FactGrid Cuneiform Discovery Project: Building Linked Open Data Repositories

- **FactGrid AWCA Google Drive & Google Colab**
- **FactGrid Cuneiform AWCA GitHub Org. Repos**

"Data is a precious thing and will last longer than the systems themselves."

Tim Berners-Lee, Inventor of the World Wide Web

This project is inspired by the durability of the data preserved in the oldest writing system known to mankind, called **cuneiform**. There are approximately a half-million artifacts with cuneiform writing spread all over the planet. Many of these objects are not even photographed, let alone translated. Scholars in this field have made a number of relational text databases, in order to identify these objects housed in museums and private collections, and while these databases have helped create a system of identificaiton and textual analysis, they have yet to be **linked** together to each other and to the existing scholarship.

FactGrid is a Wikibase triplestore designed for historical research, which makes this the ideal hub for linking the existing scholarship, both primary and secondary sources, for every cuneiform artifact in publication. As a graph database (using RDF triples), this system allows for an expansive number of relationships to exist between any objects or entities, and for these relationships to be classified for greater discoverability. Additionally, the structure of a graph database provides simple solutions to duplication, disambiguation, and alternative editions of the same entity, which allows for a more comprehensive approach for linking the different datasets in various languages. Because the database is in Wiki(data) format, this provides an open forum for editing and collaborating internationally, which makes the data more robust, timely, and sustainable.

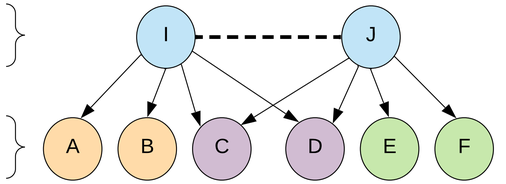
BERT & D2V

Entities I and J can be related based on shared **references**.

docs A, B, C, & D have similarities based on shared **contexts**.

Concepts of Linked Data and bibliographic coupling

Entities I and J are items with URIs, for example in Wikidata. Such items can be attested bibliographically and cited by documents C and D. The documents which cite these entities can be bibliographically coupled and their relational values can be measured by a count-frequency and cosine similarity scores.



Topic Model (LDA)

Entities I and J are assigned the same topic number. Topic weight determines directedness of edges.

Sources are more representative of a topic or genre, based on TF-IDF + LDA

The goal for this project is to link every cuneiform artifact on record to the primary and secondary sources, which make explicit reference to the given object, and to extend this referencial system to include the entities named on each artifact
(i.e. people, places, and things)

PDF (OCR)

html (no OCR)

Catalog for Drive (CSV), fields:
- (google) 'id'
- (file) 'name'
- 'mimeType'
- 'md5Checksum'
- 'webViewLink'
- 'createdTime'
- 'modifiedTime'
- 'size'

Page Splitting Orientation

Google ID integration

Batch processing

Language detection

OCR
tesseract
- raw TXT
- Google ID
- Doc ID
- Page ID
- Page count
- File name
- Catalog info

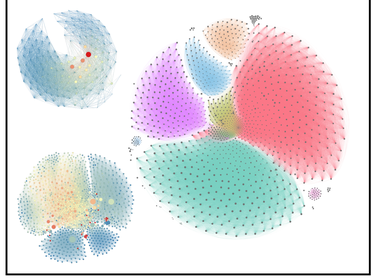
ocr-output (CSV), fields:
- Google ID
- File name
- Page ID
- Page Count
- Catalog meta
- Language(s)
  - Scripts

Trans-lations
- ocr-output
- segment
- google translate

NER
- PNs
- RNs
- GNs
- Bib

NLP networks
- TopicModel
- D2V
- W2V
- BERT+
- GPT+

Linked Data



Discovery Students

Aidan Curran	2022
Giselle Fuchs	2022
Minoo Kim	2022
Qianlin Wang	2022
Win Moe	2022
Ziyue Wang	2022
Circle Chen	2021-22
Conner Mi	2021-22
Daisy Wang	2021-22
Floyd Fang	2021-22
Kevin Gao	2021-22
Tina Chen	2021-22
Zaid Maayah	2021-22

Our Cuneiform project in FactGrid is building language support for all languages written in cuneiform, a writing system used for about 4000 years (from 3200 BCE to 50 CE). We are working with more than 350k documents to build dictionaries for these languages and social network graphs for the people, places, other entities named in these texts.

The main challenge we're working on is how to build reproducible workflows for linking four online databases of cuneiform sources (each with more than 100k documents) with two datasets of secondary sources. We're using Python notebooks (ipynb) to harmonize these open source databases, and we are linking the results using FactGrid Cuneiform, which is a triple store (or database for RDF triple statements). Our work helps us deepen our knowledge of Python for NLP and SparQL, the query language for Wikidata.

cdli

Oracc

ETCSL

BDTNS

KeiBi

AfO

URL + URI
Linking
- CDLI
  - catalog
  - text 350k
- ORACC
  - catalog
  - text 200k?
- BDTNS 150k
- ETCSL
- KeiBi
- AfO Register

PNs
names
- NER match
- ocr-output
- CDLI [ATF]
- ORACC [PN]
- edge list
- Wikidata
- FactGrid
- Import (CSV)

RNs
rulers
- NER match
- ocr-output
- CDLI [dates]
- ORACC [RN]
- Period
- Dynasty
- Wikidata
- FactGrid
- Import (CSV)

GNs
places
- NER match
- ocr-output
- CDLI [provenience]
- ORACC [GN]
- Wikidata
- FactGrid
- Import (CSV)

Bib
citations
- ocr-output
- Author
- Title
- Date
- **Pub**
- **Abb.**
- **oclc**
- **isbn**
- **doi**

zenodo

GitHub

WIKIDATA

FactGrid a database for historians

QuickStatements (CSV) import to Wikidata & FactGrid

SparQL Query for exporting whole datasets

End-to-end pipeline completion for participating databases

FactGrid:Cuneiform Project
Factgrid

FactGrid Cuneiform & Wikidata Lexemes

WIKIDATA LEXEMES
Wikidata:Lexicographical data
Wikidata

cuneiform clay tablet
Wikidata

personal name
Wikidata

historical period
Wikidata

ancient city
Wikidata

toponym
Wikidata

dynasty
Wikidata

archaeological site
Wikidata

Abbreviations for Assyriology
Ox

WIKIDATA BOOKS
Wikidata:WikiProject Books
Wikidata