# Digital Assyriology: Research Plan

Niek Veldhuis, September 2016

<http://datasci.berkeley.edu>

## Research team

Niek Veldhuis, Professor of Assyriology, Department of Near Eastern Studies. Team leader. [veldhuis@berkeley.edu](mailto:veldhuis@berkeley.edu)

Jason Kha, Junior in CS. [jasonkha@berkeley.edu](mailto:jasonkha@berkeley.edu)

Swarup Sahu, Senior in Economics. [swarupsahu@berkeley.edu](mailto:swarupsahu@berkeley.edu)

Kyle Nguyen, team manager; BIDS liaison. [kyle\_nguyen@berkeley.edu](mailto:kyle_nguyen@berkeley.edu)

Anthony Suen; BIDS collaborative. [anthonysuen@berkeley.edu](mailto:anthonysuen@berkeley.edu)

## Research Goals and Deliverables

The goal of the Digital Assyriology project is to demonstrate the feasibility and usefulness of computational methods for research in cuneiform studies. The deliverables consist of Jupyter Notebooks that give a step by step demonstration of data import, data manipulation, data analysis, and data visualization for a small number of selected problems. Preliminary and final results will be presented at a Brown Bag in Near Eastern Studies (October 19; 12PM; provisional date); at the BIDS presentation December 7th and at the American Oriental Society conference in LA (March 17-20, 2017). Finally, an article with the main results and references to the Github repository will be submitted to the Journal of Cuneiform Studies ([JCS](http://web.a.ebscohost.com/ehost/command/detail?sid=70415972-15a2-409c-a114-500caeb860a1%40sessionmgr4010&vid=0&hid=4101&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#jid=D4U&db=hlh)) Spring 2017.[[1]](#footnote-1)

## Background and Expectations

Although digital publication has become common in cuneiform studies, computational approaches are still rare and watched with some suspicion. It is important, therefore, 1) to make the coding as transparent as possible (better to add an extra line of code than to make code that is obscure) and 2) to explain, as much as possible, decisions that are made in choice of libraries, settings of parameters, including or excluding (sub)sets of data, etc. In other words: making well-considered choices and documenting every step are going to be key.

## Team meetings

The team will meet every week on Monday or Friday to discuss progress and problems and/or to work on particular issues.

## Github Repository

The Github repository is <http://github.com/niekveldhuis/Digital-Assyriology>. This repository currently contains initial work on the project.

## Data Sets

There are two main data sets: 1. Sumerian literary texts, scraped from [ETCSL](http://etcsl.orinst.ox.ac.uk/) and 2. data to be scraped from [ORACC](oracc.org). ETCSL is an archival site; the scraped data are available in the Github repository in the directory Scrape-etcsl/Cleaned. ORACC is an umbrella of a variety of active projects. Data can be downloaded and scraped with the Notebooks in the directory Scrape-Oracc. Clean data sets are provided by the team leader, but it is recommended to familiarize yourself with the original data on ETCSL and ORACC and to understand the process of scraping and data formatting in the scrapers.

## Subproject 1: Data Clustering

The Sumerian literary texts published in ETCSL are subdivided into several categories: catalogs, narrative texts, hymns, lamentations, debates (these are mostly mock-debates, such as the debate between Summer and Winter), proverbs, etc. There are also other ways of sub-dividing Sumerian literary texts; for instance: some were used extensively in scribal education; others were not; some exist in many duplicates, others are known in only one exemplar. Is it possible to detect meaningful distinctions between these literary compositions in their vocabulary?

This project has three main stages:

1. clean up the hierarchical clustering/dendrogram Notebook and make well-considered decisions on parameters etc. Dead line Oct 19th (Brown Bag, NES).

2. Cluster the texts by k-means; identify an appropriate and attractive way to visualize the clusters (some inspiration at <http://brandonrose.org/clustering>). Dead line Dec. 7th (BIDS presentation)

3. Cluster texts by their overlap (in vocabulary) with lexical texts (school texts) available in ORACC.

## Subproject 2: Topic Modeling Neo-Assyrian letters

Several thousand letters to Neo-Assyrian kings available in ORACC. These letters come from different parts of the empire and are from different periods of Neo-Assyrian history. What topics are found in these letters, how do these topics change over time, which topics are characteristic for a specific (geographical) part of the empire?

This subproject has two main stages:

1. identify the best topic-modeling library; make well-considered decisions on parameters etc. and create a first topic model. Visualize in pyLDAvis or a similar library. Dead line: Oct 19th; NES Brown Bag.

2. relate the topic model to metadata (geography, date, sender, etc.) and visualize these dependencies. Dead line: Dec. 7th (BIDS presentation).

## Subproject 3: Clean up, Improve, and Document the Scrapers

The ETCSL and ORACC scrapers provide the foundation for Digital Assyriology. The scrapers work, but may not be the most user-friendly and efficient way of doing it. A thorough review of each step, perhaps restructuring of the code, are in order. In addition, an interactive output formatter that allows a user to make choice between data elements to be scraped and how to format the output is necessary.

This subproject has two main stages:

1. create the interactive output formatter on the basis of the current functionality of the ORACC-scraper. Dead line: Oct. 17 (NES brown bag)

2. thorough review of the scraper, documenting every step for users and/or future developers. Dead line: Dec. 7th (BIDS presentation).

## Subproject 4: Explore Word2vec

Word2vec is a relatively new natural language processing technique that may be used to reconstruct the linguistic context of words (words that share common contexts are located close to each other). The project is exploratory: what are the applications of Word2vec, how can it be applied to cuneiform data.

This subproject has two main stages:

1. explore Word2vec algorithms, what are they used for, how are they implemented? Dead line Oct 19th (NES brown bag)

2. Apply Word2vec to an appropriate text corpus. Dead line: Dec 7th.

1. JCS is a print journal that appears once a year. It takes at least a year (but often longer) for an article to appear. [↑](#footnote-ref-1)