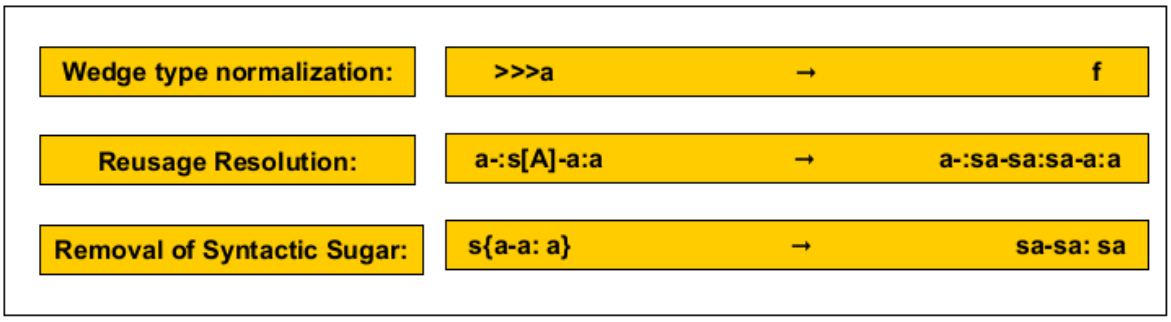


Figure 4.8: Examples of PaleoCodage normalization rules to convert human-readable PaleoCodes to PaleoCodes stored by a database. Wedge types are normalized according to the wedge description model (example 1), reusage operators are resolved to PaleoCode representations (example 2), and syntactic sugar is removed for the machine-readable representation (example 3).

While the complete set of normalization steps is documented in [Hom21], Figure 4.8 shows three examples of common normalizations which can be automatically applied to each PaleoCodage representation. The PaleoCodes described in these examples also include the **s** operator, one of the stylistic operators which change the size of individual cuneiform wedges. Here, the **s** operator serves as one example of an application for reuses and syntactic sugar in PaleoCodes.

4.2.4 Applications of PaleoCodage

The PaleoCodage encoding may be used in two main ways. The first way is an intuitive way of documenting paleographic features in a cuneiform digital scholarly edition. Every digital scholarly edition of a cuneiform text corpus will curate a sign list, which expresses the particularities of the given excavation with its specificities. If the sign list is curated manually, sign variants are drawn by hand and often added to the respective publication, e.g., as an appendix. If sign variants are captured using a computer, drawings will happen in software, and results will often be saved in image formats. Both representations will capture the shape of the cuneiform sign variants but not the relevant aspects of its structure, that is, which cuneiform wedge types are included and how they relate to one another. Modeling cuneiform signs using PaleoCodage allows for the export of cuneiform sign images and a machine-readable structured representation. The machine-readable part of the encoding is a suitable match to be included in, e.g., machine learning training sets that propose the classification of cuneiform signs. Also, encodings allow for a comparison of cuneiform signs using reproducible metrics. For example, PaleoCodes may be compared using String similarity metrics such as the Levenshtein Distance [YB07] to find similarly shaped graphemes of cuneiform signs as shown in Figure 4.9.

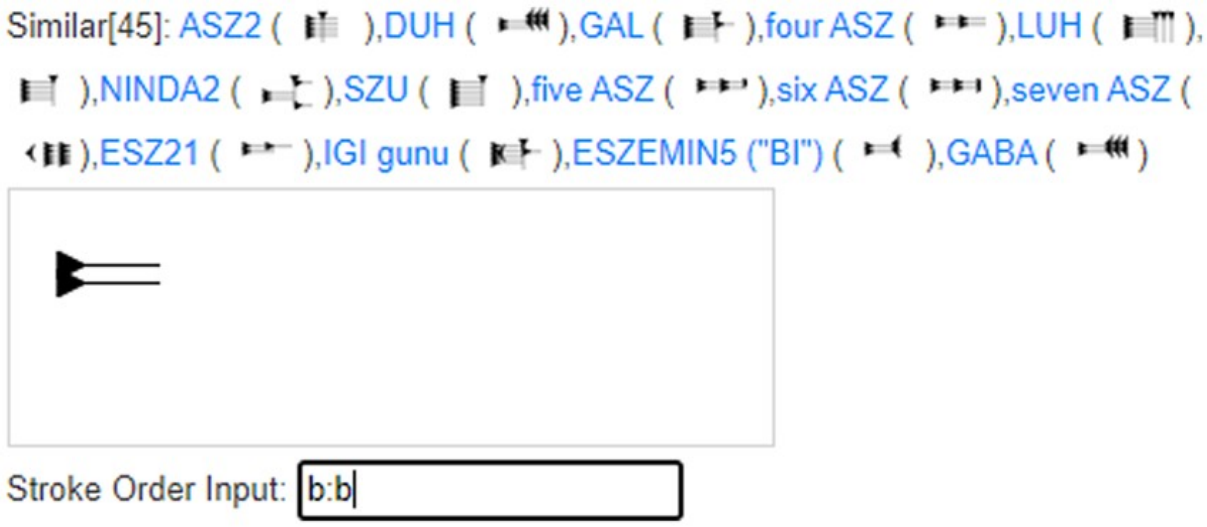


Figure 4.9: Cuneiform sign similarity using String similarity metrics based on the PaleoCodage encoding. This image shows the application of a simple substring matching metric, matching the occurrence of the PaleoCodage pattern of the cuneiform sign TAB in other encoded cuneiform signs.

However, PaleoCodage can only provide the means to encode cuneiform sign variants. In the respective research projects, scholars must determine how PaleoCodage is used in practice. Questions like: How precise should cuneiform signs be modeled, e.g., to which degree on the unit circle and to which degree does an abstraction from cuneiform glyph to grapheme need to happen are tasks to be solved by the respective scholars and by curators of cuneiform sign collections. However, assuming a consensus has been reached about how PaleoCodes should be used in the community, PaleoCodage may be used to solve a longstanding problem of reconciliation between Unicode cuneiform signs and cuneiform sign variants, as will be elaborated on in the next section.

4.2.5 Creation of PaleoCodes

This thesis mainly discusses the PaleoCodage encoding itself and sees the development of (semi-)automatic input methods for PaleoCodes as future work. While the results of this thesis feature tools that may convert a keyboard input of a PaleoCode to an SVG image, it will be desirable to create PaleoCodes from images or 3D data in the future. Such a process will be interesting for two reasons. At first, the technology required to identify PaleoCodes from image data will be helpful for image annotation software. Scholars will be able to receive a suggestion for a sign variant from image data and will be able to confirm or modify this suggestion. The result will be a more straightforward process of PaleoCodage annotation, provided the image resource is suitable enough. The second reason could be that already annotated sign corpora without PaleoCodes could be enriched with automatically generated PaleoCodes to create sign variant corpora. Scholars may be able to check the resulting PaleoCode assignment for plausibility over a whole corpus of annotations. An additional plausibility check can be performed once databases containing PaleoCodes can provide options for matching algorithms to choose from or be the basis of alternative suggestions of sign variants. For these (semi-)automatic solutions to become a reality, further research is required in the area of cuneiform wedge recognition on image and 3D media, their classification as PaleoCode wedges, and finally, the automatic generation of PaleoCodes from annotated image areas.

4.2.6 PaleoCodage for font generation

Given a set of PaleoCodes that represent cuneiform signs of a specific corpus and time period, PaleoCodage can help solve a problem that is very prevalent in cuneiform studies and which often prevents the field of Assyriology from adopting Unicode cuneiform fonts: Signs present on cuneiform tablets do not necessarily look like the standardized signs in an epoch-specific cuneiform font.

This statement is not to be understood as that cuneiform signs are stylistically different, i.e., that the shape of the individual cuneiform wedges differs to a great extent (even though that is also a concern), but that sign variants are present on cuneiform tablets which comprise a different number and a different arrangement of cuneiform wedges as compared to the variants present in cuneiform fonts. To understand why cuneiform fonts do not just include more sign variants for Unicode code points, two things are important:

1. A state-of-the-art Open Type Font (OTF) maps a Unicode code point to precisely one visual representation of this Unicode codepoint

2. The cuneiform font cannot include images of sign variants that are specific to one excavation, as they are likely unknown to the creator of the font

Because of this, usage of Unicode cuneiform is often neglected in favor of including JPEG images of the actual cuneiform signs, line arts, or just the representation of the transliteration in the Latin script. This is a significant obstacle for computer scientists, as a Unicode representation of cuneiform signs usually simplifies natural language processing tasks. At the same time, a visually accurate representation of Unicode signs might lead Assyriologists to consider a representation of cuneiform signs in Unicode as a viable option in digital scholarly editions. PaleoCodage can help to solve this problem using a process of automated font generation:

1. Creation of a set of PaleoCodes that describe the sign variants of a particular corpus of cuneiform texts

2. Assignment of unique IDs to PaleoCodes

3. Conversion of PaleoCodes to SVG or Open Type Font Paths for inclusion in the font

4. Creation of GSUB rules in the Open Type Font [BKKM11], which resolves unique IDs to PaleoCode-generated SVG images

The GSUB table of an open-type font includes substitution rules for sequences of Unicode characters. Suppose every sign variant is assigned a unique identifier composed of a sequence of Unicode characters, which are guaranteed not to be included as a reading in a transliteration. In that case, they can be encoded as substitution rules in an open-type font. For example, suppose four sign variants of the cuneiform sign **E** exist. These four sign variants might be described by a URI and possibly a shorthand sequence of Unicode signs, for example, Ev1, Ev2, Ev3. These representations may become part of GSUB table rules which can substitute these identifiers for the correct SVG graphemes. In a digital scholarly edition depicted, e.g., as a homepage, it is possible to generate these open type fonts dynamically using JavaScript so that, given a repository of PaleoCodes and a set of annotated transliterations, the generated font may overlay the to-be-substituted character sequences.

Figure 4.10: Application example of overlaying a PaleoCodage-generated font over a transliteration text. Copying the cuneiform signs either results in the sign identifier, e.g., Ev1 or e.g., on a homepage to a modified output in JavaScript.

Figure 4.10 shows an example homepage[2] that displays the possibilities of loading a PaleoCodage-generated font over a transliteration text loaded from CDLI, as described before. The generated PaleoCodage font is loaded as a Web font and substitutes paleography IDs given in the respective transliteration. This allows for an accurate representation of graphemes of the sign variants depicted on the cuneiform tablet on a homepage. The represented cuneiform Unicode signs may be copied with some added JavaScript. The generated cuneiform font, however, is portable and may be bundled with any common representation format supported by a word processor.

In conclusion, PaleoCodage can provide a more realistic representation of sign variants in various word processing applications, enabling Assyriologists to share the results of their work better. However, a prerequisite for the interoperability of said cuneiform fonts must be the definition and sharing of unique identifiers for cuneiform sign variants and their accurate representation in a linked data graph. Only then can cuneiform fonts base sign representations on a pool of sign variants curated by an authority. The following section will examine how a linked data model for signs and sign variants can be incorporated into the cuneiform linked data cloud.

4.3 Graphemon: Grapheme Model for Ontologies

The Ontolex-Lemon model [MBGG+17] is an ontology model for the representation of lexicons, including their (grammatically assigned) word forms and dictionary contents. It can be used to connect word representations to semantic descriptions of their meaning, as is shown in Figure 4.11.

[2] https://situx.github.io/PaleoCodage/fonttester2.html

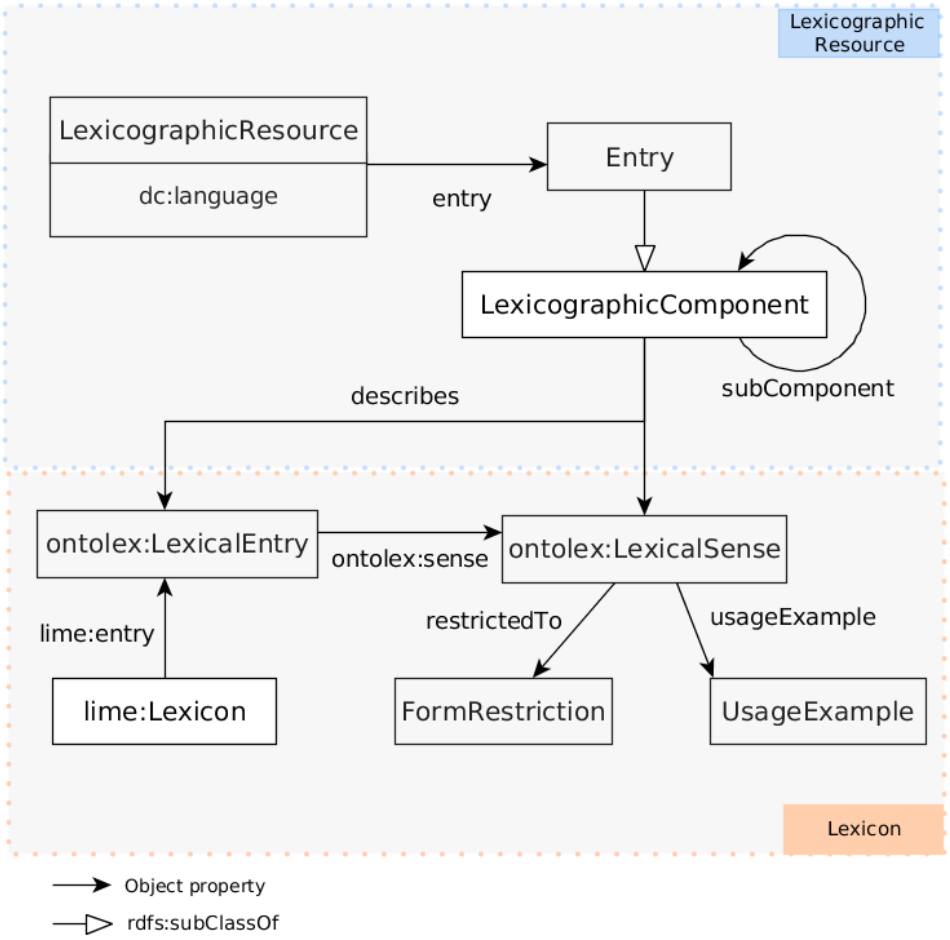


Figure 4.11: A diagram of the Ontolex-Lemon model including its most basic components [MBGG+17]

Figure 4.11 shows the representation of the Ontolex-Lemon model as defined in its official W3C community report[3]. Word forms are encountered and attested in cuneiform texts and are related to a lemon:LexicalEntry. The Lexical Entry (equivalent to a dictionary entry) is then connected to a lemon:LexicalSense linked to a machine-interpretable semantic concept that can depict a unique meaning associated with the Lexical Entry. Lexical Entries can be grouped by language to represent, e.g., Sumerograms in Akkadian, i.e., words in different languages from the main language of a given text. Given this ontology representation, it is possible to encode cuneiform words and attestations in a linked data dictionary. Currently, the creation of an Ontolex-Lemon dictionary for Sumerian cuneiform is prepared in Wikidata[4], showing the relevance of the Ontolex-Lemon model for cuneiform in practice. However, to encode words only based on their transliterations would not be sufficient, considering the wide variety of paleographic sign variants that are to be expected in cuneiform words, as discussed in Section 4.2. This section, therefore, introduces a Character extension of the Ontolex-Lemon model, called Graphemon (Grapheme Model for Ontologies). This ontology model aims to describe cuneiform signs,

[3] https://www.w3.org/2019/09/lexicog/
[4] https://w.wiki/5kzt

cuneiform sign variants, and depictions thereof. It connects these to already established ontology models describing features on existing artifacts and annotation models referring to representations of said cuneiform sign variants. In doing so, the ontology model is generic enough to apply to other languages. While a detailed description of the ontology model extension has been published at the Grafematik 2022 conference [HD22], the following sections will focus mainly on its applicability in cuneiform studies and its connection to PaleoCodage.

4.3.1 Preliminary definitions

First, definitions of the most essential terms of this ontology model are presented:

Definition 18. ***Glyph***=*cidoc:TX9_Glyph*
The physical manifestation of a grapheme on a written medium.

This definition covers written glyphs on any medium and is equivalent to the concept Glyph in CIDOC [Doe05] CRMtex [FM21]. For cuneiform, this is equivalent to a single cuneiform sign depicted on a clay tablet. This cuneiform sign might be a non-standard variant. It might deviate from this standard variant because the glyph might be broken and have a different number of wedges or wedges that do not point in the expected directions.

Definition 19. ***Grapheme***=*graphemon:Grapheme*
Digital representation of relevant features of a representation of a glyph or equivalent non-written representation.

A graphemon:Grapheme consists of an idealized character form, represented by a digital representation, i.e., abstraction of the set of glyphs describing the cuneiform sign.

Definition 20. ***GraphemeManifestation***=*graphemon:GraphemeManifestation*
The manifestation of a grapheme either on a written medium or using non-written means.

A graphemon:GraphemeManifestation is a more general concept to a Glyph, as the ontology model which is developed should be generalizable to other scripts and non-written language representations [HD22].

Definition 21. ***GraphemePart***=*graphemon:GraphemePart*
A representation of a grapheme that is found as a part of some other Grapheme in the same script.

This definition relates to parts of characters found in other cuneiform characters. As [Hom21] discovered in a preliminary experiment of a cuneiform font of one time period, the reoccurrence of cuneiform sign parts in other cuneiform signs is prevalent. In this experiment, only about one-third of all cuneiform characters in the Unicode code point list were not parts of other cuneiform sign representations.

Definition 22. ***AtomicPart***=*graphemon:AtomicPart*
A representation of an atomic part out of which Graphemes are comprised.

An example of an atomic part in cuneiform languages would be a single cuneiform wedge, comprised of each cuneiform sign. The concept of an atomic part is also present in other structured scripts, such as Chinese (strokes as parts of Chinese characters), so the ontology model also applies to these scripts. Atomic parts might be assigned further types, e.g. the classifications of PaleoCodage.

It has to be noted that GraphemeParts and AtomicParts may define their own meanings. These meanings are usually derived from the pictographs they originally represented, as is illustrated in Figure 4.12.

Figure 4.12: Cuneiform sign NAG with the meaning of "to drink" which is comprised of the sign KA (mouth) and A (water). Both parts also exist as individual signs.

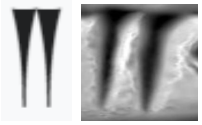
GraphemeParts could, therefore, also be considered single characters or even single words. Therefore, this classification is not script-dependent but language-dependent, as it depends on the respective language to use the individual GraphemeParts as a single word.

4.3.1.1 What constitutes a grapheme?

Now that the terminology describing graphemes and characters has been introduced, the next question is how to define and distinguish actual grapheme variants from glyph representations on a cuneiform tablet. When is a glyph on a cuneiform tablet considered a scribe stylistic choice, and when does it constitute an actual sign variant?

To answer this question, this thesis follows a set of criteria and assumes that at least one agreed-upon standard variant of a cuneiform sign exists. Let c_{vs} be this standard cuneiform sign variant of a cuneiform sign described in a sign list such as Unicode or the Borger list of cuneiform signs [Bor04]. c_{vs} is preferably the most occurring form that the respective linguistic community has agreed upon, but in the absence of agreements of this kind, the generally most occurring form of the corpus that is the subject of study would be chosen.

(a) The cuneiform sign A with its standard form once as grapheme and once as an actual occurrence in the cuneiform text HS 367, front side, column 1, line 3, sign 4



(b) The cuneiform sign A with an alternative form more common in older cuneiform texts once as grapheme and as an actual representation in the cuneiform text HS 1163, back side column 1, line 14, sign 4. This form also resembles the cuneiform sign for the number two.

Figure 4.13: Examples for cuneiform signs and sign variants

Consider again the cuneiform sign A[5], which constitutes of three vertical cuneiform wedges with at least one attested meaning of liquid water (wd:Q29053744) and is described with PaleoCode *a-a:a* shown in Figure 4.13a. A sign variant to cuneiform sign *A* is a variant that differs in at least one of the following criteria:

C.1 Amount of cuneiform wedges per type

C.2 Positioning of cuneiform wedges towards each other

C.3 Changes in the type of cuneiform wedges at their respective positions

Figure 4.13b constitutes such a variant. This example also shows that sign variants of one cuneiform sign might have the same glyph shape or grapheme definition of another cuneiform sign (in this case, the sign 2(disz) or MIN to depict, among other meanings the number 2 (wd:Q200). It is important to model these definitions as graphemes and document their occurrences in the cuneiform text to set them into context. Apart from criteria that are deemed necessary to distinguish sign variants, some criteria are useful as documentation of the individual Glyphs representing a grapheme variant:

D.1 The writing order of wedges if known and not exposing a semantic of their own

D.2 The style of cuneiform wedges themselves (e.g., cuneiform head, cuneiform stroke)

D.3 The absolute sizes of cuneiform wedges as long as their proportional size are the same

D.4 Changes in color or material on which the cuneiform wedges are imprinted unless they capture a semantic meaning

These criteria may help researchers search for similar Glyphs when solving paleography-related questions but are not criteria that would impact the assignment of a different grapheme variant.

4.3.2 Representing Graphemes in linked data

In cuneiform languages, we can see many cuneiform sign variants that need to be attested, classified, and finally set into relation to their interpretations. This section introduces the necessary extensions to the Ontolex-Lemon model to represent those.

[5] https://en.wiktionary.org/wiki/%F0%92%80%80

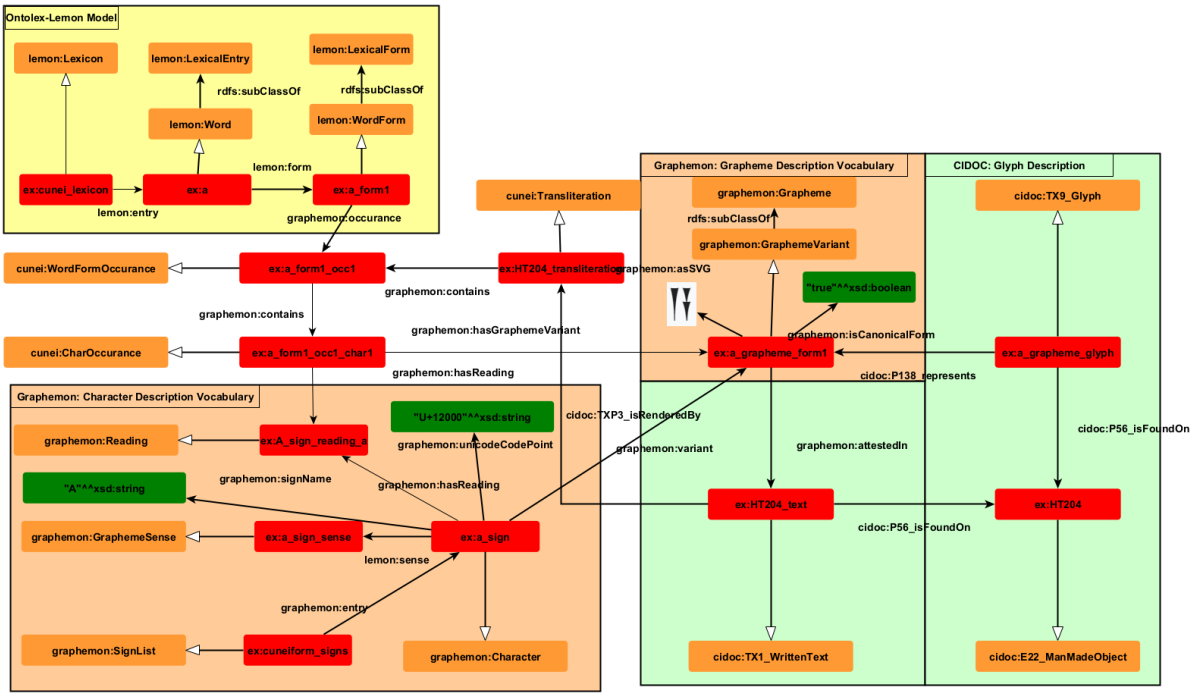


Figure 4.14: Graphemon Ontology: A character extension to the Ontolex-Lemon model to represent, among others, cuneiform signs. The vocabulary defines Graphemes linked to Glyph representations, potentially modeled using CIDOC-CRM. Characters are connected by Readings to occurrences of characters in transliteration texts. Characters, in turn, are included in word occurrences representing word forms in the Ontolex-Lemon model. Finally, characters may be linked to Graphemes to express their shape.

Figure 4.14 shows the main idea of the extension to the Ontolex-Lemon model. A new entity graphemon:Character is defined, which exhibits a representation of the abstract concept of a character, i.e., the smallest functional unit of a representation in any language and may be described, e.g., by a Unicode code point or a sign list reference. Characters might be associated with a sense representation, which describes the pictographic sense exhibited by the character's shape independent of its grapheme representations. This sense may or may not be the same as words or word forms that use this character. The graphemon:Character definition refers to graphemon:Grapheme or graphemon:GraphemeVariant, which describe the idealized representations of a cuneiform sign. Occurrences of characters in transliterations are linked to readings that connect to Character instances, as only readings in cuneiform are usually represented in transliterations.

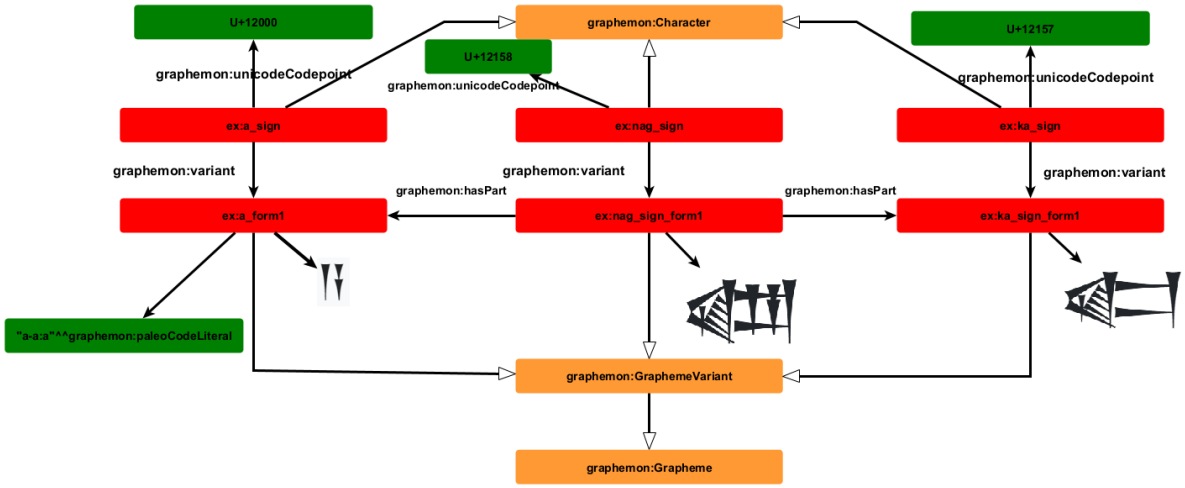


Figure 4.15: Representing the example in Figure 4.12 using the Graphemon model. Two GraphemeVariants which are parts of another Grapheme are connected using concepts in the knowledge graph. A Unicode codepoint is assigned to all Character representations and one GraphemeVariant has been assigned a PaleoCode

To illustrate the data model, Figure 4.15 shows an example of application of the Graphemon model to the initial example presented in Figure 4.12. This example shows how Graphemes can be modeled as parts of other graphemes and their relation to Unicode codepoints.

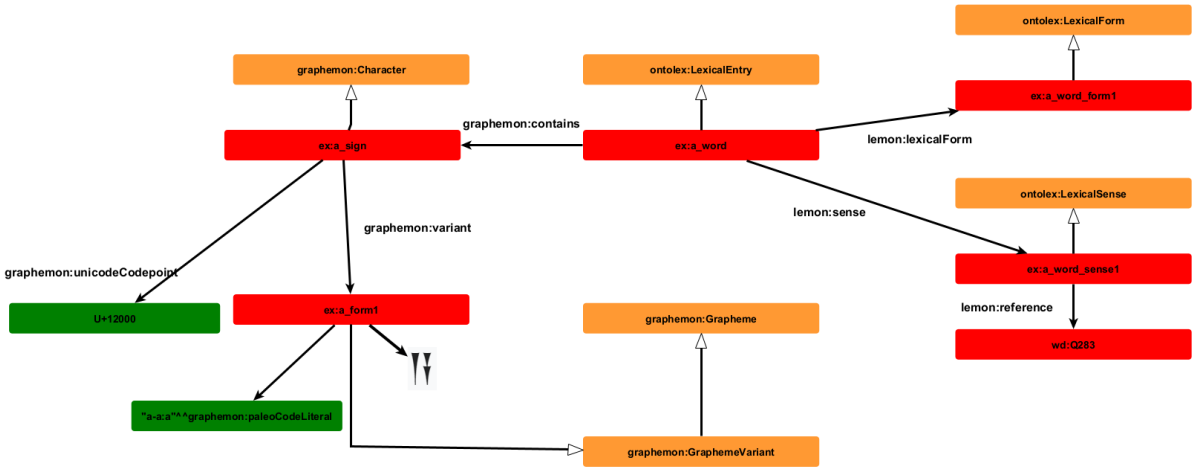


Figure 4.16: Example of the connection between an Ontolex-Lemon dictionary including the Sumerian word for water connected to the Graphemon model describing the sign water in Unicode and as a GraphemeVariant. The dictionary entry is connected to the Wikidata concept for water.

Finally, Figure 4.16 shows how Graphemon characters can be connected to Ontolex-Lemon resources. The connection via a transliteration had already been demonstrated in Figure 4.14.

Graphemes relate to these occurrences in transliteration texts (to be exemplified in the next chapter) and link to glyph representations on cuneiform tablets, which annotations may exemplify, e.g., image media. Each so-defined grapheme may now exhibit certain characteristics. It might

- represent its meaning (irrespective of its context)
- be used by one or many (compound) lexemes to represent their distinct meaning
- include character parts that may or may not exhibit their meaning
- be represented using an arbitrary number of character variants

A grapheme variant is a representation of a grapheme. A grapheme may or may not exhibit an arbitrary amount of grapheme variants, and a language-practicing community may or may not classify some of these grapheme variants as standard (canonical) variants of a specific time and epoch. However, if grapheme variants exist, then the possibility of attestation of these grapheme variants also exists. It will be up to scholars to provide links to actual occurrences of representations of these graphemes in image media representations of cuneiform tablets and on transliteration contents.

4.3.2.1 Encoding information in Grapheme representations

In certain publications, the grapheme representation might provide semantic information about the Glyphs they represent.

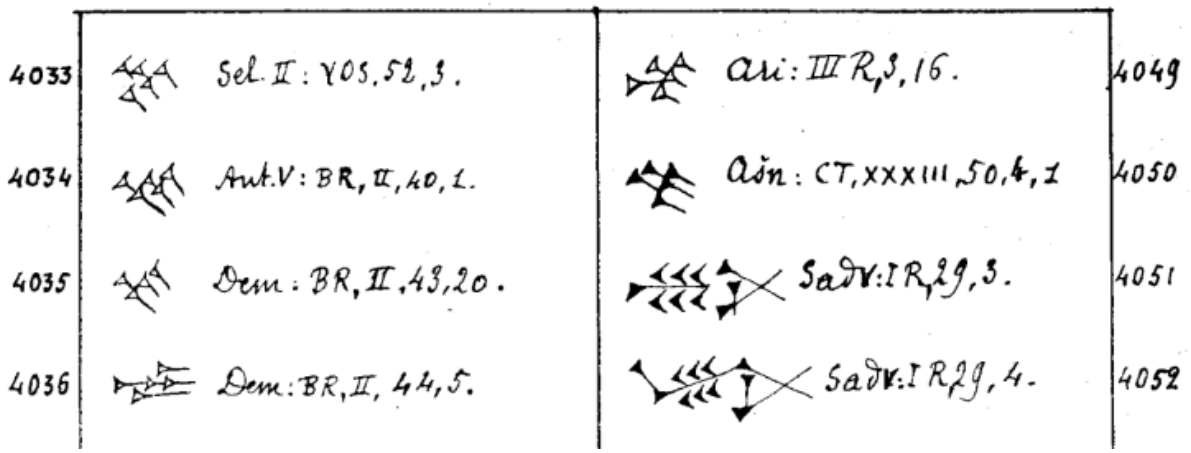


Figure 4.17: Excerpt from [Fos26] page 75: Cuneiform signs documented here, which are completely black, represent cuneiform signs found on a stone surface. Other cuneiform signs are found on a clay surface.

Figure 4.17 shows an excerpt from [Fos26] in which the individual drawings of cuneiform graphemes depict the semantic information of whether the documented grapheme was (only) found on clay or stone materials. This information can help to further classify graphemes in the knowledge graph and verify if the claims present in relevant literature are backed up by data (associated Glyphs). Are graphemes that are claimed to be only present in stone materials also found in clay and vice versa? How accurate are the collections of cuneiform grapheme variants in light of more recent research in the last century?

4.3.3 Encoding grapheme etymology and similarity

As discussed in Section 4.2, cuneiform signs are subject to massive changes in appearance over the centuries. Therefore, similar to capturing word etymologies, the etymologies of cuneiform signs depicted as graphemes are valuable information to add to the knowledge base to be constructed. Graphemon can capture these etymological relations, which are attestations by individual scholars. To encode this information, Graphemon extends lemonETY [Kha18], a vocabulary that extends the Ontolex-Lemon model to capture the etymology of words. This model defines three main concepts, lety:Cognate, lety:etymon, and lety:Derivative, which are reused in the Graphemon model. Adding an etymology component to individual cuneiform sign representations allows modeling etymology not only on cuneiform words, which is already possible with lemonETY, but will allow for a deeper computational understanding of the history of the cuneiform signs themselves, similar to Chinese. As attested in the relevant literature, the etymology of cuneiform signs has been broadly researched, so the creation of etymological relations such as shown in Figure 4.18 would greatly enrich the digital understanding of cuneiform sign variants. The question of whether an atypical sign variant for the time in which a text is written has been observed, e.g., a sign that makes sense semantically but is in an atypical shape for the time period, could then be simply queried from a knowledge graph representation of the sign etymology data.

4.3.3.1 Similarity vs. Etymology

Apart from a human judgment about how characters evolved, graphemes may also be compared by metrics derived from statistics and executed on grapheme or glyph representations. Suppose graphemes are encoded using the PaleoCodage encoding. In that case, any String similarity metric may be used to compare the PaleoCode representations of the respective grapheme to all others in the knowledge graph. If graphemes are represented as images, image similarity metrics may use to achieve the same, and these could also be applied to glyph images to achieve a similar result. Encoding similarities on a grapheme or glyph level in Graphemon requires the definition of new classes and properties that represent the similarity metrics and classes and properties capable of representing the results of said metrics.

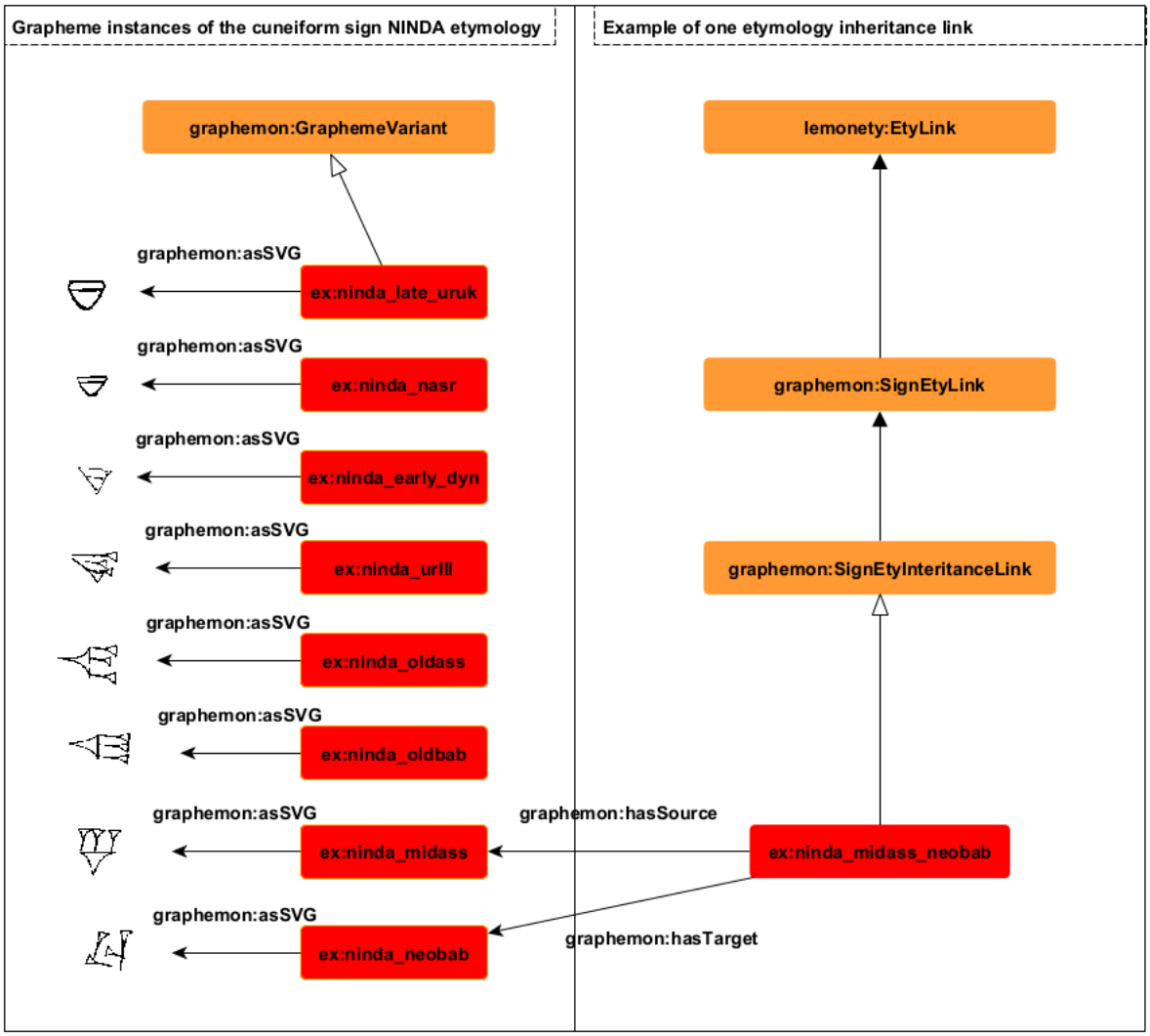


Figure 4.18: Etymology of a cuneiform sign using the Graphemon ontology model: Example using one inheritance link for illustration (some links are omitted for brevity). Here, scholars have determined that one sign variant is the predecessor of the next. References to scholarly publications, sign variant occurrences, time periods, and findspots could further substantiate these claims in the knowledge graph. Styles in this graphic follow Table A.1

Figure 4.19: Representation of similarities using the Graphemon ontology model: Two graphemon:GraphemeVariant are compared using a String Similarity metric based on the Levenshtein Distance Algorithm, modeled with the help of the fno Function Ontology Vocabulary [DSDV20] and the om Units of Measurement Ontology [RVAT13]. The result of the comparison is a graphemon:SimilarityLink instance, with the metric result and a link to the original targets of comparison. Styles in this graphic follow Table A.1

Figure 4.19 shows how to encode relations based on similarity metrics rather than on etymological relations into the knowledge graph. A String-based similarity metric has been used in this example to compare the PaleoCodage representation of two Grapheme Variants. Further thinkable comparisons could target the grapheme SVG representation or even glyph representations linked to the grapheme representations. The result of each comparison yields a new subgraph as the one shown in Figure 4.19. Awareness of both possibilities is crucial to represent relations between grapheme representations. Combining the two might lead to intriguing research questions in computer science and Assyriology.

Listing 4.1: A sample query which allows to query cuneiform sign graphemes which are similar to a given grapheme variant under consideration of a similarity algorithm and a similarity value threshold

```
SELECT ?graphemevar_sim ?glyphimage ?simvar WHERE {
    ?graphemevar_link graphemon:similarityLinkSource ex:mygraphemevar ;
        graphemon:similarityLinkTarget ?graphemevar_sim ;
        graphemon:linkBasedOn graphemon:PaleoCodageComparisonMetric ;
        om:hasNumericalValue ?simvalue .
    ?graphemevar_sim graphemon:asSVG ?glyphimage .
    FILTER(?simvalue>0.8)
}
```

Listing 4.1 selects all graphemes above a given similarity threshold of a chosen similarity score. This allows assyriologists to find similar grapheme variants of cuneiform signs for the signs currently being examined and generate similarity statements within the respective text corpus they investigate.

4.3.4 A cuneiform sign variant registry

This section describes a cuneiform sign variant registry based on linked open data that can be established with the help of the Graphemon ontology model described previously. A cuneiform sign variant registry is a registration service for cuneiform sign variants. Researchers may consult a cuneiform sign registry to find predefined cuneiform sign variants and identify sign variants they are currently facing on their respective cuneiform tablets. The data model of such a registry can be the Graphemon model itself, as it saves an arbitrary number of cuneiform signs and sign variants with appropriate metadata. Similar to an Ontolex-Lemon dictionary, the registry should add new entries, search for existing entries using different means, and be accessible for other tools wanting to reuse the respective information. This registry allows the digitization and consolidation of a central task of every digital scholarly editions: The creation of the paleographic sign list. Instead of creating this list on paper, sign variants can be created in PaleoCodage. They can be sent to the cuneiform sign variant registry to find either the sign variant itself or a similar sign variant based on etymology, text metadata (e.g., the time period of the text to be annotated), or on a chosen similarity metric. A confirmation of the found sign variant or the creation of a new sign variant to be submitted to the sign registry will result from the annotation process.

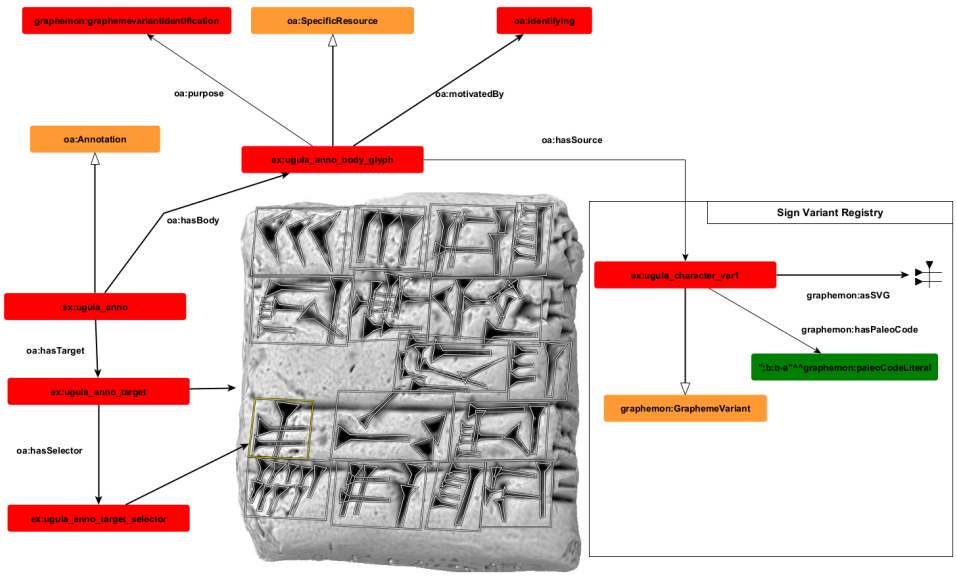


Figure 4.20: Connection between an image annotation and the graph of a cuneiform sign registry as advocated by the Graphemon model: The sign with reading "ugula" is annotated as an image annotation and linked to an already existing representation of the grapheme variant in the cuneiform sign variant registry. Styles in this graphic follow Table A.1

Figure 4.20 shows that a sign variant registry may also become part of image annotation processes, as discussed in the previous chapters. Already registered sign variants may simply be annotated, as URIs exist to identify them uniquely. However, as image media are not always available for annotation purposes, it seems equally prudent to examine already existing transliteration formats for their compatibility with paleographic annotations.

4.4 Paleographic extensions in Transliteration formats

This section examines how the paleographic information discussed in the previous sections can be included in already-established transliteration formats for cuneiform and how this information relates to cuneiform sign lists, words, and dictionaries. To achieve the integration of paleographic information, annotations on transliteration contents must first be examined to establish a state-of-the-art. Next, annotation contents should be defined and applied to create paleographically enriched versions of cuneiform transliterations.

Annotations on transliteration content

As established previously, annotations on images and 3D media are only possible in some transliteration formats. However, even text annotations are not always a possibility. The class of ATF formats only allows for the modification of the transliteration representation; that is, new syntax elements in the transliteration format can lead to indicators of annotation content (e.g., **a#** as a broken sign "a" as opposed to the sign a which is not broken **a**. Therefore, in ATF-based formats, the introduction of sign variants warrants a significant change in the overall transliteration style, which is likely not to be accepted by cuneiform scholars. A realization of a modified ATF-Format which could incorporate Paleography is the Paleographic ATF [Hom21] (P-ATF) format.

Listing 4.2: Example of P-ATF with the description of paleographic features in the transliteration content. Readings in the transliteration are appended with a version suffix delineating the grapheme variant

```
1 Valid ATF: nam-usz2
2 P-ATF: nam_v2-usz2_v3
```

Listing 4.2 shows a small example of the word namusz (death) (wd:L709382) represented in P-ATF. It becomes apparent that P-ATF arguably changes the whole nature of the representation of cuneiform transliteration, as a given version suffix modifies every reading. A less intrusive way of annotating paleographic information in transliteration information would be to create text annotations with a link to Grapheme variants represented in the Semantic Web, using, e.g., the W3C Web Annotation data model, which was introduced in Section 3.5.1.

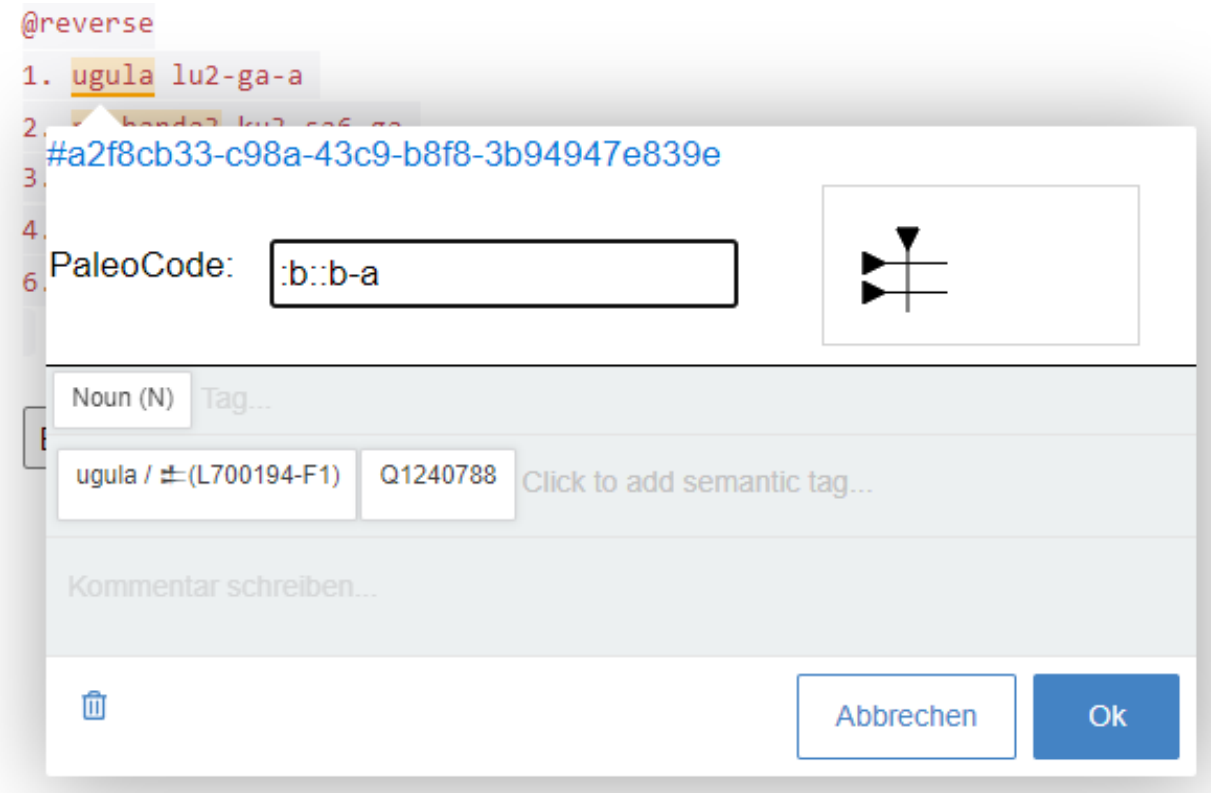


Figure 4.21: Annotation example on transliterations referencing a Grapheme URI by linking it to a corresponding PaleoCode in the background

Text annotations describing the paleography of the cuneiform signs (cf. Figure 4.21) are one solution to adding paleographic information to traditional transliterations. However, they still provide an additional sidecar file and software to be processed. In addition, a more preferential way of annotation would be an annotation on image media (cf. Figure 4.20), which is unfortunately not always available in digital form. An ideal transliteration format would include references to paleography on image annotations if available, provide the possibility to annotate paleographic particularities with reference to the transliteration, and link to different media representations of parts of interest of representations of the cuneiform tablet in question. Even more importantly, the ideal transliteration format should be enabled to represent and retain links to annotations on image media with direct reference to transliteration content, i.e., refer image crops to indices in the transliteration. Currently, no transliteration format can provide this functionality, which leaves annotations of paleographic data to sidecar files that may need to be loaded in software capable of displaying the transliterations and annotations together. Because of these reasons, the next chapter will investigate the foundations of a unified ontology model and an enhanced transliteration format, which will overcome the problems addressed here.

4.5 Summary and Discussion

This chapter has created the foundations for encoding machine-readable paleographic information of cuneiform signs and beyond in a linked data model. The PaleoCodage encoding system (cf. Section 4.2) proposes a machine-readable encoding of the actual shapes of cuneiform signs as they appear on the cuneiform tablet, thereby formalizing the representation of cuneiform signs to make these signs reproducible. This allows PaleoCodes to be used by Assyriologists to normalize depictions of cuneiform signs. PaleoCodes may also be used as classification targets for machine learning classifications or as features for training classifiers.

However, arguably, the most potential lies in collecting and inventorying PaleoCodage-modeled cuneiform sign variants across time and space. It must be stressed that even though the PaleoCodage encoding has been tested on an arguably large test corpus of cuneiform signs from one time period, future work should examine whether all possible cuneiform signs in most periods can be represented. At the same time, tools that can create Line Art from PaleoCodes might interest the Assyriology community. Using the Graphemon ontology model, the paleographic descriptions can be organized, compared to word occurrences in given texts and connected to the image and 3D annotation instances introduced in Chapter 3.

Graphemon even allows for creating paleography-aware inter-language relations. In the past, the cuneiform script has been analyzed per language by different scholars specializing in interpreting the respective language. A dialogue between scholars of different communities is not always happening despite their preoccupation with the same script. It could be further strengthened if tools can suggest paleographic features that lead to a better understanding of cuneiform texts in their historical context.

However, the degree of standardization, i.e., which occurrence of a cuneiform sign is to be assigned a specific sign variant and the amount, location, and angle of different cuneiform wedges, as well as the reconstruction of possibly missing cuneiform wedges from context or the single cuneiform sign is still a process of interpretation. This process depends on a community of Assyriologists who find value in this form of representation and on the interpretation of the individual scholar judging the content of a cuneiform clay tablet. While machine learning algorithms might be enhanced with paleographic information and facts from a knowledge graph, further research needs to determine the accuracy of this additional information in machine learning classification tasks. Getting the Assyriology community engaged in creating PaleoCodes and/or referencing existing PaleoCode descriptions in annotations needs to be incentivized and made more accessible using appropriate tools.

Integrating PaleoCodage and possibly other sign description languages, such as Gottstein codes in a linked data graph using the Graphemon model, can enhance the scientific discourse about sign variants and establish standardization across research disciplines.

After describing how to integrate 3D model representations of cuneiform tablets, their annotations, and how to create a digital representation of paleography, the next chapter will answer the question of how to represent the cuneiform tablet itself alongside transliterations and how to interconnect these representations to annotations and dictionary data.

Chapter 5

Language Resources and classification

This chapter includes contributions to language resources to identify cuneiform characters and enable natural language processing on transliterated cuneiform texts. However, at first, Section 5.1 describes a holistic ontology model which describes all elements of a transliteration of a cuneiform tablet and its links to paleography, dictionary data, and all aforementioned media resources. This model serves as a backbone for data extraction for several tasks in OCR and Natural Language Processing [NOMC11] (NLP) applications and is used by JTF-LD (cf. Section 5.2), a new linked-data based transliteration format which aims to mitigate the shortcomings of current transliteration formats illustrated in Section 5.2.3. Finally, the new ontology model is applied as the means of data provided for the training data of a machine learning experiment, MaiCuBeDa in Section 5.3, which showcases the advantages of linked data provision and eventual feedback of classification results into the knowledge graph. Finally, in this context, perspectives of the usage of crowdsourcing for data verification are discussed in Section 5.3.4.

5.1 A holistic ontology model for cuneiform resources

This section describes the Cuneiform Ontology model, which is a vocabulary that connects cuneiform artifacts with the previously established vocabularies for paleography and 3D representation in this thesis.

5.1.1 Representation of cuneiform artifacts

At first, cuneiform artifacts need to be represented in the ontology model. Cuneiform artifacts may be cuneiform tablets or further artifacts on which cuneiform script has been written. Instances classified as cuneiform artifacts are considered subclasses of the class cidoc:E22_ManMadeObject. This follows a best practice of reusage of vocabularies defined by the CIDOC-CRM model and de-facto makes this part of the ontology model an extension of the CIDOC-CRM model for cuneiform objects, as practiced by the CDLI.

For brevity, the following descriptions are based on the most common cuneiform artifact, the cuneiform tablet, whose parameters are formally defined as follows:

Definition 23. ***Cuneiform Tablet***=*cunei:Tablet*
$CuneiformTablet\ ct = \{s_0, s_1, ...s_i, ...s_n\} \in \mathbb{CT} s_i \in \mathbb{S}, i \in \mathbb{N}_0, n = |ct|$
$Surface\ s = (co_0, co_1 ... co_i, ... co_n) \in \mathbb{S},\ co_i \in \mathbb{CO},\ i \in \mathbb{N}_0, n = |s|$
$Column\ c = (l_0, l_1 ... l_i, ... l_n) \in \mathbb{CO},\ l_i \in \mathbb{L},\ i \in \mathbb{N}_0, n = |s|$
$Line\ l = (w_0, w_1 ... w_i ... w_n) w_i \in \mathbb{W},\ i \in \mathbb{N}_0,\ n = |l|$
$Word\ w = (c_0, c_1 ... c_i ... c_n) c_i \in \mathbb{C},\ i \in \mathbb{N}_0,\ n = |w|$
$Character\ c = (cw_0, cw_1 ... cw_i ... cw_n) cw_i \in \mathbb{CW}, i \in \mathbb{N}_0, n = |c|$
$Wedge\ cw = \{a, b, c, d, e, f, w\}$

Definition 23 describes a cuneiform tablet as an object with six different surfaces (obverse, reverse, top, bottom, left, right). Each surface of a cuneiform tablet may include an arbitrary number of columns. Columns include a set of lines. Lines include an arbitrary number of words. Words include an arbitrary number of cuneiform signs; cuneiform signs include an arbitrary number of cuneiform wedges. Columns may be omitted for cuneiform tablet surfaces which do not feature these.

This description will fit the most common shape of a cuneiform tablet. Other shapes will include more or fewer surface areas but retain the proposed hierarchical structure of tablet-surface-line-word-characters-wedge. Cuneiform artifacts are expected to define a set of metadata which, in the case of the cuneiform ontology, are derived from cuneiform repositories such as the CDLI and include at least:

- Findspot of the cuneiform tablet: Modeled in the GeoSPARQL vocabulary with the extension for spatial reference system modeling (Section 3.4)
- Material the cuneiform tablet is composed of: Usually clay
- Genres/Subgenres[1]: Classification of the cuneiform tablet by text content
- Time Period[2]: A time period according to a given chronology (usually Middle Chronology [RR11]) - modeled using a URI describing the time period either in CDLI or Pleiades
- References to the collection and museum location (if known)

Depending on the research project content, more information may be appended to the instance of the cuneiform artifact. Likely candidates of available data are data about the archaeological context, which could be modeled using the CIDOC CRM archaeo model [BMT08]. However, further extensions could include results from the material science experiments [ABD+21] to solidify and extend the claims of the material of the cuneiform tablet. This thesis focuses on the metadata descriptions for standard cuneiform data repositories such as CDLI. It, therefore, only aims to provide the possibility for other, more specialized data to be included.

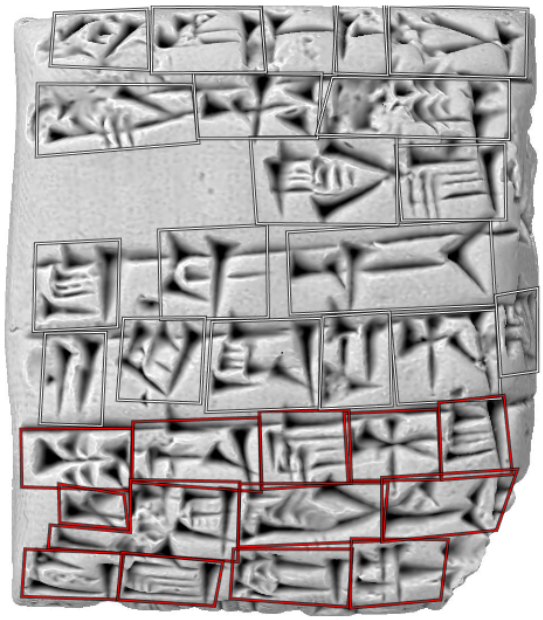
[1] https://cdli.earth/genres
[2] https://cdli.earth/periods

5.1.1.1 Subdivision of cuneiform artifact contents

Once the cuneiform artifact and its properties are described as linked data instances (e.g., `http://www.example.org/cuneiform/mytablet`), the modeling of the cuneiform artifact continues with detailed descriptions of its surfaces.

Definition 24. ***Surface*** *The surface of a cuneiform tablet is defined as any flat area with optional textual or image contents*

Each surface is assigned its URI (e.g., `http://www.example.org/cuneiform/mytablet_obverse` for the obverse surface), which is connected to the cuneiform artifact with the cunei:hasSurface property. Surfaces themselves may be described using their surface structure and state of preservation by a respective metric, e.g., 50% destroyed, and often include content to be exemplified in linked open data. It needs to be stated that the classification of the surface itself already needs an interpretation by Assyriologists, as it is not in all cases immediately apparent which surface is, e.g., the obverse or reverse surface. As with cuneiform artifacts, various media may describe surfaces, e.g., image references. A common content of a cuneiform tablet surface is an inscription that, written from left to right, starts at the top position of the respective surface but does not necessarily end at the boundary of the surface.



(a) Cuneiform tablet HS 1087 reverse with text written around the edges of the cuneiform tablet and example of a cuneiform line which spans more than one sign row (line 5 reverse in the transliteration P134399 is highlighted in red)

(b) Right side of cuneiform tablet HS 1087, which shows cuneiform signs written around the edge of the cuneiform tablet

Figure 5.1: Examples of cuneiform tablets surfaces displaying typical challenges for annotation

Figures 5.1a and 5.1b show examples of surfaces of a cuneiform tablet in which text is written around the right edge of the cuneiform tablet and in which the determination of a line is not as easy. This means that the right surface of the cuneiform tablet contains cuneiform signs, which have to be treated as the continuation of cuneiform text on the front side in a later interpretation step. Consequently, describing a cuneiform sign on a cuneiform tablet needs both the attestation to a line in the cuneiform text and the attestation on the surface of the cuneiform tablet itself may be subject to interpretation. To make matters worse, in the linked data graph, we are also dependent on a consistent naming of each of the cuneiform signs in the form of URIs, ideally, a URI of the form `http://www.example.org/mytablet_obverse_line1_char1`, encompassing the exact information it represents. So the following question is to be solved:

> How many lines/words/signs are on the cuneiform artifact, and how should they be named?

5.1.1.2 Aligning interpretations with URIs

Different cuneiform scholars may have differing opinions on the positioning of the lines or the interpretations of the cuneiform signs within a line. The opinions of these respective scholars are usually captured in transliterations which can be harvested for creating URIs in the knowledge graph. In essence, for each line attested in the transliteration, an instance of cunei:Line and cunei:TransliterationLine can be created in the knowledge graph, for example as `http://www.example.org/mytablet_obverse_line1` for line 1 on the obverse side of a fictional cuneiform tablet TAB1. The information conveyed here is not the position of the line or the extent of the line on the cuneiform tablet's surface but rather the statement:

> "Scholar XYZ has stated that there is a line with index 1 on the obverse surface of the cuneiform tablet TAB1"

To make these statements more tangible, they need to be supported by annotations on image media. An annotation on an image media, as exemplified in Section 3.5.2 conveys:

> "Scholar XYZ states that: On this image media which depicts cuneiform tablet TAB1, there is an area described by a bounding polygon which depicts the line with index 1 on the obverse surface of the cuneiform tablet TAB1"

Transferred to a knowledge graph, each image annotation can point us to a URI representing a particular cuneiform line, sign, or word, thereby asserting its existence.

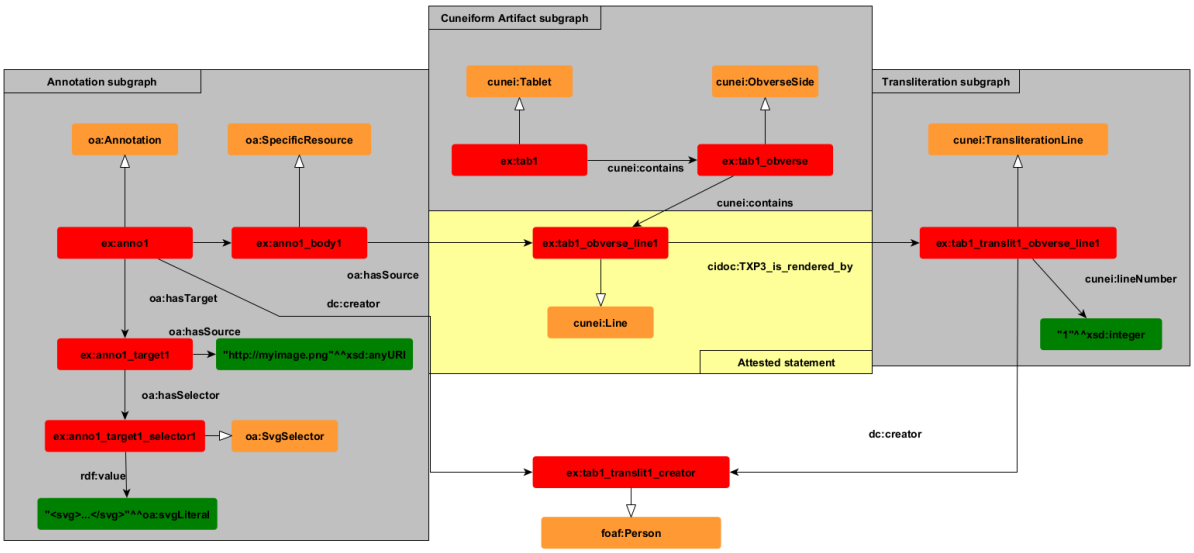


Figure 5.2: Knowledge graph representation of a cuneiform line described by an image annotation and a transliteration assertion by the same author. Styles in this graphic follow Table A.1

Figure 5.2 shows a practical example of this URI assertion. The URI ex:tab1_obverse_line1 only exists because it is referenced by two assertions - one from a transliteration representation and one from an image annotation. Whether these assertions mean the same area on the cuneiform tablet is not defined in this part of the graph. Rather, the graph only includes the statement that the same author claimed the annotations on the image and the transliteration. In the same way, different authors may make different claims about what obverse line 1 might be on the particular cuneiform tablet. This assertion needs to be done by adding author statements to the knowledge graph, which, in addition to Figure 5.2 claim their equality.

This approach allows for the following advantages when building the knowledge graph:

1. A set of interpretations of annotations and transliterations describing the same cuneiform artifact may be represented simultaneously without judgment of correctness

2. Interpretations of this kind define claims in the cuneiform artifact knowledge graph

3. Claims by individual researchers may be easily compared and contrasted by simply querying different versions of the annotation and/or transliteration graph

In that sense, part of the interpretation of the cuneiform tablet's contents is completely up to the researcher's discretion, and individual interpretations of the cuneiform tablet's contents could be compared and aligned.

5.1.1.3 Adding provenance to annotation graphs

Transliterations and annotations are by no means static. As seen in many cuneiform data repositories such as CDLI and ORACC, corrections to existing transliterations are submitted regularly as the scientific discourse evolves, and new findings might demand transliterations to be corrected. To reflect a provenance in the knowledge graph, the provenance ontology model [LSM13] is an extension to the Cuneiform Ontology. However, the question to be solved in the knowledge graph is which provenance level is best applied: On a transliteration level, word level, or sign level?

The preferred solution is to create a changeset between previous and current transliteration versions, which is then reflected in the knowledge graph. Changesets are very well understood in computer science, as they are the basis of many version control systems, such as Git or Subversion [PCSF08] (SVN), which are commonplace in software development. People in software development distinguish between major and minor software releases and major and minor contributions to a source code. The same principles can be applied to transliterations and annotations.

Transliterations are worked on iteratively, during a longer period of time, e.g., in a digital scholarly edition project, and are often worked on by not a single individual. In practice, transliterations are first released in the respective publication and, after some time, might be corrected in subsequent publications - corrigenda. This process is not dissimilar to a software release. One could treat the publication of a transliteration as a major release in a software release cycle and minor corrections as minor releases. Following this reasoning, this thesis proposes to directly integrate minor changes in the knowledge graph, that is, to update individual word occurrence instances with a new interpretation by the same author. Any other changes, such as the creation of a new transliteration from the older source material, or a completely new interpretation, will be treated as a major revision and result in a new subgraph for the transliteration, possibly derived from a previous transliteration.

This setup enables the use of a Git versioning system for the creation of the transliteration, capture changes to the transliterations in Git, and use Git release management tools for release creation, i.e., the creation of a new transliteration subgraph. After a first release in the Git environment, minor changes can be tracked as new branches and subsequently possibly squashed and converted to an RDF provenance hierarchy for the individual changes RDF elements. Optionally, the editing history to receive the first major transliteration version could be saved as minor revisions if that interests the respective scholars. In conclusion, versioning transliterations allows us to capture the decisions of scholars in the creation process of transliteration using a minor revision process in the creation process and allows for the representation of major and minor revisions with different graph patterns.

While not tested in the scope of this thesis, a conversion from a Git history to a provenance history seems straightforward to implement so that a version control system could be applied to transliterations and annotations.

5.1.1.4 Media depicting cuneiform tablet features

The cuneiform artifact might be depicted in various media such as images, 3D models, line art, etc. To integrate these media in the data model, one could use generic properties such as foaf:image to link to an image URL. However, as we deal with cuneiform artifacts of a specific nature, not all image media should be expected to be connected to the instances of the cuneiform artifact itself. Typically, images would be associated with depicting a cuneiform surface; e.g., more specific properties have been defined to address images that target surfaces explicitly. Hence, a cuneiform tablet that is fully documented will link the instance of the cuneiform artifact directly to all image and 3D media representations. Annotations on these representations will link to features extracted from interpretations (e.g., lines and character occurrences) via annotations. In this way, the cuneiform artifact is described in every viable granularity deemed necessary by the researcher.

5.1.2 Representation of transliterations

A transliteration of a cuneiform tablet is an interpretation of the textual contents of the surfaces of the cuneiform tablet created at the time of analyzing the cuneiform tablet. The creation of transliteration typically involves two stages.

1. The identification of the cuneiform signs on the cuneiform tablet and their sign names/Unicode code points

2. The attestation of readings to the previously identified cuneiform signs

In technical terms, step 1 is equivalent to creating a mapping between a part of the cuneiform tablet - the sign - to a unique natural language representation of a Unicode code point: The sign name. Step 2 could be technically understood as the annotation of an appropriate reading depending on a few parameters such as context, spelling, and judgment of the individual Assyriologist. To represent transliterations in the linked data model, the components of the transliteration need to be represented as follows:

CO.1 Perceived textual surfaces which may or may not equal actual surface designations

CO.2 Line observations which may or may not equal the actual line occurrences

CO.3 Word observations depending on the justification of word compositions by the scientist

CO.4 Character classifications depending on the observation of the expert

CO.5 Choice of readings for the cuneiform transliteration depending on the cuneiform target language

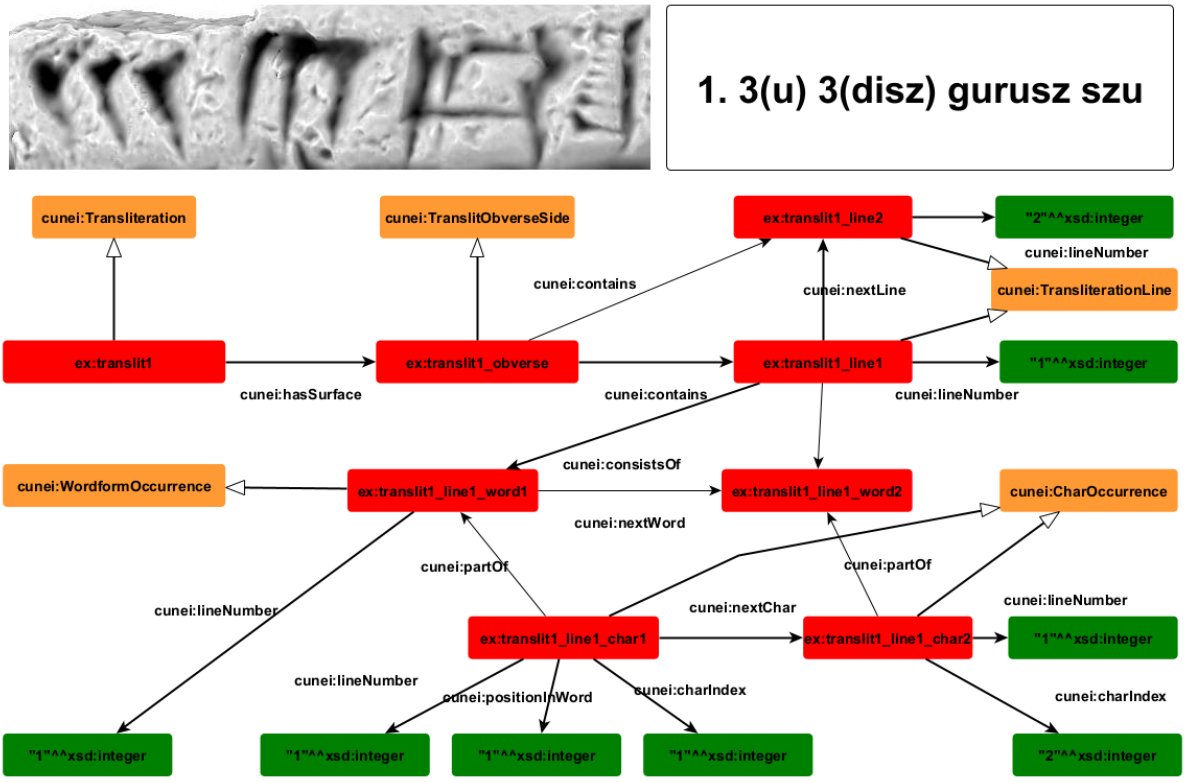


Figure 5.3: Representation of the transliteration of the first two cuneiform signs and words of the obverse side of cuneiform tablet HS1174 as interpreted by contributors of the CDLI: The transliteration instance (cunei:Transliteration), an instance of the interpretation of the obverse side (cunei:TranslitObverseSide), two interpretations of lines (cunei:TransliterationLine), one cunei:WordformOccurrence), consisting of one cunei:CharOccurrence, provide the structure of the interpretation of the transliteration as perceived by the researcher. Styles in this graphic follow Table A.1

Figure 5.3 shows all instances that describe the structure of the transliteration in a linked data graph, that is, components CO.1-CO.3. Components CO.4-CO.5 are to be derived from the readings the researcher described in the transliteration content, which will be shown in the following.

5.1.3 Inclusion of readings, dictionaries, and paleography

A linked data representation of a transliteration would not be complete without references to dictionary resources and sign registries (if available). To that end, as already hinted at in Section 4.3, positions in the transliteration need to be referenced to representations of the readings of the respective cuneiform signs and words they represent. In the Ontolex-Lemon model, a reading of a cuneiform word is treated as the lexical form of the lexeme it represents. The reading of a cuneiform word can be unambiguously mapped to a list of Unicode code points using correspondence lists such as Nuolenna[3]. With sufficiently available dictionary resources, links to actual WordForm representations will be

[3] https://github.com/tosaja/Nuolenna/blob/master/sign_list.txt

possible directly in the future. On a sign level, readings of cunei:CharOccurrence may be represented as instances of and in the future also using the Wikidata Lexeme graph. Hence, every cunei:CharOccurrence represented in the transliteration is associated with an instance of graphemon:GraphemeReading.

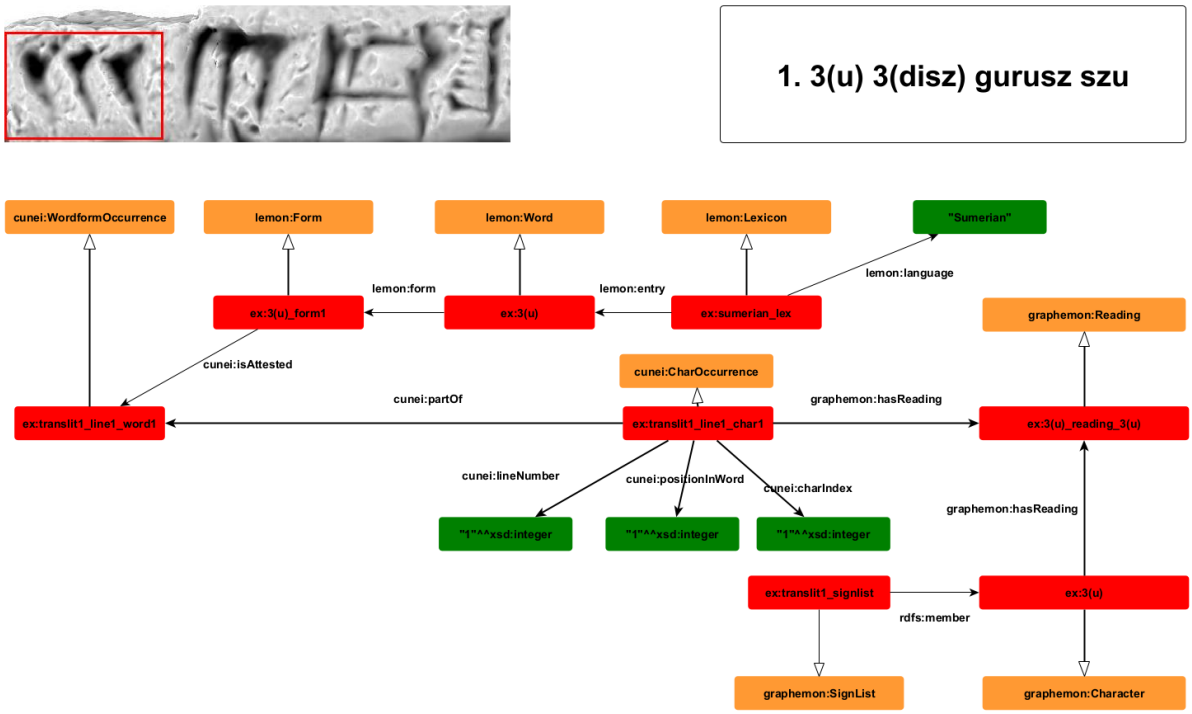


Figure 5.4: Connnection of a representation of the first word of HS1174 in a transliteration to Lexicons modeled in the Ontolex-Lemon vocabulary and to Characters modeled using the Graphemon vocabulary: Each occurrence of a character or word can be linked to its wordform, which may be enriched with linguistic information. Each character may be linked to information from a sign registry. GraphemeVariant information cannot be inferred from the transliteration but only from annotations. Styles in this graphic follow Table A.1

Figure 5.4 shows how to connect transliteration contents to the dictionary and sign list resources. In an ideal case, these resources already exist and can be linked automatically, as all occurring word forms can be matched in the knowledge graph. However, even with only the information given in a transliteration, generating parts of the graph is already possible. CharOccurrences and their respective character representations can be inferred automatically from readings, WordformOccurrences may also be inferred automatically, and Wordform instances may be created. Manual work would be required to connect these instances to dictionary resources. To create better integrations between transliterations and further resources such as dictionaries, grapheme representations, image media, and sign lists, annotations must link them.

5.1.4 Annotations with respect to transliterations

This section discusses which annotations are expected to be present in transliterations to fully access the transliteration text semantically and linguistically. This thesis proposes annotations on a word level in transliterations to take advantage of the following possibilities:

- Resolve disambiguations of WordformOccurences by annotating which wordform is meant
- Annotation of word senses
- Annotation of part of speech tags
- Annotation of tags that are used in the current research project context (e.g., classifications of named entities in the given text)

In this way, annotations in transliterations help to create connections from the knowledge graph representing a transliteration to the knowledge graph of an applicable dictionary and the knowledge graph of possible sign readings. Furthermore, it is important to link transliteration contents, i.e., interpretations, to annotations on other image media. In this way, a computer can aggregate image information and connect it to information included in the transliteration or its annotations.

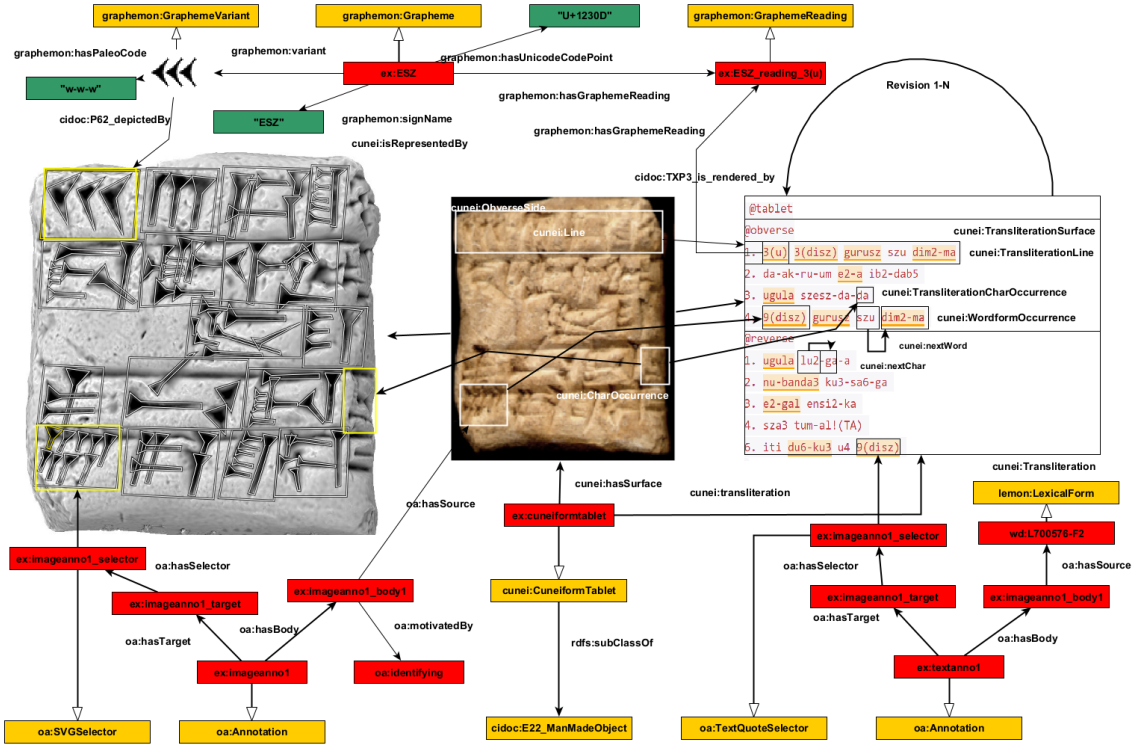


Figure 5.5: Full example of the holistic graph model using the example of HS1174 with annotation content on 2D renderings , the abstract classes of the cuneiform tablet in the center, with interconnections in the knowledge graph between different elements on the cuneiform tablet representations and with connections to dictionary and sign list resources via annotations

While Figure 5.2 has already shown that image annotations and transliteration content provide the basis for the identifiers of lines in the knowledge graph, Figure 5.5 shows a complete example of one cuneiform word connected via annotations to all aforementioned knowledge graphs. This knowledge graph representation is the basis for the linked data representation of any cuneiform artifact interpreted by a cuneiform scholar. In the following, four application examples highlight aspects of said knowledge graph and show its applicability for different languages and epochs of cuneiform writing.

5.1.5 Application example: Modeling of four cuneiform tablets

As an application example of the holistic ontology model, four cuneiform tablets have been selected to be modeled. Each cuneiform tablet modeling also emphasizes at least one of the previously discussed ontology models. To show the applicability of the ontology model for different cuneiform tablet types, the cuneiform tablets were selected to contain cuneiform script written in different cuneiform languages, the cuneiform tablets are from different time periods, and they resemble different writing styles within the same language, if possible. All knowledge graph representations have been serialized in HTML on Github.

(a) Cuneiform Tablet HS1174 **(b)** Cuneiform Tablet HT073195 **(c)** Cuneiform Tablet O.O147 **(d)** Cuneiform Tablet TCH92

Figure 5.6: Cuneiform Tablets used for testing the Cuneiform Ontology model in this chapter

The first cuneiform tablet is cuneiform tablet HS1174 (cf. Figure 5.6a) included in the HeiCuBeDa Hilprecht Collection [MB19]. This cuneiform tablet is attested in the UrIII time period (ca. 2100-2000BC), written in Sumerian cuneiform and of administrative nature. The second cuneiform tablet is the tablet HT073195 (cf. Figure 5.6b), an administrative tablet from the Haft Tappeh Collection written in Akkadian cuneiform. It is attested in the Middle Elamite time period (ca. 1300-1000BC) The third tablet is O.O147 (cf. Figure 5.6c), which contains the Akkadian text from the Old Babylonian period and has been provided from the corpus of the Cune-IIIF-ORM project[4] research project. This cuneiform tablet comes with various 3D renderings and, as such, can provide many targets for single annotations in both 3D and 2D. The fourth tablet, TCH92 (cf.

[4] https://www.kmkg-mrah.be/en/scientific-research/cune-iiif-orm

Figure 5.6d), is a tablet with the Sumerian cuneiform text of an administrative nature but in a different, more non-formal writing style than a cuneiform tablet. In the following, each of these cuneiform tablets is taken as a showcase for different aspects of the ontology model. Finally, synergies of modeling these four cuneiform tablets gained from the data model are explored.

5.1.5.1 HS1174: Modelling Paleography

Cuneiform tablet HS1174[5] highlights paleographic data modeling. The cuneiform tablet includes an obverse and a reverse side and has been annotated on 3D renderings. The 2D annotations have been converted to 3D. However, in addition to this conversion, also single wedges have been annotated on the cuneiform tablets' obverse side. Hence, each wedge annotation points to a unique URI in the knowledge graph describing this particular wedge according to the Graphemon model explained in Section 4.3.2.1. Figure 5.6a shows a screenshot of the annotated cuneiform tablet HS1174 in which each visible cuneiform wedge is annotated in addition to every sign. In the linked data model, not only the single wedges modeled but also grapheme variants have been created to represent the sign variants available on the cuneiform tablet. Selected grapheme variants include a PaleoCode, which uniquely describes the structure of the grapheme variants and can therefore serve as a basis for further examination.

5.1.5.2 HT073195: Creation of dictionary resources

On cuneiform tablet HT073195[6] only the obverse surface contains cuneiform signs. It, therefore, has a limited sign and word form vocabulary. However, this small vocabulary makes it ideal for demonstrating the integration of dictionary information resources into the Cuneiform Ontology model. Each word form occurrence of the cuneiform tablet is linked to a word representation modeled in the Ontolex-Lemon model in a customized dictionary created for only this cuneiform tablet instance. Hence, this cuneiform tablet includes its linked data dictionary linked to the word form occurrences present on the cuneiform tablet and links to a customized sign list created for all Unicode cuneiform signs. The tablet modeling does not include grapheme variants.

5.1.5.3 O1.147: Text annotations in Akkadian

The cuneiform tablet O1.147[7] is used to showcase annotations on transliteration content which are done in a different language to the transliteration content. As this cuneiform tablet depicts an Akkadian text, the transliteration of this tablet may involve additional steps.

[5] https://situx.github.io/cuneiformontology/examples/hs1174/HS1174/index.html
[6] https://situx.github.io/cuneiformontology/examples/ht073195/HT073195/index.html
[7] https://situx.github.io/cuneiformontology/examples/o147/O147/index.html

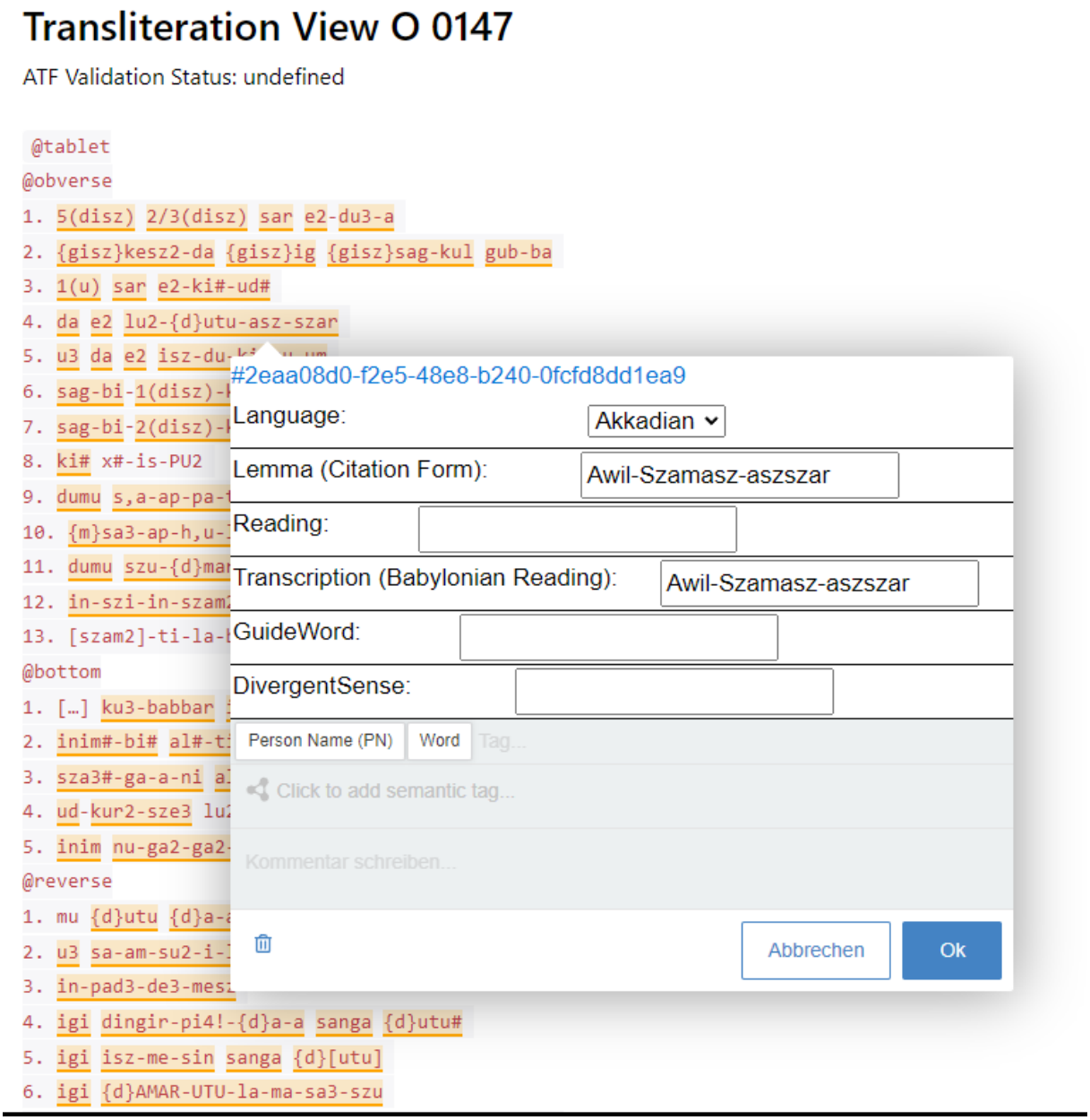


Figure 5.7: Example of annotating the Sumerian transliteration of a word (lu2-{d}utu-asz-szar), by interpreting it using an Akkadian reading (Awil-Szamasz-aszszar) in text O.0147

Step one, like on any cuneiform tablet, identifies the sign on the tablet. In this case, as shown in Figure 5.7, a transliteration based on the correct Sumerian reading is created. However, as the text is to be interpreted in Akkadian, the Akkadian reading is superimposed on the Sumerian reading interpretation using an annotation. In essence, the annotation here does not annotate the transliteration text; rather, it annotates the Akkadian reading, which is not present in the pure transliteration representation. This way of interpretation is not uncommon, as many languages use cuneiform scripts syllab-

ically. In this case, the knowledge graph structure will still follow the structure of the transliteration without annotations, as the Akkadian readings are attested on a word and not on a sign level. The Akkadian reading may be queried from the knowledge graph by querying the respective transliteration annotation contents.

5.1.5.4 TCH92: Sumerian text with casual writing style

The cuneiform tablet TCH92[8] contains Sumerian cuneiform text but with a more casual writing style than tablet HS1174, which was purely administrative. This use case is interesting regarding the cuneiform sign annotation, which can be taken here. Annotated cuneiform signs will differ greatly from their counterparts on the Sumerian tablet HS1174 introduced previously and may provide a valuable asset for machine learning algorithms. Even if stylistically very different, the cuneiform signs do not necessarily differ in shapes, which is relevant to the PaleoCodage encoding. This makes the tablet a good example of the necessity of modeling grapheme paleography in an encoding, independent of their representations as glyphs on the cuneiform tablet. In addition, this tablet, in particular, is a showcase for implementing 3D annotations. Every annotation on the cuneiform tablet has been replicated as 3D bounding cuboids, which are interlinked to the transliteration representations as proposed in Section 5.1.2

5.2 JTF-LD: A linked data-based transliteration format

In the last sections and the last chapter, the need to include accurate information about the cuneiform signs and the cuneiform words, senses, annotations, and metadata has been exemplified. As this thesis advocates, this information must be provided in controlled vocabularies and ontologies to be of value to both the computer science and Assyriological communities. However, even if the aforementioned information is given as linked data, it might be scattered across many different files, which may need to be found and merged by an application to be useful.

This is in contrast to the work of other digital edition projects, such as projects in TEI/XML, which include as much information about the artifact in the digital edition file in one place. This also better fits the workflow of an Assyriologist, as transliterations with their respective annotated components can be shared easily in one single file without the need to crawl information from different resources. The solution can be a data format incorporating annotation data, transliteration data, and links to definitions of cuneiform signs.

In this sense, this chapter proposes JTF-LD as an extension and linked data-friendly alternative to the already defined JSON Transliteration Format (JTF) introduced in Section 2.2.3. At first, JTF is extended by a JSON-LD context and further terms to encode annotations and paleographic data. The context is available using the namespace http://www.purl.org/cuneiform/contexts/jtfld.json and can be included in any JTF serialization.

Listing 5.1: JTF-LD example based on the transliteration of a text passage

[8] https://situx.github.io/cuneiformontology/examples/tch92/TCH92/index.html

```
#JTF
{"_class":"object", "type":"tablet","children":[
    {"_class":"surface","type":"obverse","children":[
        {"_class":"column","name":"1","children":[
            {"_class":"line","name":"1","children":[
                {"_class":"sequence","type":"short","children":[
                    {"_class":"chr","value":"a"}
                ]}
            ]}
        ]}
    ]}
]}

#JTF-LD
{
"@context": "http://www.purl.org/cuneiform/contexts/jtfld.json",
 "@graph":
 {"_class":"object", "@id":"tab1", "@type":"tablet", "children":[
  {"_class":"surface", "@id":"tab1_side_1","@type":"obverse",
     ↪ "children":[
    {"_class":"column", "@id":"tab1_side_1_column_1",
       ↪ "@type":"column", "name":"1", "children":[
      {"_class":"line", "@id":"tab1_side_1_column_1_line_1",
         ↪ "@type":"line", "name":"1", "children":[
        {"_class":"sequence",
           ↪ "@id":"tab1_side_1_column_1_line_1_seq_1",
           ↪ "@type":"short", "children":[
          {"_class":"chr",
             ↪ "@id":"tab1_side_1_column_1_line_1_seq_1_char_1",
          "@type":"Glyph", "type":"U+1200", "value":"a",
          "grapheme":"GRAPHEMEID",
          "glyphrep":"GLYPHREPRESENTATIONLINK","children":[
        {"_class":"wed",
        "@id":"tab1_side_1_column_1_line_1_seq_1_char_1_wedge_1",
        "@type":"wedgetype_a"}
          ]}
        ]}
      ]}
    ]}
  ]}
]}}
```

Listing 5.1 shows the application of the JTF-LD context using one cuneiform clay tablet. Like JTF, the format is hierarchical. It consists of JSON Objects, including parts of the cuneiform tablets' surfaces and their subdivisions into surfaces, columns, lines, sequences, and characters and newly added, singular wedges in order of their occurrence in the PaleoCode. It proposes two modifications to the original JTF model:

- Usage of JSON-LD identifiers @type and @id to identify instances and classes
- Integration of Part Of Speech Tagging vocabularies as URIs
- Integration of Semantic Annotation information
- Paleographic, metadata, and annotation extensions

5.2.1 JTF-LD Paleographic Extension

The JTF-LD paleographic extension adds a unique local or global cuneiform sign identifier to the character sections of the JTF-LD representation (cf. Listing 5.2). If the JTF-LD graph is to be treated as a local graph, i.e., as its own document, these identifiers may be comprised of unique IDs generated for the document exclusively. However, a better way would be to match the grapheme representation used to match the respective cuneiform sign in a cuneiform sign registry based on the Graphemon model as elaborated in Chapter 4. The Grapheme may also be defined using a character description language such as PaleoCodage in JTF and matched once the JTF file is processed at its target repository. Another important inclusion on the sign level is the URI of the reading of the respective sign and its Unicode value. Further information to be added on the level of a character is all annotation information conducted on image data.

Listing 5.2: JTF-LD with image annotation information on the sign level

```
1 {"_class":"chr",
    ↪ "@id":"tab1_side_1_column_1_line_1_seq_1_char_1",
2 "@type":"Glyph", "type":"U+1200", "value":"a",
3 "grapheme":"GRAPHEMEID",
4 "glyphrep":"GLYPHREPRESENTATIONLINK"}
```

5.2.2 JTF-LD Semantic and Metadata Extension

Semantic annotations in cuneiform transliteration texts provide the means to express unique statements about the meaning of word forms in the transliteration text. In JTF-LD, annotations on the word level are attached to the word JSON objects in the JSON object hierarchy.

Listing 5.3: JTF-LD with image annotation information on the sign level

```
1 {"_class":"sequence", "@id":"tab1_side_1_column_1_line_1_seq_1",
    ↪ "@type":"short",
2 "pos":["olia:noun"],"semantic":[{"rel":"type","wd:Q146"],
3 "reading":{"value":"Awilum","language":"akk"},
4 "children":[
5 ...
6 ]}
```

Listing 5.3 shows text annotations included on a word level directly merged into the JSON object. Semantic annotations, part of speech tags, and differing readings, as suggested before may be attached to any wordform occurrence.

5.2.3 JTF-LD Annotation Encoding

Finally, to allow annotations on different image media to be included in JTF-LD, it provides support for annotations in the W3C Web Annotation data model. It relies on the assumption that JTF-LD also allows encoding annotations created in the web annotation data model, such as in Chapter 3.

Listing 5.4: Example of an annotation of a cuneiform sign in the web annotation data model encoded in a JTF representation

```
{
...
{"_class":"sequence", "@id":"tab1_side_1_column_1_line_1_seq_1",
   ↪ "@type":"short", "children":[
  {"_class":"chr",
     ↪ "@id":"tab1_side_1_column_1_line_1_seq_1_char_1",
  "@type":"Glyph", "type":"U+1200", "value":"a",
  "grapheme":"GRAPHEMEID",
  "glyphrep":"GLYPHREPRESENTATIONLINK"
  "isRepresentedBy":{
  { "@type":"Annotation",
    "hasBody":{
        "@type":"SpecificResource",
        "value":"tab1_side_1_column_1_line_1_seq_1_char_1",
    },
    "hasTarget":"http://...."
    }
}]}
...
}
```

Listing 5.4 shows an example of encoding an image annotation in JTF. The content of the image annotation does not deviate from an image annotation in the Cuneiform Annotator application but is merely entered in the appropriate JSON encoding. Annotation targets, such as image media, may be accessible online and must be resolved by the viewer application. Alternatively, they might be bundled with the JTF-LD file using relative paths.

5.2.4 JTF-LD-supported bundling of transliterations

JTF-LD may be used to create bundles of files that support the creation of transliteration packages, for example, in ZIP files of the following contents:

- Image media (2D and 3D)
- Transliteration in JTF-LD with relative links to image media in the same ZIP file
- Annotations included in the JTF-LD encoding
- Creation of a customized Unicode cuneiform font for the respective tablet based on annotated PaleoCodage annotations as suggested in Section 4.2.6

- Creation of local glossaries/dictionaries in linked open data within JTF-LD, depending on the completeness of the annotation content

While specific software needs to be developed to view all of the aforementioned components in company, it can serve as a unified format for distributing these transliterations.

From a linked open data perspective, a fully annotated JTF-LD file can be directly included in any triple store implementation. With properly hosted image data outside of the JTF-LD bundle, JTF-LD may be used in digital edition environments as an interoperable format between transliteration provision and linked data provision. In the next section, an application case of data that could be saved in this new JTF-LD format exemplifies the advantages of the new linked data provision in practice.

5.2.5 Implications of JTF-LD usage

JTF-LD, as a JSON-based linked open data representation, provides the advantages of a self-contained data format for scholars and the benefits of a data format that can be directly imported into a linked open data database, e.g., a triple store. For scholars, a format like JTF-LD can sustain already established workflows, such as easily sharing transliteration documents. Even though transliteration variants may be preserved within JTF-LD documents so that the scholar may have maximum flexibility in the expression of the document, if it contains enough annotation information, it can provide enough information also to support digital applications. A disadvantage of the JTF-LD format is that it is not immediately human-readable like other transliteration formats. A user needs a viewer application to view its contents in a comparable way to common transliteration formats or needs to convert JTF-LD content to one of the other standard transliteration formats with the loss of certain information. However, one could argue that hosting transliteration data as JTF-LD provides more advantages despite requiring the needs of a viewer application since, as compared to other transliteration formats, different media annotations and links may be directly embedded. Assuming all referenced media are accessible to the scholar operating the JTF-LD viewer, the scholar may have better opportunities to gain scientific insights from the respective source material.

5.3 MaiCuBeDa

As was established in Chapter 2, the first step of analyzing a cuneiform text using a computer is to recognize the cuneiform signs on an appropriate medium, such as a 3D scan of an image or a 3D rendering.

Machine-learning algorithms need sample data to build training sets for different machine-learning approaches to recognize cuneiform characters from these mediums automatically. In related work, [DKMO20] conducted cuneiform sign recognition on extracted and automatically segmented photos gained from the Cuneiform Digital Library Initiative (CDLI) [Eng16]. However, photos are less precise in picturing cuneiform signs than, e.g., 3D renderings, which can often more accurately depict cuneiform sign variants because of their customizability of lightning and depth. In addition, despite the collection of over 3000 cuneiform signs in the corpus of [DKMO20], the authors still remark:

> Deep learning requires large amounts of training data in the form of bounding boxes around cuneiform signs, which are not readily available and costly to obtain in cuneiform script.

To overcome this situation, this thesis contributes the Mainz Cuneiform Benchmark Dataset (MaiCuBeDa) [MH23], a dataset of about 30.000 image annotations on renderings of cuneiform 3D scans provided by the Heidelberg Cuneiform Benchmark Dataset [Mar19] (HeiCuBeDa). These annotations were conducted following the annotation model described in Chapter 3 and constitute a practical application case shown in this thesis. MaiCuBeDa is described using a knowledge graph with the components described in the previous sections. This allows for a targeted selection of annotation images for specific image classification tasks, such as sign recognition, and to leverage additional information from the linked open data cloud as machine learning features.

The following describes the MaiCuBeDa dataset and its extraction process from the linked data graph. Some sample classifications on the extracted data show the potential of using linked data technologies for machine learning dataset generation and the first classifications on renderings of 3D images. Finally, a linked-data-based evaluation method of the classification results using crowdsourcing is presented in Section 6.1.5.3.

5.3.1 Data collection and preparation

MaiCuBeDa consists of annotations of cuneiform signs on renderings of the 3D models of the Hilprecht Collection. The annotations, which are represented in the W3C Web Annotation Data Model in RDF were created using the Cuneiform Annotator (cf. Figure 6.3). This JavaScript-based image annotation tool will be discussed in detail in Section 6.1.3. Out of the 1988 cuneiform tablet 3D scans available in the Hilprecht Collection, about 500 transliterations were attested in the Cuneiform Digital Library Initiative (CDLI). Signs have been annotated according to the transliterations given and saved as one JSON-LD file per cuneiform tablet surface as a basis for the knowledge graph. Transliteration data are converted to the Section 5.1 and linked to 2D annotations using their annotation information. That is, each annotated cuneiform sign identified by the tablets name, surface, line, and character index relative to its line is associated with a URI comprised of this information, e.g., `http://example.org/cuneiform/hs1174_front_line1_char1_glyph` is linked to a corresponding URI in the transliteration, e.g., `http://example.org/cuneiform/hs1174_transliteration1_front_line1_char1`. The same is done on a word level and a line level.

In the next step, a post-processing script analyses the annotation information and crops 2D images for lines, cuneiform signs, and words. It creates a knowledge graph representation according to the holistic ontology model introduced in Section 5.1. Finally, additional metadata about the cuneiform tablet artifacts, such as the tablet's genre, language, and time period, is added to the knowledge graph as attributes of the cuneiform artifact itself. The knowledge graph can now be published online or used locally to find relevant data for creating machine learning features.

Listing 5.5: SPARQL query to select 3D rendering images of cuneiform signs from the Old Babylonian time period which are not damaged

```
SELECT DISTINCT ?signimage ?sign ?signunicode ?signname WHERE {
    ?sign rdf:type cidoc:TX9_Glyph ;
          cunei:isRepresentedBy ?signimage ;
          graphemon:hasUnicodeCodePoint ?signunicode ;
          graphemon:signName ?signname ;
          cunei:isDamaged "false"^^xsd:boolean ;
          cidoc:TXP8_is_component_of ?sign_writtentext .
    ?sign_writtentext cidoc:P56_found_on ?tablet .
    ?tablet dc:temporal cunei:OldBabylonian .
}
```

Listing 5.5 shows how the knowledge graph can extract cuneiform sign images of a certain time period with specific features. In the case of Listing 5.5, one might think about a corpus of images to be trained to recognize cuneiform signs on 3D renderings of Old Babylonian texts. Another application case might be to extract only sign images that occur sufficiently often for a machine learning classification, e.g., only consider classifying cuneiform sign identifiers with at least 10 occurrences. The following assumes that all classification datasets are extracted and/or prepared using results from a knowledge graph query. The actual queries are omitted for brevity.

5.3.2 Data corpus definition

The MaiCuBeDa corpus at the time of writing consisted of the following annotation data:

- Reading of the sign in the transliteration and its Unicode code point
- Sign Index
- Line Index
- A tag set consisting of annotation vocabularies described in Chapter 3
- Metadata derived from the cuneiform artifact

However, this thesis focuses on the benefits of linked open data for cuneiform studies. Hence, the focus of classifications using MaiCuBeDa data is on basic classifications, which can be created by considering mainly the image data with limited metadata. Future work should fully exploit more characteristics of the cuneiform knowledge graph for this purpose.

Listing 5.6: MaiCuBeDa Machine Learning Dataset in the Attribute-Relation File Format (ARFF) specification

```
1 @Relation
2 @ATTRIBUTE filename string
3 @ATTRIBUTE class {1(disz), A, AN....}
4 @data
5 d_1_HS_1100_06_back.jpg,AN
```

```
6 2(disz)_1_HS_1137_06_back.jpg,MIN_(2)
7 e3_1_HS_1100_06_back.jpg,UD_(BABBAR)+DU+
```

Listing 5.6 shows the general structure of the machine learning dataset for sign recognition. A file path to an extracted image annotation from the knowledge graph and its classification label is given. The set of classification labels is adjusted if, e.g., the time period of the given cuneiform sign image is to be classified. The filename includes further potential machine learning features, such as the transliteration, position, surface, and cuneiform tablet name, all of which could be included as machine learning features in the machine learning vector but are left for future work in the context of this thesis.

The classification target is not taken from the transliteration, as many readings per cuneiform sign exist. Instead, as attested in the Unicode proposal, the sign name is used to form the class name for the machine learning dataset. In some cases, one expression in a transliteration relates to more than one cuneiform sign, like in the case of the transliteration **e3**. **e3** is therefore mapped to one classification of the first sign name **UD_(BABBAR)**, connected to the second sign name **DU**.

Hence, targets for sign classification consist of either a single cuneiform sign or a combination of cuneiform signs as mandated by transliteration. These sign combinations are so common that they have been assigned their own reading. Therefore, they are considered their own class in the training dataset. However, this is just one possible interpretation of sign classification. Treating sign combinations of cuneiform signs as separate signs would be an equally valid interpretation. Comparative works in the future could reveal if there are significant differences between the treatment of cuneiform signs in classification tasks. The Weka EdgeHistogramFilter[9] was used to extract 80 features from single cuneiform sign images as a preprocessing step for the actual classifications.

Considering the available data, two classification tasks are conducted in this chapter, which needs two training data sets. The first classification task tries to classify the correct cuneiform sign (combination) as a Sign name / Unicode identifier from a given corpus of cuneiform sign images presented as cropped images and embedded in a knowledge graph. The training dataset had the following parameters:

Instances	27589
Classes	256
Minimum Instance Count per Class	10 (sign GESZ2)
Maximum Instance Count per Class	1008 (sign AN)

Table 5.1: Training dataset for sign recognition extracted from the knowledge graph. Only cuneiform sign (combinations) with at least 10 instances were extracted for as the training data set

The second classification task tries to determine the time period of a cuneiform sign depicted in the given image. The training dataset had the following parameters:

[9] https://github.com/mmayo888/ImageFilter/blob/master/ImageFilter/src/weka/filters/unsupervised/instance/imagefilter/EdgeHistogramFilter.java

Instances	28226
Classes	7
Minimum Instance Count per Class	444 (Old Akkadian)
Minimum Instance Count per Class	17766 (Ur III)

Table 5.2: Training dataset for time period classification task. Instances are mapped to one of seven available periods, as extracted from the metadata of the cuneiform artifacts in the knowledge graph. The minimum instance count per class is 444 instances per time period

Further classifications that account for paleographic features, part of speech tags, or semantic classifications of text contents are exciting new prospects enabled by the knowledge graph. They are left for future work.

5.3.3 Machine Learning Experiments and Discussion

A set of different machine learning algorithms was chosen for each classification task to test the bandwidth of different classification tasks. This section presents the results of the machine learning classifications performed on the extracted parts of the knowledge graph. All machine learning classifications have been performed using the Weka machine learning toolkit [SF16] version 3.8.5 using the default settings of the classifiers of this toolkit. Machine learning parameters have been added to the table where appropriate. Optimizing the parameters and evaluating more sophisticated machine learning approaches is considered out of the scope of this work. However, the following results might indicate a baseline upon which further approaches might improve once the image corpus is released publicly.

Algorithm	Precision	Recall	F-Score	F-Score training set=test set
J48	13%	17%	15%	72%
NaiveBayes	29%	30%	30%	35%
IBk (LinearNNSearch)	33%	32%	33%	99%
Random Forest	14%	14%	14%	99%
Logistic Regression	31%	34%	32%	38%

Table 5.3: Classification results: Sign classification task using the respective classifiers and a 10-fold cross-validation approach and a comparison classification where the training set equals the test set

The results in Table 5.3 show that about a third of cuneiform signs can be classified correctly with the simple classifications attempted. These preliminary results indicate that further features apart from purely image-based features might be needed to classify single cuneiform signs more reliably. Considerations in this direction can be the context of the cuneiform sign or the classifications of cuneiform signs in specific periods or modeled with specific PaleoCodes. However, more sign annotations, in total, would benefit a better classification as the training set extracted from the knowledge graph, as the number of instances for the different signs in the MaiCuBeDa corpus vary significantly. The dataset is, therefore, currently unbalanced in this regard. On top of that, each class

contains possibly differing grapheme variants, which were annotated but not yet classified in the dataset. As discussed in Section 4.3.1.1, there are signs whose shape is the same but represent different Unicode code points. These are classifications with which the machine learning approach will also have problems. In sum, MaiCuBeDa is one example of extracting data from a knowledge graph to reach these conclusions. The shortcomings identified in this experiment and the call for more data coverage can likely be mitigated with further additions to said knowledge graph by Assyriologists.

Algorithm	Precision	Recall	F-Score	F-Score training set=test set
J48	51%	51%	51%	90%
NaiveBayes	56%	46%	50%	50.3%
IBk (LinearNNSearch)	60%	63%	61%	100%
RandomForest	64%	64%	64%	99%
Logistic Regression	60%	65%	58%	58%

Table 5.4: Classification results: Time period classification task using the respective classifiers and a 10-fold cross-validation approach and a comparison classification where the training set equals the test set

Time period classifications on the extracted MaiCuBeDa set work more reliably (cf. Table 5.4), most likely because there are fewer classes to classify. The ambiguity among these classes is not as big as in the sign classification set in which many sign variants from different periods must be classified. Still, classifications are by no means perfect. One reason is that periods besides UrIII are comparatively underrepresented in the MaiCuBeDa corpus. Consequently, classification results in the other periods perform worse than UrIII classifications, leading to the results we can observe in Table 5.4. Two main takeaways exist here. At first, the need to enhance the MaiCuBeDa corpus with more annotations in different periods is exemplified. Secondly, a better, more equal representation of the corpus could be achieved by limiting the amount of UrIII instances that could be attempted in a future experiment. Linked data technologies would allow monitoring the progress of the MaiCuBeDa dataset by querying the status of annotations for specific periods and by indicating to researchers which periods are more urgently needed to be annotated.

To conclude: MaiCuBeDa has highlighted the importance of quality-assured data for machine learning purposes and the need to monitor their While this result is interesting to be improved upon, it can only limitedly express the problems with classifying cuneiform signs from image corpus data. The infusion of expert knowledge and the procurement of further well-annotated training instances by Assyriologists would seem prudent to improve machine learning classifications further.

5.3.4 Crowdsourcing Evaluation

While the machine learning evaluation in Section 5.3.3 can give information about the performance of the supervised machine learning task, it must be acknowledged that the data used as the foundation of the machine learning task are not necessarily quality-assured - as a computer scientist collected this data. In addition, texts entered into the

Cuneiform Digital Library Initiative (CDLI) repository might be error-prone. They may have been entered by students with less experience than the individual author or may consist of disputed signs in the research community. One possible way to overcome this problem is to prepare the machine learning data and, by extension, the knowledge graph, which represents possible machine learning data with results from crowdsourcing tasks, as illustrated also by related work [LNMFRS+20]. To that end, options to incorporate crowdsourcing have been explored by creating different crowdsourcing test sets from the given graph data. The test sets have been published on the crowdsourcing platform Zooniverse[10] and were discussed with students and senior researchers in Assyriology. In the following, the results of these discussions with respect to improving machine learning training data are briefly discussed.

5.3.4.1 Crowdsourcing: The expert crowd

The first crowdsourcing test set asked whether a cuneiform sign has been classified correctly (cf. Figure 5.8). To identify a cuneiform sign in a crowdsourcing task involving humans, these humans need to be experts in their respective fields, i.e., at least Assyriology students of a certain semester.

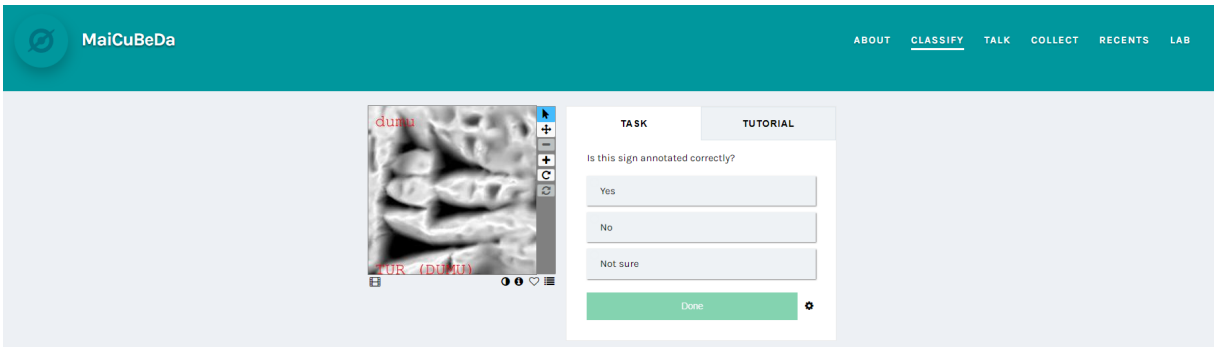


Figure 5.8: Zooniverse crowdsourcing expert task example: Is the cuneiform sign correctly annotated?

Participants were presented with an image of the cuneiform sign, which has been used in the machine learning classification task, and an image overlay that reveals the sign name and the sign reading in the context of the transliteration. The result of this classification might be correct, incorrect, or unknown. It might be used to correct annotations and can, in its entirety, capture a degree of uncertainty in the community of Assyriologists to identify the particular sign. As a machine learning feature, it can be treated as a confidence score, which may be calculated from the percentage of correctly assigned votes set based on the percentage value of the negatively assigned votes. A target audience for this kind of crowdsourcing task could be a conference of cuneiform scholars or the audience of lectures on cuneiform studies. Depending on the lecture type, Crowdsourcing

[10] https://www.zooniverse.org/projects/situx/maicubeda/

tasks may also be considered educational material to be discussed with students in a plenum. However, in practice, even experts needed a line context to conclude satisfactorily about the authenticity of an annotation. This improvement can easily be added to the crowdsourcing dataset and provided in MaiCuBeDa.

5.3.4.2 Crowdsourcing: The layman crowd

The second crowdsourcing test set asks whether two cuneiform signs are identical. The first cuneiform sign is extracted from the MaiCuBeDa corpus; the second cuneiform sign is extracted from a standardized cuneiform font for the time period of the cuneiform tablet.

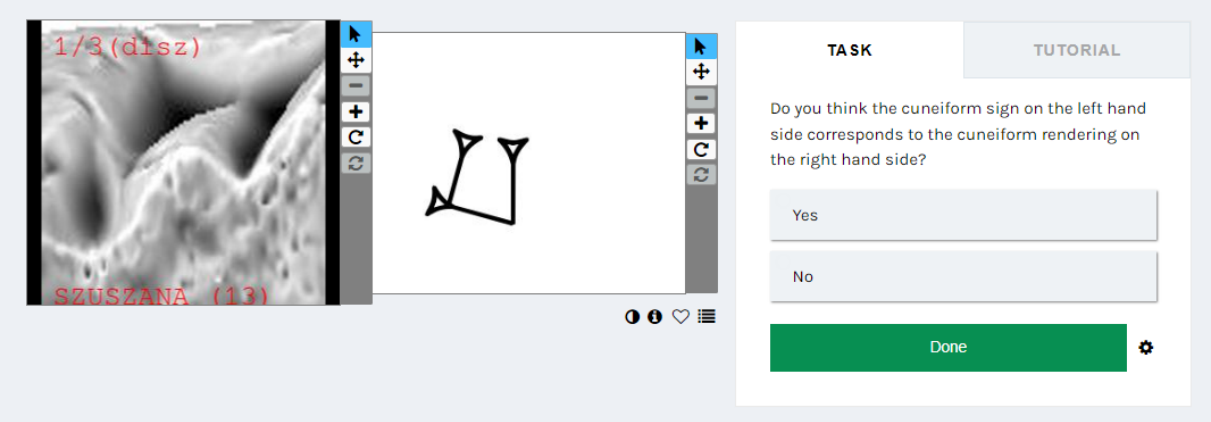


Figure 5.9: Zooniverse crowdsourcing layman task example: Is this cuneiform sign the same as in the font image?

Figure 5.9 shows an example of this type of classification, which, due to its nature, can also be solved by a layman. The result of this classification is likely to either reflect mistakes in the classification of the source material or uncover cuneiform sign variants that do not look like the de facto standardized cuneiform sign in the given time period-specific font. Therefore, it can be a valuable hint to Assyriologists on how to improve their annotation data or whether some cuneiform sign variants need a more close investigation by experts.

5.3.4.3 Crowdsourcing Paleography

The third crowdsourcing test set asked users to repaint the cuneiform wedges of an image.

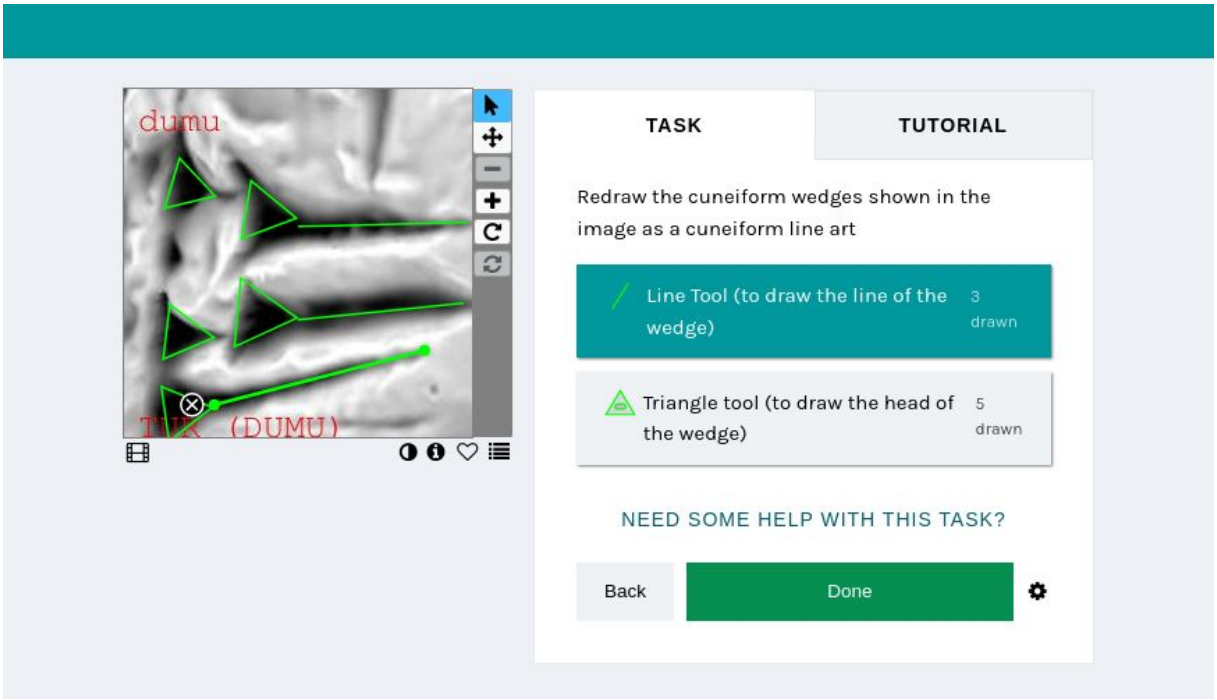


Figure 5.10: Zooniverse crowdsourcing paleography task example: Repaint the cuneiform sign for paleographic studies

Figure 5.10 shows an example of this task. Users can use two drawing tools to repaint wedges on images displaying cuneiform signs. The triangle drawing tool allows drawing the head of a cuneiform wedge or, in the case of the Winkelhaken wedge, the entire wedge. The line drawing tool is used to draw the wedge line beginning from the wedge head to the end of the wedge. According to either Gottstein or Paleocodage, the wedge type may be inferred from the coordinates of the overlayed drawings. The results of this crowdsourcing task may be used for the following purposes:

1. Results may be converted to PaleoCodes and may be shown to scholars who may judge upon their categorization
2. Results may be used to verify PaleoCodes, which have been used as machine-learning features
3. Results may be used to document perceived variations of cuneiform signs by scholars to further the scientific discourse
4. Results may be added to the knowledge graph and treated as data quality metrics similar to the proposed method in Section 3.3.2

Outsourcing the creation of PaleoCodes to a crowdsourcing audience seems like an interesting prospect to explore, especially since paleographic annotations are very costly to obtain in research projects.

5.3.5 MaiCuBeDa Future Perspectives

MaiCuBeDa is suitable to be published as a standardized image classification corpus created on 3D renderings once it covers a sufficient amount of cuneiform signs and possibly sign variants from sufficiently many time periods of cuneiform history. In the future, the initial experiments conducted in this thesis should be followed up in appropriate settings, such as workshops at Machine Learning conferences that compete for the best machine learning algorithms for image recognition. Also, machine learning experiments on related media, such as images of cuneiform signs originating from the same spatio-temporal contexts, would be suitable comparison benchmarks to MaiCuBeDa.

Finally, another future perspective for MaiCuBeDa is that of a "living" corpus. This means that apart from regular stable releases, MaiCuBeDa can and should be extended by aggregating more online cuneiform image resources. Cuneiform scholars should be encouraged to create annotations on already existing 3D rendering data, which could be incorporated into an instance of the cuneiform annotator tool. With the integration of crowdsourcing approaches mentioned in Section 6.1.5.3, the Assyriology community is enabled to contribute to a better quality of the MaiCuBeDa dataset and, in the long run, also to a better quality of the machine learning dataset and machine learning algorithms relying thereon upon.

5.4 Summary and Discussion

This chapter introduced a holistic ontology model (cf. Section 5.1), which integrates the paleographic components and 3D mesh components and annotations on various mediums introduced in the last chapters with a digital representation of the cuneiform artifacts components and its transliterations. This missing model allows one to access and index each element inside a transliteration and relate it to statements made on equivalent representations in another medium. In this way, computers may conclude cuneiform signs or whole words from various representations and interpretations, as a unifying data model to relate these has been defined. A showcase of four cuneiform tablet representations explained how the different components interact.

Next, Section 5.2 shows how the holistic ontology model combined with the other introduced vocabularies may be used to create a singular transliteration file that Assyriologists may share and which can be converted to other more familiar transliteration representations with possible information loss. This unified JTF-LD format allows the seamless export and sharing of transliterations and artifact information in a single file. At the same time, it can act as an input for linked data repositories. It may be bundled with other media resources and custom-generated fonts to represent local digital edition instances. While the format provides these advantages, software that supports the displaying and parsing of JTF-LD still needs to be developed, or software that converts JTF-LD to other already established formats will be needed - without the objective to replace JTF-LD, but rather to supplement a JTF-LD data export.

Finally, Section 5.3 showed the first application of the holistic ontology model and its features for a machine learning application. The MaiCuBeDa corpus could be created by several targeted SPARQL queries to the defined knowledge base and result in several training sets for machine learning purposes, the results of which could be added to the knowledge graph again to help further the automatic detection of cuneiform signs from 3D renderings. Also, MaiCuBeDa applies several techniques introduced in the previous sections. It leverages annotations in 2D and 3D akin to the defined annotation model, it can utilize the PaleoCodage encoding to enhance machine learning classifications, and it benefits from the knowledge graph connection between the different mediums to generate more specific features for classifications. Yet, for this thesis, only baseline classifications have been conducted. Further classifications will be left to future work when more annotations of cuneiform signs in other periods have been completed. The classification results of MaiCuBeDa also again highlighted the demand for quality-assured digital scholarly edition data for machine learning tasks or the validation of already existing datasets. Applying crowdsourcing tasks for validation seems like a helpful prospect to further pursue in upcoming research.

In the next chapter, a new digital workflow for Assyriologists with a toolset built upon linked open data resources will be introduced, which shows the feasibility of creating the linked open data resources deemed necessary and valuable in this and the previous chapters.

Chapter 6

Results and Discussion

The last chapters have laid the foundations for the representation of features of cuneiform tablets on different mediums, from the cuneiform artifacts over the surfaces of the cuneiform tablet down to individual wedges of cuneiform signs found on a cuneiform tablet and their interpretations.

This chapter proposes the appropriate tools and interfaces to let Assyriologists create these kinds of data in a familiar environment. Section 6.1 introduces the outline of a set of tools called the CuneiformWorkbench, which allows Assyriologists to create transliterations from previously prepared image media and generates (linked) data exports that are usable by different research communities. The CuneiformWorkbench comprises different modules, which will be described in the following. These modules can, to a certain extent, be used independently and are partially implemented at the time of writing. Each module takes advantage of the linked open data architecture introduced previously. Finally, Section 6.2 will discuss the usefulness of all presented data models, tools, and software for the respective research communities.

6.1 The CuneiformWorkbench: A digital edition environment

This section describes a digital edition framework called the CuneiformWorkbench. This framework should enable researchers to create digital scholarly editions and connect the artifacts of these digital editions. First, general considerations for creating the CuneiformWorkbench are discussed. As an environment that should support the creation of a digital scholarly edition, it should be built to accommodate the following thoughts:

1. Create research data that can be reused in other data repositories
2. Provide a coherent view of the digital edition data
3. Provide tool-assisted support for all tasks that should be tackled in the process of the digital scholarly edition
4. Harvest already existing linked open data resources to facilitate the transliteration process

5. Create statistics on the local corpus of cuneiform texts, which allow for first interpretations of the context of the text corpus

6. Provide a self-contained, static, publishable digital scholarly edition

7. Generate data products from the digital edition corpus in community-accepted publishable formats

To achieve this, the goal of the CuneiformWorkbench implementation should be compatible with the principles of FAIR data, that is, to produce data in open standards, using interoperable vocabularies, and with the possibility to reuse these data in different research communities and to add these to different already existing research data repositories. Next, the anticipated users of the CuneiformWorkbench, mainly Assyriologists, should be empowered to use the CuneiformWorkbench without the need to understand or create the digital formats exported from its scope.

6.1.1 Architecture

The CuneiformWorkbench is designed as a web application that builds upon a Git repository as a data backend. The application is designed to be run in a Gitlab environment as a Gitlab page but could be easily adjusted to other Git environments, such as Github. The reasoning behind such a setup is that at the end of the digital scholarly edition, we can expect a set of data objects, which typically evolve in the creation process of the digital edition. These evolving digital objects are best captured in a versioning system like Git. In addition, Git makes it easy to assign creators and contributors to the various created digital objects.

As exemplified in Section 5.1.1.3, all Git history metadata could also be converted to RDF vocabularies so that the research data can be captured in the Git repository. Its entire creation history can be represented in RDF. In that sense, a setup like this already goes beyond the scope of a typical digital scholarly edition, which mostly only publishes the final results without detailed individual contributions during the creation process.

6.1.1.1 Data preparation

The CuneiformWorkbench is thought of as a digital edition environment and, due to the limitations of a Git repository concerning space, is likely unsuitable to host large amounts of image data or 3D models. Not only are the current capacities of Git repositories usually exceeded by the space that 3D models use, but other media, such as 3D models, renderings, or photos, need different workflows of preparation which, depending on the setting of the project, are expected to be completed before the process of interpretation starts. Hence, they can be expected to be hosted (publicly or privately, but accessible) in, e.g., image repositories before the CuneiformWorkbench environment is used.

This aligns with common practices in Assyriology, such as excavating cuneiform tablets and documenting the individual artifacts, which produces images or 3D scans before philologists interpret the corpus of individual cuneiform texts. Building on this premise, the CuneiformWorkbench expects a fully documented and possibly published corpus of source media (image and/or 3D data) before the philologist begins the work with the CuneiformWorkbench as such. Hence, the CuneiformWorkbench expects a set of media URLs to be present before work on the actual transliterations and annotations can begin.

This requires a digital edition project to take care of image media data first or iteratively in its lifetime. At first, 3D models would need to be created using a 3D scanner and with appropriate documentation of the 3D models during the scanning process, as exemplified in Chapter 3. The 3D models would need to be post-processed using software such as Gigamesh, and renderings of these would need to be created using the metadata description model described in Chapter 3. Photos would need to be hosted in some webspace, preferably using an image web service such as IIIF [C+19], with persistent URIs to integrate them into the CuneiformWorkbench environment.

Figure 6.1: Overview screen of a CuneiformWorkbench instance providing access to cuneiform tablets present in previously prepared image media

Once image media have been set up (cf. Figure 6.1), the CuneiformWorkbench environment can be initialized with links to these image media resources. Researchers can then begin creating transliterations in a centralized web view. Usually, the transliteration will be created first from an image or 3D representation. After the transliteration, the researcher is expected to create annotations at their discretion, usually related to the research project's goal. For example, a common goal for researchers might be to mark person names, as their exploitation, later on, might reveal networks of persons who interacted throughout the text contents. Other goals might be the linguistic and semantic annotation of a set of selected texts. With a sufficiently populated knowledge graph, the number of annotations that need to be created can be decreased, as frequently occurring

words might be unambiguously picked up by preprocessing tasks leveraging the local or global cuneiform knowledge graph. In other cases, words to be annotated might be asked of the user for clarification of their sense or grammatical content - an interesting area of research once sufficient knowledge graph resources are available.

Annotations should also be exploitable within the cuneiform workbench environment to automate tasks requiring significant manual input during the research project context. In particular, this concerns the annotation of paleographic particularities and their aggregation in project-specific sign lists. To automate this task, a Graphemon-assisted Paleography module as part of the CuneiformWorkbench will be described in the following.

6.1.2 Paleography Module

The Paleography module allows the definition of cuneiform sign variants within the CuneiformWorkbench environment. A cuneiform sign variant may be defined using the PaleoCodage JavaScript tool [Hom20, HH20] by modeling the sign variant as a grapheme with a PaleoCode. Suppose a global cuneiform sign registry is available. In that case, the PaleoCode may be used to find similarly modeled cuneiform signs (using one of PaleoCodage's proposed similarity metrics and possible further criteria) from a knowledge graph to allow the cuneiform researcher to confirm an existing sign variant or add a new sign variant if this one is considered a unique occurrence. If the sign is registered with a globally unique identifier present in the connected sign registry knowledge base (e.g., Wikidata) the Paleography module saves a reference to the identifier in a locally created knowledge graph of cuneiform sign variants, which is stored in the CuneiformWorkbench Git repository (simply an added RDF file for this purpose).

This local sign registry allows researchers to access graphemes of already encountered signs and is linked to annotations on cuneiform tablet images on which these signs may be reused. Suppose a new sign is encountered, and a grapheme for this sign is modeled using PaleoCodage. In that case, the researcher may create an instance in the local paleography registry and, depending on the regulations of the cuneiform sign registry, prepare it for submission in the global sign registry.

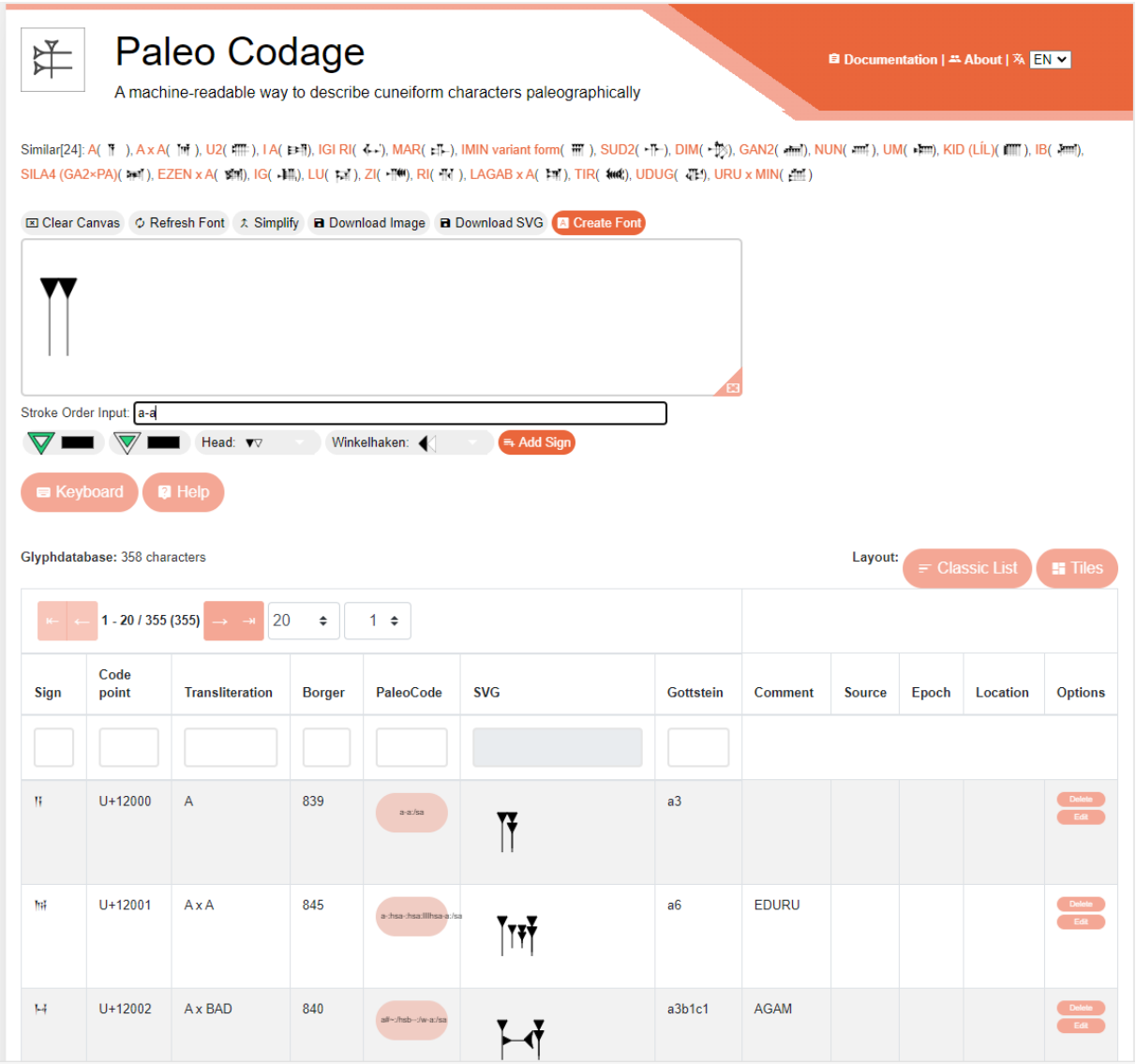


Figure 6.2: Test page of a partially implemented Cuneiform Sign registry in JS and RDF based on [HH20]. Cuneiform signs can be entered and visualized and saved with metadata in the local registry. A synchronization step with a potential global sign registry could be applied.

Figure 6.2 shows a partially implemented cuneiform sign registry based on PaleoCodage, including sign metadata. This sign list could ideally be created from annotations in image media to identify cuneiform signs on the tablet. In this way, annotations document the tablet on which media they are created and contribute to further data products such as sign lists, which can also register their occurrences in the digital edition corpus. Hence, an annotation framework must also be part of the CuneiformWorkbench environment and is presented in the next section.

6.1.3 The Cuneiform Annotator

The Cuneiform Annotator [HMB23] is an annotation tool for cuneiform image resources, which allows the creation of image annotations following the W3C Web Annotation Data Model and the annotation vocabulary introduced in Section 3.5.5. The annotator is implemented as a part of the CuneiformWorkbench environment and is also available as a standalone tool[1] deployed directly on Gitlab. The annotator uses links to image media data to display images or 3D resources on the left-hand side of the user's viewpoint. It allows a user to enter a transliteration text on the right-hand side of their view. Annotations are stored per image in the Git repository and can be instantly retrieved by the Git(lab) page for displaying or modification. This allows easy access for users in Assyriology and guarantees data saved in interoperable data formats.

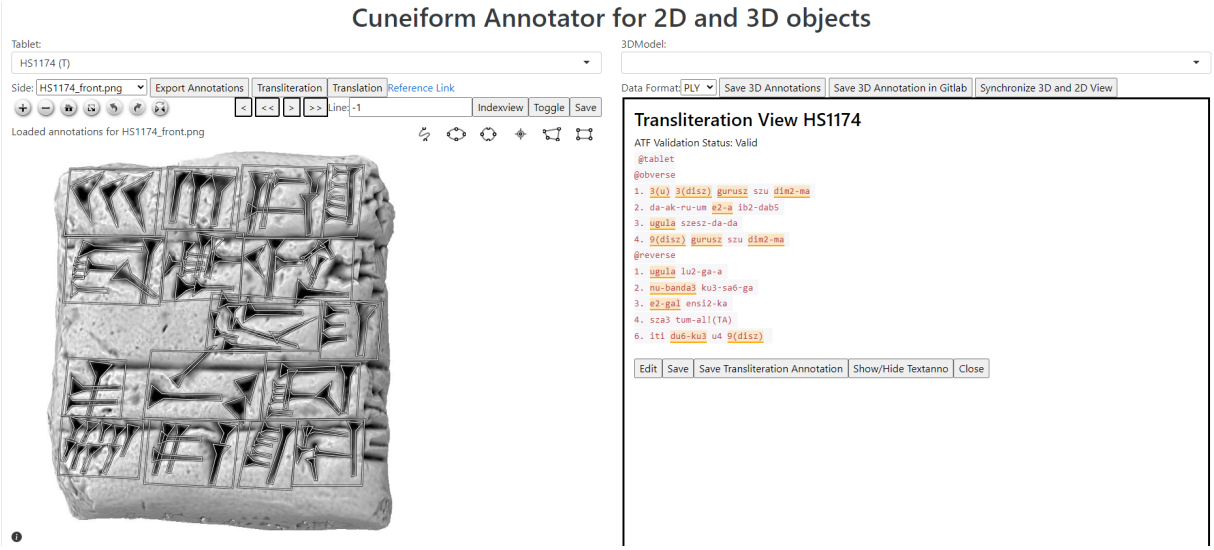


Figure 6.3: Annotations in the Cuneiform Annotator: Sign and wedge annotations on the left-hand side and a transliteration text with text annotations on the right-hand side on renderings of cuneiform tablet HS 1174

Figure 6.3 shows an example of the Cuneiform Annotator, on the HS1174 tablet. In this example, the annotator creates bounding boxes around signs of the cuneiform tablet and annotates each wedge. A wedge annotation includes the wedge type according to PaleoCodage or Gottstein. A sign annotation contains the reading of the transliteration, its index in the transliteration, a possible rotation value, a tagset to describe eventual damages or the sign, a free text comment field, and an optional PaleoCode of the annotated sign. With a cuneiform sign registry or integrated into the CuneiformWorkbench environment, the CuneiformAnnotator could be used as a source of sign data for the Paleography module described in Section 6.1.2.

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

Researchers may annotate the images of cuneiform tablets with the correct PaleoCode. The additional annotation information will become the basis of paleographic entries, which will be verified in the paleography dialog. In this sense, annotating cuneiform signs on tablet images can highlight interesting image data, linking image data visually to transliteration contents and iteratively grow the sign list needed in digital scholarly editions projects involving cuneiform.

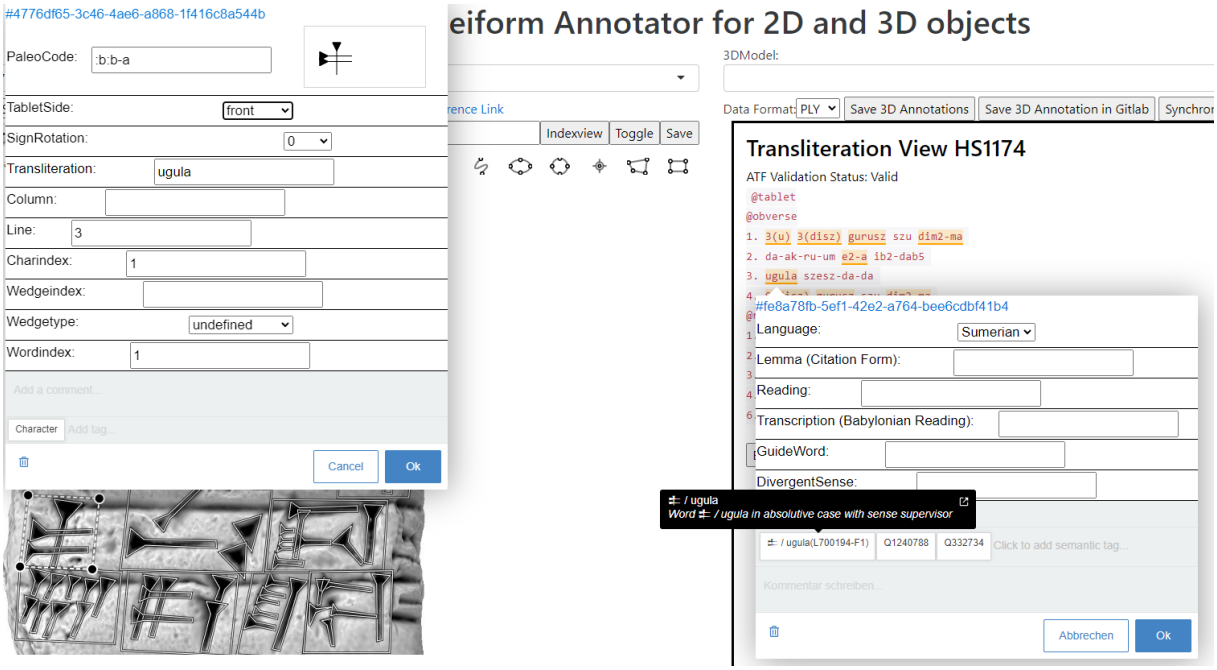


Figure 6.4: Annotation contents of the image and textual annotations: On the left, the cuneiform sign "ugula" is annotated on the 2D rendering, assigned the *PaleoCode :b:b-a* and indexed. On the right, in the transliteration, the word form wd:L700194-F1 (ugula) of lemma wd:L700194 (ugula) with the sense wd:L700194-S2 (overseer) linked to the Wikidata concept wd:Q1240788 (supervisor) is annotated. In the same way, any linguistic annotation attached to the word form wd:L700194-F1 (ugula), here the wd:Q332734 (absolutive case) can be added.

Figure 6.4 shows that the Cuneiform Annotator can also be used to create text annotations, which may or may not be linked to a dictionary knowledge graph. This allows the researcher to create linguistic and semantic annotations next to the image media of the cuneiform tablet. In the future, the cuneiform annotator could be enabled to remember textual or image annotations from an arbitrary knowledge graph to allow the researcher to apply existing annotation contents of the same or similar words in the text. This allows for better annotation consistency and paves the way to research automated pre-annotation approaches, in which results could be visualized and tested in the cuneiform annotator environment.

Finally, the cuneiform annotator allows researchers to check if annotations have been created and linked correctly. Several highlighting options allow one to hover over transliteration parts and image parts and highlight the other components, respectively. To achieve correct indexing in image annotations, highlighting individual lines based on the annotation content is possible.

Figure 6.5: Cuneiform Display and integration of annotations created by the Cuneiform Annotator in the CuneiformWorkbench. Annotations are displayed read-only on the overview page of a cuneiform tablet.

Figure 6.5 shows the integration of the cuneiform annotator in the cuneiform workbench environment. The Cuneiform Workbench is an extension of the Cuneiform Annotator, as it aggregates necessary statistics and data exports and provides web views beyond a simple annotation view. A showcase of created statistics in the Cuneiform Workbench is discussed in the following.

6.1.4 Representation of corpus statistics

Creating linked open data about all artifacts of a cuneiform digital edition is a necessary prospect but neglects aspects of the analysis of the digital edition corpus at hand. The CuneiformWorkbench module for statistics aims to provide some basic corpus statistics, which are also encoded in the linked open data graph. The encoding of this information warrants the extension of the ontology model by a new class cunei:Corpus, which is a representation of the collection of texts of the digital edition as a whole. Corpus instances are subject to the same versioning scheme as individual texts. Apart from an initial corpus instance, a new corpus instance will be generated when a new release is created. This corpus's description follows an aggregation of the included text's metadata and annotation data, which can be automatically created using postprocessing scripts in the CuneiformWorkbench environment.

6.1.4.1 Examples of common statistics

While the Assyriology community provides repositories that provide some basic statistical metrics mainly used for finding texts of specific categories or words in certain texts, there is no consensus on suitable statistics for data science or digital scholarly editions. Often, these are not part of a digital scholarly edition because they do not constitute the main focus - the texts being investigated. However, statistics can be essential to substantiate specific interpretations about the whole text corpus, can be a suitable way to communicate research results of the digital edition, or be a criterion of an algorithm to select a certain corpus to select data for a machine learning task.

With digital scholarly editions becoming more prominent, the question of which every edition project expects statistical metrics will likely be a point of discussion in the upcoming years. The CuneiformWorkbench, at the time of writing, implements specific basic statistics that were motivated from the lens of a computer scientist. These metric results are also included in the ontology model and describe a corpus instance at a certain time.

The first kind of statistical metrics are derived from the text corpus alone and include the frequency of lemmas, phrases, and word forms so that they can be compared to other corpus instances of a global knowledge graph. When paleographic information is added to the knowledge graph, statistics become more interesting to analyze for humans and even when exploited, e.g., as machine learning features. They can also incorporate statistics about word forms used in paleographic writings. In the following, example metrics of the aforementioned categories are presented with a hint in which contexts they could be applied:

Metric	Category	Sample Value
Avg. number of Words/Signs per surface	ML, AS	15
Number of Damages per tablet/Damage degree	AS, NLP	5 - 100% damage
Tablet size/Text content	AS, 3D	20x20x20cm/60 signs
Tablets per find spot	AC, AS	10 per spot
Annotated words per tablet	AC, AS	75%
Sign Variants per sign	ML, AS	ESZ: 2 variants
Similar signs	ML, AS, 3D	A similar to MIN

Table 6.1: Examples of statistical metrics for reuse in the CuneiformWorkbench environment or other application contexts (ML=Machine Learning, NLP=Natural Language Processing, AS=Assyriology, 3D=3D Processing, AC=Archaeology)

Table 6.1 shows a selection of example metrics that can be gained from the knowledge graph, from a data export, and which are partly accessible in the CuneiformWorkbench as visualizations. Dataset selection approaches for machine learning profit from statistics that allow a targeted selection of training data (e.g., large cuneiform tablets or containing a certain amount of text characterized by metadata and annotations). Assyriologists may inquire about the completeness of their digital scholarly edition data during their

creation and the similarities and contexts of different sign variants throughout the given corpus of cuneiform texts. To support and export all aforementioned statistical data, the next section will introduce the anticipated data exports and their further usage in data repositories.

6.1.5 Data exports using Continuous Integration

When publishing data on digital editions of cuneiform corpora, the digital artifacts' persistent, secure, and long-term storage should be prioritized. However, as data should be accessible and usable by a variety of research communities, the exports of the CuneiformWorkbench need to be able to reflect this bandwidth for its maximum efficiency, and many different kinds of data exports can be considered as shown in Figure 6.6.

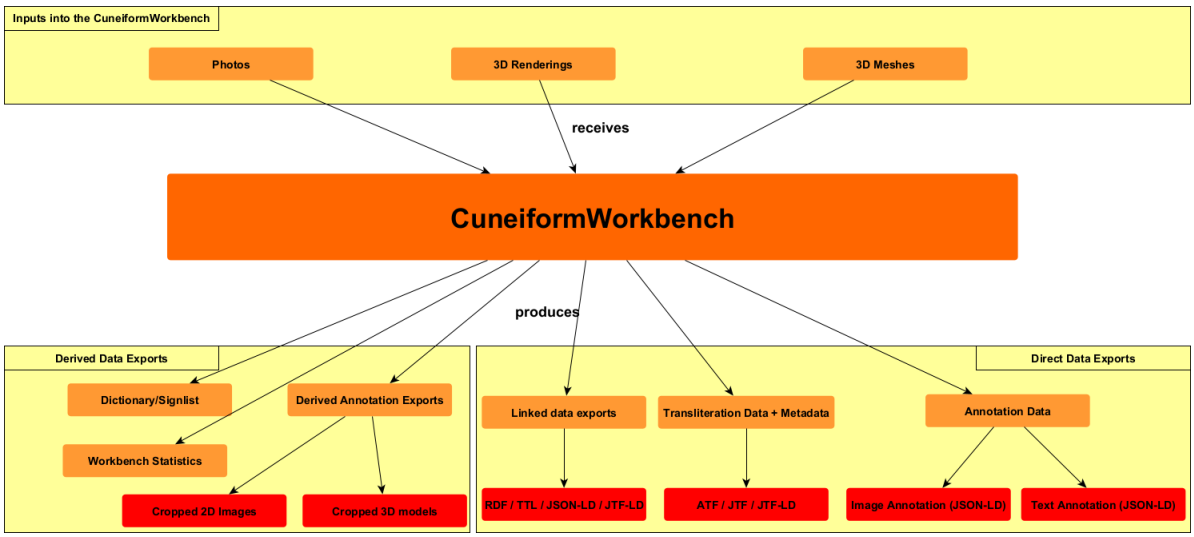


Figure 6.6: Overview of data exports from the CuneiformWorkbench environment. Exports are derived from imports, pick up interpretations in the cuneiform workbench, and generate exports either from directly interpreted information (e.g., transliteration data) or derived information (e.g., dictionary data)

Hence, Section 6.1.5.1 elaborates on the kind of data different targeting research communities anticipate and how the CuneiformWorkbench generates these. Next, Section 6.1.5.3 describes how feedback from the respective researchers can be integrated into the CuneiformWorkbench research environment. Finally, the Section 6.1.5.4 describes the publication of the knowledge graph and the data exports it links to. Section 6.1.5.5 illustrates how the different exports are used in a practical example of the Haft Tappeh project.

6.1.5.1 Data exports for research communities

The CuneiformWorkbench considers exports for different research communities but, first and foremost, provides exports in linked open data. Annotations, transliterations, paleographic descriptions, and data objects form a localized linked data cloud of the individual digital edition, which can be hosted in any triple-store database. In addition, given enough web space, postprocessing scripts, such as the SPARQLing Unicorn QGIS plugin [TH22] might create HTML visualizations of the individual data items as shown in the showcases in Chapter 5. Derived versions of the knowledge graph are beneficial for various research communities. A summary of data that has been considered for export includes:

- Geospatial data concerning the findspots (GeoJSON [BDD+16])
- Data for corpus linguistics in CoNLL-U or CoNLL-RDF [CF17]
- Exports of transliterations in various formats (e.g., ATF, JTF)
- Dictionary data derived from the digital edition text corpus (e.g., in JSON-LD)

These exports allow for the representation of corpus contents on maps, the use of data in natural language processing contexts, to provide of transliteration formats for Assyriologists, and finally, the creation of dictionary data for the reusage in cuneiform-related applications.

6.1.5.2 Export of the digital edition

Apart from exporting the individual data products of the digital edition, the digital edition itself should be accessible as its data product. The idea is to use a non-editable static version of the CuneiformWorkbench environment, which can be hosted as a static website. This static website will require little to no maintenance by the individual researcher and can be hosted on any webspace after the end of a digital edition project. Since it is linked to already published media sources, it can be easily included in established data repositories such as Zenodo or hosted on an appropriate web space.

6.1.5.3 Crowdsourcing approaches

The enrichment of digital image corpus resources and crowdsourcing approaches such as the ones introduced in Section 6.1.5.3 can be supported by the CuneiformWorkbench as data exports, as a data export of cropped images and configuration files for crowdsourcing platforms may support postprocessing scripts. In Section 6.1.5.3, exports have been generated for the particular tasks of sign recognition and time period classification. Hence, researchers may decide to verify their work after a digital edition project or during the digital edition process on a local crowdsourcing instance in the same institute. Results from crowdsourcing are, in theory, fully convertible to the knowledge graph model. Therefore, they may be displayed as error hints in the CuneiformWorkbench environment. This, however, has not been tested at the time of writing, mainly because of a lack of a sufficiently large crowdsourcing community.

6.1.5.4 Linked data export

There are different ways linked open data can be published on the internet. The first way is to publish linked open data using a SPARQL endpoint, i.e., in a triple store database. This is only a sustainable way of publishing if the SPARQL endpoint is run by an organization, such as a library, that guarantees the hosting of this data for a longer period of time. For most digital scholarly edition projects, this is something that cannot be guaranteed in the funding or duration of the project's lifetime. A second option is a publication as a data dump in a repository such as Zenodo[2]. While this guarantees that the data are persistently stored, access to these data might not be as linked data applications expect. Data cannot be queried; it can only be downloaded in bulk and then processed locally, which is against the initial idea of linked open data. Data in this way is also not easily discoverable and accessible by Assyriologists. A third solution, which the cuneiform workbench supports, is a data export of linked open data as targeted RDF files which could be hosted as triple pattern fragments [VVH+16], in Solid Pods, [MSH+16] or simply as a data dump which could be directly accessed with Javascript libraries [HG12]. The HTML export solution of the CuneiformWorkbench provides the latter option, and the other options can be used at the user's discretion.

6.1.5.5 Data publication

To apply the previously mentioned data export in a real-world use case, the Haft Tappeh project should serve as an example of a digital scholarly edition created using the Cuneiform-Workbench system. In the Haft Tappeh project, the CuneiformWorkbench has been tested as a digital edition environment and produced the previously mentioned data products. Based on these experiences, the following data publication recommendations have been created:

- Publication of image media and 3D models in repositories that support long-term archiving, for example, at a university library[3]
- Publication of transliteration data in C-ATF to be added in respective repositories such as the CDLI or ORACC
- Publication of JTF(-LD) transliterations
- Publication of further data exports in community-led repositories or long-term archives such as Zenodo
- Creation of a static homepage instance of the CuneiformWorkbench instance, which links to all aforementioned media files
- Creation of linked open data documentation and files in HTML

[2] https://zenodo.org
[3] https://heidicon.ub.uni-heidelberg.de

In this way, the consistency of the digital scholarly edition, including its generated data exports, is preserved in a static homepage, an instance of the CuneiformWorkbench without the possibility of editing, which may be hosted on any webspace without a high degree of maintenance and independent of web services that need to be operative. The data publication, meanwhile, can be accessed in single files in the repositories best suited for the type of data in question. Including all data exports in different long-term storage repositories also means that both the digital scholarly edition and all of its data objects are citable independently and can be assigned independent creator and contributor information. This allows for an easy reusage and author attribution of any part of the digital scholarly edition in traditional scholarly publications discussing its content. Finally, the linked open data export guarantees machine-readable accessibility of the data in question across different media and data formats.

6.2 Discussion

The previous chapters have defined the data models and ways to digitize the essential elements of a digital edition: The 3D scan, 3D renderings, photos, transliterations, and finally, data products to be generated as parts of the digital edition process, such as statistics. This chapter has introduced tools that enable Assyriologists to create data reusable by various research communities. This section would like to discuss the suitability of the proposed toolchain, technologies, and workflow for the daily work of an Assyriologist in Section 6.2.1, the appropriateness and potentials of a linked open data edition in this way, and finally summarize the work of this thesis with respect to the research questions posed in Section 1.3.

6.2.1 Digital edition workflow

This section wants to reflect on the implications of these digital tools for the workflow of an Assyriologist and the computer science community. The first question to be answered is whether this toolset can be used in a digital scholarly edition setting for Assyriologists. Image data is usually available based on experiences from the Haft Tappeh project and further correspondence with Assyriologists. Still, using the CuneiformWorkbench toolset requires a new approach to cuneiform scholarly edition data. While in an analog workflow, researchers were free to use any available media and create the edition on paper, the digital workflow requires approaching and preparing image media in a first step before creating the transliteration. In the following step, the transliteration can be created in the CuneiformWorkbench web application or in the first step, the Cuneiform Annotator application. However, the creation of the transliteration in the CuneiformWorkbench will go beyond an analog creation of a transliteration, as annotations on image media and in the transliteration itself need to create.

This is because this additional work is always within the scope of a digital edition project, as this additional time might have to be deducted from the time researchers need to analyze the entire corpus of texts. Indeed, so as not to overburden researchers, tasks such as annotations and the creation of transliteration data should be streamlined to a certain extent. This can be achieved with technologies introduced in this thesis, in particular using the following mechanisms:

1. Annotation of cuneiform signs on 2D renderings only, if necessary calculations of 3D annotations from 2D renderings
2. Autocompletion in annotation dialogs of text and image data by taking advantage of data stored in the linked open data cloud
3. Exploration auto automated pre-annotation technologies using the data of the knowledge graph

Creating image annotations on only one medium is one way of streamlining the work under the condition that these annotations are transferable to other media. In addition, image annotations can fulfill multiple tasks at once - the documentation of cuneiform signs on the medium and the description, the identification of the cuneiform sign, and the documentation of its paleography with a PaleoCode. Documenting a PaleoCode becomes much easier if repositories of cuneiform sign variants allow an autocompletion in the annotation dialog based on a half-entered PaleoCode or sign identifier. As sign variants tend to repeat throughout a corpus of cuneiform signs from the same location, it can be expected that the number of paleographic variants to document decreases throughout the creation time of a digital edition as more and more paleographic sign variants are already registered in the system or a global paleographic sign registry. In this way, autocompletion tasks can simplify the work. When annotating the transliteration, already known lexemes registered in Wikidata can also help prefill textual annotation contents such as part of speech annotations recognized from a word's morphological structure.

It can be assumed that most words annotated in a transliteration are well-known vocabularies found in already-established dictionaries. Therefore, they might also be available for pre-annotation from a linked open data dictionary. Words that do not fit this category are arguably the words that are of particular interest for the Assyriologist, as they comprise, for example, person names, unknown locations or commodities, and a possible small subset of new or still unknown words - possibly also the interesting words to report about as a result of a digital scholarly edition.

Following this line of thought, the linked open data cloud can also be exploited for a possible (semi-)automated pre-annotation: An automated algorithm might be used to annotate the transliteration based on heuristics, which consider the linguistic classification and its most likely meaning. The algorithm might ask the Assyriologist for confirmation if many options exist or choose the most likely option, prompting the Assyriologist for correction when doing final corrections on the edition corpus. An appropriate algorithm might also suggest image annotations given enough training data, such as MaiCuBeDa, in the future. A suggestion algorithm for images might:

- Create bounding boxes based on the estimated extent of cuneiform signs found on the image

- An assumed classification of the annotated cuneiform sign based on a pre-trained classifier, e.g., using MaiCuBeDa as a training corpus

Considering these prospects, it can be assumed that the change from a traditional cuneiform edition to a digital scholarly edition provides an overhead of work that is likely to be mitigated as future research advances. On the other hand, this overhead of work also provides further chances for Assyriologists to interpret better the text corpus they are working on. Hence, it might be argued that these tasks are in their interest.

6.2.2 Digital scholarly editions as linked open data

Digital scholarly editions as linked open data provide new opportunities compared to previous representations of digital scholarly editions. The main advantage of a representation of digital scholarly editions in this way is the ability to represent each individual part of the digital scholarly editions as single referencable and possible citable objects. This provides maximum flexibility for other researchers to engage with the given edition data. Instead of describing particularities of cuneiform tablets in text, they can be directly referenced digitally in terms of annotations by anyone. In addition, the reusage of vocabularies and the potential to link new digital scholarly editions to unique identifiers of words, signs, and word forms and their potential discovery in knowledge graphs allows digital scholarly editions to be seen as a part of the cuneiform linked open data cloud, i.e., as a part of a continuously growing machine-readable corpus of knowledge concerning cuneiform artifacts. At the same time, the characteristics of the digital scholarly edition are still present. The linked open data cloud defines the components of the digital scholarly edition in terms of linked data items and serializations of the digital scholarly edition contents; for example, the cuneiform Workbench allows researchers to browse the content of the digital scholarly edition in a user-friendly way. Referring back to the criteria of a digital scholarly edition by [Sah16], the data model supports the representation of cuneiform tablets in various media, the addition of critical information about various media and comments about these, mainly as annotations and metadata of the different documents.

6.2.3 The potentials of the cuneiform linked open data cloud

Linking digital scholarly edition data together and seeing digital scholarly editions of cuneiform texts as parts of the linked open data cloud will allow various new research questions to be tackled primarily because this work has provided the data models for their representations.

Research questions in paleography, for example, about the distribution of cuneiform sign variants, seem like a major point of research in the upcoming years. Paleographic investigations might uncover specific writing behaviors in spatio-temporal contexts but also require a great effort to annotate previously undocumented sign variants in data repositories. Image corpora like MaiCuBeDa might set a first example here, as they might lay the foundation for defining a large subset of suitable sign variants.

In computer science, the strife to automate the processing of cuneiform text artifacts is primarily data-driven. The cuneiform linked open data cloud will likely benefit all natural language processing approaches (e.g., machine translation), the modeling of sign recognition tasks in 2D and 3D, and all experiments that rely on a sufficient amount of corpora data. Suppose the computer science discipline aims to provide machine learning-driven automatic natural language processing pipelines from image media to eventual translations. In that case, they will likely improve with more and better quality-assured training data.

In the long run, linked open data could be enriched with interpretations gained from machine learning results and work in combination with OCR and augmented reality tasks. As a vision into the future, researchers could envision mobile phone apps that, similar to applications like Google Lens, can capture and maybe even translate freshly excavated cuneiform tablet texts. The effort to be taken in various parts of the OCR and NLP pipelines is still enormous, but this target seems worth striving for.

Chapter 7

Conclusions and Outlook

The work presented in this thesis served a main common goal: The definition and consolidation of data models and formats to represent all common facets of cuneiform clay tablets as a research medium. The results have provided the research communities with various linked data models to ensure cuneiform research data interoperability independent of different media used to represent cuneiform data. With these data models, the interoperability and cooperation between research disciplines dealing with cuneiform data are on a good path to becoming easier and, therefore, more accessible even for researchers who cannot read the cuneiform script independently. This work will likely motivate various research groups to discover cuneiform studies as application cases simply because more data can potentially be accessed for interesting research questions. To achieve this goal, this thesis answered it's four defining research questions as illustrated in the following:

RQ1: Paleographic representation

RQ1 asked about the representation of paleographic features of cuneiform texts in linked open data, i.e., to create a paleography cuneiform linked open data cloud. CON1 solves this research question by defining the PaleoCodage encoding and the Graphemon ontology. The PaleoCodage encoding provides a way to model cuneiform signs from its atomic components, the single wedges, to model their similarity and to display them appropriately for researchers using font technologies. Together with the Graphemon ontology, enabling the representation and classification of features modeled in PaleoCodage, they provide the means to represent paleographic particularities of cuneiform signs and their variants in linked open data and interlink these to other digital representations of cuneiform tablets. This allows Assyriologists to document cuneiform sign variants, computer scientists to reuse sign abstractions in classification algorithms, and computational linguists to include the aspects of paleography in their work in natural language processing applications.

RQ2: Annotation model

RQ2 asked for an annotation model to represent particularities of cuneiform artifacts in different media representations. To that end, CON2 allows the representation of 3D meshes with MeshSPARQL, including spatial references with GeoCRS and metadata of their creation using a third ontology model, as a prerequisite to allowing annotations on 3D media. MeshSPARQL acts as a common core vocabulary that enables capturing specifics of the 3D representations describing contents on the 3D model in terms of annotations and finally evaluates and saves provenance and data quality metrics of 3D meshes for computer scientists. The annotation model consists of an annotation vocabulary that may be used in 2D and 3D representations and the definition of 3D annotations on 3D media. The latter also involves the transformation of 2D into 3D annotations under specific circumstances and a way to transfer annotations between rescans of 3D meshes. These contributions allow to generalize annotations on 3D mediums beyond their applications on 3D meshes and provide a first way to integrate 3D meshes into the linked open data cloud. Using the Cuneiform Annotator application, this model has already been applied by a tool in practice.

RQ3: Holistic ontology model

RQ3 asked for an interconnected linked open data graph model that could contribute to classification approaches for machine learning algorithms. This question is answered by CON3, which includes the definition of the Cuneiform Ontology model, a holistic ontology model to combine all discussed media of cuneiform artifact representation in this thesis. The application example on four cuneiform tablets of different time epochs and languages showed its applicability to not only different cuneiform writing styles but also different time epochs and language contents. Complemented by JTF-LD, it provides the core for a knowledge graph from which classification approaches may benefit. This was exemplified by presenting MaiCuBeDa, which used the knowledge graph for extracting essential knowledge to crop specific cuneiform 2D images for sign classification and resourced crowdsourcing classifications back into the knowledge graph.

RQ4: Provision of interoperable research data

RQ4 asked to define and interpret the creation of the data of the defined data models with appropriate tools in the workflow of an Assyriologist, the most reliable source for quality-assured data in Assyriology. As suggested in CON4, the CuneiformWorkbench and its accompanying toolset consisting of the Cuneiform Annotator and the paleographic sign registries showed the application of an appropriate data collection of linked open data in practice. Further, it showed the advantages of this linked data-based digital scholarly edition and its applicability in research contexts, highlighting its advantages and disadvantages. One conclusion to this research question is that despite a significant amount of newly defined tasks concerning annotations and further documentation, the workflow of an Assyriologist is enhanced and not overly hindered. However, the usefulness of the technologies presented depends, as in many linked open data and computer science settings on the amount of data that has been collected and can be reused for the greater

benefit of everyone. To that end, the full potential of, e.g., annotation contents, which may be accessible from the linked open data cloud, could, at the time of writing, not be accessed. It will depend on the adoption of linked open data technologies in Assyriology whether this potential is used.

7.1 Outlook

In the future, much effort should be put into developing research data infrastructures that can implement many of the concepts presented in this thesis. This effort can be seen two-fold. At first, a standardization process will need to be created, similar to OGC or W3C standardization approaches, which would need to be acknowledged and supported by sufficient people from the (digital) Assyriology community. This will allow us to officially recommend and possibly adapt certain technologies, as advocated in this thesis or elsewhere, to give research data infrastructures the foundations for hosting their cuneiform-related data.

At the same time, we can expect neighboring standardization efforts to bring forward results that are usable to the cuneiform scholar community. The Ontolex-Lemon working group is likely to standardize the representation of graphemes, as this concerns a variety of languages and scripts, even apart from cuneiform. The GeoSPARQL working group is keen on creating standards for the representation of 3D data and CRS so that the ontology models proposed in this thesis can help create recommendations adjacent to cuneiform objects. Following the standardization phase, an adoption phase is likely to occur; that is, research data repositories need to adopt these standards, data repositories need to be extended by missing pieces of the linked open data graph, and software supporting these standards should be further developed with the workflow of Assyriologists in mind. At this time, we can expect research questions in computer science and computational linguistics to provide better results that can directly benefit the work of an Assyriologist in their toolchain.

Given better-grounded data foundations for training sets, classifications concerning the cuneiform natural language processing toolchain will likely improve and provide better suggestions for automating digital scholarly edition tasks. In such an environment, Assyriologists can focus more on interpretation tasks and less on repetitive tasks. At the same time, they can provide statistically grounded results that will improve the scientific discourse in Assyriology and beyond.

Appendix A

Linked Data Terminology

This appendix is a short documentation of graphics used in ontology diagrams in this thesis. The following elements are used:

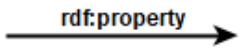
Terminology	URI	Representation	Comment
Ontology Class	owl:Class		Classification
Individual	owl:NamedIndividual		Data instance
Literal	Textual Value + Type		Typed textual content
Property relation	rdf:property		Node to node relation
Type property relation	rdf:type		Instance to class relation

Table A.1: Representations of terminology in ontology model graphics

Appendix B

PaleoCodage Examples

This section gives further examples of using the PaleoCodage encoding to model specific cuneiform signs, further illustrating its capabilities. In addition to remarkable signs, Table B.1 will feature examples of PaleoCodes using additional operators, which have only been described in the accompanied publication [Hom21].

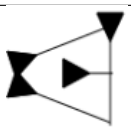
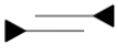
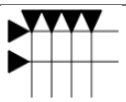
Sign Name	PaleoCode	Image	Comment
A	**a-a:sa**		Sign A with the last wedge slightly moved down / and made smaller
BA	**<(25)d:::::/>(25)c-#;/llsb-:a**		Sign BA with 25° rotation operators **<(25)** and **>(25)** and made smaller
IDIM	**!b:b**		Sign IDIM making use of the inversion operator **!** to invert the direction of the horizontal wedge **b**
U2	**B::B a-a-a-a**		Sign U2 with two long horizontal wedges **B** and one uplifted wedge **a** with the up operator
URI	**sb:sb:sb:sb: :sb:sb:sb:sb- :::f;f- a;a:::-sa;sa-llsa;llsa**		Sign URI with many small long horizontal wedges **sb**, smaller **a** wedges decreasing in size and two **f** wedges
ZA	**sa::sa-sa::sa**		Sign ZA including four vertical wedges of small sizes on top of each other

Table B.1: PaleoCodage Examples highlighting more elaborate examples and further operators

Acronyms

2D Two-dimensional. 25, 28, 33, 37, 42–44, 46, 95, 104, 118, 122

3D Three-dimensional. v, vi, 1, 3, 7, 10–12, 16, 20–33, 35, 37–40, 42, 44–47, 50, 74, 76, 77, 87, 88, 93–96, 103, 104, 106, 107, 110, 116–118, 122, 123

AD Anno Domini. 15

API Application Programming Interface. 3

ARFF Attribute-Relation File Format. 96

ASCII American Standard Code for Information Interchange [Cer69]. 13, 14, 127

ATF ASCII Transliteration Format. 13, 15, 74, 115, 129

BC Before Christ. 15

CDLI Cuneiform Digital Library Initiative. 2, 3, 11, 15, 16, 62, 77, 78, 82, 94, 95, 100, 116

CIDOC Comité international pour la documentation [Doe05]. 64, 78, 127

CIDOC-CRM CIDOC CRM [Doe05]. 11, 28, 77

CoNLL-RDF Conference on Natural Language Learning Format in RDF [CF17]. 115

CoNLL-U Conference on Natural Language Learning Format with Universal Dependencies [BV22]. 115

CRM Conceptual Reference Model. 127

CRMtex Conceptual Reference Model for the study of ancient texts [FM21]. 64

CRS Coordinate Reference System [Tob64]. 25, 123

CS Coordinate System. 23

DE-9IM Dimensionally Extended 9-Intersection Model [Str08]. 25

DWG Domain Working Group. 46

EPSG European Petroleum Survey Group Geodesy [NS08]. 30, 32

ETCSL Electronic Text Corpus of Sumerian Literature [Rob98]. 13

EXIF Exchangeable image file format [Tac01]. 11

FAIR Findable Accessible Interoperable Reusable [WDA+16]. 6, 9, 106

GeoSPARQL Geographic SPARQL Protocol And RDF Query Language ([BK12]). v, 10, 24, 25, 33, 45, 78, 123

GIS Geographic Information System ([Cha17]). 129

GSUB Glyph Substitution Table. 61

HeiCuBeDa Heidelberg Cuneiform Benchmark Dataset [Mar19]. 87, 95

HTML Hypertext Markup Language [BLC95]. 4, 28, 30, 115, 116

ID Identifier. 34, 38, 61, 62

IIIF International Image Interoperability Framework. 107

IRI Internationalized Resource Identifier [DS05]. 10

JPEG Joint Photographic Experts Group Interchange Format [Wal92]. 36, 52, 61

JS JavaScript [WBE20]. ix, 109

JSON JavaScript Object Notation [Bra17]. 13, 90–93, 128

JSON-LD JSON as linked data [LCK20]. 44, 90, 92, 95, 115

JTF JSON Transliteration Format. 13, 90–93, 115, 128

JTF-LD JTF as linked data (Section 5.2). 77, 90–94, 103, 122

KanjiVG Kanji Vector Graphics (`https://kanjivg.tagaini.net`). 52

Lemon Lexicon Model for Ontologies [MBGG+17]. vii, 63, 67, 70, 73, 128

lemonETY Lemon Etymology [Kha18]. 70

LLOD Linguistic Linked Open Data [CCMG20]. 17

LOD Linked Open Data [BK11]. 9, 10

MaiCuBeDa Mainz Cuneiform Benchmark Dataset (Section 5.3). 8, 77, 95, 96, 98, 99, 101, 103, 104, 118, 119, 122

MeshSPARQL Mesh SPARQL Protocol And RDF Query Language (Section 3.2). 24–26, 45, 122

MOCCI The Munich Open-access Cuneiform Corpus Initiative [RW18]. 3

NLP Natural Language Processing [NOMC11]. 77

OCR Optical Character Recognition [MNY99]. 1, 5, 77, E

OGC Open Geospatial Consortium [vR13]. 30, 46, 123

OliA Ontologies of Linguistic Annotation [CS15]. 17

ORACC Open Richly Annotated Cuneiform Corpus. 2, 3, 15, 19, 82, 116

OTF Open Type Font. 61, 135

OWL Web Ontology Language [Gro12]. 10

P-ATF Paleographic ATF [Hom21]. 74

PCA Principal Component Analysis. 40, 41, 47

PDF Portable Document Format [BCASMV93]. 4

PLY Polygon File Format (`http://paulbourke.net/dataformats/ply/`). 25

PROJ Projection Library (`https://proj.org`). 31

QGIS Quantum GIS [Gra13]. 115

RDF Resource Description Framework [Pan09]. vi, ix, 9, 10, 19, 32, 82, 95, 106, 108, 109, 116, 127–129

RDFS Resource Description Framework Schema [BG14]. 10

RINAP The Royal Inscriptions of the Neo-Assyrian Period [Fra11]. 19

SHACL Shape Constraint Language [KK17]. 10

SPARQL SPARQL Protocol And RDF Query Language [SH13]. 10, 45, 46, 95, 104, 116, 128, 129

SRS Spatial Reference System [Lot15]. 25

SVG Scalable Vector Graphics [Qui03]. 36, 37, 52, 58, 61, 72

SVN Subversion [PCSF08]. 82

TEI Text Encoding Initiative [IV95]. 129

TEI/XML TEI/XML Format. 13, 19, 90

URI Uniform Resource Identifier [BLFM05]. vi, 10, 25, 36, 39, 61, 78–81, 88, 92, 95, 107, 125

URL Uniform Resource Locator [BLMM94]. 83, 107

W3C World Wide Web Consortium [Bro15]. vi, 9, 17, 35–37, 44, 46, 63, 74, 95, 110, 123

WKT Well-known text [Lot15]. 31, 37, 46

WWW World Wide Web [BLCL+94]. 9

XML Extensible Markup Language [BPSM97]. 13, 129

XMP Extensible Metadata Platform [Ado04]. 11

Glossary

3D rendering An image rendered using a specific algorithm that represents a surface of a 3D object. In cuneiform studies, a rendering commonly represents the surface of a cuneiform tablet. The appearance of the rendering may be customized depending on the rendering algorithm. v, 2

Akkadian Akkadian is the oldest documented Semitic language and used the cuneiform script as its medium of writing [Hue18]. 12, 16, 17

Akkadogram An Akkadian word, often a noun, used in one of the other attested languages written in cuneiform [KY16]. 16, 133

Area of interest An area on a cuneiform tablet that is of interest to be exploited digitally. vi, 41

Assyriologist A person researching in the area of Assyriology. 1–3, 13, 21, 76, 122, 123

Assyriology Assyriology is the science of the archaeological, anthropological, and linguistic study of the cultures of Assyria and ancient Mesopotamia. This includes the conduction and documentation of artifacts found at excavations, the decipherment of cuneiform artifacts and their interpretation. 1, 14, 72, 76, 123, 131

character A unit of information roughly corresponding to a grapheme, a symbol, or another written or non-written form of a natural language. Commonly, characters are represented using Unicode Codepoints.. 10, 14–16, 50, 52, 135

cognate Describes a relationship between similar words or characters. 70

computational linguistics Computational linguistics is a field of computer science that deals with the study of computational approaches with regard to linguistic questions. 1, 4

crowdsourcing A collaborative effort of a group of dispersed participants which contribute to solving a task, often presented as an online activity. ix, 100–103, 115

Cuneiform Annotator An standalone JavaScript-based annotation tool which forms the core of the CuneiformWorkbench digital edition environment. ix, 110, 111, 117, 122

cuneiform tablet The primary medium of the writing of the cuneiform text, typically made out of clay. v, 7, 12, 50

CuneiformWorkbench A digital edition environment based on Gitlab technologies introduced in Chapter 6. ix, xi, 8, 105–108, 110, 112–117, 122

data quality Data which are fit for use by a data consumer, that is, data which is formatted in such a way and contains content in such a way that a particular objective of the consumer can be achieved. v, 28, 29

data silo A data system incapable of interacting with other similarly structured data systems or prohibited from doing so. 1, 3

digital scholarly edition A scholarly edition which is guided by digital paradigms in their theory, method, and practice [Sah16]. 1, 2, 7, 11, 16, 18–20, 52, 59, 73, 82, 105, 106, 113, 117, 119

etymology The study of this history of word forms. Etymology describes the changes of (cuneiform) characters across history and the possible implications of these changes in semantic or other ways. vii, 70, 71

etymon A word, morpheme or character from which a later word, morpheme or character is derived. 70

F-Score In statistical analysis, the F-Score is calculated as the harmonic mean of the Precision and Recall. 98, 99

firing hole A hole on a cuneiform tablet which is dependent on the process of burning the cuneiform clay. 41

gazetteer A gazetteer is a directory of geographical names used in conjunction with geo positions and additional information about these geographical locations. In cuneiform studies, gazetteers are used to locate find spots of cuneiform tablets and their excavation contexts. 1

Git A distributed version control system which is the basis of the platforms Gitlab and Github. 82, 106

glyph A specific shape that represents a character. In cuneiform studies, a cuneiform sign that is physically present on a cuneiform tablet would be described as a Glyph. 50, 70

grapheme A grapheme is the smallest semantic unit of a script of a given language. 50, 51, 64–66

Graphemon Grapheme Model for Ontologies. An ontology model to describe paleographic elements. viii, 63, 70, 73, 76

Hilprecht Collection The Hilprecht Collection of Babylonian antiquities (Hilprecht Sammlung) is a collection of archaeological objects. It encompasses about 3300 objects with about 3000 cuneiform texts of all relevant epochs of history. 87, 95

Hittite Hittite is an extinct Indo-European language written in the cuneiform script. 16

ideographic An ideographic representation is a symbol representing an idea often understood across languages.. 53

JavaScript A programming language and one of the key technologies for the web [WBE20]. 61, 108

Lemma A form of a Lexeme which represents the canonical form and is often used as a dictionary entry's headword. 133

Lexeme An abstract unit of lexical meaning that a set of words follow. The words are related through inflection. The canonical form of a Lexeme is commonly described as a Lemma. 17, 18, 50, 133

Line Art A (manual) or digital drawing of the surface of a cuneiform tablet that is used as documentation or to highlight important aspects of said surface. v, 2, 3, 12, 16, 22, 52, 76

logogram A logogram is a character that represents a word. In many cuneiform languages, logograms are imported from other cuneiform languages such as in the case of Sumerograms or Akkadograms. 13

machine learning Machine learning describes a field of computer science that builds on methods that leverage data to improve the performance on a set of usually pattern recognition tasks [Zho21]. 5

machine learning feature A measurable property of a machine learning dataset which represents a characteristic necessary for the machine learning algorithm to arrive at its decision. 97

metadata Data which provides information about other data. 26

metric A measurement of a property of a data or software. 24, 25

Middle Eastern Studies The study of the history, culture, and socioeconomic of the Middle East. 1

NaiveBayes A simple classifier which applies the Bayes' theorem with strong independence assumptions between features. 98, 99

Old Persian Old Persian is one of the two directly attested Old Iranian languages. It used a syllabic cuneiform script as its writing system.. 16

Ontolex An abbreviated name for vocabularies for lexical resources created by the W3C Ontology-Lexica Community Group (https://github.com/ontolex). vii, 62, 63, 67, 70, 73

ontology An ontology in computer science encompasses a representation of a domain of discourse by describing sets of concepts and properties representing it. v, vii, viii, 7, 8, 11, 22–24, 30–32, 41, 45, 46, 49, 62–65, 71–73, 75–77, 82, 87, 88, 95, 103, 104, 112, 113, 121–123

PaleoCodage An encoding to represent cuneiform graphemes in a machine-readable way, defined in Chapter 4. vii, xi, 49, 55–62, 64, 65, 70, 72, 73, 76, 104, 108, 121

PaleoCode A code in the PaleoCodage notation typically describes one cuneiform sign variant. vii, 57, 58, 66, 70, 76, 110, 111, 118

Paleography The study of historic handwriting of its shape, cultural and spatio-temporal contexts. 49

pictographic A pictographic representation is a symbol that visually resembles a physical object.. 53

POSTag Part of Speech Tag: A linguistic annotation of a word, typically addressing the characteristic of its wordform. 17

Precision The number of true positive results of a given classification divided by the number of all positive results given by the classifier. 98, 99, 132

provenance Provenance in computer science describes the process to record the history of a given dataset, including its owners, creation, and creation process. 21, 26, 28, 29, 45

raster image A two-dimensional picture which is described by a matrix of pixels, whereas each value of the given image matrix represents a color in a given color encoding system (typically RGB) and which may be georeferenced to represent a surface on a given spheroid. 10

Recall The number of true positive results divided by the number of results that should have been identified as positive. 98, 99, 132

ruling A ruling on a cuneiform tablet surface, often representing the boundary of a given line. 41

scholarly text edition A person who specializes in writing and copying texts. 3

Scribe A person who specializes in writing and copying texts. 1

seal A seal, typically a cylinder seal, is used to roll an impression of cuneiform signs or image content onto a clay surface (such as a cuneiform tablet's surface). 41

sense A sense or lexical sense represents the lexical meaning of a Lexical Entry or Lexeme. In the Ontolex-Lemon model a sense consists of a gloss text and a link to a concept.. 67

Sign Name The name of a cuneiform sign as attested in the Unicode definition. 50

similarity metric A metric which describes a degree of similarity between two data objects, often textual Strings. 70

spatial Data Spatial data describes data which is related to a spatial location. 10

structured data Data which is presented in a well-defined format and follows a given data model. 9

Sumerian Sumerian is the language of ancient Sumer, the first language to be written in the cuneiform script and a language isolate by classification [Jag10] . 12, 16, 17

Sumerogram A Sumerian word, often a noun, used in one of the other attested languages written in cuneiform [KY16]. 16, 133

text corpus A language resource that consists of texts that may or may not be categorized, of the same language, or annotated with additional meta information. 1

Transliteration A transliteration is a conversion of one script to another script by mapping signs or languages using one or many predefined mapping standards. A transliteration in cuneiform studies involves the conversion of one or many cuneiform signs to a reading of these in the Latin script in one of the available transliteration styles. v, 1–3

treebank A treebank is a part of speech tag annotated text corpus that is comprised of optional syntactic and semantic annotations. 17

Unicode An information technology standard which defines a consistent encoding for the mapping of text expressed in the world's writing systems [Kor06]. 4, 10, 14–16, 61, 65

Unicode Codepoint A Unicode code point is a standardized mapping of a numerical value representing a specific character used for character encodings.. vi, 50, 51, 131

UrIII The UrIII dynasty is the last dynasty of the Sumerian culture approx. 2112-2004BC. 99

vector graphics A type of computer graphic in which the appearance of the graphic is derived from geometric primitives on a cartesian plane, e.g., from lines, curves, points, polygons [Gan08]. 10

Web font A web font is a special font customized for embedding into web spaces. The base of a web font is usually an Open Type Font (OTF) or a true type font. 62

wedge The basic stoke-like element of the cuneiform writing system. 11, 12

Winkelhaken A type of cuneiform wedge which is comprised of the wedge head only [Hom21]. xi, 54–56, 102

word glossary A glossary of a corpus of words in cuneiform studies. A glossary typically includes words, word forms, translations, and references to texts including the particular word. 1

List of Namespaces

Bibliography

[ABD+21] Mehwish Alam, Henk Birkholz, Danilo Dessì, Christoph Eberl, Heike Fliegl, Peter Gumbsch, Philipp von Hartrott, Lutz Mädler, Markus Niebel, Harald Sack, et al. Ontology modelling for materials science experiments. In *SEMANTICS Posters&Demos*, 2021. URL https://ceur-ws.org/Vol-2941/paper11.pdf. cited p. 78

[Ado04] XMP Adobe. extensible metadata platform, 2004. cited p. 11, 130

[Atw08] ES Atwell. Development of tag sets for part-of-speech tagging. In A Lüdeling and M Kyto, editors, *Corpus Linguistics: An International Handbook, Volume 1*, volume 1, pages 501 – 526. Walter de Gruyter, December 2008. URL https://eprints.whiterose.ac.uk/81781/. (c) 2008, Walter de Gruyter. Reproduced with permission from the publisher. cited p. 17

[AVL22] Shahed Bassam Almobydeen, José R.R. Viqueira, and Manuel Lama. Geosparql query support for scientific raster array data. *Computers & Geosciences*, 159:105023, 2022. doi:10.1016/j.cageo.2021.105023. cited p. 10

[BC03] Tom Bishop and Richard Cook. A specification for cdl character description language. In *Glyph and Typesetting Workshop*, 2003. URL https://www.unicode.org/L2/L2003/03404-cdl-spec.pdf. cited p. 52

[BCASMV93] Tim Bienz, Richard Cohn, and Calif.) Adobe Systems (Mountain View. *Portable document format reference manual*. Citeseer, 1993. cited p. 4, 129

[BCG17] George Bruseker, Nicola Carboni, and Anaïs Guillem. *Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM*, pages 93–131. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-65370-9$_6$. cited p. 11

[BDD+16] H. Butler, M. Daly, A. Doyle, Sean Gillies, T. Schaub, and Stefan Hagen. The GeoJSON Format. RFC 7946, August 2016. doi:10.17487/RFC7946. cited p. 115

[BE95] Marshall Bern and David Eppstein. Mesh generation and optimal triangulation. In *Computing in Euclidean geometry*, pages 47–123. World Scientific, 1995. doi:10.1142/9789812831699_0003. cited p. 45

[BG14] Dan Brickley and Ramanathan Guha. RDF schema 1.1. W3C recommendation, W3C, February 2014. https://www.w3.org/TR/2014/REC-rdf-schema-20140225/. cited p. 10, 129

[BGH+88] J Buurman, N Grimal, M Hainsworth, J Hallof, and D van der Plas. Inventaire des signes hiéroglyphiques en vue de leur saisie informatique–informatique et egyptologie 2. *Institut de France, Paris,*, 1988. cited p. 56

[BGP00] J.A. Black, A.R. George, and J.N. Postgate. *A Concise Dictionary of Akkadian*. Santag : Arbeiten und Untersuchungen zur Keilschriftkunde. Harrassowitz, 2000. URL https://books.google.de/books?id=-qIuVCsRb98C. cited p. 1

[BHL+21] Toby Burrows, Matthew Holford, David Lewis, Andrew Morrison, Kevin Page, and Athanasios Velios. *Transforming TEI Manuscript Descriptions into RDF Graphs*, volume 15, pages 143–154. BoD, Norderstedt, 2021. URL https://kups.ub.uni-koeln.de/55231/. cited p. 19

[BK11] Florian Bauer and Martin Kaltenböck. Linked open data: The essentials. *Edition mono/monochrom, Vienna*, 710, 2011. cited p. 3, 9, 128

[BK12] Robert Battle and Dave Kolas. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4):355–370, 2012. doi:10.3233/SW-2012-0065. cited p. 10, 128

[BKKM11] Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, and Akira Maeda. A study of traditional mongolian script encodings and rendering: Use of unicode in opentype fonts. *Int. J. Asian Lang. Process.*, 21(1):23–44, 2011. URL https://colips.org/journals/volume21/21.1.3-Biligsaikhan.pdf. cited p. 61

[BLC95] Tim Berners-Lee and Daniel W. Connolly. Hypertext Markup Language - 2.0. Technical Report 1866, Internet Engineering Task Force, November 1995. doi:10.17487/RFC1866. cited p. 4, 128

[BLCL+94] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. The world-wide web. *Commun. ACM*, 37(8):76–82, aug 1994. doi:10.1145/179606.179671. cited p. 9, 130

[BLFM05] Tim Berners-Lee, Roy T. Fielding, and Larry M Masinter. Uniform Resource Identifier (URI): Generic Syntax. Technical Report 3986, Internet Engineering Task Force, January 2005. doi:10.17487/RFC3986. cited p. 10, 129

[BLHL+01] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001. URL https://www.scientificamerican.com/article/the-semantic-web/. cited p. 9

[BLMM94] Tim Berners-Lee, Larry M Masinter, and Mark P. McCahill. Uniform Resource Locators (URL). Technical Report 1738, Internet Engineering Task Force, December 1994. doi:10.17487/RFC1738. cited p. 130

[BM20] Bartosz Bogacz and Hubert Mara. Period classification of 3d cuneiform tablets with geometric neural networks. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 246–251, 2020. doi:10.1109/ICFHR2020.2020.00053. cited p. 22

[BMT08] Ceri Binding, Keith May, and Douglas Tudhope. Semantic interoperability in archaeological datasets: Data mapping and extraction via the cidoc crm. In Birte Christensen-Dalsgaard, Donatella Castelli, Bolette Ammitzbøll Jurik, and Joan Lippincott, editors, *Research and Advanced Technology for Digital Libraries*, pages 280–290, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. doi:10.1007/978-3-540-87599-4_30. cited p. 78

[Bor04] R. Borger. *Mesopotamisches Zeichenlexikon*. Alter Orient und Altes Testament. Ugarit-Verlag, 2004. URL https://books.google.de/books?id=0kGkZwEACAAJ. cited p. 51, 65

[BPSM97] Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen. Extensible markup language. *World Wide Web J.*, 2(4):29–66, November 1997. doi:10.5555/274784.273625. cited p. 130

[Bra17] Tim Bray. The JavaScript Object Notation (JSON) Data Interchange Format. Technical Report 8259, Internet Engineering Task Force, December 2017. doi:10.17487/RFC8259. cited p. 13, 128

[Bro15] Terrence A Brooks. World wide web consortium (w3c). *Encyclopedia of library and information sciences*, pages 5695–5699, 2015. doi:10.1081/E-ELIS3-120044744/world-wide-web-consortium-w3c-terrence-brooks. cited p. 9, 130

[BV22] Luca Brigada Villa. UDeasy: a tool for querying treebanks in CoNLL-U format. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10)*, pages 16–19, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.cmlc-1.3. cited p. 127

[C+19] IIIF Consortium et al. International image interoperability framework, 2019. cited p. 107

[Cat83] Kevin J Cathcart. Edward hincks (1792-1866) and the decipherment of cuneiform writing. *Proceedings of the Irish Biblical Association*, 7:24–43, 1983. cited p. 1

[CCMG20] Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. Linguistic linked open data cloud. In *Linguistic Linked Data: Representation, Generation and Applications*, pages 29–41. Springer International Publishing, Cham, 2020. doi:10.1007/978-3-030-30225-2_3. cited p. 17, 19, 128

[Cer69] Vinton G Cerf. Rfc0020: Ascii format for network interchange. Technical Report 20, Internet Engineering Task Force, October 1969. doi:10.17487/RFC0020. cited p. 127

[CF17] Christian Chiarcos and Christian Fäth. Conll-rdf: Linked corpora done in an nlp-friendly way. In *International Conference on Language, Data and Knowledge*, pages 74–88. Springer, 2017. doi:10.1007/978-3-319-59888-8_6. cited p. 115, 127

[CH22] Nicholas J. Car and Timo Homburg. Geosparql 1.1: Motivations, details and applications of the decadal update to the most important geospatial lod standard. *ISPRS International Journal of Geo-Information*, 11(2), February 2022. doi:10.3390/ijgi11020117. cited p. 10

[Cha17] Kang-Tsung Chang. *Geographic Information System*, pages 1–9. John Wiley & Sons, Ltd, 2017. doi:10.1002/9781118786352.wbieg0152. cited p. 128

[Con22] Unicode Consortium. Cuneiform for unicode 15.0. Technical report, 2022. URL https://www.unicode.org/charts/PDF/U12000.pdf. cited p. 50

[CRH22] Anja Cramer, Laura Raddatz, and Timo Homburg. 3dcap-md-gen, December 2022. doi:10.5281/zenodo.7468298. cited p. 28

[CS15] Christian Chiarcos and Maria Sukhareva. Olia - ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386, 2015. doi:10.3233/SW-140167. cited p. 17, 129

[CWL14] Richard Cyganiak, David Wood, and Markus Lanthaler. RDF 1.1 concepts and abstract syntax. W3C recommendation, W3C, February 2014. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/. cited p. 9

[CYS17] Paolo Ciccarese, Benjamin Young, and Robert Sanderson. Web annotation data model. W3C recommendation, W3C, February 2017. https://www.w3.org/TR/2017/REC-annotation-model-20170223/. cited p. 35

[DB96] Peter T Daniels and William Bright. *The world's writing systems*. Oxford University Press on Demand, 1996. doi:10.2307/416027. cited p. 1

[DFMMO18] Francesco Di Filippo, Massimo Maiocchi, Lucid Milano, and Renzo Orsini. The "ebla digital archives" project: how to deal with methodological and operational issues in the development of cuneiform texts repositories. *Archeologia e Calcolatori*, 29:117–142, 2018. doi:10.19282/ac.29.2018.14. cited p. 13, 15

[DFS+11] Patrick Dengler, Jun Fujisawa, Doug Schepers, Jonathan Watt, Anthony Grasso, Chris Lilley, Cameron McCormack, Dean Jackson, Jon Ferraiolo, and Erik Dahlström. Scalable vector graphics (SVG) 1.1 (second edition). W3C recommendation, W3C, August 2011. https://www.w3.org/TR/2011/REC-SVG11-20110816/. cited p. 36

[DKMO20] Tobias Dencker, Pablo Klinkisch, Stefan M. Maul, and Björn Ommer. Deep learning of cuneiform sign detection with weak supervision using transliteration alignment. *PLOS ONE*, 15(12):1–21, 12 2020. doi:10.1371/journal.pone.0243039. cited p. 94

[DMF20] Martin Doerr, Francesca Murano, and Achille Felicetti. Definition of the crmtex, 2020. cited p. 11

[Doe05] Martin Doerr. The CIDOC CRM, an Ontological Approach to Schema Heterogeneity. In Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors, *Semantic Interoperability and Integration*, volume 4391 of *Dagstuhl Seminar Proceedings (DagSemProc)*, pages 1–5, Dagstuhl, Germany, 2005. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/DagSemProc.04391.22. cited p. 11, 28, 64, 127

[DS05] Martin J. Dürst and Michel Suignard. Internationalized Resource Identifiers (IRIs). Technical Report 3987, Internet Engineering Task Force, January 2005. doi:10.17487/RFC3987. cited p. 10, 128

[DSDV20] Ben De Meester, Tom Seymoens, Anastasia Dimou, and Ruben Verborgh. Implementation-independent function reuse. *Future Generation Computer Systems*, 110:946–959, 2020. doi:10.1016/j.future.2019.10.006. cited p. viii, 72

[Ell02] John L. Ellison. *A paleographic study of the alphabetic cuneiform texts from Ras Shamra /Ugarit*. PhD thesis, Harvard University, 2002. URL https://www.proquest.com/dissertations-theses/paleographic-study-alphabetic-cuneiform-texts-ras/docview/305527804/se-2. cited p. 49

[EMW+32] Erich Ebeling, Bruno Meissner, Ernst F Weidner, Wolfram von Soden, Dietz Otto Edzard, and Michael P Streck. *Reallexikon der Assyriologie und vorderasiatischen Archäologie*. de Gruyter, 1932. cited p. 56, 57

[Eng16] Robert K Englund. The cuneiform digital library initiative: Dl in dh, 2016. doi:10.7916/D8KD1XWN. cited p. 3, 94

[FM21] Achille Felicetti and Francesca Murano. Semantic modeling of textual entities: The crmtex model and the ontological description of ancient texts. *Umanistica Digitale*, 5(11):163–175, Jan. 2021. doi:10.6092/issn.2532-8816/13674. cited p. 64, 127

[Fos26] Charles Fossey. *Manuel d'assyriologie: Évolution des cunéiformes*, volume 2. E. Leroux, 1926. cited p. vii, 51, 69, 70

[Fox22] Daniel A Foxvog. Elementary sumerian glossary (revised 2022). *Cuneiform Digital Library Initiative Preprints*, 2022. URL http://cdli.earth/pubs/cdlp/cdlp0003_20220412.pdf. cited p. 1

[Fra11] Grant Frame. *The royal inscriptions of the Neo-Assyrian period*. Eisenbrauns, 2011. cited p. 19, 129

[FŞA+20] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler. Introduction: What is a knowledge graph? In *Knowledge Graphs: Methodology, Tools and Selected Use Cases*, pages 1–10. Springer International Publishing, Cham, 2020. doi:10.1007/978-3-030-37439-6_1. cited p. 9

[Fun19] William PM Funk. Etymologies of chinese hànzì and japanese kanji: Explanations on liùshū and rikusho. *Chinese Language Teaching Methodology and Technology*, 2(2):54, 2019. URL https://engagedscholarship.csuohio.edu/cltmt/vol2/iss2/6. cited p. 53

[Gab10] Hans Walter Gabler. Theorizing the digital scholarly edition. *Literature Compass*, 7(2):43–56, 2010. doi:10.1111/j.1741-4113.2009.00675.x. cited p. 1

[Gan08] Vijay Gandhi. *Vector Data*, pages 1217–1221. Springer US, Boston, MA, 2008. doi:10.1007/978-0-387-35973-1_1438. cited p. 135

[Got13] Norbert Gottstein. Ein stringentes identifikations- und suchsystem für keilschriftzeichen. *Mitteilungen der Deutschen Orient-Gesellschaft zu Berlin*, 145, 2013. URL http://www.orient-gesellschaft.de/repositorium/MDOG/MDOG_145.pdf. cited p. vi, 54

[Goz13] R Gozzoli. Hieroglyphic text processors, manuel de codage, unicode, and lexicography. *Text, Languages & Information Technology in Egyptology, Ægyptiaca Leodiensia*, 9:89–101, 2013. cited p. 56

[GP20] Daniel Garijo and María Poveda-Villalón. Best practices for implementing FAIR vocabularies and ontologies on the web. *CoRR*, abs/2003.13084, 2020, 2003.13084. URL https://arxiv.org/abs/2003.13084. cited p. 10

[Gra13] Anita Graser. *Learning QGIS 2.0*. Packt Publishing Ltd, 2013. cited p. 129

[Gro12] W3C OWL Working Group. OWL 2 web ontology language document overview (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-overview-20121211/. cited p. 10, 129

[H+11] John Herring et al. Opengis® implementation standard for geographic information-simple feature access-part 1: Common architecture [corrigendum]. Technical report, 2011. URL https://portal.ogc.org/files/?artifact_id=25355. cited p. 37

[HAI21] Eero Hyvonen, Riccardo Albertoni, and Antoine Isaac. Introducing the data quality vocabulary (dqv). *Semant. Web*, 12(1):81–97, jan 2021. doi:10.3233/SW-200382. cited p. v, 29

[Ham04] Eric Hamilton. Jpeg file interchange format. Technical report, 2004. URL https://www.w3.org/Graphics/JPEG/jfif3.pdf. cited p. 36

[HC16] Timo Homburg and Christian Chiarcos. Word segmentation for akkadian cuneiform. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May 2016. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/816_Paper.pdf. cited p. 4

[HCRM21] Timo Homburg, Anja Cramer, Laura Raddatz, and Hubert Mara. Metadata schema and ontology for capturing and processing of 3d cultural heritage objects. *Heritage Science*, July 2021. doi:10.1186/s40494-021-00561-w. cited p. v, 21, 26, 27, 45

[HD22] Timo Homburg and Thierry Declerck. Towards the integration of cuneiform in the ontolex-lemon framework. In *Proceedings of Grapholinguistics in the 21st Century, 2022*, volume 9, pages 265–297. Fluxus Editions, 2022. doi:10.36824/2022-graf-homb. cited p. 64

[HDE17] Gerald Hiebel, Martin Doerr, and Øyvind Eide. Crmgeo: A spatiotemporal extension of cidoc-crm. *International Journal on Digital Libraries*, 18(4):271–279, 2017. doi:10.1007/s00799-016-0192-4. cited p. 11

[HG12] Antonio G Hernández and MNM GARCıA. A javascript rdf store and application library for linked data client applications. In *Devtracks of the, WWW2012, conference. Lyon, France*. Citeseer, 2012. cited p. 116

[HG14] Andreas Harth and Yolanda Gil. Geospatial data integration with linked data and provenance tracking. In *W3C/OGC Linking Geospatial Data Workshop*, pages 1–5. Citeseer, 2014. URL http://www.planet-data.eu/sites/default/files/publications/lgd14_submission_54.pdf. cited p. 10

[HH20] Marc Häuser and Timo Homburg. Paleocodage tool version 2, 2020. cited p. ix, 108, 109

[HKAA22] Timo Homburg, Frans Knibbe, Nathalie Abadie, and Joseph Abhayaratna. Ogc crs ontology, 2022. URL https://github.com/opengeospatial/ontology-crs/blob/master/discussion-paper.md. OGC Discussion paper: In draft. cited p. 30

[HMB21] Timo Homburg, Hubert Mara, and Kai-Christian Bruhn. Cuneiform in the LOD cloud: Connecting 2D and 3D representations of philological objects with linguistic concepts, November 2021. doi:10.5281/zenodo.5749763. cited p. 28, 39, 44

[HMB23] Timo Homburg, Hubert Mara, and Kai-Christian Bruhn. The cuneiform annotator: Linked data based annotation of cuneiform artifacts. 2023. cited p. 110

[Hom20] Timo Homburg. situx/paleocodage: Paleocodage js tool prerelease, October 2020. doi:10.5281/zenodo.4068426. cited p. 108

[Hom21] Timo Homburg. Paleocodage - enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding. *Digital Scholarship in the Humanities*, 36(Supplement 2):ii127–ii154, November 2021. doi:10.1093/llc/fqab038. cited p. 55, 56, 59, 64, 74, 126, 129, 135

[HS13] Steven Harris and Andy Seaborne. SPARQL 1.1 query language. W3C recommendation, W3C, March 2013. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/. cited p. 10

[HSJ20] Timo Homburg, Steffen Staab, and Daniel Janke. Geosparql+: Syntax, semantics and system for integrated querying of graph, raster and vector data. In *The Semantic Web – ISWC 2020*. Springer, November 2020. doi:10.1007/978-3-030-62419-4_15. Best Student Paper Award. cited p. 10, 24

[Hue18] John Huehnergard. *A grammar of Akkadian*. Brill, 2018. URL https://brill.com/display/title/38184. cited p. 1, 131

[HZBM22] Timo Homburg, Robert Zwick, Kai-Christian Bruhn, and Hubert Mara. 3d data derivatives of the haft tappeh processing pipeline. *CDLI Journal*, oct 2022. URL https://cdli.earth/articles/cdlj/2022-1. cited p. vi, 21, 38, 41

[HZMB22] Timo Homburg, Robert Zwick, Hubert Mara, and Kai-Christian Bruhn. Annotated 3d-models of cuneiform tablets. *Journal of Open Archaeology Data*, 10(4), may 2022. doi:10.5334/joad.92. cited p. 21, 42, 44

[IV95] Nancy Ide and Jean Véronis. *Text encoding initiative: Background and contexts*, volume 29. Springer Science & Business Media, 1995. doi:10.1007/978-94-011-0325-1. cited p. 13, 129

[Jag10] Bram Jagersma. *A descriptive grammar of Sumerian*. PhD thesis, Leiden University, 2010. URL https://scholarlypublications.universiteitleiden.nl/handle/1887/16107. cited p. 1, 135

[Jef98] K Jeffery. Metadata: an overview and some issues. *Ercim News*, 35, 1998. cited p. 26

[JHS21a] Milos Jovanovik, Timo Homburg, and Mirko Spasić. A geosparql compliance benchmark. *ISPRS International Journal of Geo-Information*, 10(7), July 2021. doi:10.3390/ijgi10070487. cited p. 10

[JHS21b] Milos Jovanovik, Timo Homburg, and Mirko Spasić. Software for the geosparql compliance benchmark. *Software Impacts*, 8:100071, March 2021. doi:10.1016/j.simpa.2021.100071. cited p. 10

[KCD+03] S. Kumar, J. Cohen, D. Duncan, J. Cooper, and D. Snyder. Digital preservation of ancient cuneiform tablets using 3d-scanning. In *3D Digital Imaging and Modeling, International Conference on*, page 326, Los Alamitos, CA, USA, oct 2003. IEEE Computer Society. doi:10.1109/IM.2003.1240266. cited p. 21

[Kha18] Anas Fahad Khan. Towards the representation of etymological data on the semantic web. *Information*, 9(12), 2018. doi:10.3390/info9120304. cited p. 70, 128

[KK17] Dimitris Kontokostas and Holger Knublauch. Shapes constraint language (SHACL). W3C recommendation, W3C, July 2017. https://www.w3.org/TR/2017/REC-shacl-20170720/. cited p. 10, 129

[Klo07] Alwin Kloekhorst. *Etymological Dictionary of the Hittite Inherited Lexicon*. Brill, Leiden, Niederlande, 2007. URL https://brill.com/view/title/12608. cited p. 1

[Kor06] J.K. Korpela. *Unicode Explained*. Internationalize documents, programs, and web sites. O'Reilly Media, 2006. URL https://books.google.de/books?id=lxndiWaFMvMC. cited p. 135

[KY16] Maksim Kudrinski and Ilya Yakubovich. Sumerograms and akkadograms in hittite: Ideograms, logograms, allograms, or heterograms? *Altorientalische Forschungen*, 43(1-2):53–66, 2016. doi:10.1515/aofo-2016-0018. cited p. 131, 135

[Lab02] René Labat. *Manuel d'épigraphie akkadienne. Signes. Syllabaire, Idéogrammes*. Geuthner, 2002. cited p. vi, 51, 53

[Lau14] Olivier Lauffenburger. *Hittite grammar*. CreateSpace Independent Publishing Platform, 2014. URL https://books.google.de/books?id=44KvoQEACAAJ. cited p. 1

[LBC17] Bernadette Farias Loscio, Caroline Burle, and Newton Calegari. Data on the web best practices. W3C recommendation, W3C, January 2017. URL https://www.w3.org/TR/2017/REC-dwbp-20170131/. cited p. 10

[LBHL15] Yudong Liu, Clinton Burkhart, James Hearne, and Liang Luo. Enhancing Sumerian lemmatization by unsupervised named-entity recognition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1446–1451, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi:10.3115/v1/N15-1167. cited p. 4

[LCK20] Dave Longley, Pierre-Antoine Champin, and Gregg Kellogg. JSON-ld 1.1. W3C recommendation, W3C, July 2020. https://www.w3.org/TR/2020/REC-json-ld11-20200716/. cited p. 128

[LGK+20] Anna-Lena Lamprecht, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, Stephanie van de Sandt, Jon Ison, Paula Andrea Martinez, Peter McQuilton, Alfonso Valencia, Jennifer Harrow, Fotis Psomopoulos, Josep Ll. Gelpi, Neil Chue Hong, Carole Goble, and Salvador Capella-Gutierrez. Towards fair principles for research software. *Data Science*, 3:37–59, 2020. doi:10.3233/DS-190026. 1. cited p. 6

[LNMFRS+20] Martín López-Nores, Juan Luis Montero-Fenollós, Marta Rodríguez-Sampayo, José Juan Pazos-Arias, Silvia González-Soutelo, and Susana Reboreda-Morillo. Cuneiforce: Involving the crowd in the annotation of unread mesopotamian cuneiform tablets through a gamified design. In Ilias O. Pappas, Patrick Mikalef, Yogesh K. Dwivedi, Letizia Jaccheri, John Krogstie, and Matti Mäntymäki, editors, *Digital Transformation for a Sustainable Society in the 21st Century*, pages 158–163, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-39634-3_14. cited p. 100

[Lot15] Roger Lott. Geographic information-well-known text representation of coordinate reference systems. Technical report, Open Geospatial Consortium, 2015. URL http://docs.opengeospatial.org/is/18-010r7/18-010r7.html. cited p. 30, 31, 46, 129, 130

[LSM13] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-o: The PROV ontology. W3C recommendation, W3C, April 2013. https://www.w3.org/TR/2013/REC-prov-o-20130430/. cited p. 28, 82

[MAB+19] John P McCrae, Andrejs Abele, Paul Buitelaar, Richard Cyganiak, Anja Jentzsch, Vladimir Andryushechkin, and J Debattista. The linked open data cloud, 2019. URL https://lod-cloud.net. cited p. 10

[Mar19] Hubert Mara. HeiCuBeDa Hilprecht - Heidelberg Cuneiform Benchmark Dataset for the Hilprecht Collection, 2019. doi:10.11588/data/IE8CCN. cited p. 21, 95, 128

[MB19] Hubert Mara and Bartosz Bogacz. Breaking the code on broken tablets: The learning challenge for annotated cuneiform script in normalized 2d and 3d datasets. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 148–153, 2019. doi:10.1109/ICDAR.2019.00032. cited p. 22, 87

[MBGG+17] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf. cited p. vii, 17, 62, 63, 128

[MH23] Hubert Mara and Timo Homburg. MaiCuBeDa Hilprecht - Mainz Cuneiform Benchmark Dataset for the Hilprecht Collection, 2023. doi:10.11588/data/QSNIQ2. cited p. 95

[MKJB10] Hubert Mara, Susanne Krömker, Stefan Jakob, and Bernd Breuckmann. GigaMesh and Gilgamesh 3D Multiscale Integral Invariant Cuneiform Character Extraction. In Alessandro Artusi, Morwena Joly, Genevieve Lucet, Denis Pitzalis, and Alejandro Ribes, editors, *VAST: International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*. The Eurographics Association, 2010. doi:10.2312/VAST/VAST10/131-138. cited p. 24, 28

[ML17] Seyed Muhammad Hossein Mousavi and Vyacheslav Lyashenko. Extracting old persian cuneiform font out of noisy images (handwritten or inscription). In *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*, pages 241–246. IEEE, 2017. doi:10.1109/IranianMVIP.2017.8342358. cited p. 52

[MNY99] S. Mori, H. Nishida, and H. Yamada. *Optical Character Recognition*. Wiley Series in Microwave and Optical Engineering. Wiley, 1999. URL https://books.google.de/books?id=fyaoQgAACAAJ. cited p. 1, 129

[MPSP12] Boris Motik, Peter Patel-Schneider, and Bijan Parsia. OWL 2 web ontology language structural specification and functional-style syntax (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/. cited p. 10

[MR93] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993. doi:10.1016/0098-3004(93)90090-R. cited p. 40

[MSH+16] Essam Mansour, Andrei Vlad Sambra, Sandro Hawke, Maged Zereba, Sarven Capadisli, Abdurrahman Ghanem, Ashraf Aboulnaga, and Tim Berners-Lee. A demonstration of the solid platform for social web applications. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 223–226, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. doi:10.1145/2872518.2890529. cited p. 116

[NALM20] Stavros Nousias, Gerasimos Arvanitis, Aris S. Lalos, and Konstantinos Moustakas. Mesh saliency detection using convolutional neural networks. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020. doi:10.1109/ICME46284.2020.9102796. cited p. 41

[NF18] Terhi Nurmikko-Fuller. Publishing sumerian literature on the semantic web. In *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, pages 336–364. Brill, 2018. URL http://www.jstor.org/stable/10.1163/j.ctv4v349g.20. cited p. 19

[Nic17] Franco Niccolucci. Documenting archaeological science with cidoc crm. *International Journal on Digital Libraries*, 18(3):223–231, 2017. doi:10.1007/s00799-016-0199-x. cited p. 11

[NOMC11] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011. doi:10.1136/amiajnl-2011-000464. cited p. 77, 129

[NP10] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1023. cited p. 17

[NS08] R. Nicolai and G. Simensen. The new epsg geodetic parameter registry. In *70th EAGE Conference and Exhibition incorporating SPE EUROPEC 2008*. European Association of Geoscientists & Engineers, 2008. doi:10.3997/2214-4609.20147655. cited p. 32, 128

[NTM+22] Nicholas J. Car, Timo Homburg, Matthew Perry, John Herring, Frans Knibbe, Simon J.D. Cox, Joseph Abhayaratna, and Mathias Bonduel. OGC GeoSPARQL - A Geographic Query Language for RDF Data. OGC Implementation Standard OGC 11-052r4, Open Geospatial Consortium, 2022. URL http://www.opengis.net/doc/IS/geosparql/1.1. cited p. 10

[Pan09] Jeff Z. Pan. Resource Description Framework. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 71–90. Springer, December 2009. doi:10.1007/978-3-540-92673-3. cited p. 9, 129

[Pan15] Strahil V. Panayotov. The gottstein system implemented on a digital middle and neo-assyrian palaeography. *Cuneiform Digital Library Notes*, 17, 2015. URL https://cdli.earth/articles/cdln/2015-17. cited p. 54

[PCSF08] C Pilato, Ben Collins-Sussman, and Brian Fitzpatrick. Version control with subversion. 2008. cited p. 82, 129

[PDM12] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf. cited p. 17

[PH12] Matthew Perry and John Herring. OGC GeoSPARQL - A Geographic Query Language for RDF Data. OGC Standard, Open Geospatial Consortium, Wayland, MA, USA, 2012. https://www.ogc.org/standards/geosparql. cited p. 10, 22

[Pie16] Elena Pierazzo. *Digital scholarly editing: Theories, models and methods*. Routledge, 2016. cited p. 1

[Píš05] Karel Píška. Fonts for neo-assyrian cuneiform. In *Proceedings of the EuroTEX Conference (Paperless TEX)*, pages 20–24. Citeseer, 2005. cited p. 52

[Pop16] Olga V Popova. Cuneiform orthography of the stops in alalah vii akkadian. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 106(1):62–90, 2016. doi:10.1515/za-2016-0006. cited p. 49

[PSCPP20] Ravneet Punia, Niko Schenk, Christian Chiarcos, and Émilie Pagé-Perron. Towards the first machine translation system for Sumerian transliterations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3454–3460, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.308. cited p. 4

[Pí12] Karel Píška. Creating cuneiform fonts with metatype1 and fontforge. River Valley TV, 2012. doi:10.5446/30777. cited p. 52

[Qui03] Antoine Quint. Scalable vector graphics. *IEEE MultiMedia*, 10(3):99–102, jul 2003. doi:10.1109/MMUL.2003.1218261. cited p. 129

[Rat19] Rune Rattenborg. Cuneiform site index (csi): A gazetteer of find-spots for cuneiform texts in the eastern mediterranean and the middle east, 2019. URL http://ancientworldonline.blogspot.com/2019/12/cuneiform-site-index-csi-gazetteer-of.html. cited p. 1

[Rofrm-e1] Elisa Roßberger. A middle bronze ii cylinder seal of north syrian style from tel shimron (jezreel valley). *Bulletin of the American Schools of Oriental Research*, 385:119–130, 2021. doi:10.1086/712486. cited p. 42

[Rob98] Eleanor Robson. The electronic text corpus of sumerian literature. -, 1998. URL https://www.worldcat.org/de/title/40902216. cited p. 13, 128

[RR11] Karen Radner and Eleanor Robson. *The Oxford Handbook of Cuneiform Culture*. Oxford University Press, 09 2011. doi:10.1093/oxfordhb/9780199557301.001.0001. cited p. 1, 15, 78

[RVAT13] Hajo Rijgersberg, Mark Van Assem, and Jan Top. Ontology of units of measure and related concepts. *Semantic Web*, 4(1):3–13, 2013. doi:10.3233/SW-2012-0069. cited p. v, viii, 31, 72

[RW18] Karen Radner and Frauke Weiershäuser. Neue wege zu antiken texten. *Akademie aktuell. Zeitschrift der bayerischen Akademie der Wissenschaften*, 2018(1):32–37, 2018. doi:10.5282/ubm/epub.57509. cited p. 3, 129

[Sah16] Patrick Sahle. What is a scholarly digital edition? *Digital scholarly editing: Theories and practices*, 1:19–39, 2016. URL https://books.openedition.org/obp/3397. cited p. 18, 19, 119, 132

[Sch14] Desmond Schmidt. Towards an interoperable digital scholarly edition. *Journal of the Text Encoding Initiative*, (7), 2014. doi:10.4000/jtei.979. cited p. 1

[SF16] Tony C Smith and Eibe Frank. Introducing machine learning concepts with weka. In *Statistical genomics*, pages 353–378. Springer, 2016. doi:10.1007/978-1-4939-3578-9$_1$7. cited p. 98

[SH07] Benjamin Studevent-Hickman. *The ninety-degree rotation of the cuneiform script*. na, 2007. cited p. 49

[SH10] Sebastian Speiser and Andreas Harth. Taking the lids off data silos. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, New York, NY, USA, 2010. Association for Computing Machinery. doi:10.1145/1839707.1839761. cited p. 1

[SH13] Andy Seaborne and Steven Harris. SPARQL 1.1 query language. W3C recommendation, W3C, March 2013. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/. cited p. 10, 129

[SR18] Annamaria De Santis and Irene Rossi, editors. *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*. De Gruyter Open Poland, Warsaw, Poland, 2018. doi:10.1515/9783110607208. cited p. 3

[Str08] Christian Strobl. Dimensionally extended nine-intersection model (de-9im). In Shashi Shekhar and Hui Xiong, editors, *Encyclopedia of GIS*, pages 240–245. Springer, 2008. doi:10.1007/978-0-387-35973-1_298. cited p. 25, 127

[Stu08] Edgar H Sturtevant. *A Hittite glossary: words of known or conjectured meaning with Sumerian and Akkadian words occurring in Hittite texts*. Wipf and Stock Publishers, 2008. cited p. 1

[TAA14] Raphaël Troncy, Ghislain Auguste Atemezing, and Nathalie Abadie. Modeling Geometry and Reference Systems on the Web of Data. In *Linking Geospatial Data Workshop*, London, United Kingdom, December 2014. URL https://hal.archives-ouvertes.fr/hal-02398638. cited p. 30

[Tac01] Tsurozoh Tachibanaya. Description of exif file format, 2001. URL http://park2.wakwak.com/tsuruzoh/Computer/Digicams/exif-e.html. cited p. 11, 128

[TH22] Florian Thiery and Timo Homburg. SPARQLing Unicorn QGIS Plugin, 2022. doi:10.5281/zenodo.3786814. cited p. 115

[Tob64] Jr Leo W Tobin. Coordinate reference system, apr 1964. URL https://patents.google.com/patent/US3131292A/. US Patent 3,131,292. cited p. 127

[TR14] Steve Tinney and Eleanor Robson. Oracc: The open richly annotated cuneiform corpus, 2014. cited p. 3

[vdBBT+19] Linda van den Brink, Payam Barnaghi, Jeremy Tandy, Ghislain Atemezing, Rob Atkinson, Byron Cochrane, Yasmin Fathy, Raúl García Castro, Armin Haller, Andreas Harth, et al. Best practices for publishing, retrieving, and using spatial data on the web. *Semantic Web*, 10(1):95–114, 2019. doi:10.3233/SW-180305. cited p. 10

[VdM15] Marc Van de Mieroop. *A history of the ancient Near East, ca. 3000-323 BC*. John Wiley & Sons, 2015. URL https://www.wiley.com/en-us/A+History+of+the+Ancient+Near+East,+ca+3000+323+BC,+3rd+Edition-p-9781118718162. cited p. 1

[VK14] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014. doi:10.1145/2629489. cited p. 17

[VP22] Cristina Vertan and Christian Prager. From inscription to semi-automatic annotation of Maya hieroglyphic texts. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 114–118, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lt4hala-1.16. cited p. 4

[vR13] Eric van Rees. Open geospatial consortium (ogc). *Geoinformatics*, 16(8):28, 2013. cited p. 129

[VVH+16] Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert. Triple pattern fragments: A low-cost knowledge graph interface for the web. *Journal of Web Semantics*, 37-38:184–206, 2016. doi:10.1016/j.websem.2016.03.003. cited p. 116

[Wal92] G.K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992. doi:10.1109/30.125072. cited p. 128

[War06] Frank Warmerdam. Coordinate systems: Proj. 4, epsg and ogc wkt. *Presentation at the FOSS4G2006-Free And Open Source Software for Geoinformatics*, 2006. URL http://2006.foss4g.org/getFile.py/accessf802.pdf. cited p. 31

[WBE20] Allen Wirfs-Brock and Brendan Eich. Javascript: The first 20 years. *Proc. ACM Program. Lang.*, 4(HOPL), jun 2020. doi:10.1145/3386327. cited p. 128, 133

[WDA+16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, Mar 2016. doi:10.1038/sdata.2016.18. cited p. 6, 9, 128

[Wee18] Mark Weeden. *Current Research in Cuneiform Palaeography. Proceedings of the Workshop organised and the 60th Rencontre Assyriologique Internationale Warsaw 2014, edited by Elena Devecchi, Gerfrid G. W. Müller, and Jana Mynářová*, volume 77. 2018. doi:10.1086/696827. cited p. 49

[WFA+01] Sandra Isobel Woolley, Nicholas J. Flowers, Theodoros N. Arvanitis, Alasdair Livingstone, Tom R. Davis, and John Ellison. Three-dimensional capture, representation, and manipulation of Cuneiform tablets. In Brian D. Corner, Joseph H. Nurre, and Roy P. Pargas, editors, *Three-Dimensional Image Capture and Applications IV*, volume 4298, pages 103 – 110. International Society for Optics and Photonics, SPIE, 2001. doi:10.1117/12.424894. cited p. 21

[Win70] Gernot L. Windfuhr. The cuneiform signs of ugarit. *Journal of Near Eastern Studies*, 29(1):48–51, 1970. doi:10.1086/372042. cited p. 49

[WK00] Stuart L Weibel and Traugott Koch. The dublin core metadata initiative. *D-lib magazine*, 6(12):1082–9873, 2000. cited p. 26

[YB07] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007. doi:10.1109/TPAMI.2007.1078. cited p. 59

[Zho21] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021. doi:10.1007/978-981-15-1967-3. cited p. 133

Timo Homburg

Curriculum Vitae

https://www.situx.github.io
situx
0000-0002-9499-5840
Ww8a7E0AAAAJ
Wikidata: Q67624599

Education (4)

2012–2015 **M.Sc. Computer Science**, *Goethe University*, Frankfurt am Main, Germany, *Grade 1.0 (A)*
Minor: Chinese Studies, Thesis Title: Methods for word segmentation of non-alphabetic scripts

Jul 2013–Aug 2013 **Summer School: Doing Business in Africa**, *Polytechnic of Namibia*, Windhoek, Namibia, Grade: B
Business Summer School at Polytechnic of Namibia

Jul 2010–Aug 2010 **Summer School: China Cultural Exchange**, *Shanghai International Studies University*, Shanghai, China
Summer School for learning Chinese

2008–2012 **B.Sc. Computer Science**, *RheinMain University of Applied Sciences*, Wiesbaden, Germany, *Grade 1.4 (A)*
Fokus: Database systems, Thesis Title: Development of an Android application for the status display of measuring stations of the German Weather Service

Work experience (7)

Oct 2017–Dez 2019 **Systemadministrator**, *Hochschule Mainz*, Mainz, Germany, Industry: Research
Systemadministrator of the research group i3mainz

Apr 2015–Current **Research Associate**, *Hochschule Mainz*, Mainz, Germany, Industry: Research
Research areas: GIS, Spatial Humanities, Computational Humanities, Standardisation, Linked open data and Geospatial Semantic Web

May 2013–Mar 2015 **Student Assistant**, *Forschungsgruppe Wirtschaftsinformatik/Goethe Universität*, Frankfurt, Germany, Bereich: Forschung
Working in the research project EVER

Nov 2011–Feb 2012 **Student Assistant**, *MeteoSolutions GmbH*, Darmstadt, Germany, Industry: Software Engineering
Development of an Android application for monitoring weather stations as part of my bachelor thesis

Mar 2011–Aug 2011 **Trainee**, *Sirona Dental Systems Foshan Co. Ltd.*, Foshan, China, Industry: Medical Devices
Investigation of the possibility of connecting a Syteline 8 ERP system with an Oracle Agile PLM system, development of a database for the management of individual parts for the production of dental chairs

Feb 2011 **Student Assistant**, *Laboratory of Software Engineering*, Wiesbaden, Germany, Industry: Research
Part of a feasibility study for the German Weather Service (DWD)

2010–2011 **Student Assistant**, *DOPSY Laboratory for Distributed Systems*, Wiesbaden, Germany, Industry: Research
Development of a content management system for the laboratory's homepage

2007–2008 **Alternative Service**, *Main-Kinzig Hospital*, Gelnhausen, Germany, Industry: IT Systems
IT support and administration work in a hospital

Teaching experience (5x Supervisor, 10x Lecturer)

Mar 2025 - Jul 2025 **Lecturer**, *Mainz University Of Applied Sciences*, Mainz, Germany
Software engineering and interactive visualization

Okt 2024 - Jan 2025 **Supervisor**, *3x Practical project,* *Mainz University Of Applied Sciences*, Mainz, Germany
Supervision of practical projects for three students in the field of 3D annotation software, 3D quality evaluation and RDF databases for cuneiform seals

Okt 2023 - Mai 2024 **Supervisor**, *Master Thesis Marc Häuser,* *Mainz University Of Applied Sciences*, Mainz, Germany
Automated recognition of cuneiform characters and their variants

Mai 2023 - Okt 2023 **Supervisor**, *Master Thesis Robert Zwick,* *Mainz University Of Applied Sciences*, Mainz, Germany
Classification of cuneiform signs using machine learning algorithms

Okt 2021 - Mar 2022 **Supervisor**, *Mastera Thesis Dennis Gottwald,* *Mainz University Of Applied Sciences*, Mainz, Germany
Solid Digital Humanities: Approaches to data sovereignty and decentralization using Solid Pods, using the example of research data on Terra Sigillata objects

Okt 2021 - Jan 2022 **Supervisor**, *Practical project Robert Zwick,* *Mainz University Of Applied Sciences*, Mainz, Germany
Supervision of practical semester on 3D annotation techniques with results of a research paper

Apr 2021– Jul 2021 **Lecturer**, *Johannes Gutenberg University*, Mainz, Germany
Teaching practice in the course: Digital Methods in antiquity research

Oct 2020– Feb 2021 **Lecturer**, *Mainz University Of Applied Sciences*, Mainz, Germany
Teaching practice in the course: "Application-based software development"

Oct 2019– Jan 2020 **Supervisor**, *Praxisprojekt Marc Häuser,* *Mainz University Of Applied Sciences*, Mainz, Germany
Supervision of a practical semester on encoding cuneiform characters

Oct 2019– Feb 2020 **Lecturer**, *Mainz University Of Applied Sciences*, Mainz, Germany
Teaching practice in the course: "Application-based software development"

Oct 2018– Feb 2019 **Lecturer**, *Mainz University Of Applied Sciences*, Mainz, Germany
Teaching practice in the course: "Application-based software development"

Oct 2017– Jul 2019 **Lecturer**, *RheinMain University Of Applied Sciences*, Wiesbaden, Germany, 4 courses (evaluated)
Teaching experience in courses: "Introduction to Computer Science", "Algorithms and Data Structures", "Programming Methods and Techniques"

Awards (7)

Sep 2023 GCH2023 Best Paper Award

Nov 2020 ISWC Best Student Paper Award

Oct 2020 FIG Paper Of The Month

Apr 2017 WebIST Best Student Paper Award

Jul 2016 Valedictorian Award (Jahrgangsbester) Computer Science Master Of Science

Mar 2015 DARIAH-DE Digital Humanities Award 2015

Jun 2012 Valedictorian Award (Jahrgangsbester) Computer Science Bachelor Of Science

Jun 2012 PROMOS-Scholarship for studying abroad

Conferences with contributions (24)

- ArchaeoFOSS 2020 Conference (Publication and Talk)
- AGIT 2019 (Publication and Talk)
- CAA 2022 (Contribution to a workshop)
- BIS: Business Information Systems Conference 2018, 2019 (Workshop on Data Quality)
- CAA-DE Webcast 2021 (Workshop on Linked Open Data)
- Coding Da-Vinci Rhein Main 2018 (Workshop on Wikidata)
- Digital Humanities Conferences 2017, 2019, 2020 (Publications and lectures)
- Digital Humanities im deutschsprachigen Raum (DHd) 2015, 2018, 2019 (Publications and lectures)
- Digital Humanities Summit 2015 (Poster + *DARIAH-DE Digital Humanities Award*)
- European Semantic Web Conference 2021,2022,2024 (Publication 2021 and 2022-2024 GeoLD Workshop)
- FIG 2020 Konferenz (Talk and Publication)
- FOSSGIS Konferenzen 2020, 2024 (Publications and lectures)
- International Conference on Computer Vision (ICCV) 2023 (Contribution)
- 10. Internationales Colloquium der Deutschen Orient-Gesellschaft 2019 (Talk)
- INSPIRE Conference 2016 (Poster presentation)
- International Semantic Web Conference 2020 (Publication, Talk and *Best Student Paper Award*)
- KI 2014: 37th German Conference on Artificial Intelligence PUK Workshop (Contributor)
- Linked Pasts Konferenz 2021 (Poster)
- Location Based Services Conferences 2018, 2019 (Publications and talk)
- Language Resources and Evaluation Conference (LREC) 2016 (Publikation und Poster)
- Open Science Festival Mainz 2024 (Workshop about Geospatial Linked Open Data)
- NFDI4Objects Community Meeting 2024 (Poster about Linked Open Data Tools)
- OGC Technical Committee Meetings (several) (Presentations on GeoSPARQL)
- WebIST 2017 (Contribution and *Best Student Paper Award*)

Reviewer functions for journals and conferences (46)

46 reviews for a total of 37 publications documented in Web Of Science:

- Journal: it Information Technology 1x (2024)
- Journal: International Journal of Digital Earth 1x (2023)
- Journal: International Journal for Geo-Information 3x (since 2019)
- Journal: Heritage Science 4x (since 2021)
- Journal: Sustainability 1x (2022)
- Journal: Applied Sciences 1x (2022)
- Journal: Digital Scholarship in the Humanities 1x (since 2020)
- 2022 and 2024: Reviewer for two papers each of the Workshop on Geospatial Linked Open Data
- since 2017: Reviewer for the Digital Humanities Konferenz
- since 2019: Reviewer for the conference Digital Humanities in the German Speaking Area (DHd)

Publication list (peer-reviewed) (83)

Articles

[HSB25] Timo Homburg, Steffen Staab und Frank Boochs. "QPredict: Using low quality volunteered geospatial data to evaluate high quality authority data: A case study on building footprint data". In: *ACM Transactions on Spatial Algorithms and Systems* (Jan. 2025). DOI: 10.1145/3715910.

[Gor+24] Shai Gordin u. a. In: *it - Information Technology* 66.1 (2024), S. 1–3. DOI: doi:10.1515/itit-2024-2001. URL: https://doi.org/10.1515/itit-2024-2001.

[Ham+24] Hendrik Hameeuw u. a. "Preparing multi-layered visualisations of Old Babylonian cuneiform tablets for a machine learning OCR training model towards automated sign recognition". In: *it - Information Technology* (Jan. 2024). DOI: 10.1515/itit-2023-0063. URL: https://doi.org/10.1515/itit-2023-0063.

[Tan+23] Jeremy Tandy u. a. "Spatial Data on the Web Best Practices: W3C Working Group Note". In: (Sep. 2023). URL: https://www.w3.org/TR/sdw-bp/.

[CH22] Nicholas J. Car und Timo Homburg. "GeoSPARQL 1.1: Motivations, Details and Applications of the Decadal Update to the Most Important Geospatial LOD Standard". In: *ISPRS International Journal of Geo-Information* 11.2 (Feb. 2022). ISSN: 2220-9964. DOI: 10.3390/ijgi11020117. URL: https://www.mdpi.com/2220-9964/11/2/117.

[Hom+22a] Timo Homburg u. a. "3D Data Derivatives of the Haft Tappeh Processing Pipeline". In: *CDLI Journal* (Okt. 2022). URL: https://cdli.earth/articles/cdlj/2022-1.

[Hom+22b] Timo Homburg u. a. "Annotated 3D-Models of Cuneiform Tablets". In: *Journal of Open Archaeology Data* 10.4 (Mai 2022). ISSN: 2049-1565. DOI: 10.5334/joad.92. URL: https://openarchaeologydata.metajnl.com/articles/10.5334/joad.92/.

[Thi+22] Florian Thiery u. a. "Software Review for the Software PointSamplingTool". In: *Archäologische Informationen* 44 (Nov. 2022). URL: https://dguf.de/fileadmin/AI/archinf-ev_thiery-etal.pdf.

[Hom21] Timo Homburg. "PaleoCodage - Enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding". In: *Digital Scholarship in the Humanities* 36.Supplement$_2$ (Nov. 2021), S. ii127–ii154. ISSN: 2055-7671. DOI: 10.1093/llc/fqab038. URL: https://academic.oup.com/dsh/article/36/Supplement_2/ii127/6421811.

[Hom+21b] Timo Homburg u. a. "Diskussionsbeitrag - Handreichung zur Rezension von Forschungssoftware in den Altertumswissenschaften / Impulse - Recommendations for the review of archaeological research software". In: *Archäologische Informationen* (Jan. 2021). ISSN: 0341-2873. DOI: 10.11588/ai.2020.1.81422. URL: https://www.dguf.de/fileadmin/AI/archinf-ev_homburg-etal.pdf.

[Hom+21c] Timo Homburg u. a. "Metadata Schema and Ontology for Capturing and Processing of 3D Cultural Heritage Objects". Englisch. In: *Heritage Science* (Juli 2021). ISSN: 2050-7445. DOI: 10.1186/s40494-021-00561-w. URL: https://www.nature.com/articles/s40494-021-00561-w.

[JHS21a] Milos Jovanovik, Timo Homburg und Mirko Spasić. "A GeoSPARQL Compliance Benchmark". In: *ISPRS International Journal of Geo-Information* 10.7 (Juli 2021). ISSN: 2220-9964. DOI: 10.3390/ijgi10070487. URL: https://www.mdpi.com/2220-9964/10/7/487.

[JHS21c] Milos Jovanovik, Timo Homburg und Mirko Spasić. "Software for the GeoSPARQL compliance benchmark". In: *Software Impacts* 8 (März 2021), S. 100071. ISSN: 2665-9638. DOI: 10.1016/j.simpa.2021.100071. URL: https://www.sciencedirect.com/science/article/pii/S2665963821000191.

[Thi+21] Florian Thiery u. a. “SPARQLing Geodesy for Cultural Heritage – New Opportunities for Publishing and Analysing Volunteered Linked (Geo-)Data”. Englisch. In: *FIG Journal* (Mai 2021). Update paper of the publication of the previous year. ISSN: 2307-4086. URL: https://www.fig.net/resources/publications/prj/showpeerreviewpaper.asp?pubid=11032.

[Hom20a] Timo Homburg. “Connecting Semantic Situation Descriptions with Data Quality Evaluations—Towards a Framework of Automatic Thematic Map Evaluation”. Englisch. In: *Information* (Nov. 2020). ISSN: 2078-2489. DOI: 10.3390/info11110532. URL: https://doi.org/10.3390/info11110532.

[HB19] Timo Homburg und Frank Boochs. “Situation-Dependent Data Quality Analysis for Geospatial Data Using Semantic Technologies”. Englisch. In: *Lecture Notes in Business Information Processing* (Jan. 2019). Hrsg. von Witold Abramowicz und Adrian Paschke, S. 566–578. DOI: 10.1007/978-3-030-04849-5_49. URL: https://link.springer.com/chapter/10.1007/978-3-030-04849-5%5C_49.

[HN19] Timo Homburg und Pascal Neis. “Linked Data & VGI - Eine komparative Qualitätsanalyse für Deutschland, Österreich und die Schweiz auf Basis von Wikidata und OpenStreetMap”. Englisch. In: *AGIT Journal* (Juli 2019). ISSN: 2364-9283. DOI: 10.14627/537669013.

[HSW19] Timo Homburg, Sebastian Steppan und Falk Wuerriehausen. “Semantic Data integration and quality assurance of thematic maps in the German geographic authority”. Englisch. In: *Lecture Notes in Business Information Processing* (Juni 2019). DOI: 10.1007/978-3-030-36691-9_46. URL: https://link.springer.com/chapter/10.1007/978-3-030-36691-9_46.

[Pru+19] Claire Prudhomme u. a. “Interpretation and automatic integration of geospatial data into the Semantic Web”. Englisch. In: *Computing* (Feb. 2019), S. 1–27. ISSN: 1436-5057. DOI: 10.1007/s00607-019-00701-y. URL: https://doi.org/10.1007/s00607-019-00701-y.

[Hom+17] Timo Homburg u. a. “Integration, Quality Assurance and Usage of Geospatial Data with Semantic Tools”. In: *gis.Science - Die Zeitschrift fur Geoinformatik* 3 (Sep. 2017), S. 91–96. ISSN: 1869-9391. URL: https://gispoint.de/artikelarchiv/gis/2017/gisscience-ausgabe-32017/4201-integration-quality-assurance-and-usage-of-geospatial-data-with-semantic-tools-i-integration-bewertung-und-nutzung-heterogener-datenquellen-mittels-semantischer-werkzeuge-i.html.

Contributions to conference proceedings

[Hom24a] Timo Homburg. "GeoWebAnnotations: Extending the W3C Web Annotation Data Model for geospatial data". In: *Proceedings of the 6th International Workshop on Geospatial Linked Data 2024 co-located with 21st Extended Semantic Web Conference (ESWC 2024)*. Aug. 2024. URL: https://ceur-ws.org/Vol-3724/paper5.pdf.

[Hom24b] Timo Homburg. "PaleOrdia: Semantically Describing (Cuneiform) Paleography using Paleographic Linked Open Data". In: *Proceedings of the First International Workshop of Semantic Digital Humanities (SemDH 2024) co-located with the European Semantic Web Conference 2024 (ESWC 2024)*. Juli 2024. URL: https://ceur-ws.org/Vol-3724/short1.pdf.

[Hom+24b] Timo Homburg u. a. "The Case for a standardised CRS ontology". In: *Proceedings of the 6th International Workshop on Geospatial Linked Data 2024 co-located with 21st Extended Semantic Web Conference (ESWC 2024)*. Aug. 2024. URL: https://ceur-ws.org/Vol-3724/paper3.pdf.

[TH24b] Florian Thiery und Timo Homburg. "The SPARQL Unicorn Ontology documentation: Exposing RDF geodata using static GeoAPIs". In: *FOSSGIS Tagungsband*. März 2024, S. 82–93. DOI: 10.5281/zenodo.10570985.

[Thi+24] Florian Thiery u. a. "3D und FDM: Beispiele für semantische 3D-Annota- tion und -Modellierung im Datenqualifizierungspro- zess in Konsortien der Nationalen Forschungsdaten- infrastruktur (NFDI)". In: *Beiträge der Oldenburger 3D-Tage und des BIMtages 2024*. Juni 2024, S. 84–93. ISBN: 978-3-87907-750-2. URL: https://www.vde-verlag.de/buecher/537750/photogrammetrie-laserscanning-optische-3d-messtechnik.html.

[Hom+23a] Timo Homburg u. a. "Forschungssoftware rezensieren: Konzeption, Durchführung und Umsetzung". In: Zenodo, März 2023. DOI: 10.5281/zenodo.7688632. URL: https://doi.org/10.5281/zenodo.7688632.

[SHM23a] Ernst Stötzner, Timo Homburg und Hubert Mara. "CNN Based Cuneiform Sign Detection Learned from Annotated 3D Renderings and Mapped Photographs with Illumination Augmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Okt. 2023, S. 1680–1688. DOI: 10.1109/ICCVW60793.2023.00183. URL: https://openaccess.thecvf.com/content/ICCV2023W/e-Heritage/papers/Stotzner_CNN_Based_Cuneiform_Sign_Detection_Learned_from_Annotated_3D_Renderings_ICCVW_2023_paper.pdf.

[Stö+23] Ernst Stötzner u. a. "R-CNN based PolygonalWedge Detection Learned from Annotated 3D Renderings and Mapped Photographs of Open Data Cuneiform Tablets". In: *Eurographics Workshop on Graphics and Cultural Heritage*. Hrsg. von Alberto Bucciero u. a. Best Paper Award. The Eurographics Association, Sep. 2023. ISBN: 978-3-03868-217-2. DOI: 10.2312/gch.20231157.

[Hab+22] David Habgood u. a. "Implementation and Compliance Benchmarking of a DGGS-enabled, GeoSPARQL-aware Triplestore". In: *Geospatial Linked Data Workshop 2022*. Hrsg. von Timo Homburg u. a. Bd. Vol-3157. CEUR-WS, 2022. URL: https://CEUR-WS.org/Vol-3157/paper7.pdf.

[HD22] Timo Homburg und Thierry Declerck. "Towards the Integration of Cuneiform in the OntoLex-Lemon Framework". In: 2022. ISBN: 978-2-487055-04-9. DOI: 10.36824/2022-graf1. URL: https://www.fluxus-editions.fr/gla9.php.

[CH21] Nicholas J. Car und Timo Homburg. "GeoSPARQL 1.1: an almost decadal update to the most important geospatial LOD standard". In: *Geospatial Linked Data Workshop 2021*. Hrsg. von Beyza Yaman u. a. Bd. Vol-2977. Hersonissos, Greece: CEUR-WS, Okt. 2021, S. 26–33. URL: http://ceur-ws.org/Vol-2977/paper4.pdf.

[HT21] Timo Homburg und Florian Thiery. "Little Minions and SPARQL Unicorns as tools for archaeology". In: *ArcheoFOSS XIV 2020: Open Software, Hardware, Processes, Data and Formats in Archaeological Research*. Archaeopress Publishing Ltd, Nov. 2021. ISBN: 978-1-80327-124-8. URL: https://www.archaeopress.com/Archaeopress/Products/9781803271248.

[TSH21] Florian Thiery, Sophie-Charlotte Schmidt und Timo Homburg. "SPARQLing Publication of Irish Ogham Stones as LOD". In: *ArcheoFOSS XIV 2020: Open Software, Hardware, Processes, Data and Formats in Archaeological Research*. Archaeopress Publishing Ltd, Nov. 2021. ISBN: 978-1-80327-124-8. URL: https://www.archaeopress.com/Archaeopress/Products/9781803271248.

[Bog+20] Julian Bogdani u. a. "Book of Abstracts. ArcheoFOSS International Conference 2020". In: Zenodo, Aug. 2020. DOI: 10.5281/zenodo.4002961. URL: https://doi.org/10.5281/zenodo.4002961.

[Hom20c] Timo Homburg. "Mind the gap: Filling gaps in cuneiform tablets using Machine Learning Algorithms". In: *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts*. Hrsg. von Laura Estill und Jennifer Guiliano. Juli 2020. URL: https://dh2020.adho.org/wp-content/uploads/2020/07/151%5C_MindthegapFillinggapsincuneiformtabletsusingMachineLearningA
html.

[Hom20e] Timo Homburg. "Towards Paleographic Linked Open Data (PLOD): A general vocabulary to describe paleographic features". In: *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts*. Hrsg. von Laura Estill und Jennifer Guiliano. Juli 2020. URL: https://dh2020.adho.org/wp-content/uploads/2020/07/369%5C_TowardsPaleographicLinkedOpenDataPLODAgeneral
html.

[HSJ20] Timo Homburg, Steffen Staab und Daniel Janke. "GeoSPARQL+: Syntax, Semantics and System for Integrated Querying of Graph, Raster and Vector Data". Englisch. In: Best Student Paper Award. Springer, Nov. 2020. DOI: 10.1007/978-3-030-62419-4_15. URL: https://link.springer.com/chapter/10.1007/978-3-030-62419-4_15.

[TH20a] Florian Thiery und Timo Homburg. "Linked Ogham Stones - Semantische Modellierung und prototypische Analyse irischer Ogham-Inschriften". Englisch. In: Paderborn, Germany, März 2020. DOI: 10.5281/zenodo.3697060.

[TH20b] Florian Thiery und Timo Homburg. "QGIS - A SPARQLing Unicorn? Eine Einführung in Linked Open Geodata zur Integration von RDF in QGIS Plugins". Englisch. In: Freiburg, Germany, März 2020. DOI: 10.5281/zenodo.3706962.

[Thi+20] Florian Thiery u. a. "SPARQLing Geodesy for Cultural Heritage – New Opportunities for Publishing and Analysing Volunteered Linked (Geo-)Data". Englisch. In: FIG Article Of the Month October 2020. Amsterdam, Netherlands, Mai 2020. ISBN: 978-87-92853-93-6. DOI: 10.5281/zenodo.3766154. URL: https://fig.net/resources/monthly_articles/2020/Thiery_etal_October_2020.asp.

[Hom19a] Timo Homburg. "Paleo Codage - A machine-readable way to describe cuneiform characters paleographically". Englisch. In: Utrecht, Netherlands, Juli 2019. DOI: 10.34894/QAVLOY. URL: https://dev.clariah.nl/files/dh2019/boa/0259.html.

[Hom19c] Timo Homburg. "Towards Creating A Best Practice Digital Processing Pipeline For Cuneiform Languages". Englisch. In: Utrecht, Netherlands, Juli 2019. URL: https://dev.clariah.nl/files/dh2019/boa/1204.html.

[Hom+19] Timo Homburg u. a. "Evaluating linked data location based services using the example of Stolpersteine". Englisch. In: Vienna, Austria, Nov. 2019. ISBN: 978-3-030-36690-2. DOI: 10.5194/ica-adv-2-7-2019. URL: https://www.adv-cartogr-giscience-int-cartogr-assoc.net/2/7/2019/ica-adv-2-7-2019.pdf.

[SUH19] Magdalena Scherl, Martin Unold und Timo Homburg. "Ein unscharfer Suchalgorithmus für Transkriptionen von arabischen Ortsnamen". In: Frankfurt, Germany, März 2019. ISBN: 978-3-00-062166-6. DOI: 10.5281/zenodo.2596095. URL: https://dhd2019.org/programm/do/1400-1530/session-bestimmen-und-identifizieren/vortrag-182/.

[Hom18] Timo Homburg. "Semantische Extraktion auf antiken Schriften am Beispiel von Keilschriftsprachen mithilfe semantischer Wörterbücher". In: *Extended Abstract Digital Humanities im deutschsprachigen Raum (DHd 2018)*. Feb. 2018. ISBN: 978-3-946275-02-2. URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

[HPB18] Timo Homburg, Claire Prudhomme und Frank Boochs. "Semantic Geographic Information System: Integration and management of heterogeneous geodata". Englisch. In: *Fachaustausch Geoinformation 2018*. Nov. 2018. DOI: 10.13140/RG.2.2.35246.56645. URL: https://i3mainz.hs-mainz.de/sites/default/files/public/data/conference_poster_3.pdf.

[Hom+18] Timo Homburg u. a. "Map Change Prediction for Quality Assurance". Englisch. In: *LBS 2018*. Hrsg. von Peter Kiefer u. a. ETH Zurich, Jan. 2018, S. 194–200. DOI: 10.3929. URL: https://doi.org/10.3929/ethz-b-000225617.

[Hom17] Timo Homburg. "POSTagging and Semantic Dictionary Creation for Hittite Cuneiform". Englisch. In: *Digital Humanities 2017, DH 2017, Conference Abstracts, McGill University & Université de Montréal, Montréal, Canada, August 8-11, 2017*. Hrsg. von Rhian Lewis u. a. Alliance of Digital Humanities Organizations. Montréal, Canada: Alliance of Digital Humanities Organizations (ADHO), Aug. 2017. URL: https://dh2017.adho.org/abstracts/139/139.pdf.

[Pru+17] Claire Prudhomme u. a. "Automatic Integration of Spatial Data into the Semantic Web". Englisch. In: *Proceedings of the 13th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,* Best Student Paper Award. INSTICC. Porto, Portugal: SciTePress, Apr. 2017, S. 107–115. ISBN: 978-989-758-246-2. DOI: 10.5220/0006306601070115. URL: http://www.scitepress.org/digitalLibrary/PublicationsDetail.aspx?ID=9PVXQr5fDjQ=%5C&t=1.

[HC16] Timo Homburg und Christian Chiarcos. "Word Segmentation for Akkadian Cuneiform". Englisch. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Hrsg. von Nicoletta Calzolari u. a. Portorož, Slovenia: European Language Resources Association (ELRA), Mai 2016. ISBN: 978-2-9517408-9-1. URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/816_Paper.pdf.

[Hom+16] Timo Homburg u. a. "Interpreting Heterogeneous Geospatial Data Using Semantic Web Technologies". Englisch. In: *International Conference on Computational Science and Its Applications*. Hrsg. von Osvaldo Gervasi u. a. Springer International Publishing. Beijing, China, Juli 2016. Kap. 19, S. 240–255. ISBN: 978-3-319-42110-0. DOI: 10.1007/978-3-319-42111-7. URL: https://link.springer.com/chapter/10.1007/978-3-319-42111-7_19.

[WHM16] Falk Würriehausen, Timo Homburg und Hartmut Müller. "Using an INSPIRE Ontology to Support Spatial Data Interoperability". Englisch. In: Barcelona, Spain, Sep. 2016. URL: https://inspire.ec.europa.eu/events/conferences/inspire_2016/pdfs/2016_psessions/28%20WEDNESDAY_PSESSIONS_H3_14.00-15.30_____28_H3_14.15_188_Presentation.pdf.

[HSM14] Timo Homburg, Pol Schumacher und Mirjam Minor. "Towards workflow planning based on semantic eligibility". Englisch. In: *The 37th German Conference on Artificial Intelligence*. Stuttgart, Germany, Sep. 2014. URL: http://wi.cs.uni-frankfurt.de/webdav/publications/puk2014.pdf.

Conference proceedings and special issues as editor

[Gor+] Shai Gordin u. a., Hrsg. *Insights into Digital Ancient Near Eastern Studies*. To be published end 2024. DeGruyter.

[Hom+24a] Timo Homburg u. a., Hrsg. *Proceedings of the 6th International Workshop on Geospatial Linked Data 2024* (Hersonissos, Greece). Bd. Vol-3743. CEUR-WS, Aug. 2024. URL: https://CEUR-WS.org/Vol-3743/.

[Hom+23b] *From an Analog to a Digital Workflow: An Introductory Approach to Digital Editions in Assyriology*. CDLI Journal, Dez. 2023. URL: https://cdli.earth/articles/cdlb/2023-4.

[Hom+22d] Timo Homburg u. a., Hrsg. *Proceedings of the 5th International Workshop on Geospatial Linked Data 2022* (Hersonissos, Greece). Bd. Vol-3157. CEUR-WS, Juni 2022. URL: https://CEUR-WS.org/Vol-3157/.

Published Research software

[HM24] Timo Homburg und Hubert Mara. *Cuneur - Cuneiform Annotator*. Annotationsprogramm für 3D Renderings. 2022-2024.

[Mar+24] Hubert Mara u. a. *Gigamesh Software Framework*. 3D Mesh processing software. 2012-2024. URL: https://gigamesh.eu.

[HT24] Timo Homburg und Florian Thiery. *SPARQL Unicorn Ontology Documentation*. Tool for the documentation of linked open data dumps as linked open usable data dump deployments. 2024. URL: https://github.com/sparqlunicorn/sparqlunicornGoesGIS-ontdoc.

[JHM24] Milos Jovanovik, Timo Homburg und Spasic Mirko. *GeoSPARQL Compliance Benchmark*. Benchmark to test GeoSPARQL Compliance. 2024. URL: https://github.com/OpenLinkSoftware/GeoSPARQLBenchmark.

[TH24a] Florian Thiery und Timo Homburg. *SPARQLing Unicorn QGIS Plugin*. QGIS Plugin zum Zugriff auf Linked Open Data Ressourcen. 2024. URL: https://plugins.qgis.org/plugins/sparqlunicorn/.

[CRH22] Anja Cramer, Laura Raddatz und Timo Homburg. *3dcap-md-gen*. Version 0.1.3. Scripts for capturing 3D scanning metadata of specific scanning software. The results are converted to conform to an ontology model. Dez. 2022. DOI: 10.5281/zenodo.4566044.

[Hom20b] Timo Homburg. *GeoPubby*. Version 0.5. Linked Data Browser with a specific focus on geodata in JavaScript in Java. Okt. 2020. URL: https://github.com/i3mainz/geopubby/.

[Hom20d] Timo Homburg. *SemanticWFS*. Version 0.5. Prototypical implementation of a OGC API Features Services with a linked open data backend. Okt. 2020. URL: https://github.com/i3mainz/semanticwfs.

[Tim20] Timo. *situx/PaleoCodage: PaleoCodage JS Tool Prerelease*. Version v0.1-beta. JavaScript Implementierung des PaleoCodage Encodings zur Beschreibung von Keilschriftzeichenvarianten. Okt. 2020. DOI: 10.5281/zenodo.4068426. URL: https://doi.org/10.5281/zenodo.4068426.

Published Research datasets

[MH23] Hubert Mara und Timo Homburg. *MaiCuBeDa Hilprecht - Mainz Cuneiform Benchmark Dataset for the Hilprecht Collection*. Version V1. MaiCuBeDa Benchmark Dataset of cuneiform sign annotations for machine learning applications. 2023. DOI: 10.11588/data/QSNIQ2. URL: https://doi.org/10.11588/data/QSNIQ2.

[Hom+22c] Timo Homburg u. a. *Annotated 3D-Models of cuneiform tablets*. Experimente der Einbindung von Annotationen in Formaten von 3D Modellen. Ergebnisse des Praxisprojektes von Robert Zwick. Zenodo, Apr. 2022. DOI: 10.5281/zenodo.6506560. URL: https://doi.org/10.5281/zenodo.6506560.

[Hom+21a] Timo Homburg u. a. *Application Case 1: Metadata for capturing and processing 3d scans of a ship and stone*. Sample data for Beispieldaten für die Metadatenbeschreibung von dem Erfassungsprozess von 3D Modellen. Zenodo, Feb. 2021. DOI: 10.5281/zenodo.4428389. URL: https://doi.org/10.5281/zenodo.4428389.

Further publications (standards, white papers, preprints, posters and presentations)

[Nic+23] Nicholas J. Car u. a. *OGC GeoSPARQL - A Geographic Query Language for RDF Data*. OGC Implementation Standard. Version 1.1. 2023. URL: http://www.opengis.net/doc/IS/geosparql/1.1.

[SHM23b] Ernst Stötzner, Timo Homburg und Hubert Mara. *CNN based Cuneiform Sign Detection Learned from Annotated 3D Renderings and Mapped Photographs with Illumination Augmentation (Preprint)*. Aug. 2023. arXiv: 2308.11277 [cs.CV]. URL: https://arxiv.org/abs/2308.11277.

[HMB21] Timo Homburg, Hubert Mara und Kai-Christian Bruhn. *Cuneiform in the LOD cloud: Connecting 2D and 3D representations of philological objects with linguistic concepts*. Poster at Linked Pasts Conference. Nov. 2021. DOI: 10.5281/zenodo.5749763. URL: https://doi.org/10.5281/zenodo.5749763.

[JHS21b] Milos Jovanovik, Timo Homburg und Mirko Spasić. *A GeoSPARQL Compliance Benchmark - Preprint*. Englisch. Feb. 2021. arXiv: 2102.06139 [cs.DB]. URL: https://arxiv.org/pdf/2102.06139.

[Abh+20a] Joseph Abhayaratna u. a. *OGC GeoSPARQL 2.0 SWG Charter*. Englisch. Aug. 2020. URL: https://portal.ogc.org/files/?artifact_id=94480.

[Abh+20b] Joseph Abhayaratna u. a. *White paper: OGC Benefits of Representing Spatial Data Using Semantic and Graph Technologies*. Englisch. OGC White Paper. Okt. 2020. URL: http://docs.ogc.org/wp/19-078r1/19-078r1.html.

[HT20] Timo Homburg und Florian Thiery. *Linked Open Geodata in GIS? Ein Überblick über Linked Geodata Open Source Software*. Presentation. Juli 2020. DOI: 10.5281/zenodo.3931262.

[BHZ19] Tim Brandes, Timo Homburg und Ali Zalaghi. *Die Keilschrifttexte aus Haft Tappeh – Ein Werkstattbericht*. Presentation at ICDOG 2019. Presentation. Mainz, Germany, Apr. 2019. URL: https://converia.uni-mainz.de/frontend/index.php?folder_id=488.

[Hom19b] Timo Homburg. *Querying spatial data in the SemanticGIS project - Towards a new version of GeoSPARQL?* Englisch. 111th OGC Technical Committee. Presentation. Leuven, Belgium, Juni 2019. DOI: 10.13140/RG.2.2.25171.32807. URL: https://www.researchgate.net/publication/333674385_Querying_spatial_data_in_the_SemanticGIS_project_-_Towards_a_new_version_of_GeoSPARQL.

[Hom15a] Timo Homburg. *Verfahren zur Wortsegmentierung nicht-alphabetischer Schriften*. DARIAH-DE Digital Humanities Award 2015. Berlin, Germany, März 2015. URL: https://de.dariah.eu/documents/61689/82910/46_dhaward.pdf/82a466e7-436c-4637-b076-d05171ebd90f.

[Hom15b] Timo Homburg. *Verfahren zur Wortsegmentierung nichtalphabetischer Schriften*. Valedictorian Award Computer Science Master of Science. Frankfurt, Germany, März 2015. URL: https://www.researchgate.net/publication/312605727_Verfahren_zur_Wortsegmentierung_nichtalphabetischer_Schriften.

[Hom+15] Timo Homburg u. a. *Learning Cuneiform The Modern Way*. Graz, Austria, Feb. 2015. URL: http://gams.uni-graz.at/o:dhd2015.p.55.

[Asi+13] A Asir u. a. *WikiNect: Gestisches Schreiben für kinetische Museumswikis*. Frankfurt, Germany, 2013. URL: https://hucompute.org/applications/wikinect/.

[Hom12] Timo Homburg. *Entwicklung einer Androidanwendung zur Zustandsanzeige von Messstationen des Deutschen Wetterdienstes*. Valedictorian Award Computer Science Bachelor of Science. Wiesbaden, Germany, Feb. 2012. URL: https://hds.hebis.de/hsrm/Record/HEB305977989.