

**White Paper**

Grant number: HAA-256224-17

"Cuneiform Digital Library Initiative Framework Update"

Principal Investigator :

Dr. Robert K. Englund - University of California Los Angeles

Submitted: September 30th, 2020

# Project Activities

Our project involved mainly the development of NLP processing algorithms and the web development of an advanced web platform employed to offer open access to various datasets, while providing services such as advanced search and processing access points.

We have shared tasks based on expertise with the Machine translation and Automated Analysis of Cuneiform Languages (grant number HJ-253601-17) where we handled preparing the data converters we have implemented in our infrastructure which they required to process their data while they provided expertise for designing the interface of the web platform we were building, which we were in agreement they would use to display the data they produced.

A major change in our project was caused by arduous negotiations with our File Maker contractor. After months of unfruitful discussions, we decided to work with another contractor who performed the tasks assigned successfully.

## Accomplishments

### Data & Database

The goal of this objective was the creation of an optimized relational schema for the database and connection with the PHP Framework and Filemaker, all these objectives were achieved.

### Code core

Early in the project, we decided to abandon the idea of using the Drupal CMS for this iteration of the CDLI framework, based on the fact that Drupal cannot handle database replication which was essential in our infrastructure design. Instead, we built the part of the site we thought would use Drupal into our PHP Framework, CakePHP.

We have successfully replicated most of the current functionalities of the CDLI platform into the new core. Some fine tuning is still required, for example, testing is required for the collections images bundles download and our search engine still needs optimization for some search fields.

Because the MTAAC team handled the design part of our interface based on their local expertise, we have taken on the development of data converters which were to be implemented in the new Framework. The various tools we have prepared are the following:

- C-ATF to CDLI-CoNLL converter
- ATF to TEI converter
- CDLI-CoNLL to CoNLL-U converter
- CoNLL-U to Brat standalone converter
- Brat standalone to CDLI-CoNLL converter
- CDLI version of PyOracc, an ATF format and content checker
- JTF ATF format and content checker

# Interface Concept, Design, and Implementation

The conception of our interface and its design was handled by a member of the MTAAC project team which has extensive experience in interface and user experience. We conducted extensive research in the usage of our current platform which informed every decision we took in preparing the design. The implementation is completed for existing pages and all these pages are WCAG compliant to the AA level or will be in the next weeks.

## Documentation

In collaboration with the MTAAC project, we have set up a documentation website which documents most of the newest data formats in use at the cdi. As we finalize the search engine and the interface, we will integrate standalone documentations that were produced to accompany each feature into our documentation website for a more coherent and easy to consult central documentation.

## Servers

We have moved our whole code infrastructure to Gitlab in a public repository, under an open license. The deployment of the platform is now done using Docker and Ansible. Our backups are running as planned to all partner locations (Compute Canada, Max Planck Institute for the History of Science in Berlin and at the University of Oxford) and we will deploy mirrors of our website once the new platform will be ready to replace the current CDLI Website.

## Future developpement

Some work remains to finish up the Framework, essentially integrating various components of which the developpement is finished, fine tune the search engine, etc. Through its Oxford University Co-PI the CDLI will soon hire a Devops and a Developers, the first in order to finish up any tasks related to de deployment and backup infrastructure, and the second to assist in the finalization of the development of the platform.

We currency have three volunteers working continuously on our code base and as December approaches, the influx of Google Summer of Code prospective participants will bring a few more. 2020 is the first year for which we were able to sustain the retention of volunteers offering continuous support to the project. It took us four years to get here and we expect our base to keep growing as we have developed retention expertise.

# Audiences

## Current audiences

The current CDLI platform accommodates on average 2500 sessions per month for 20k page views. Only 25% are new visitors as the bulk of our audience consists of researchers. 26% of our traffic comes from the US, but Germany comes second at 13%, followed by France, Japan, and the UK. In the last month we have received traffic from 62 different countries. Our users almost invariably search for an artifact or a group of artifacts based on its metadata and inscription. When we release the new platform, the users will be able to perform the same activities but will have access to a lot more information about the artifacts, inscriptions and annotations and in many more formats.

Some of our repositories in Github are very active, for instance our main data repo has generally between 1 to 3 visitors every day (<https://github.com/cdli-gh/data/graphs/traffic>). Our Gitlab main repository, which contains the code for our new web platform, is active everyday with volunteers and senior personnel opening and closing issues, committing improvements, etc.

## Prospective Audiences

As we have worked on making our web platform more accessible, we hope to reach a larger and more varied public and not only colleagues in our field and serious enthusiasts. Before embarking in the redesign of the interface, we have conducted a large scale survey with our users to have a better understanding of who we were serving. Based on the participants' answers, we were able to identify many accessibility problems in our current interface which we addressed in our new design. Therefore, we expect that fringe users will be able to get the information they need much more easily and thus use the site more often, also anyone who is differently abled might see their access need met.

It is well known that there are few consumers of Linked Data at this time but we assume that, as the tools to consume linked data become more refined, more consumers will be ready to ingest our data. To this effect, we will be ready to serve them with our API and SPARQL endpoint. Our offer is in line with parallel developments in other philologies, e.g., for Latin (<https://lila-erc.eu/>), poetry (<https://postdata.linhd.uned.es/>), and the humanities in general (for topics such as geospatial identifiers [<https://pelagios.org/>], chronology [<https://chronontology.dainst.org/>], coins [<http://nomisma.org/>] or museums [<https://pro.europeana.eu/page/sparql>], or archives [<https://www.timemachine.eu/time-machine-a-pan-european-digitisation-and-processing-infrastructure/>]) where Linked Data solutions are applied to provide machine-readable data for their respective communities and for integrating their respective data. In the long term, an accelerating integration of resources and data is to be anticipated, as we are currently seeing in the development of the language sciences (<http://linguistic-lod.org/>), where global networks have been embracing the linked data paradigm to integrate their resources, especially in lexicography (for example, the Global WordNet Association now provides an RDF/OntoLex-based Interlingual Index to interlink lexical networks at a world-wide scale, <http://globalwordnet.org/>). With more

and more communities using Linked Data to consolidate resources in their respective fields, chances of creating links between them are also emerging, and this is what we expect for the foreseeable future.

## Evaluation

The project was not formally evaluated, although we expect to set in place a more formal feedback system to track closely the needs of our audiences at more regular intervals.

## Continuation of the Project

CDLI is a long lived initiative, it has been running for 22 years now, offering essential services for researchers principally in the field of Assyriology. With this grant provided by the NEH, CDLI was able to renew it's code base and rehaul it's infrastructure. We expect this iteration of the platform will easily last 7-10 years, if not more, before requiring major maintenance work. Regular minor maintenance will be constantly performed.

The changes we have made to our development ecosystem now provides access to our code by anyone interested in contributing, while we already have multiple volunteers working with us at this time, who joined during the funding period. This would not have been possible with our old infrastructure. We also have become independent of our current hosts, meaning we could deploy the CDLI platform on any virtual machine around the world with low resources requirements, this increases drastically the sustainability of the CDLI.

This project also made it possible to modularly integrate services to our platform so as technology develops, we will easily be able to integrate and deploy adapted interfaces, tools and data for our community. For instance, we collaborate with a group that produces a 3D viewer and 3D viewable digital files. It is easy now for any of our contributors to taking on the update of this viewer when required. External parties can also submit requests to merge their code into our code base.

Through our collaboration with the MTAAC project, we have access to new resources at Compute Canada to host a mirror of our web server and backups. These resources help us both preserve our data while sustaining our web presence, guaranteed until 2023, and renewable every three years.

In all cases, we are committed to finalize the new CDLI interface which will be released in alpha version in October 2020 (<https://cdli.utoronto.ca>). Users will be able to employ both platforms while we finish up fine-tuning the new platform. The new platform should replace the CDLI production interface, which will stop being updated around September 2021.

In the next five years, CDLI services will be extended to cater for the requirements of the Akkadian language. This long-term and major research project will likely be a Frankfurt / University of Oxford collaboration. In the shorter term, we expect to start a collaboration between these same universities to finalize the implementation of an infrastructure to manage

vocabulary-related data for the Sumerian language. This infrastructure will be language-agnostic so we will be able to employ it in our Akkadian language project.

Aside from academic projects, we also plan to continue our work with volunteers, e.g., in the context of future Google Summer of Code projects, and as part of the regular data curation process at CDLI, then also extended to annotations and translations.

## Long Term Impact

CDLI, the main provider for metadata, images and textual data of cuneiform texts, has been gathering data for more than 20 years in the hope of a project like this. CDLI will be forever transformed by the new addition of services, data types and formats through this framework update. As such, its offer to various audiences has drastically evolved. In the first place, the new CDLI interface will serve a much larger public (see the details in the “Audiences” section) through the enlarged data pool we will offer, through the more accessible human interface, and through the various machine and human data and service points we have set up. But the largest impact we hope will be in showing researchers in our field how linguistic data can be presented, modelled and shared, both in open access and linked formats, and how it can be processed efficiently using cutting edge technology. The field of Assyriology has always been somewhat digital but with a lack of quantitative approaches and more involved digital projects. There is currently a shift in our field, with a renewed interest in digital approaches, and we expect our work to orient this shift towards best practices as computational linguistics can and has taught us.

## Award Products

The main product of this award is the code base of our new web platform which is accessible here: <https://gitlab.com/cdli/framework>. The converters we have developed are on Github:

- C-ATF to CDLI-CoNLL converter <https://github.com/cdli-gh/atf2conll-converter>
- CDLI-CoNLL to CoNLL-U converter <https://github.com/cdli-gh/CDLI-CoNLL-to-CoNLLU-Converter>
- CoNLL-U to Brat standalone converter <https://github.com/cdli-gh/conllu.py>
- Brat standalone to CDLI-CoNLL converter [https://github.com/cdli-gh/brat\\_to\\_cdli\\_conll\\_converter](https://github.com/cdli-gh/brat_to_cdli_conll_converter)
- CDLI version of PyOracc, An ATF format checker <https://github.com/cdli-gh/pyoracc> (unfinished)
- JTF ATF checker <https://github.com/cdli-gh/JTF>

The documentation website which was elaborated by the MTAAC project team but which we have augmented and which we will be maintaining on the long term:

<https://github.com/cdli-gh/cdli-gh.github.io>.

*This report was written by Émilie Pagé-Perron with important input from the Machine Translation and Automated Analysis of Cuneiform Languages final performance report and white paper. This white paper is identical to the final performance report.*