Digital Humanities for Classical Arabic
Applications for historians and philologists

Alicia González Martínez

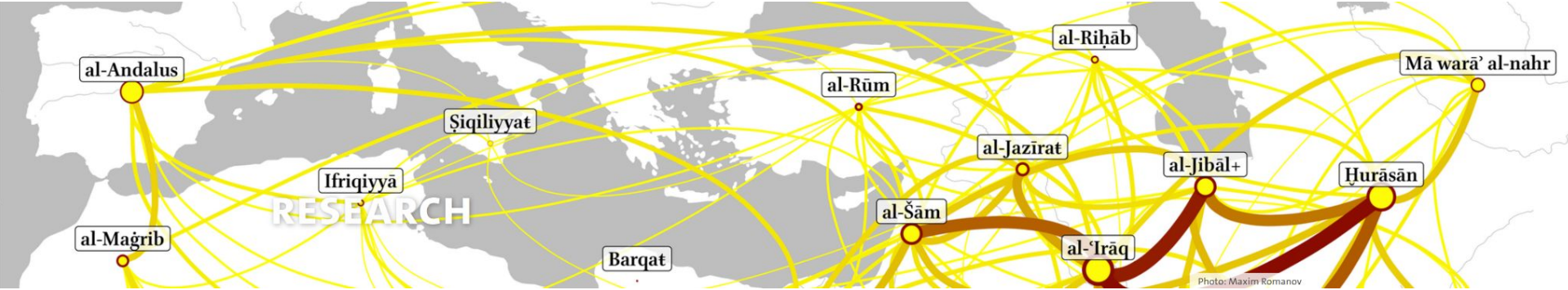
The Evolution of Islamic Societies
(c. 600-1600 CE)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

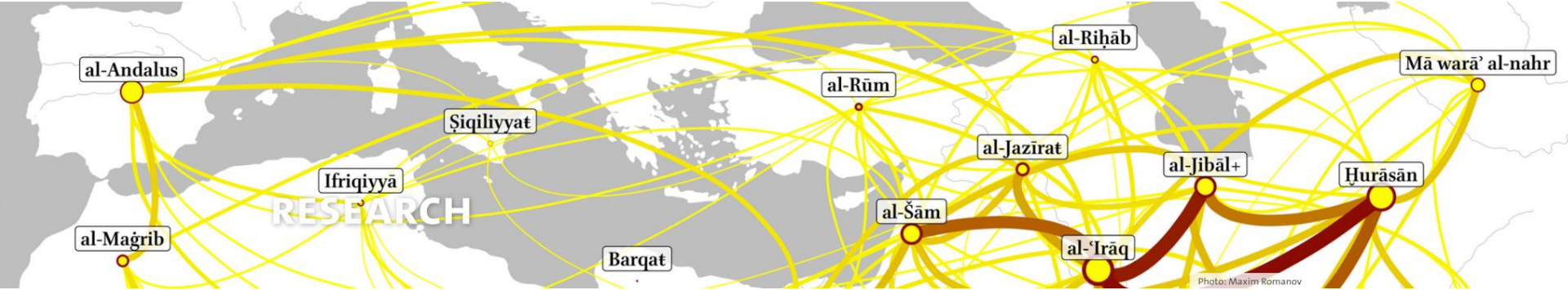
EIS1600

Universidad Autónoma de Madrid

The Evolution of Islamic Societies (c. 600-1600 CE)



The Evolution of Islamic Societies (c. 600-1600 CE)



100 million words +400,000 biographical records

What is digital humanities and how can it help us?

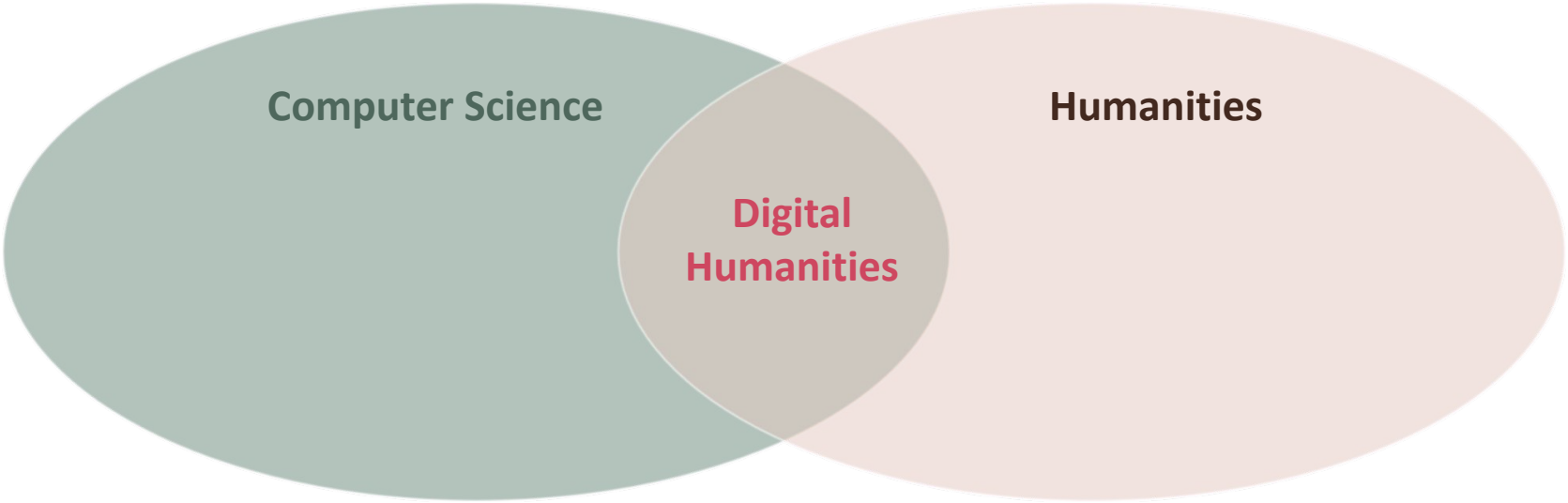
What do we need for applying digital humanities?

Resources and tools for Classical Arabic

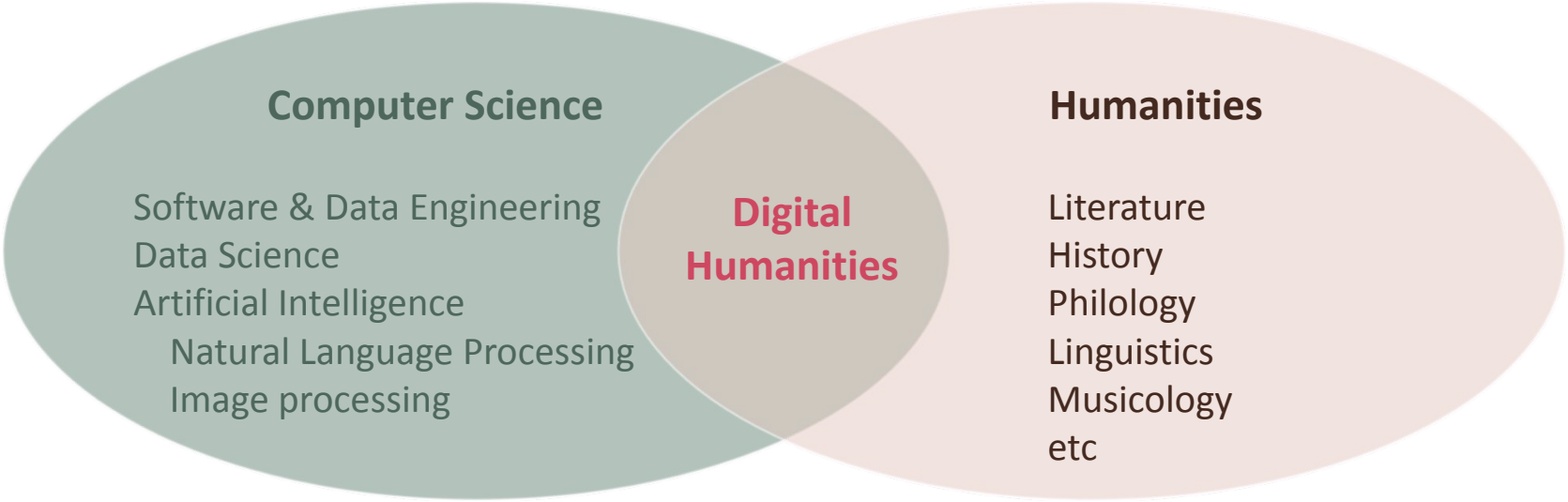
What is digital humanities and how can it help us?

Digital Humanities is a discipline that applies computational methods to study any field in the humanities

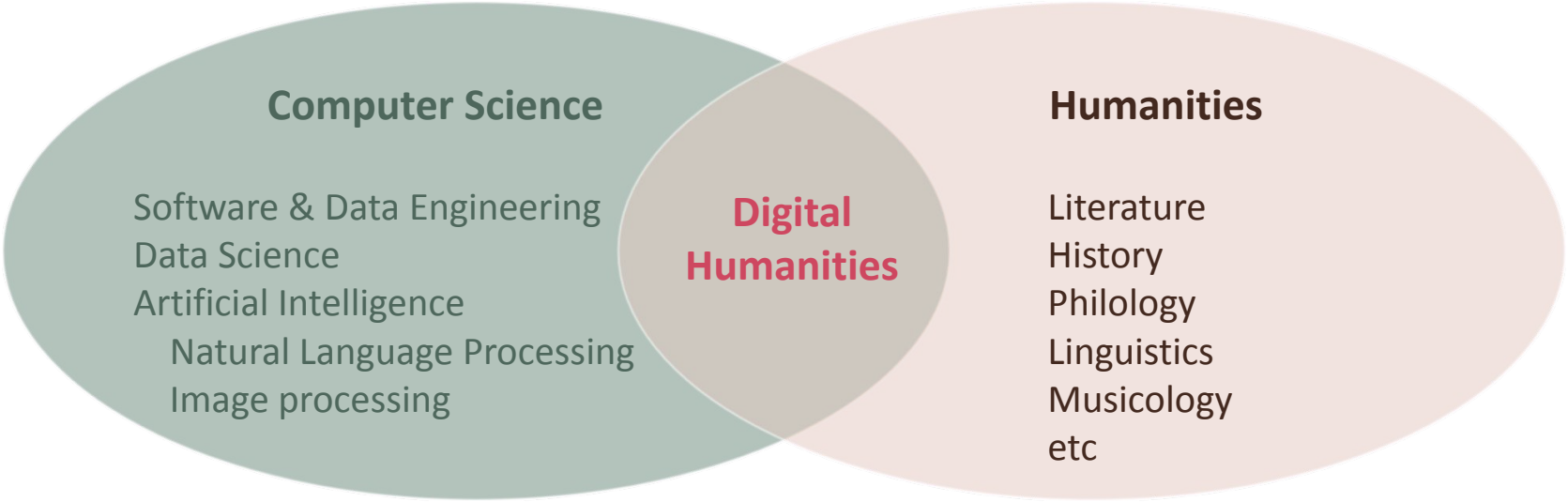
Digital Humanities is a discipline that applies computational methods to study any field in the humanities



Digital Humanities is a discipline that applies computational methods to study any field in the humanities

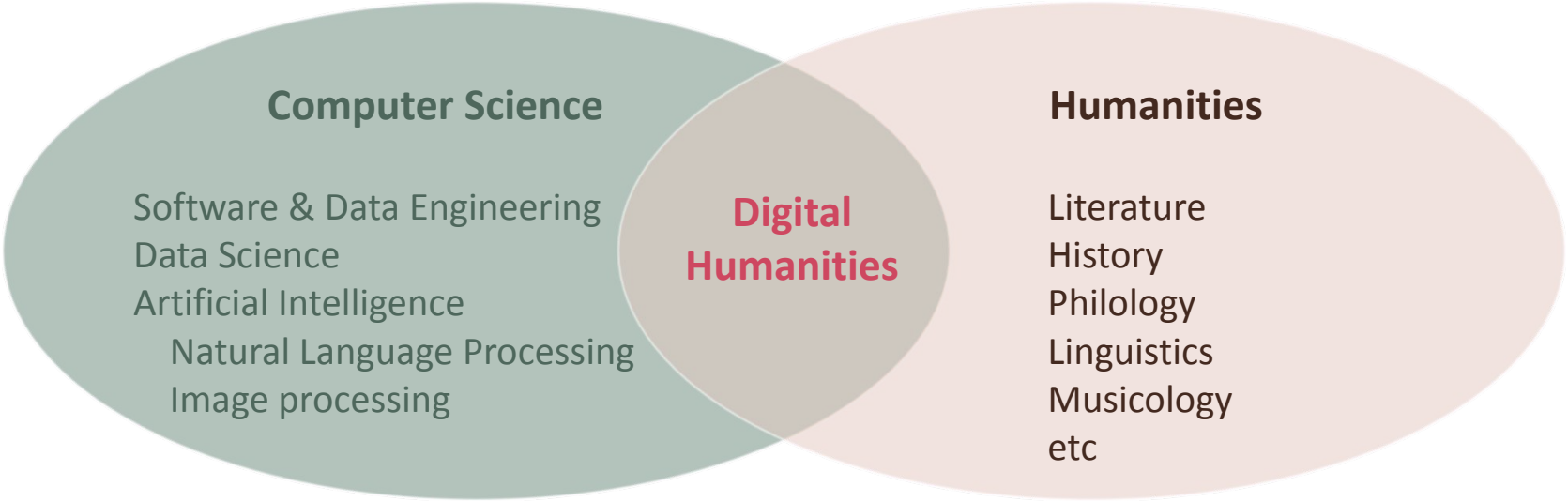


Digital Humanities is a discipline that applies computational methods to study any field in the humanities



digital resources *e.g. database of textual heritage*

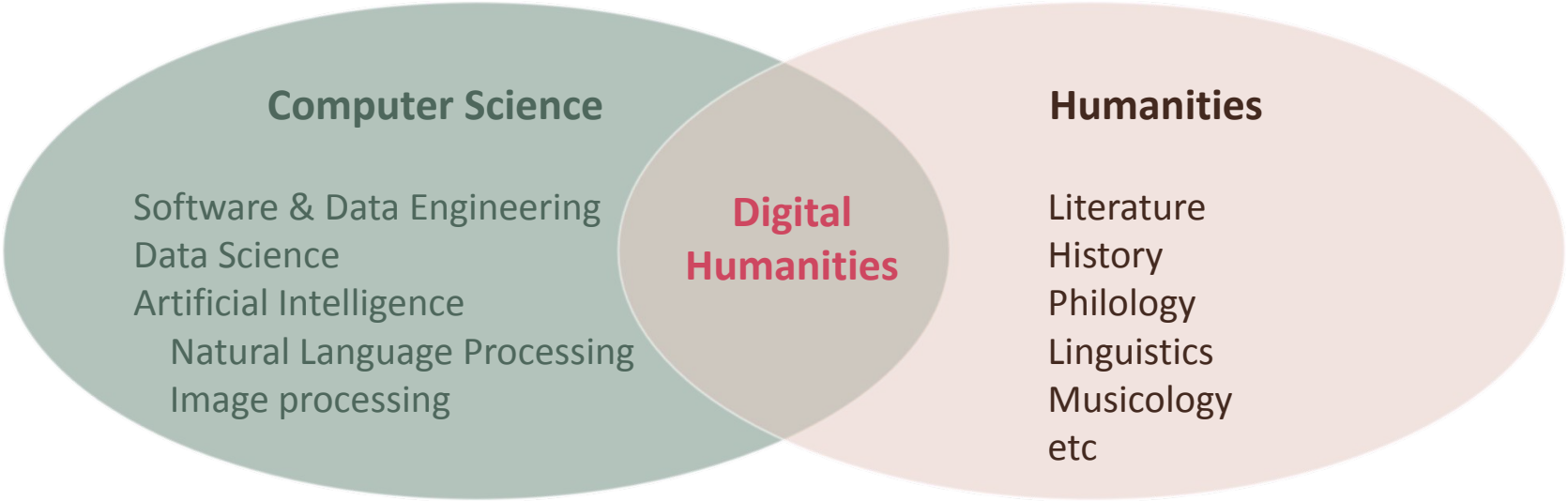
Digital Humanities is a discipline that applies computational methods to study any field in the humanities



digital resources *e.g. database of textual heritage*

digital tools *e.g. query interface*

Digital Humanities is a discipline that applies computational methods to study any field in the humanities

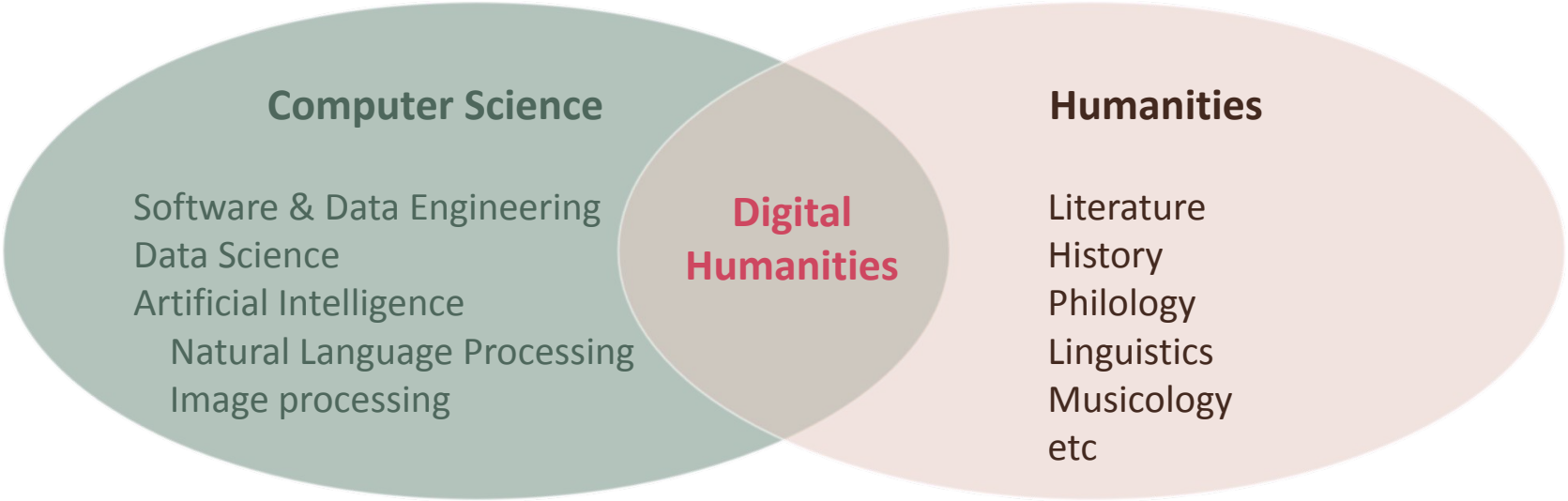


digital resources *e.g. database of textual heritage*

digital tools *e.g. query interface*

digital methods *e.g. topic modelling*

Digital Humanities is a discipline that applies computational methods to study any field in the humanities

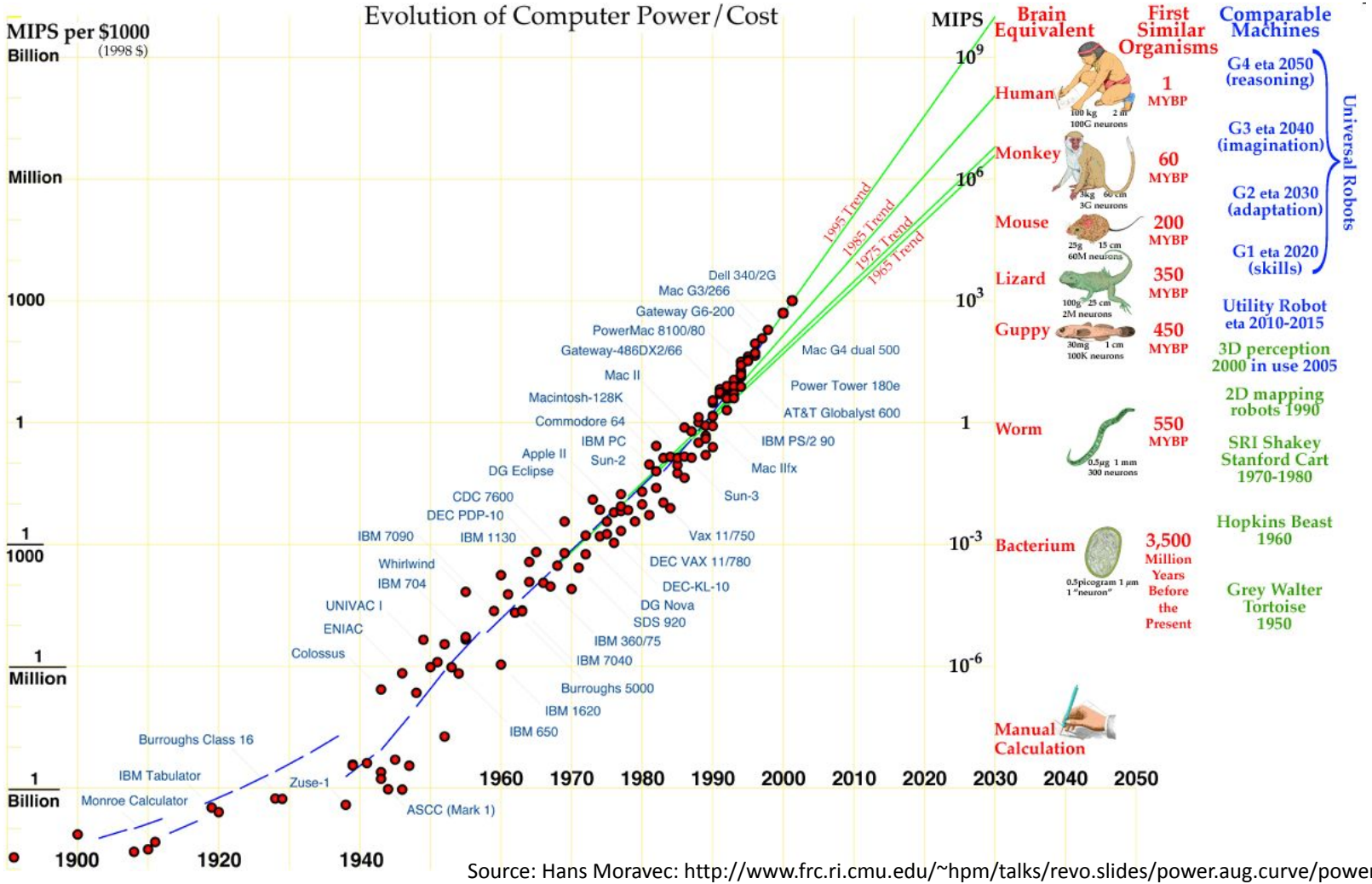


digital resources *e.g. database of textual heritage*

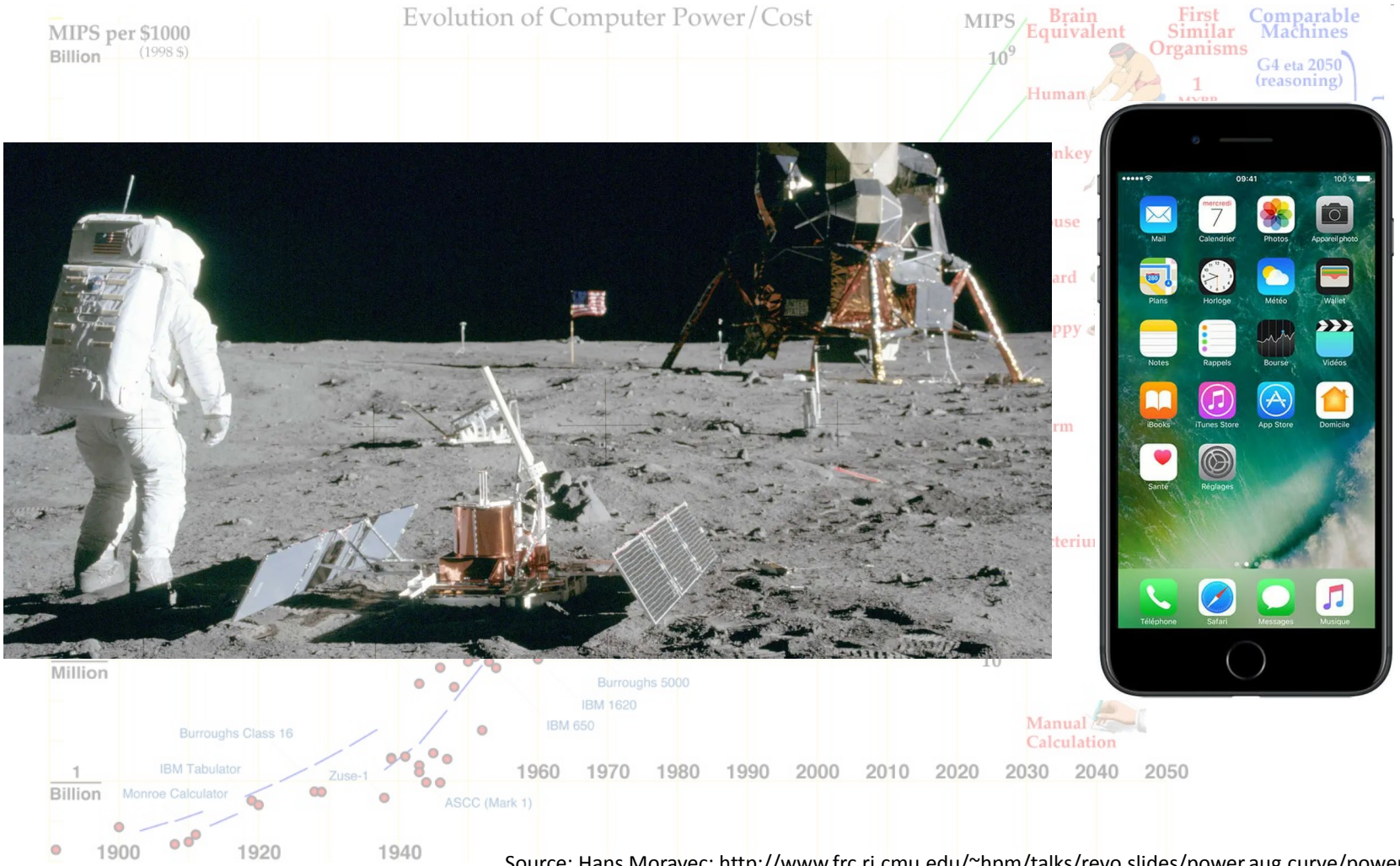
digital tools *e.g. query interface*

digital methods *e.g. topic modelling*

Digital Humanities methods does NOT replace traditional methods, but support them



Source: Hans Moravec: http://www.frc.ri.cmu.edu/~hpm/talks/revo.slides/power.aug.curve/power.aug.gif



Source: Hans Moravec: http://www.frc.ri.cmu.edu/~hpm/talks/revo.slides/power.aug.curve/power.aug.gif

Speed up traditional methods

Speed up traditional methods

find a term

Speed up traditional methods

find a term

مُحَمَّدٌ

Mehmet

Mahoma

muḥammad

Speed up traditional methods

find a term

مُحَمَّدٌ

Mehmet

Mahoma

muḥammad

find a similar text

manually feasible?

Speed up traditional methods

find a term

مُحَمَّدٌ

Mehmet

Mahoma

muḥammad

find a similar text

manually feasible?

cosine similarity measure

95% certainty

Speed up traditional methods

find a term

محمد

Mehmet

Mahoma

muḥammad

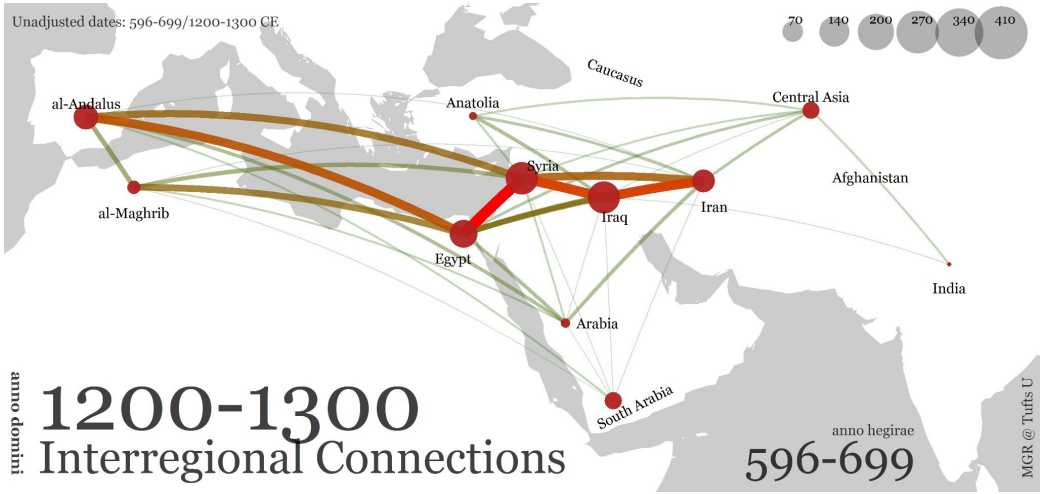
find a similar text

manually feasible?

cosine similarity measure

95% certainty

Unforeseen outcomes



Massive migrations of the 13th century CE. Source: Maxim Romanov, https://maximromanov.github.io/2017/10-14.html

What do we need for applying digital humanities?

Reader Bot

Q&A to Ibn Hazm's «Treatise on the Hierarchy of Sciences» ابن حزم رسالة في مراتب علوم

Ask

ما رأي ابن حزم في المنطق؟

ابن حزم يعتبر المنطق واحدًا من العلوم الهامة التي يجب على الباحث في الحقائق أن يدرسها. يشدد على أهمية فهم حدود المنطق وعلم الأجناس والأنواع والأسماء المفردة والقضايا والمقدمات والقرائن والنتائج. يعتبر أن هذا العلم يساعد الشخص على التفريق بين الحقائق والأباطيل ويمكنه من تجنب الشغب والتلاعب. بالإضافة إلى ذلك، يشجع على دراسة الطبيعة وعوارض الجو وتركيب العناصر والحيوانات والنباتات والمعادن. يعتبر أن دراسة هذه العلوم تساعد الشخص على فهم صنعة الخالق وتأثيره وتأليف الأعضاء واختيار المدبر وحكمته وقدرته.

- Information Retrieval

  Digitisation and text encoding

- Natural Language Processing & Deep Learning
- Data Engineering & Database Management
- Backend development & Web development
- Server Administration

- Information Retrieval
- Digitisation and text encoding
- Natural Language Processing & Deep Learning
- Data Engineering & Database Management
- Backend development & Web development
- Server Administration
- Image Processing
- Machine Learning and Data Science

Resources and tools for Classical Arabic

مختبر كامل

CAMeL Lab

The Computational Approaches to Modeling Language (CAMeL) Lab develops applications for artificial intelligence tasks, specifically focusing on natural language processing, computational linguistics, and data science.

مختبر كامل

CAMeL Lab

The Computational Approaches to Modeling Language (CAMeL) Lab develops applications for artificial intelligence tasks, specifically focusing on natural language processing, computational linguistics, and data science.

Arabic Natural Language Processing **Machine Translation** **Text Analytics** **Dialogue Systems**

مختبر كامل

CAMeL Lab

The Computational Approaches to Modeling Language (CAMeL) Lab develops applications for artificial intelligence tasks, specifically focusing on natural language processing, computational linguistics, and data science.

Arabic Natural Language Processing **Machine Translation** **Text Analytics** **Dialogue Systems**

أدوات كامل

CAMeL Tools

Camelira: An Arabic Multi-Dialect Morphological Disambiguator

About

Camelira is a web-based Arabic multi-dialect morphological disambiguation tool that covers four major variants of Arabic: Modern Standard Arabic, Egyptian, Gulf, and Levantine. Camelira is supported by the Python open-source toolkit Camel Tools. For more, see Obeid et al. (2022).

الدولة السعودية الأولى هي دولةٌ قامت في شبه الجزيرة العربية

Copy URL | Submit | Auto-detect

Analyzing as Egyptian (auto-detected)

Diacritized/POS | Tokenized | Lemmatized

الدَّولَة	السَّعُودِيَّة	الأُولَى	هِيَّ	دَولَة	قامَت	فِي	شِبْه	الجَزِيرَة	العَرَبِيَّة
Noun	Adjective	Numerical Adjective	Pronoun	Noun	Verb	Preposition	Noun	Proper Noun	Adjective

الدَّولَة

POS: Noun
Gender: Feminine
Number: Singular
State: Definite
Proclitic 0: Determiner
Gloss: state; country; states; countries

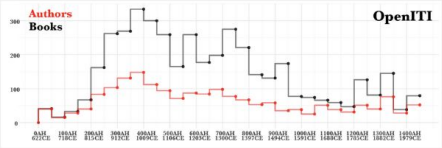
1.000	–	الدَّولَة	[دَولَة]	noun	ال / DET + دَوْل / NOUN + ةَ/ NSUFF_FEM_SG	state;country;states;countries
1.000	–	الدَّوْلَة	[دَوْلَة]	noun	ال / DET + دَوْل / NOUN + ةَ/ NSUFF_FEM_SG	state;country;states;countries
1.000	–	الدَّوْلَة	[دَوْلَة]	noun	ال / DET + دَوْل / NOUN + ةَ/ NSUFF_FEM_SG	state;country_[SAMA]
1.000	–	الدَّوْلَة	[دَوْلَة]	noun	ال / DET + دَوْل / NOUN + ةَ/ NSUFF_FEM_SG	state;country_[SAMA]
0.937	–	الدَّوْلَةِ	[دَوْلَة]	noun	ال / DET + دَوْل / NOUN + ةَ/ NSUFF_FEM_SG + ِ/ CASE_DEF_GEN	state;coun
0.937	–	الدَّوْلَةُ	[دَوْلَة]	noun	ال / DET + دَوْل / NOUN + ةَ/ NSUFF_FEM_SG + ُ/ CASE_DEF_NOM	state;cour

https://camelira.abudhabi.nyu.edu

Open Islamicate Texts Initiative

OpenITI Corpus

by OpenITI December 18, 2023



The OpenITI corpus is a open-access and machine-actionable collection of Persian and Arabic texts.

The large number of texts in our collection are in various stages of being transformed into standards-compliant and metadata-enriched scholarly corpus texts.

Most OpenITI corpus texts are built upon digital texts we obtained from Shamilah, *al-Jāmiʿ al-Kabīr*, Maktabat al-Shiʿa, Ganjoor, and other online collections that have varying levels of fidelity to the original print versions (due to manual or automatic—i.e., OCR—transcription errors). In the coming years, however, we will be dramatically scaling up the number of new digital texts that we add to the corpus through our own OCR process developed in the OpenITI AOCP project.

https://openiti.org/projects/OpenITI%20Corpus.html

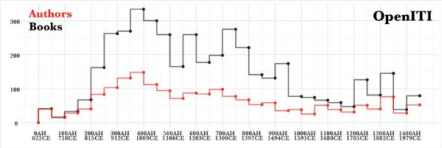
Open Islamicate Texts Initiative

OpenITI Corpus

2.2 billion words!

by OpenITI December 18, 2023

The OpenITI corpus is a open-access and machine-actionable collection of Persian and Arabic texts.



The large number of texts in our collection are in various stages of being transformed into standards-compliant and metadata-enriched scholarly corpus texts.

Most OpenITI corpus texts are built upon digital texts we obtained from Shamilah, *al-Jāmiʿ al-Kabīr*, Maktabat al-Shiʿa, Ganjoor, and other online collections that have varying levels of fidelity to the original print versions (due to manual or automatic—i.e., OCR—transcription errors). In the coming years, however, we will be dramatically scaling up the number of new digital texts that we add to the corpus through our own OCR process developed in the OpenITI AOCP project.

https://openiti.org/projects/OpenITI%20Corpus.html

OpenITI_NgramReader

OpenITI_NgramReader

The **NgramReader** charts chronological frequencies of words and phrases, using the data of the OpenITI corpus (8,462 texts; 1,048,219,219 tokens). It allows one to combine different morphological forms as well as to explore classes of objects. By default, the search is performed on ~95% of this data, since some of the texts are still not cleaned of their paraeditorial elements (407 texts, 60 mln tokens); you can choose to run the search on the complete data (option `keep all texts (100%)`).

Filename Prefix (optional)

transmissionTerms_

Ngram Group 1

#ḥaddaṯa-nā #[wf]?Hdvn[Ay]

Ngram Group 2

#aḫbara-nā #[wf]?Axbrn[Ay]

Ngram Group 3

#anbaʿa-nā #[wf]?AnbAn[Ay]

Show Results | All Data for Results as RDS

Ngram types:

◉ unigrams ○ bigrams ○ trigrams

Corpus selection:

○ keep all texts (100%) ◉ remove problematic (~95%)

Chronological Ngrams Smoothing

0.1 0.2 0.3

0.1 0.12 0.14 0.16 0.18 0.2 0.22 0.24 0.26 0.28 0.3

The default value of 0.2 appears to be a good option for chronological ngrams; you can adjust this value within the provided window. Smoothing helps to reveal underlying trends and patterns in data by reducing noise and providing a more accurate and visually clear representation. However, it also transforms the original data. The smallest value (0.1) will provide the graph without any alteration of the initial data.

General Instructions

Like *Google Ngram Viewer* (https://books.google.com/ngrams), **[OpenITI] NgramReader** charts diachronic frequencies of words and phrases, using the data of the OpenITI corpus. Unlike `Google Ngram Viewer` , however, it allows one to combine different morphological forms of the same lexical items together as well as to explore classes of objects. Why to combine forms? Arabic morphology is complex and the same word can appear in a large variety of forms: for example, *kitāb*, *al-kitāb*, *wa-kitāb*, *wa-l-kitāb* are instances of the same lemma and one might want to combine all or only some forms into a single entity. This approach also allows one to create thematic clusters of words (or, classes). For example, one can combine *Baġdād* and *Madīnat al-salām* in order to get all mentions of the ʿAbbāsid capital; or, to combine together all cities of Ḫurāsān in order to gauge frequencies of references to Ḫurāsān in general.

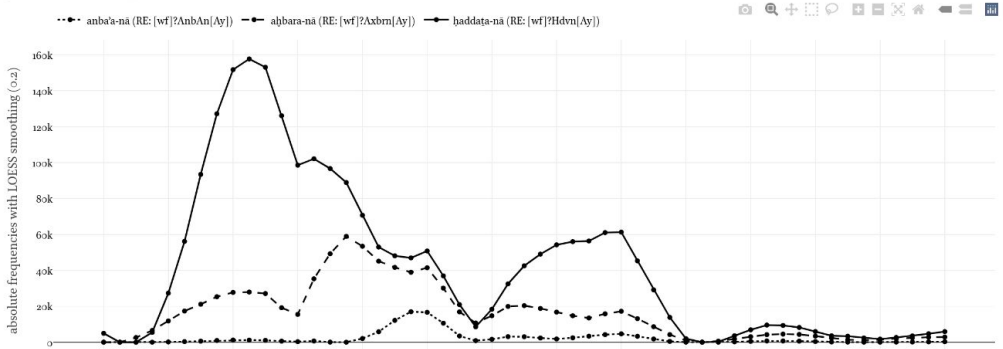
SEARCHES Syntax for the searches is as follows: `#ItemForTheLegend #FirstSearchItem #SecondSearchItem #NthSearchItem` , that is each item must begin with `#` . `#ItemForTheLegend` is not a search item but a string that you want to show on the legend of the graph. Thus, if you were to look for mentions of Baġdād, your search line would look something like this `#Baġdād #bgdAd #bbgdAd #wbgdAd #wbbgdAd` . Try searching for these tokens in one line and in separate lines to see the difference. *Regular expressions*. The search line also supports `regular expressions` which make things simpler and more robust. For example, `#Baġdād #bgdAd #bbgdAd #wbgdAd #wbbgdAd` can be also written more concisely as `#Baġdād #[wb]?bgdAd` . Simplified Buckwalter transliteration is used in the NgramReader: `ء = c` , `ا = A` , `إ = A` , `أ = A` , `آ = A` , `ب = b` , `ة = o` , `ت = t` , `ث = v` , `ج = j` , `ح = H` , `خ = x` , `د = d` , `ذ = V` , `ر = r` , `ز = z` , `س = s` , `ش = E` , `ص = S` , `ض = D` , `ط = T` , `ظ = Z` , `ع = C` , `غ = g` , `ف = f` , `ق = q` , `ك = k` , `ل = l` , `م = m` , `ن = n` , `ه = h` , `ؤ = c` , `و = w` , `ى = y` , `ئ = c` , `ي = y` .

FILENAME PREFIX. You can use this option to automatically assign a specific prefix to the results that you may want to download. You can download data for the main search results, graphs as well as data for all the summaries that are generated for each search.

TYPES OF NGRAMS. You can search for unigrams (`1`), bigrams (`2`), or trigrams (`3`). Make sure to select appropriate **ngram Type**. By default, unigrams are activated. If you search for bigrams or trigram, use `underscores` "_" instead of spaces, i.e. *kataba ilay-hi* should be transliterated as `ktb_Alyh` . Note that you can only search one type of ngrams at a time. In most cases, it does not make sense to combine ngrams of different length in the same search, since frequencies of unigrams are usually significantly higher than those of bigrams, and the frequencies of bigrams usually significantly higher than those of trigrams.

Graphs of Relative and Absolute Frequencies of Ngrams

Download Data



https://eis1600.aai.uni-hamburg.de/shiny/OpenITI_NgramReader/

al-Ṯurayyā Project

This is a new working version of the al-Ṯurayyā project which currently includes the gazetteer (*al-Ṯurayyā Gazetteer*, or *al-Thurayyā Gazetteer*), and the geospatial model of the early Islamic world. Both parts of the project are still under development.

Gazetteer: The gazetteer currently includes over 2,000 toponyms and almost as many route sections georeferenced from Georgette Cornu's *Atlas du monde arabo-islamique à l'époque classique: IXe-Xe siècles* (Leiden: Brill, 1983). Tabs relevant to the gazetteer are as follows:

? is the current tab with the general information about the project;

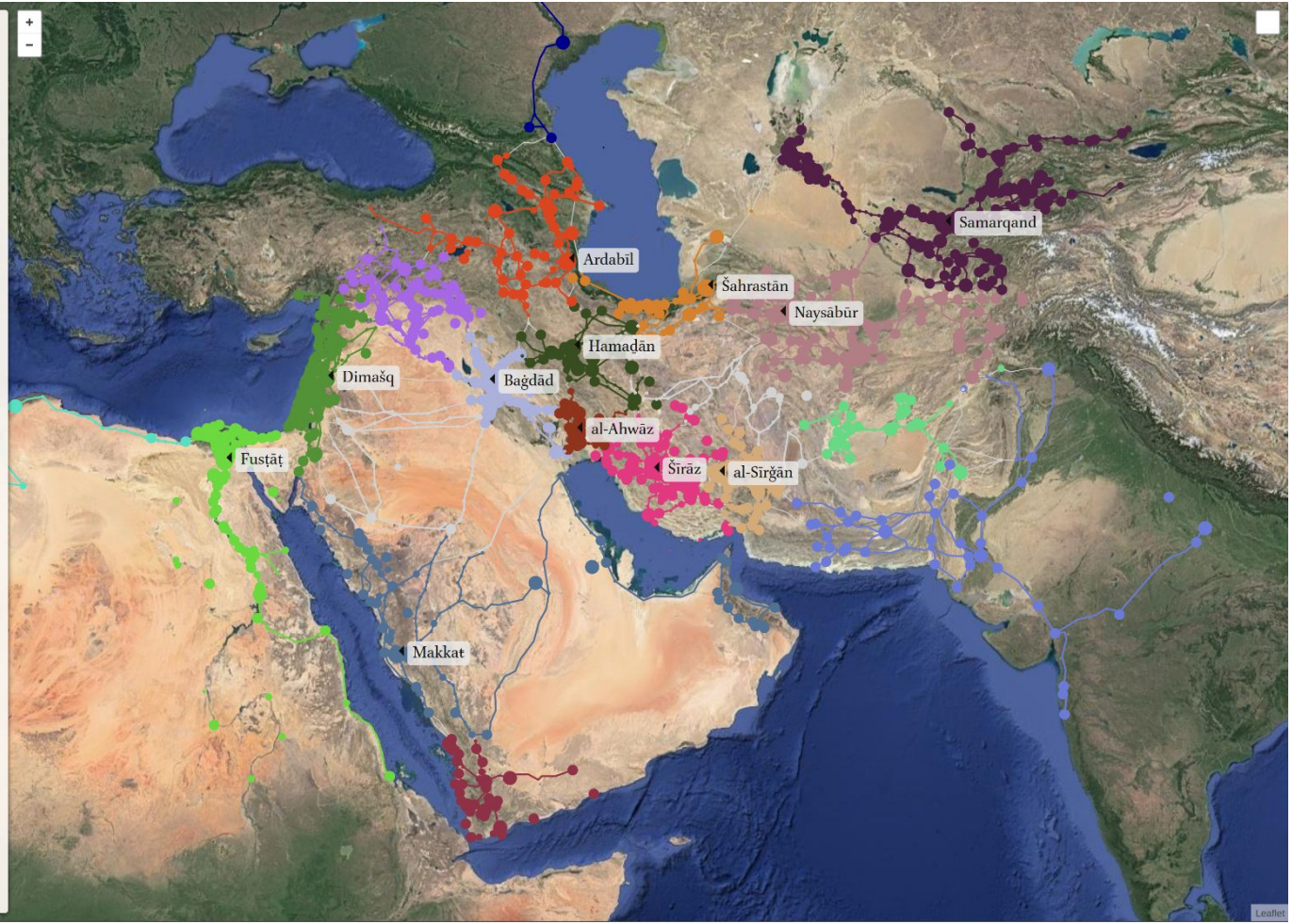
is the 'Technical information' of a selected toponym (URI, coord_certainty, language, names, region_URI, source, top_type), which is used for placing it on the map;

is the description(s) of a selected toponym from Arabic sources (at the moment, only al-Ḥimyarī's *Rawḍ al-miʿṭār*). Records from primary sources are matched automatically, with the % of the match shown in parenthesis.

Search panel. Since the gazetteer currently does not include English versions of placenames, one must search for Arabic names: for example, Dimashq instead of Damascus. One can use Arabic or simplified transliteration (LOC transliteration scheme).

The *previous version of the gazetteer* can be found here. You can browse this version by clicking on any toponym marker. The popup will show the toponym both in Arabic script and transliterated (On transliteration scheme, see below). The popup also offers a selection of possible sources on a toponym in question. You can check *Arabic Sources*: currently, al-Samʿānī's *Kitāb al-ansāb* and Yāqūt's *Muʿjam al-buldān*. The Gazetteer shows only exact matches, which means that in some cases there will not be any entry at all, while in other cases there may be more than one and they may refer to other places with the same name. You can also check if there is information on a toponym in question in Brill's *Encyclopaedia of Islam*, *Pleiades*, and *Wikipedia*.

Geospatial model currently consists of a two main modules (*work in progress*) which plot 1) routes and itineraries of various complexity; and 2) networks of reachable places from



https://althurayya.github.io

Gracias