Expanding Digital Assyriology With Open Access and Machine Learning

A Cuneiform Digital Library Initiative White Paper for the Global Philology Project

Émilie Pagé-Perron

Global Philology Open Conference

February 2017

Keywords: Digital Assyriology, Open Data, Machine Learning, Universal Access, Semantic Web

Abstract: http://dx.doi.org/10.17613/M6MP5H

CDLI Website: http://cdli.ucla.edu/
CDLI Collaborators: http://cdli.ucla.edu/?q=associates-staff
Executive Summary

After giving a short introduction about the Cuneiform Digital Library Initiative, this paper looks at some important aspects of the state of natural language processing (NLP) for cuneiform languages, highlighting some of the roadblocks to the development of tools and techniques stemming from the particularities of the languages and material sources. Next discussed are challenges that arise from domain culture and practice. Ongoing projects at the CDLI are touched on, demonstrating how the avenues of research explored are meant to resolve some issues. In the final section, concrete solutions for moving forward as a field are proposed.

Overall, digital Assyriology has a solid base of interest and tools; projects dating back more than 20 years demonstrate the discipline’s advances in cyber scholarship. But there are roadblocks that prevent the field from exploiting the sort of digital tools that are available to scholars of English, other modern languages and some ancient languages such as ancient Greek and Latin. Cuneiform languages prove to be a challenge for NLP because of their ideo-syllabic writing system and their rich morphology. But before attempting NLP, other barriers impede research, such as access to the physical artifacts, be it in person or in digital form. Dissemination of information concerning texts is also hindered by the publishing tradition and a partial embargo against unprovenienced artifacts. Other barriers arise such as access to curated data and research results. The almost zero attention paid to accessibility practice in web design constitutes an additional challenge to knowledge democratization.

The Global Philology project is an excellent means to begin removing these barriers and augment the capacity and capability of Assyriology and other small disciplines (so called Orchideenfächer). Some of these goals are:

- Shared encoding scheme for cuneiform languages
- Shared NLP toolkit
- Integrating machine learning into research methods
- Multi-disciplinary collaboration
- Linked Open Data

Among these, the most promising option is using Linked Open Data, but in order to have meaningful large datasets to share, machine learning must be integrated into research methodologies. Additionally, the Classical Language Toolkit has invited Assyriologists to extend the tool so it can process cuneiform languages. These opportunities cannot be exploited in the absence of multidisciplinary collaboration.

About the CDLI

The Cuneiform Digital Library Initiative\(^1\) (CDLI) offers a set of tools and hosts the largest collection of images, drawings, metadata and textual information from artifacts inscribed in cuneiform and related writing systems. The artifacts in question are mainly clay tablets\(^2\) inscribed in ancient Iraq and adjacent regions from the mid-4th millennium BC to the end of the first millennium. These primary sources, of which 90% are administrative documents, are rich in

\(^1\) [http://cdli.ucla.edu/](http://cdli.ucla.edu/), directed by Robert K. Englund, UCLA, and Jürgen Renn, MPIWG-Berlin, both of whom have been very supportive in the creation of this white paper for the Global Philology Project.

\(^2\) For an example, see this Ur III tablet from Umma: [http://cdli.ucla.edu/P101049](http://cdli.ucla.edu/P101049).
information about the economy, history, social setting, religion and science of the ancient cultures that produced them.

The main objectives of the CDLI are compiling and preserving information about all artifacts inscribed with cuneiform, providing this information for viewing and reuse, and advancing research in the field. We estimate the number of cuneiform artifacts to be between 500,000 and 600,000, spread around the world in public and private collections. The CDLI currently hosts 320,000 entries, each of which has a set of metadata recording information such as host institution, provenience, period, related publications, and others. Textual data include transliterations, transcriptions, translations and further annotation. The visual aspect of artifacts is also documented through varying photographic and scanning methods. In addition to the scholarly journals it publishes, most of the services offered by the CDLI are specialized presentation of data harvested from the databases and analyzed, for instance the Ur III period calendars view and the scores page for compositions, although the main services are advanced search facilities and data archives.

The CDLI first focused on 4th and 3rd millennium sources, but soon extended its breadth to the whole of the cuneiform repertoire. Since its early days, it has been an international project with strong ties in Germany, the country of its inception, now hosted at UCLA. Assisted by grants from the NSF, NEH, IMLS, the Mellon foundation and the John Fell Fund, the CDLI entered into collaborations with most major museums to digitize their cuneiform holdings. Staff have conducted missions in Europe, the Middle East, and North America to document small and large collections. Through systematic efforts, the CDLI database and image bank continue to grow every year.

2 Digital Assyriology and Natural Language Processing

Given its long history of methodology grounded in German philological tradition, Assyriology is surprisingly one of the fields that took an early digital turn. Several cuneiform research projects began as personal files and databases that grew and developed into funded research tools, mostly geared toward the preparation of print publications by privileged collaborators. Many years before the major current online cuneiform databases became available to research, Giorgio Buccelati in Los Angeles encoded Ebla texts, Simo Parpola in Chicago and Helsinki the neo-Assyrian corpus, and Marcel Sigrist in Jerusalem the Ur III accounts. Projects promoting Open Access and Open Data, however, are largely the result of a Berlin collaboration among researchers from Near Eastern Studies at the Free University and the Max Planck Institute for Human Development (working together with a number of scholars associated with Gregory Crane’s Perseus Project), and subsequent cooperation with European and American partners, above all Manuel Molina of the Spanish National Research Council, Madrid, and Steve Tinney of the University of Pennsylvania.

---

3 Among others: traditional photography, 2D scanning, photogrammetry and reflectance transformation imaging. This will be discussed further in the "Materiality" section under "Sources: Limitations & Problems".
4 http://cdli.ucla.edu/?q=publications
5 http://cdli.ucla.edu/tools/ur3months/month.html
6 http://cdli.ucla.edu/tools/scores/partitur-index.html
7 http://cdli.ucla.edu/search/
8 http://cdli.ucla.edu/?q=downloads
Standardization
Despite the now nearly three decades of collaborative work, there is actually no unified encoding for the transliteration of cuneiform languages. The better known schemes are C-ATF\(^9\) and ORACC ATF\(^{10}\). When character assignments overlap exactly (e.g., š ⇔ sz), they can be easy to convert, but the mix of semantic and structural notations can make it difficult to implement conversion tools. This creates ambiguities where meaning cannot be transferred or inferred when passing from one encoding system to another. Persistent problems include, for instance, numerals that can be noted specifying which sign appears in which quantity, or that only note a total, or some intermediary notation. Legacy systems often use accented characters to convey second and third homophonous signs,\(^11\) which look nice in print but are definitely not practical to parse. Broken or partly illegible signs can be noted as damaged using the hash sign after a sign, e.g.: ba#, or with square brackets, e.g.: b[a] meaning the right part of the sign is damaged. Half square brackets can also be used to indicate further the localisation of the breakage through specifying the affected quadrant. In these cases a perfect conversion from one system to another is impossible. Non-existent shared standards hinder the development of most needed shared natural language processing (NLP) tools such as the Classical Language Toolkit that, if enabled to process Sumerian and Akkadian, would facilitate our digital research.

Processing and Annotations
The current prevalent system for the generation of annotations of cuneiform texts is ORACC’s rule- and dictionary-based automated lemmatizer system that creates inline glossing, attributing linguistic information to tokens based on glossaries and manual verification. After an automated processing step, the user can decide to update the annotation associated with the occurring token, according to their personal analysis.

The limitations of rule- and dictionary-based methods are twofold. In the case of large corpora, if the text is not very homogenous, the time required to verify each line makes this type of approach unpractical. This important need for human involvement is due to the second challenge: contextual disambiguation. ORACC works with project-based glossaries that enhance results compared to using a general glossary. But to achieve a reliable unsupervised method, context needs to be restricted to smaller units such as sentence and text. The co-occurrence of words in a sentence or text,\(^12\) the tense of the verb or the part of speech (POS) of other words in the sentence etc, are all elements that artificial intelligence, just like human intelligence, can detect and use to disambiguate meaning and usage. At this time, contextual disambiguation is not used for the treatment of cuneiform languages.

Rule based and statistical information extraction has gained popularity. It is either used on lemmatized corpora, as appears to be the case in the Ancient Records of Middle Eastern Polities (ARMEP),\(^13\) an interactive map project, or on manually processed raw corpora, for

---

11 For example: sú ⇔ su₂.
12 E.g., if a specific noun is the subject of a particular verb, then the verb has this meaning; if a text treats a specific topic, say, one type of animal or grain, then the numerals should be interpreted as representing this or that counting or metrological system.
13 [http://oracc.museum.upenn.edu/armep/index.html](http://oracc.museum.upenn.edu/armep/index.html)
example in the case of Niek Veldhuis' Ur III names lists\textsuperscript{14} where names that are retrieved from
the Madrid-based BDTNS database are keyed on capitalized words and determinatives. Statistical methods can be applied using concordance tools, as in the case of Sara Brumfield
(2013), but they are also explored using topic modeling (Veldhuis 2016). All of these
approaches necessitate the intense involvement of the researcher at one stage of the process
or another. Identifying the right information requires intelligent discernment, and processing a
large number of texts requires extensive time commitment.

Machine Learning
Machine learning is mostly absent from Assyriological research. Computer scientists
sporadically procure data derived from cuneiform sources and enact experiments, such as for
example Liu et al. (2015), who with her colleagues worked on unsupervised entity recognition
and also more recently on dating tablets based on the information gathered about these entities.
In this experiment, the collaboration of an Assyriologist (Manuel Molina) who annotated the
corpus was instrumental to the success of the research (Hearne & Liu 2016). Some
experiments have also been carried out on Akkadian word segmentation, experiments that will
most certainly prove useful in the development of OCR for cuneiform texts (Homburg &
Chiarcos 2016).

3 Challenges of the sources

Materiality
Capturing volume and text
Cuneiform texts are in the great majority inscribed, or more specifically impressed, on three
dimensional, pseudo-rectangular objects of varying size, made out of clay. The inscription can
appear on all 6 surfaces and sometimes span from one side to another, wrapping around a
curved side. Although there are interesting developments in 3D technologies, some being
effectively used to model cuneiform artifacts, there is currently no large scale solution for the
digitization of the artifacts’ surfaces that can accommodate the very large number of tablets
scattered around the world. As opposed to ambitious projects that sought to create the ultimate
imaging tool for cuneiform artifacts,\textsuperscript{15} the CDLI developed a simple, cheap and fast method to
digitize large numbers of such objects using an affordable deep flatbed scanner.\textsuperscript{16} This method
made it possible to gather images of, at a minimum so far, a third of the world’s unearthed
cuneiform artifacts. Restricted access to collections can and does, however, impede the
digitization of artifacts.

Seal impressions
Administrative and legal cuneiform clay tablets were often sealed using a cylinder seal rolled on
the surface to leave an impression. Such impressions have traditionally been studied separately
from the text of the tablet on which they appear. The line art of tablets will traditionally display an
indication of where one most visible seal impression was and the legend of the seal, if there was
any, but accompanying iconography was rarely or poorly reproduced, and multiple impressions

\textsuperscript{14} \url{https://github.com/niekveldhuis/UrIII-names}
\textsuperscript{15} Some might remember the Hammurabi project \url{http://pages.jh.edu/~dighamm/version2/}
\textsuperscript{16} The actual technical procedure is noted here: \url{http://cdli.ox.ac.uk/wiki/doku.php?id=submission_guidelines}
would not be recorded.\textsuperscript{17} Seal imagery was and is mostly still studied by art historians. Unfortunately, current tablet scans do not display seal impressions as clearly as the text of the tablet on which they are impressed. Reflectance transformation technology gives better results.\textsuperscript{18} Jacob L. Dahl at the University of Oxford and his partners at the University of Southampton are currently developing new technologies for capturing also the physical cylinder seals.

Unprovenienced artifacts
The question of how to deal with unprovenienced cuneiform artifacts has created a rift among Assyriologists and archaeologists. Some associations, conferences and periodicals have chosen to limit communications about these artifacts that are now in public or private collections around the world. The point of dissent lies in the belief that work on these artifacts constitutes authentication and encourages greater looting of ancient sites in the Middle East. This debate will not be discussed further here but needless to say: artifacts that are not studied or of which the analysis cannot reach the research community cannot contribute to the generation of knowledge and as such the research community—and countries of artifact origin—suffer a lack of access to this group of primary sources.

Access to collections
Collections are not always easily accessible to scholars. That can result from the distance and means to visit a collection, but the barrier can lie on the side of the cultural heritage institution. This challenge stems from the conception of a publication and the reservation of rights of publication to certains scholars. In 2003, the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities did include a word about cultural heritage, stating: "cultural heritage shall be freely accessible and usable for scientists and the public".\textsuperscript{19} Obviously the culture shift to open access that was expected after the declaration is a slow process, even though some museums did facilitate shared data, encouraging the imaging of their collections. Some examples are the Penn museum, the British Museum and the Royal Museum of Ontario, who openly shared information about their still unpublished holdings.\textsuperscript{20}

Project access and accessibility
\textit{Data & Code}
The data and code produced within the context of digital projects in Assyriology are usually not easy to access. Access to online content can require a login and can be limited to researchers who are part of the project, and to other privileged individuals. Other obstacles are licenses and terms of use which can be extremely restrictive and thus block the user from even modifying the data after downloading a simple sample. In most other cases, data are copyrighted or under a creative common license. Luckily, some projects do offer their data in some form of open access, where the work, information and data are viewable for consultation. The data can be hard to reach: either they are simply not available for download, are not in an open format, or

\begin{footnotesize}
\item[	extsuperscript{17}] Compare the image and line art of this Ur III text from Umma: http://cdli.ucla.edu/search/archival_view.php?ObjectID=P105282
\item[	extsuperscript{18}] Compare the RTI render of the obverse of this Ur III text http://cdli.ucla.edu/search/rti/rti_view.php?loc=P142832_o (click and drag to orient the light) with the matching photo image: http://cdli.ucla.edu/dl/photo/P142832.jpg
\item[	extsuperscript{19}] https://openaccess.mpg.de/mission-statement_en
\item[	extsuperscript{20}] Unpublished Penn museum texts: https://goo.gl/PV31Gk, British Msuseum https://goo.gl/B8L2zt and the Royal Ontario Museum texts: https://goo.gl/eDmi5B
\end{footnotesize}
not available in bulk. Open access does not only equate with free access as in no royalties or fees. An open access work, as defined at the 2002 Budapest Open Access Initiative, must be with "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself".21 Open access encompasses any work that is digital (Suber 2012:9), thus placing or not removing technical barriers to data is contrary to the spirit of open access. Making available on the web some results to view is comparable to publishing a monograph in terms of the technical barriers to reusability. This is one of the major problems with our digital projects. When a project is put online it takes minimal effort to open its data but only two projects can hit a 3 on Berners-Lee’s 5-star open data scale.22 (Pagé-Perron et al. 2017b) Other projects necessitate the use of a web scraper software or other technical means that are not accessible to everyone. This is not an acceptable method of access to freely share information since it imposes a serious technical barrier which, again, is contrary to open access principles that ask for the removal of such difficulties of access.

Licensing
The element of open access that seem to be the most understood and appreciated in the field is the fact that licenses can be a barrier to access. Many projects and publications venues offer the option of using a creative common license. Moving away from copyright opens the door to reuse, especially with less restrictive licenses. But it is possible to do even better. In the last year, the CDLI has openly pledged that all new data and code produced by our research projects, that is information extracted from our transliterations, machine generated annotations and all algorithms and code created to support our research or to visualize results, will be released to the Public Domain using the CC0 license.23 CDLI is not the only organization to think that CC0 removes impediments to innovation. Just today, the Metropolitan Museum of Art in New York released their extensive photostock to the public domain.24 Removing all restrictions on works can only encourage others to reuse it legally and without fear of breaking licences. Additionally, as indicated in the CC0 FAQs25, in an academic setting, credit must still be given to the author and it is not possible to claim either ownership or authorship of public domain material because plagiarism rules do apply. Waiving this ethical and professional obligation and waiving rights on work does not mean that the product becomes authorless.

Interface (Accessibility)
Accessibility is the next big access challenge to address. Not a single project in Assyriology validates against the World Wide Web Consortium HTML26 and CSS27 checkers, and this is the absolute minimum required to start offering an accessible web portal (Juloux & Pagé-Perron 2017). The additional guidelines to help converge towards universal design, so that anyone can

21 http://www.budapestopenaccessinitiative.org/read. This highlights how these barriers actually impede the "common intellectual conversation".
22 http://5stardata.info/en/
23 https://creativecommons.org/publicdomain/zero/1.0/
24 February 8th, https://creativecommons.org/2017/02/07/met-announcement/.
25 https://wiki.creativecommons.org/wiki/CC0_FAQ
26 https://validator.w3.org/
27 https://jigsaw.w3.org/css-validator/
access content, are the Web Content Accessibility Guidelines (WCAG)\(^{28}\) that appear to be mostly unknown in the field.

### 4 Ongoing at the CDLI

**Primary Data collection**
Digitizing cuneiform artefacts, preparing transliterations and translations, digitizing line art are ongoing activities year round. Additionally, Jacob L. Dahl directs the "Seals and their Impressions in the Ancient Near East" which has three objectives:

- Develop an adequate methodology, including adapted hardware, for digitizing physical seals and seal impressions
- Digitize a number of seals and impressions using that technique
- Answer questions on origin, manufacture and recarving of seals, but also looking at matching seals and impressions, using among others, computational methods for typological analysis

**Preservation and sustainability**
Preservation and sustainability have often been overlooked in early digital projects but are now recognized as an essential component of a project's planning, as demonstrated by the requirements concerning these topics in digital humanities project proposals for different American funding bodies. If the CDLI has been able to survive for so long, it is in part due to the involvement of more than a few scholars in the initiative. This dynamism is complemented by measures taken to improve the potential longevity of the initiative and the preservation of its data. Our preservation solutions so far consist in:

- Public repositories for data & code
- Multiple international backup and service mirrors
- Codebase maintenance and update

The project's data bundles have been freely offered for download, the catalogue in Filemaker format and and the textual information (transliterations and translation) in text format. Offering the CDLI codebase on GitHub has been discussed and will be set up when we start the "Framework Update" project, hopefully in September 2017, pending funding. We have also instituted a catalogue dump in csv format in order to extend our open format data offer.

Since its early years, the CDLI has sought to extend its backup and mirroring network to offer a persistent service and to preserve copies of the data and interface in case of a disaster. Partner institutions are the Max Planck Institute for the History of Science in Berlin,\(^{29}\) also the home of co-director Jürgen Renn, and the University of Oxford,\(^{30}\) under the auspices of Jacob L. Dahl, co-PI. There is hope that this network can soon extend to Canada through Compute Canada,\(^{31}\) a supercomputing and storage service offered to Canadian researchers. This partnership would be made possible through the Ontario branch SciNet\(^{32}\) and the involvement of the University of Toronto with Heather D. Baker and Émilie Pagé-Perron.

---

\(^{28}\) [https://www.w3.org/TR/WCAG20/](https://www.w3.org/TR/WCAG20/)
\(^{30}\) [http://www.orinst.ox.ac.uk/](http://www.orinst.ox.ac.uk/)
\(^{31}\) [https://www.computecanada.ca/](https://www.computecanada.ca/)
\(^{32}\) [https://www.scinethpc.ca/](https://www.scinethpc.ca/)
In order to accommodate young research sub-projects and to give them a solid foundation, the "CDLI Framework Update Project" ("Phoenix"), starting in September pending funding, has the following main objectives:

- Make the code base more sustainable
- Document the initiative code and practice throughout
- Enable graduate students in the humanities to code extensions themselves
- Renew data displays for usability
- Become a model of best practice in web accessibility

This project will ensure that subsequent research endeavors at the CDLI can rely on a highly usable, accessible and modulable platform so new data can be easily analyzed, transformed, shared and reused.

Machine Learning

Active discussions have happened between the University of Toronto, UCLA, and Frankfurt University, leading to the organization of an international research group for the preparation of a methodology using machine learning (ML) and linked open data (LOD) to enable automated translation, information retrieval and open dissemination of information based on the processing of cuneiform transliterations (Pagé-Perron et al. 2017a).  

Administrative texts make up the bulk of the cuneiform texts we have at our disposal to study ancient Mesopotamian cultures. However, they remain mostly untranslated, making them effectively inaccessible to anyone not having been trained to work with either the Sumerian or Akkadian language of a specific time period. Additionally, they are the group of documents on which it would be the most pertinent to apply statistical methods to infer information about society and economy. This type of investigation is still in its infancy and focuses on small corpora that can be entirely processed manually.

Pending funding, CDLI will prepare a methodology to address those gaps and will experiment on a corpus of 67,000 Sumerian texts dated to the 21st century BC. The methodology and its algorithms that will result from this research project will be reusable and modifiable to be reused with other corpora.

Natural Language Processing and automated translation

As outlined in the challenges section, one of the main obstacles in the processing of cuneiform transliterations is their sheer number. This is why machine learning has to come into play, especially in the case of tasks like the translation of administrative texts which are not often carried out because they do not benefit directly the core of scholars that use these texts as primary sources for their studies. Computational linguists develop algorithms that cover the whole range of natural language processing but only a few experiments have been successfully conducted on cuneiform languages. Luckily, it is possible to adapt existing neural and statistical algorithms that are adapted to other languages. For example, to disambiguate morphology, Sak’s and colleagues’ (2007) neural techniques for Turkish will be adapted and tested, along with Rias’s and Mamani’s (2014) statistical method for Quechua. For automated translations, both statistical machine translation and neural network-based machine translation will be tested.

33 This section relies on the reference cited but also on unpublished material on which the same authors have collaborated, in addition to Robert K. Englund and Heather D. Baker.
on Sumerian and compared with results of Basque and Turkish samples. These techniques have been proven successful with different languages, including morphologically rich and ideo-syllabic languages.  

There are evident obstacles to the unsupervised annotation and translation of cuneiform texts. For instance, words that have an unknown or uncertain meaning will weaken the translation output. But context appraisal will help clarify the meaning of such words. The result of automated translation will not reach the finesse of translations prepared by scholars. But the goal is not to produce beautiful translations, it is to create a product that will provide the meaning necessary for those outside the field and for learners to understand the overall semantics and particularities of the text. Translations will be accompanied by linguistic information which will provide an additional tool to aid comprehension.

**Information Extraction**

Because the translation process requires the annotation of the transliterations (except in the case of neural networks usage), the CDLI will reuse the annotations generated to go one step further to identify concepts by using collocation algorithms, identifying and disambiguating toponyms and anthroponyms. This will be useful especially for prosopographical research. Going even further, statistical methods will be employed to track variables over time to facilitate social studies.

**Optical Character Recognition**

Although there are ongoing experiments by different research groups in OCR of cuneiform, the CDLI has a specific objective of developing a method to automatically map signs seen on artifacts’ surfaces with the appropriate signs in the textual annotation we have of these sources. At this stage, experiments for the recognition of individual lines and signs are being performed.

**Linked Open Data (LOD)**

Linked open data will be an integral part of all future research projects at the CDLI. The first step will be linking textual information, integrating Olia ontologies with linguistic annotations (Chiarcos & Sukhareva 2015) and semantic information from the text with widely used ontologies such as Pleiades for places, PeriodO for time periods, and Snap:drgn for ancient people relationships. In a second step, metadata will be aligned with the CIDOC-CRM, hopefully in collaboration with the ModRef French project. The goal of this integration is to standardize our results to permit interoperability with other projects as the most promising method to promote the reuse and aggregation of the CDLI curated data.

5 Vision and Prospects

34 For a detailed description of the method involved, from the raw transliterations to the final product, consult Pagé-Perron et al. 2017a.
35 [http://linghub.lider-project.eu/datahub/olia](http://linghub.lider-project.eu/datahub/olia)
36 [https://pleiades.stoa.org/](https://pleiades.stoa.org/)
37 [http://perio.do/](http://perio.do/)
38 [https://snapdrgn.net/](https://snapdrgn.net/)
40 [http://triplestore.modyco.fr:8080/ModRef/cdli](http://triplestore.modyco.fr:8080/ModRef/cdli)
The Institute of Museum and Library Services (IMLS)⁴¹ has been fostering discussions about the idea of a "national digital platform" which is a way of thinking, more than an actual formal structure: it is the sum of the network of digital capacity and capability of American libraries and museums (Erway et al. 2015)⁴². The IMLS has been funding projects that promise to eliminate gaps in this network in order to increase this national digital capacity and capability. The Global Philology Project, by increasing communication between scholars and proposing a global approach to the study of small disciplines and languages, is promoting a similar approach, in addition to the recognition of the advances in digital capabilities of German scholarship. What are ways to increase this capacity? In her keynote address to the National Endowment for the Humanities director’s meeting in 2015, Bethany Nowviskie made a case for the correlation between the capacity of digital projects and the ethics of care (Nowviskie 2015). In other words, showing care not only for people, but also for the artifacts we create and preserve, increases capacity.⁴³ Some examples are, for instance, integrating the public into the creation of knowledge and fostering research environments where non-tenured staff are granted due consideration but also offering research results in open access; embracing open data and linked data are other ways to achieve this growth.

Open Access (OA), Open Data and Open Source
The CDLI has been offering its textual data in open format for immediate download and in bulk for some time. The catalogue was available in FileMaker format but has just recently been made available in open format as a daily snapshot. ORACC has also recently made available much of its data to fetch in json format. It is these sort of initiatives that will hopefully speed up the dissemination of curated data between research groups and enable researchers to take a step further in the exploration of larger sets of data with computational methods. More projects need to increase reuse of data, and the same should be done for code. Again here ORACC is leading by making its codebase available on Github. Other projects can thus reuse part of the technology that was developed, stop reinventing the wheel and focus on creating novel methodologies and producing new, quality data. If Open Access is about removing barriers, Linked Open Data is about building bridges. Seeing how difficult it could be to harmonize standards for the textual data of Assyriological projects, further steps obviously should incorporate linked open data so that the product of research can be shared, reused and aggregated, and this with projects that would not normally interact with us, such as studies on languages from different places and time using a linguistic linked open data⁴⁴ ontology to annotate linguistic annotations. There are many established ontologies and vocabularies that are shared by many projects; Classics are again a model with the Pleiades Project. to share processed data sets, a practice that can be mimicked from other fields such as Classics with the Perseus Library but also a growing number of digital humanities on all topics.

Scholars are more inclined than before to publish their results in open access form, and hopefully this trend will gain in acceptance. OA works are cited more often and have a larger readership, with new reach in developing countries. The public demands access to research

---

⁴¹ [https://www.imls.gov/](https://www.imls.gov/)
⁴² Definition coined by Trevor Owens.
⁴³ This suggests that caring for our data and our audiences is linked to the success, growth and possible preservation of long term digital projects. It benefits others and the project in question. This can be achieved in numerous ways that can be combined for a maximum effect through networks of care.
works (Willinsky and Alperin 2011), and scholars should adopt an ethical and responsible approach to the dissemination of their research results as a public good, and make efforts to democratize knowledge. OA publishing is encouraged by funding agencies, by academic institutions and in Germany we have seen the project “DEAL” stand up against Elsevier in an unprecedented stance to defend the principles of open access. OA is the publishing model of tomorrow. Hopefully we can follow in the path of particle physics where virtually 100% of publications are OA.

Machine Learning
Aside from this slow cultural shift to open access, one important step that should be taken is the exploitation of machine learning for the study of ancient languages. As was demonstrated earlier, it seems inevitable to consider the integration of such techniques to overcome the current barriers in the digital exploration of the languages we study. In this we are fortunate to have colleagues from the computer science and computational linguistics domains that have already started working on algorithms to process data from cuneiform texts. The technology is ripe and accessible and more experiments on ancient languages every year show the immense potential of these types of algorithms. To make a significant advance, multi-disciplinary research groups need to form, since the success of enterprises that will bring a true impact, in the current context, cannot be achieved by computer specialists or Assyriologists on their own. The investigation of the languages, the cultures, the materiality and history of ancient peoples can be further understood by supplementing our approaches with intelligent pattern recognition, highlighting pertinent information that could hardly be detected or would be ignored using traditional means, especially in the case of larger corpora. Especially in natural language processing, machine learning can help close this gap of the "human touch" in automated translation and information extraction, taking into account the context of a word or information and intelligently overcoming ambiguities in the sources.

Next steps at the CDLI?
As discussed earlier, the CDLI has planned on rewriting its software platform and to put in place a methodology for the automated translation and information retrieval of cuneiform languages, as a beta rollout employing an important corpus of Sumerian texts of the 21st century BC. From there, it would be logical to extend the methodology to other periods and to Akkadian texts. Additionally, with the integration of linguistic linked open data to the CDLI services, the obvious next step would be to link the metadata of artifacts, and indeed this is currently being studied. The next interface for information pages, catalogue data and transliterations will conform to W3C html and css regulations and to WCAG guidelines, the practice of which will be extended to all CDLI web services.

45 https://www.projekt-deal.de/
46 https://www.leopoldina.org/fileadmin/redaktion/Publikationen/Allianz/Allianz_Verhandlungen_Elsevier_EN.pdf
6 Bibliography

Babeu, Alison 2011 "Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics. Council on Library and Information Resources

Brumfield, Sara 2013 "Imperial Methods: Using Text Mining and Social Network Analysis to Detect Regional Strategies in the Akkadian Empire." PhD Thesis, UCLA


http://www.imls.gov/assets/1/AssetManager/2015_IMLSFocusNDPReport.pdf

http://dx.doi.org/10.1017/CBO9781316161012

Hearne, James and Yudong Liu "Dating Tablets of the Ur III Period" Rencontre Assyriologique Internationale 63, Philadelphia


Juloux, Vanessa and Émilie Pagé-Perron 2017 "Enabling Universal Access of Assyriology and Humanities Digital Projects" (under review)

http://www.aclweb.org/anthology/N15-1167

Mace, Ronald L., Graeme J. Hardie, and Jaine P. Place 1990 Accessible environments: toward universal design. Raleigh, NC: Center for Accessible Housing, North Carolina State University


Pagé-Perron, Émilie, Maria Sukhareva, Ilya Khait and Christian Chiarcos 2017a "Automated Translation and Information Retrieval of Ur III Cuneiform Texts" (under review)


Robson, Eleanor 2014 "Tracing Networks of Cuneiform Scholarship with Oracc, GKAB and Google Earth" in M. T. Rutz & M. M Kersel eds. Archaeologies of Text:


Veldhuis, Niek 2016 "Computational Text Analysis: Comparing Lexical and Literary Vocabularies" *Rencontre Assyriologique Internationale* 63, Philadelphia

Willinsky, John and Juan Pablo Alperin 2011 "The academic ethics of open access to research and scholarship" *Ethics and Education* 6(3) http://www.tandfonline.com/doi/abs/10.1080/17449642.2011.632716