

Maxim Romanov

New York University — Abu Dhabi, April 10-12, 2017

Algorithmic Analysis of Premodern Arabic Biographical Collections

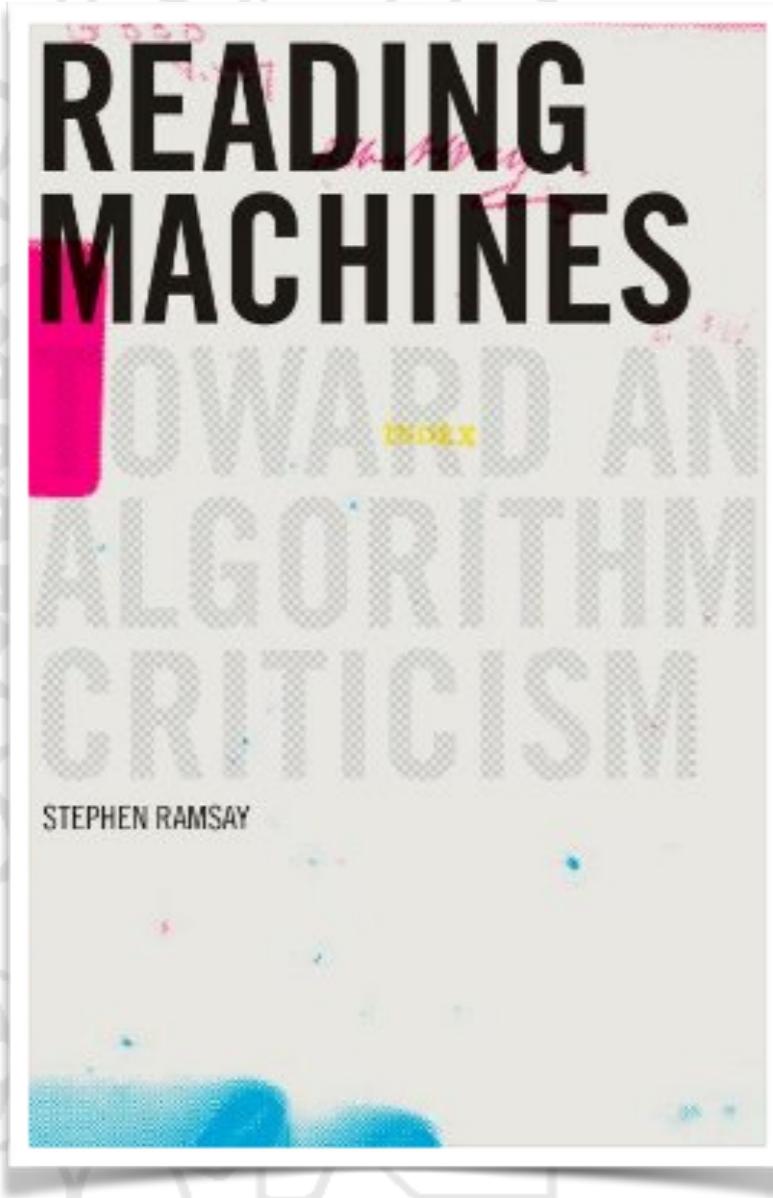
*Approach, Infrastructure,
Open Data*



Digital Humanities
UNIVERSITÄT LEIPZIG

Conceptualization

Approach: *Inspiration*



“Algorithmic criticism is easily conceived as the form of engagement that results when imperative routines are inserted into the wider constellation of texts stipulated by critical reading. But it is also to be understood as the creation of interactive programs in which readers are forced to contend not only with deformed texts, *but with the ‘how’ of those deformations.*”

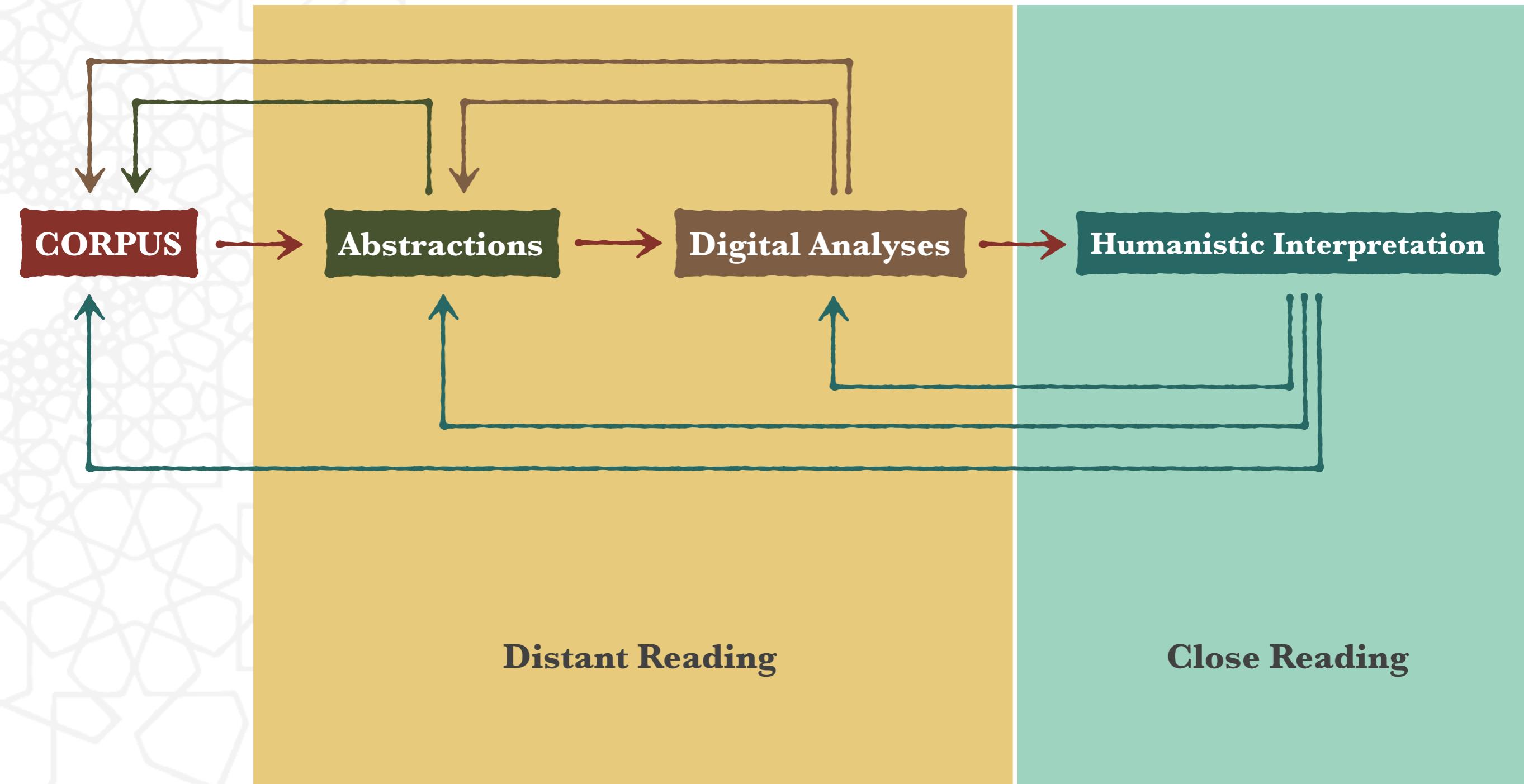
Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. 1st Edition. University of Illinois Press, 2011.

Approach: *Conceptualization*



Algorithmic analysis is a step-by-step reduction of a text in a natural language to a machine-readable abstraction which is then analyzed to discover shapes, relations and structures. This is an iterative process of engagement with texts, their abstractions, and their interpretations, where preliminary results of later steps of the loop can suggest one how to improve earlier steps to attain better results.

Approach: *Visual Conceptualization*



Abstracting: *Step by Step*

1. finding the machine-readable text of a book;
2. tagging the logical structure of the book;
3. tagging—manually and semi-automatically—relevant data in the structured text (alternatively, extracting relevant data automatically);
4. extracting and modeling tagged data;
5. visualizing and analyzing results.

Abstracting: *Structures*

• الهروي \$ #	237
• أبو سعيد إبراهيم بن طهمان بن شعيب من قرية باشان نزيل نيسابور #	238
• سافر إلى مكة ومات بها كان فقيها محدثاً توفى سنة 163 ثلاث وستين ~	239
• وصفيه صنف تفسير القرآن وكتاب الفقه وكتاب العبددين وكتاب المنافق ~	240
• المناقب ~	241

\$ al-Harawī

Abū Sa‘īd Ibrāhīm b. Ṭahmān b. Šu‘ayb, from the village of Bashan, a resident of Nishapur.

~~ He traveled to Mecca and died there. He was a jurist, transmitter of Hadith. He died in 163.

~~ ... **He composed** The Exegesis of the Qur'an, Legal hadith, The Book of Two Celebrations, The Book of Virtues.

Abstracting: Tagging & Data

```
17 ## $ Harawī L  
18 # Abū · Sa'īd · Ibrāhīm · ibn · Ṭahmān · ibn · Šu'ayb · @S01 · Harawī, · from · the · village L  
19 ~~of · @T01 · Bāšān, · a · resident · of · @T01 · Naysabūr · [Nishapur] . · He · traveled · to L  
20 ~~@T01 · Makkat · [Mecca] · and · died · there. · He · was · a · @S01 · jurist · and L  
21 ~~a · @S01 · traditionist. · He · died · in · the · @YD163 · year · one · hundred · sixty · three · .  
22 ~~He · wrote · : · Tafsīr · al-Qur'ān, · Sunan · al-fiqh, L  
23 ~~Kitāb · al-'īdayn, · Kitāb · al-manāqib. L
```

=====

id, item, category

=====

000006, 163, year_of_death

000006, Bāšān, toponym

000006, Naysabūr, toponym

000006, Makkat, toponym

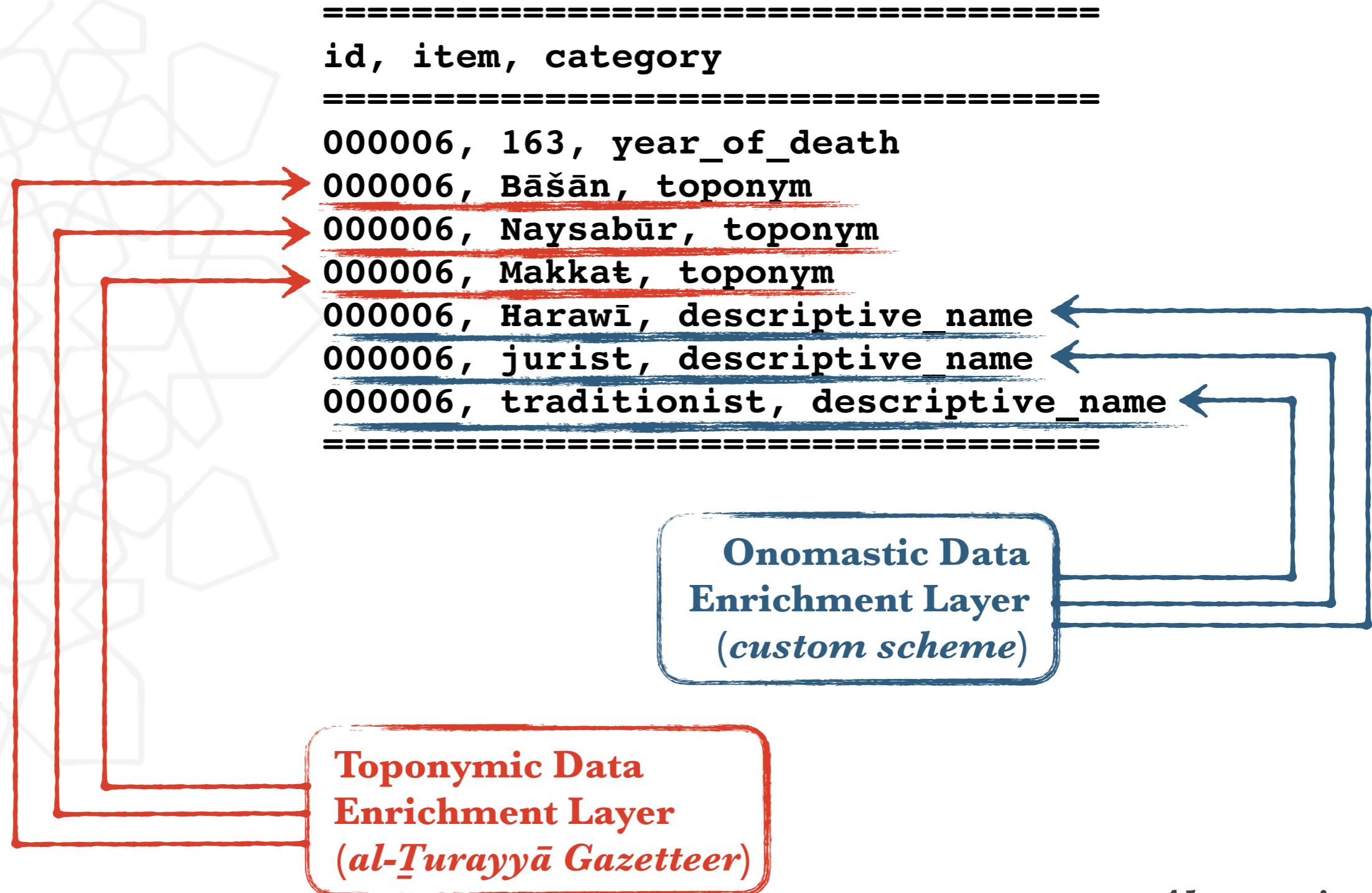
000006, Harawī, descriptive_name

000006, jurist, descriptive_name

000006, traditionist, descriptive_name

=====

Abstracting: *Enriching Data*

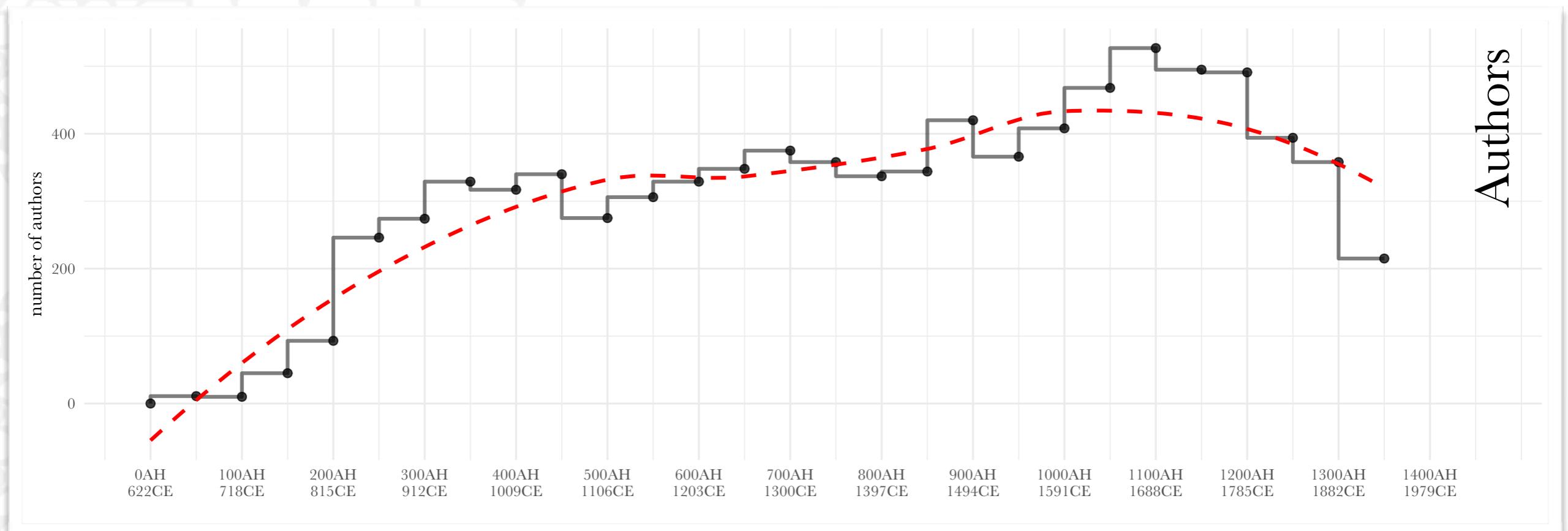


Abstraction



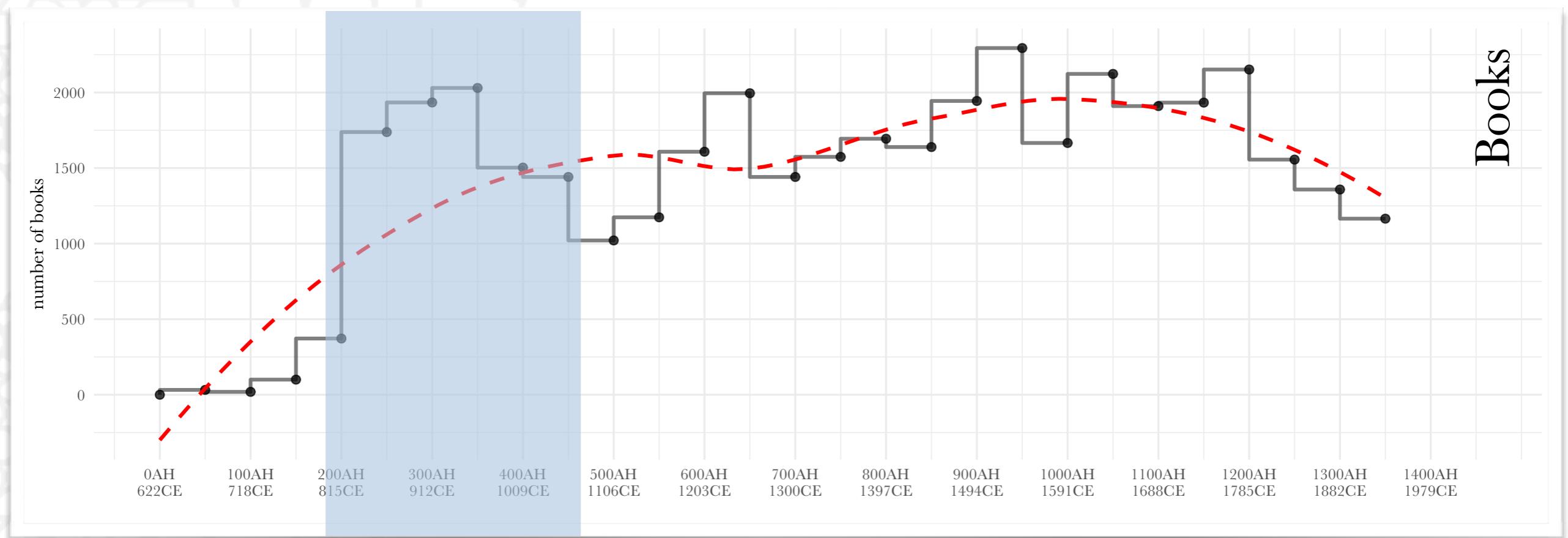
Insight I: *Cultural Production Over Time*

Insight I: *Cultural Production Over Time*



Chronological Distribution of Authors

Insight I: Cultural Production Over Time

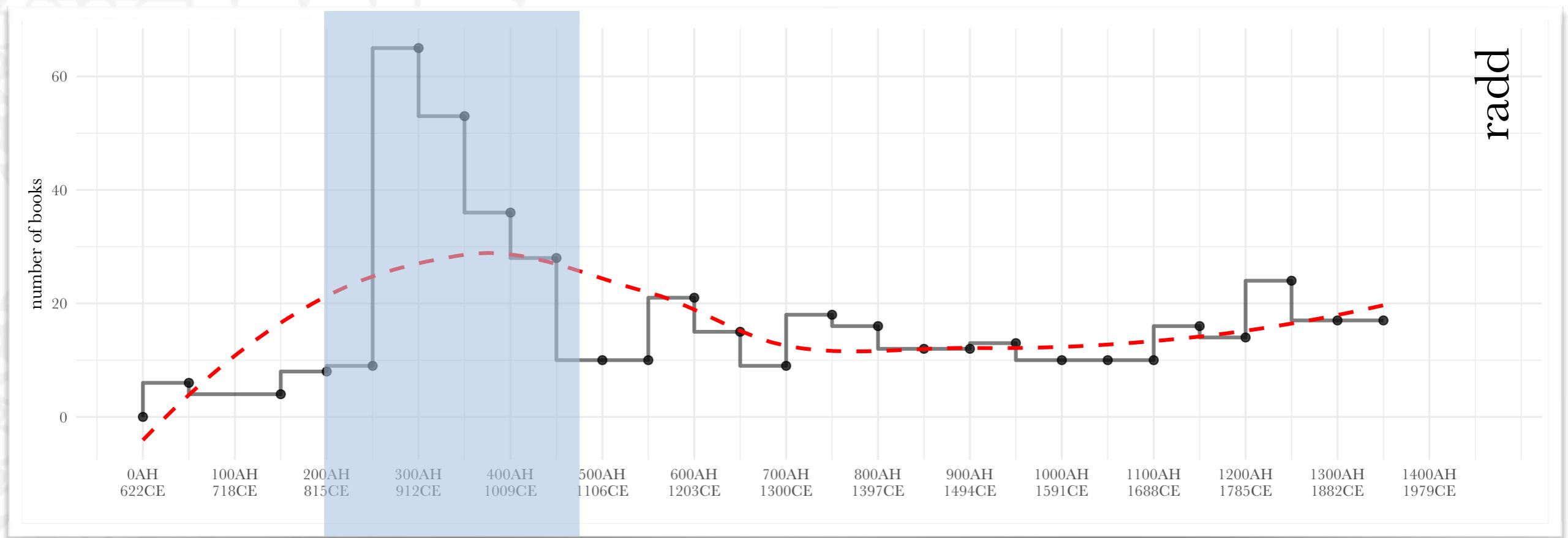


Chronological Distribution of Books

The period of:

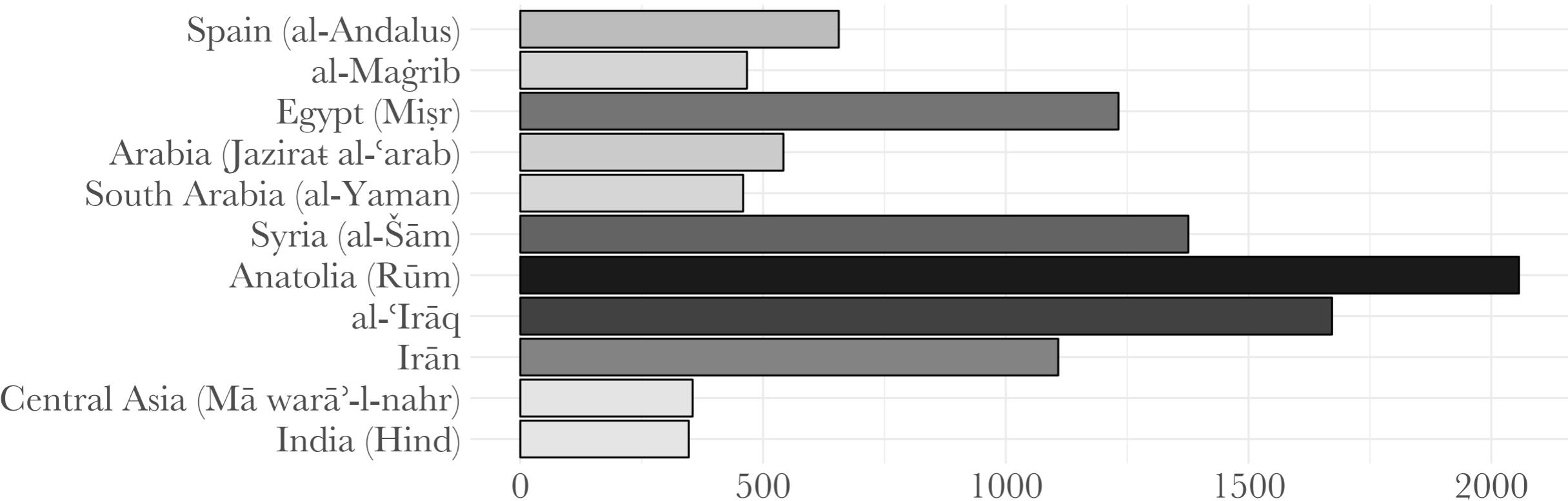
- 1) translations into Arabic
- 2) formation of the Hadith canon
- 3) formative period for Islam as religion more generally (radd)

Insight I: Cultural Production Over Time



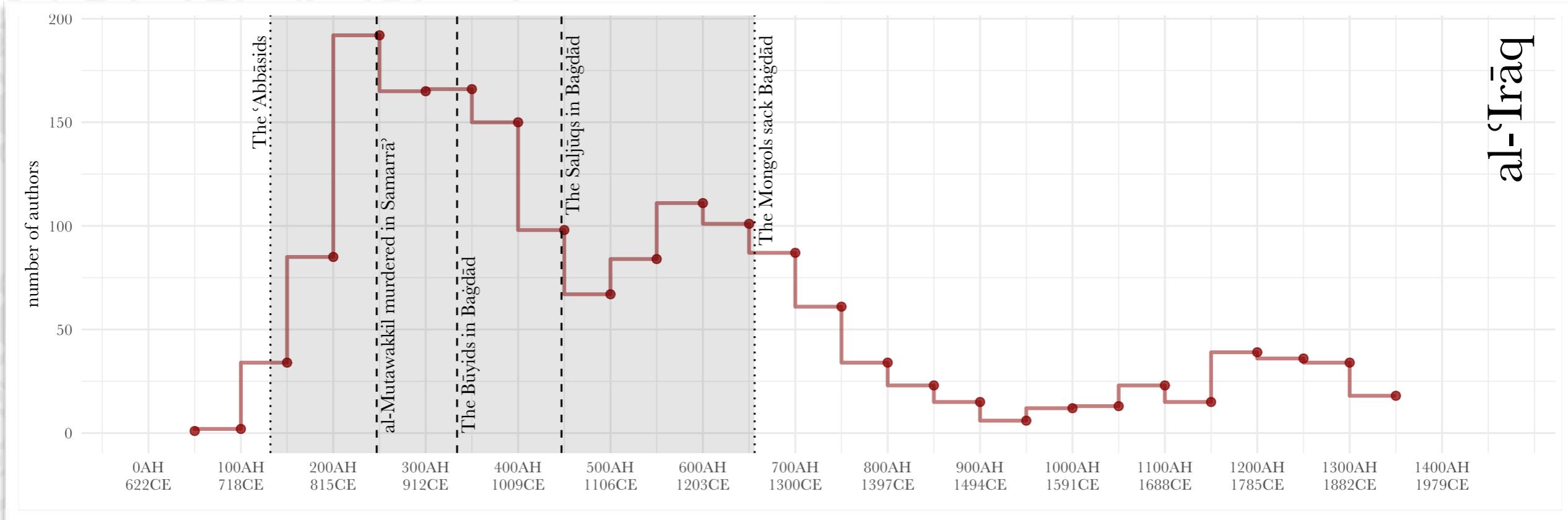
The formative period of Islam: Theological Refutations (radd)

Insight I: Cultural Production Over Time



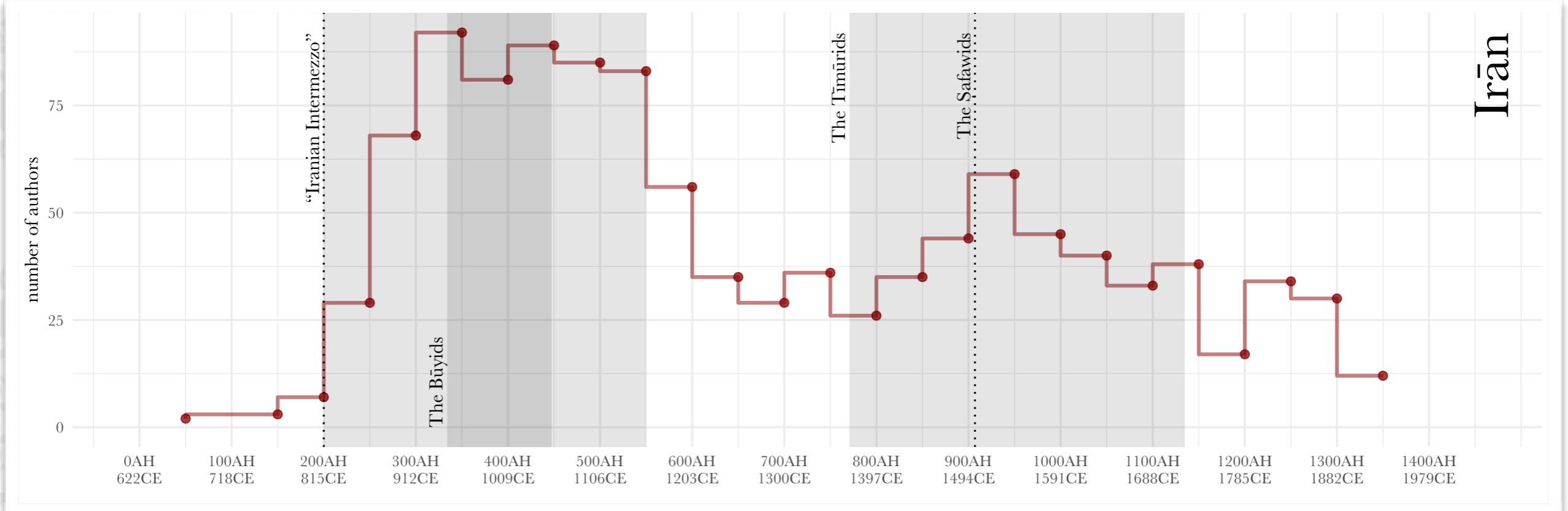
Regional Contributions

Insight I: Cultural Production Over Time



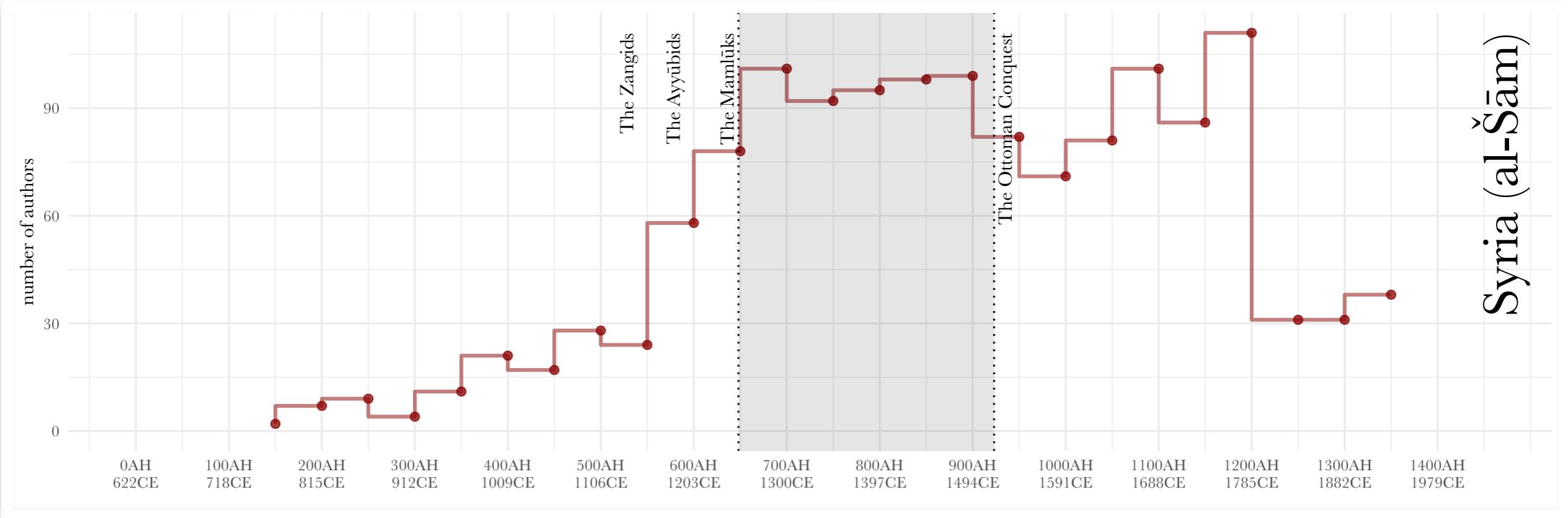
Regional Contributions

Insight I: *Cultural Production Over Time*



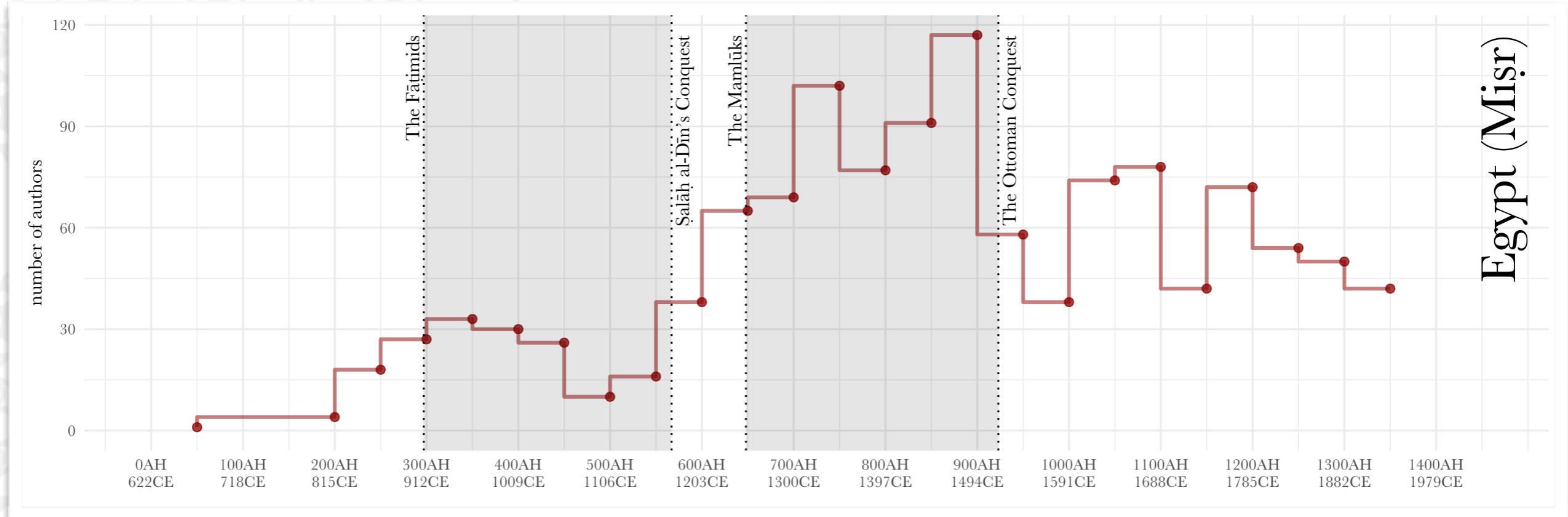
Regional Contributions

Insight I: *Cultural Production Over Time*



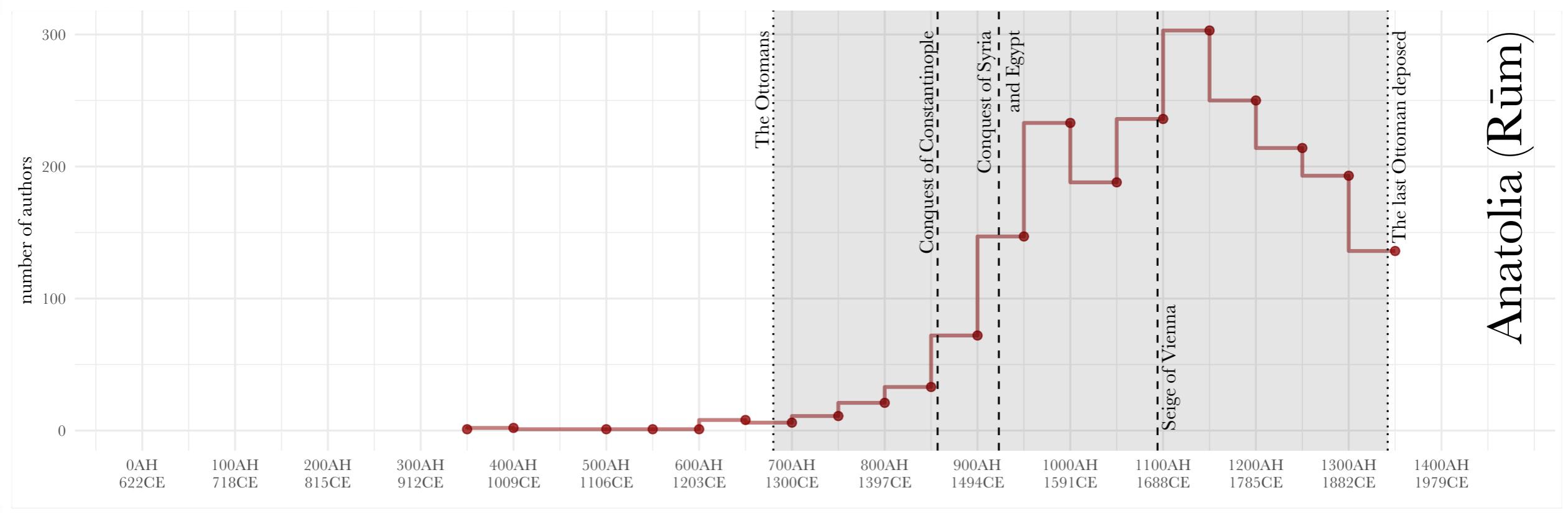
Regional Contributions

Insight I: *Cultural Production Over Time*



Regional Contributions

Insight I: *Cultural Production Over Time*



Regional Contributions



Insight II:

Cultural Connections

Insight II: Cultural Connections

17 ॥### \$ - Harawī .
18 # Abū · Sa'īd · Ibrāhīm · ibn · Ṭahmān · ibn · Šu'ayb · @S01 · Harawī , · from · the · village
19 ~~of · @T01 · Bāšān , · a · resident · of · @T01 · Naysābūr · [Nishapur] . · He · traveled · to
20 ~~@T01 · Makkat · [Mecca] · and · died · there . · He · was · a · @S01 · jurist · and
21 ~~a · @S01 · traditionist . · He · died · in · the · @YD163 · year · one · hundred · sixty · three · .
22 ~~He · wrote · : · Tafsīr · al-Qur'ān , · Sunan · al-fiqh ,
23 ~~Kitāb · al-'īdayn , · Kitāb · al-manāqib .

=====

id, item, category

=====

000006, 163, year_of_death

000006, Bāšān, toponym

000006, Naysabūr, toponym

000006, Makkat, toponym

000006, Harawī, descriptive_name

000006, jurist, descriptive_name

000006, traditionist, descriptive_name

=====

Insight II: Cultural Connections

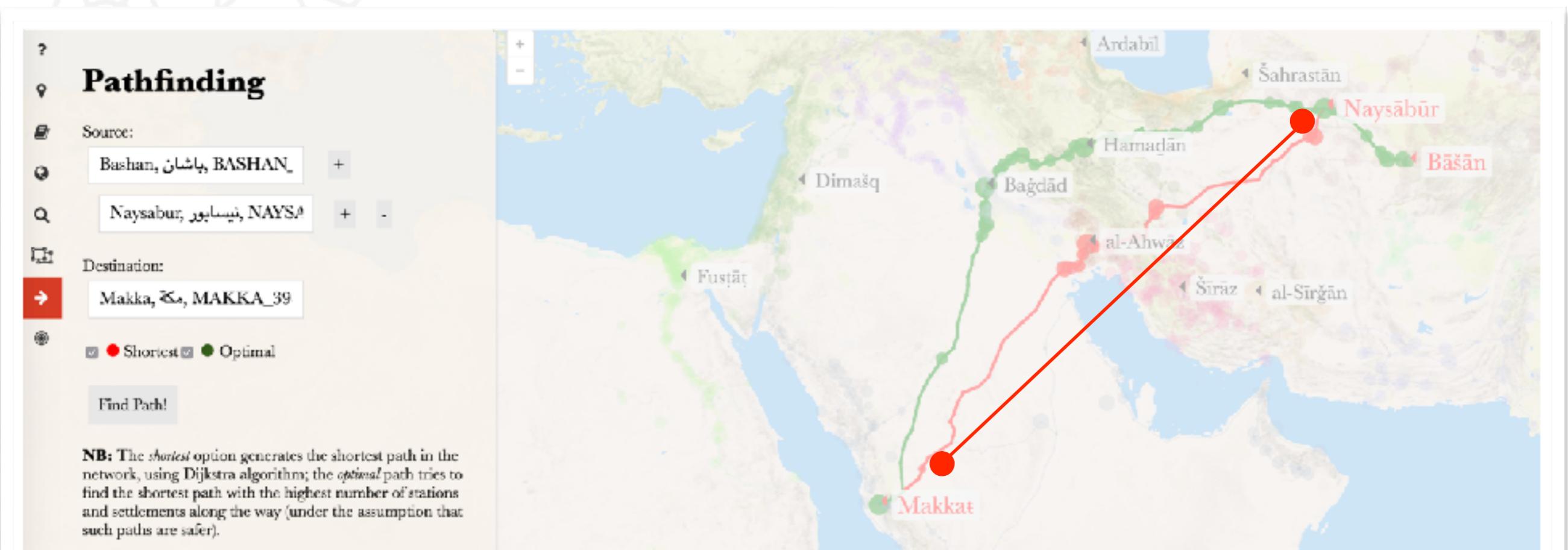
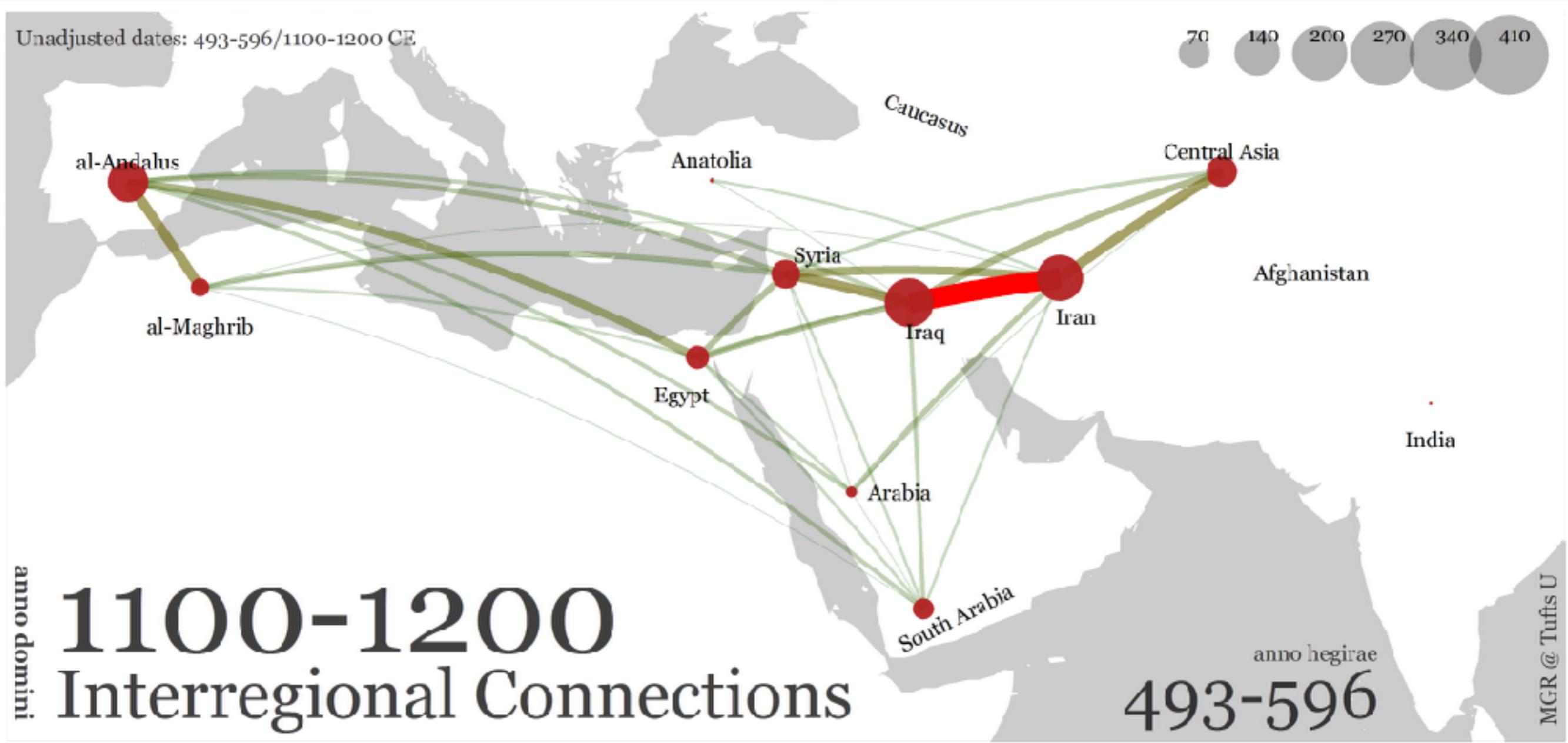


Figure 9: Geographical network of the biographee from the sample biography (using our al-Turayyā Gazetteer, (<https://althurayya.github.io/>)).

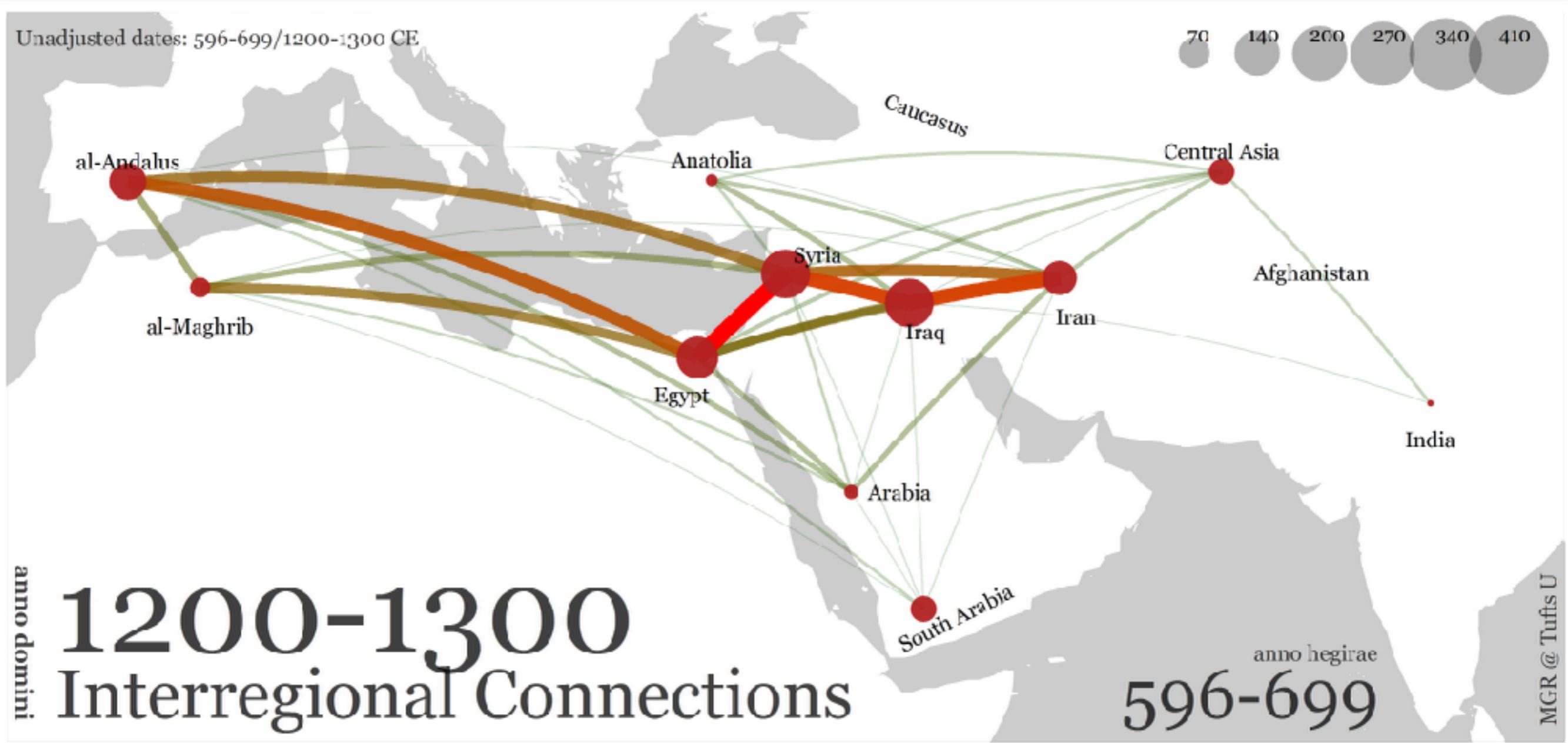
Modeling Personal Geographical Network

Insight II: *Cultural Connections*



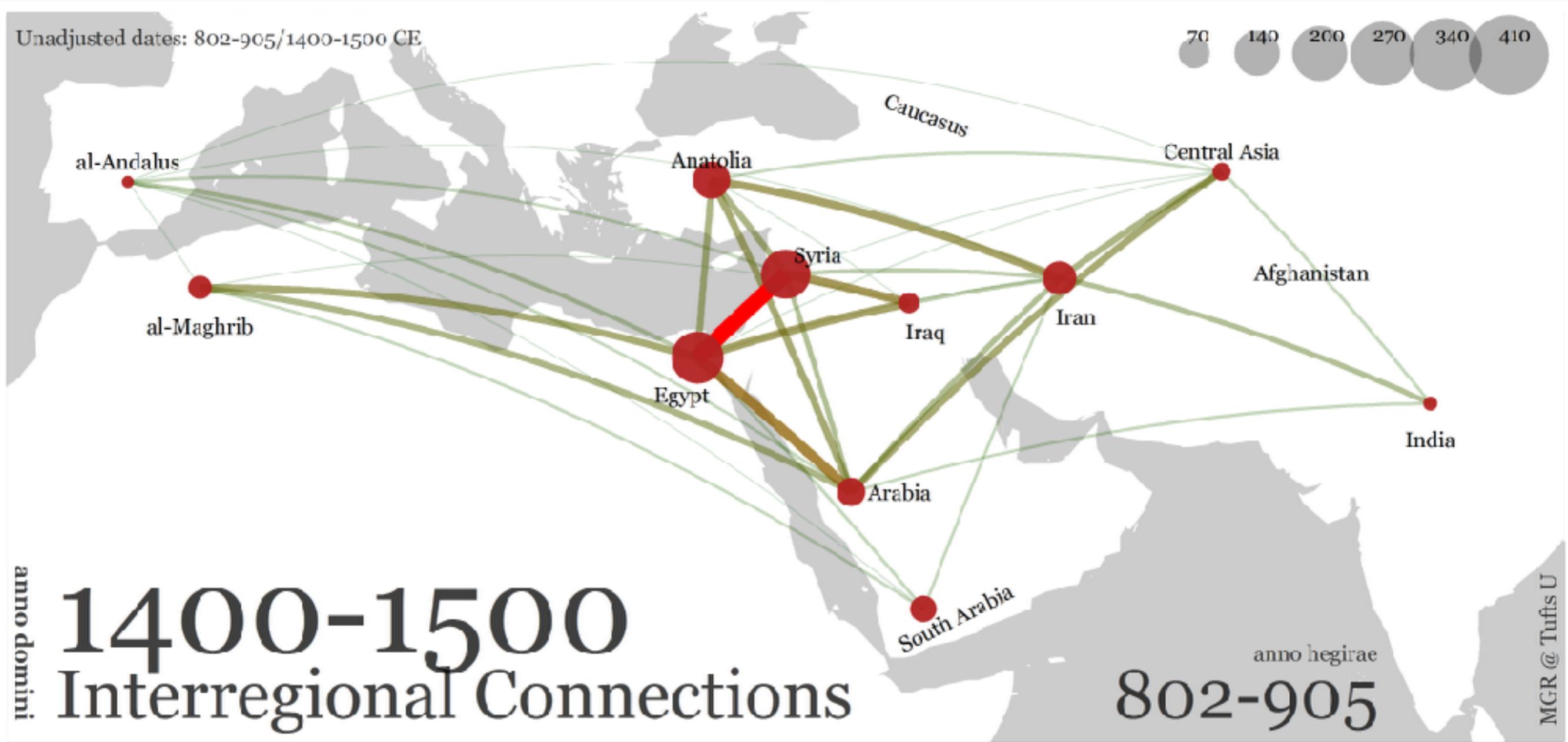
The Iraqi-Iranian Core up until the end of the 12th century CE

Insight II: Cultural Connections



Massive migrations of the 13th century CE

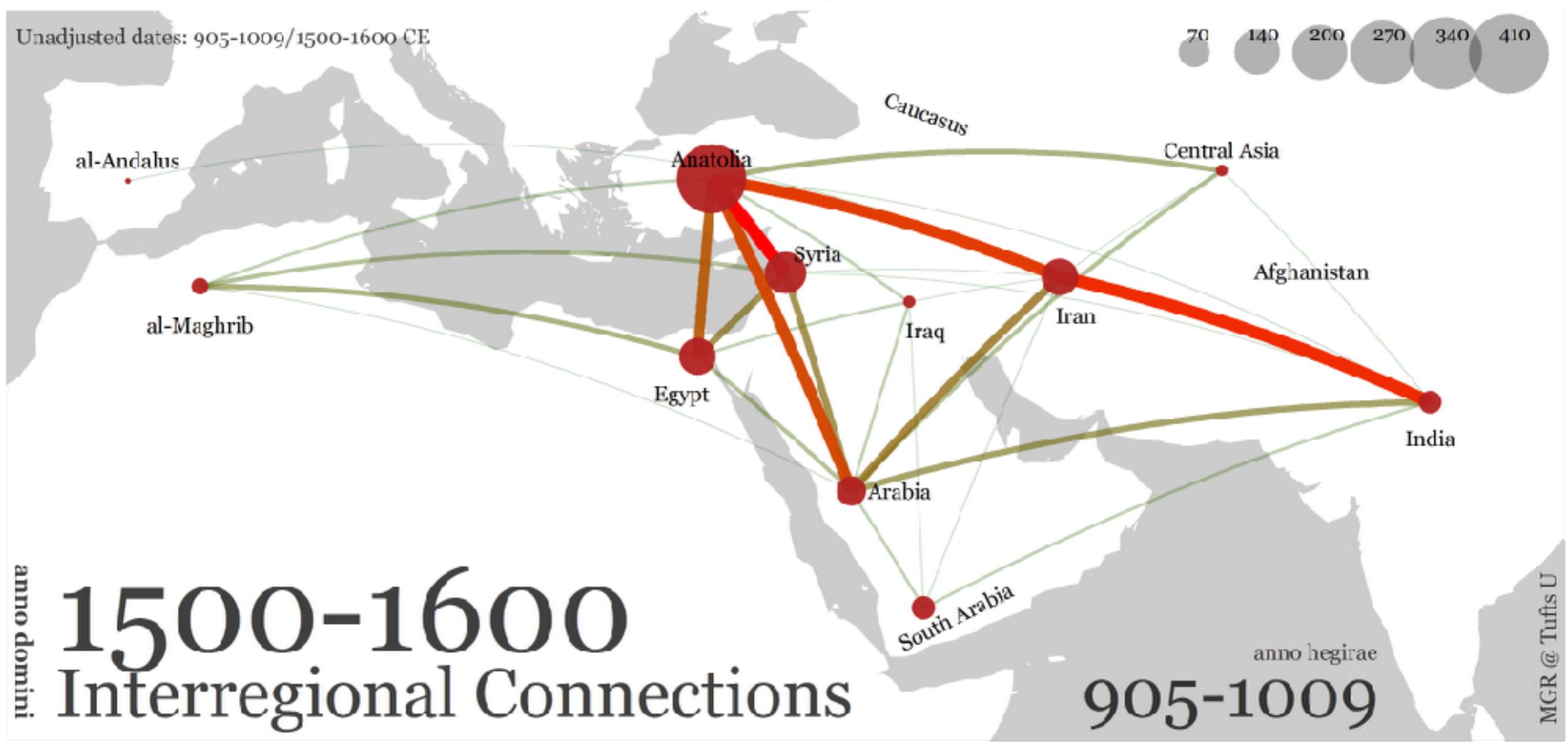
Insight II: Cultural Connections



1400-1500
Interregional Connections

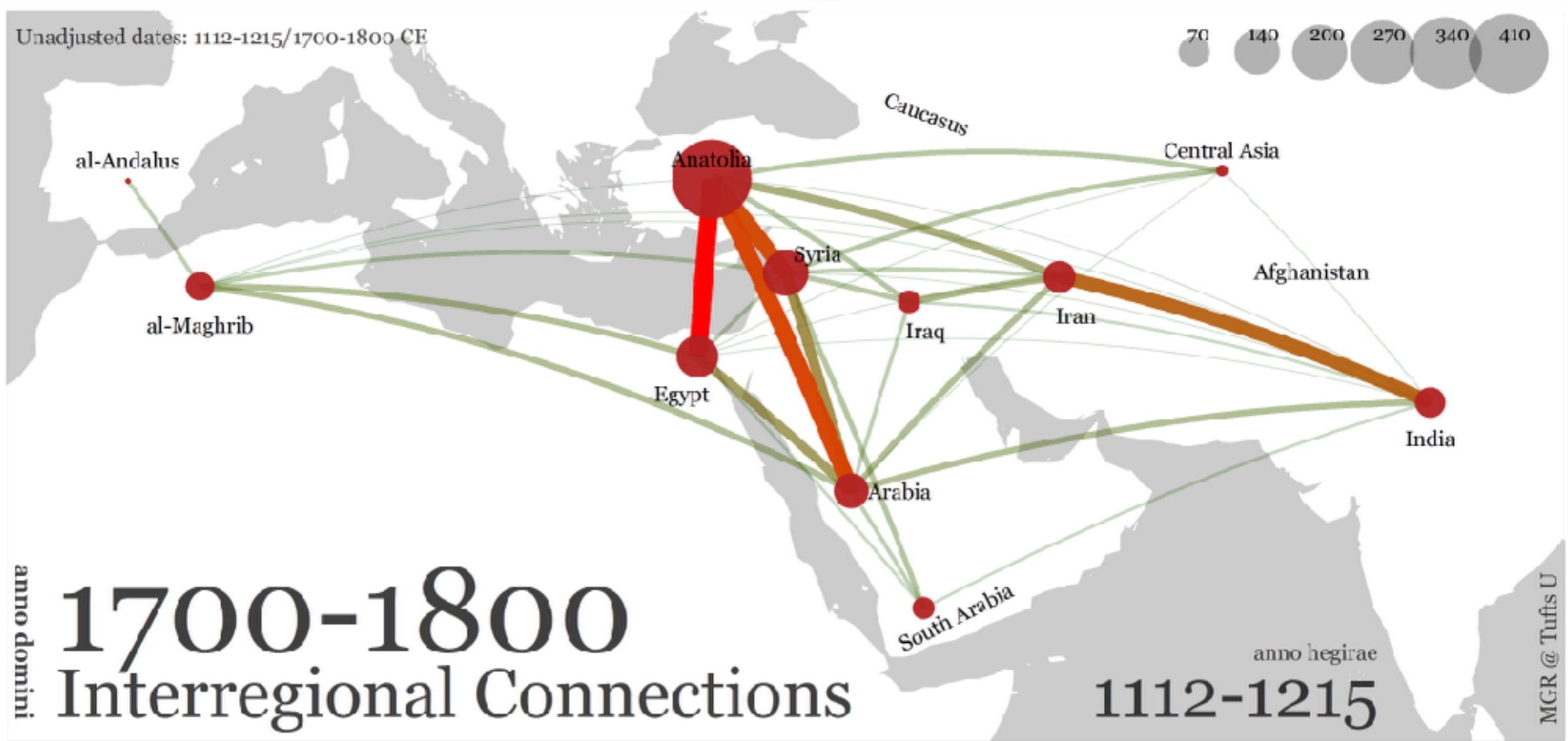
New Mamlūk Core of the 14th and 15th centuries CE

Insight II: Cultural Connections



Reconfiguration of the 16th centuries CE

Insight II: Cultural Connections



The Turco-Arabic and Indo-Iranian Cores in the 18th and 19th centuries CE



Scaling Things Up:

*OpenArabic / OpenITI
Exploratorium*

OpenArabic

OpenArabic

Description of the project and the status of development

[View the Project on GitHub](#)

OpenArabic/Annotation

Download
ZIP File

Download
TAR Ball

View On
GitHub

**OpenArabic Project (@ AvH
Lehrstuhl für Digital Humanities, U
Leipzig, led and curated by Maxim
Romanov)**

Contents

- [General Description](#)
- [Prospects and Progress](#)
- [Text Description Tags](#)
- [Preliminary Analysis of Categories of Texts](#)
- [Folder structure](#)
- [General description of the workflow with mARkdown](#)
- [Status Report](#)
- [List of books by centuries](#)
- [Statistics on the corpus](#)
- [Summary statistics on the lengths of texts in the corpus](#)
- [Texts by length \(duplicates excluded\)](#)
- [Texts in chronological order \(duplicates excluded\)](#)
- [Chronological Distribution of Texts - up until 1930 \(5,467 texts, 726,946,794 words\)](#)
- [Forms, Themes, Genres \(provisional assessment\)](#)

General Description

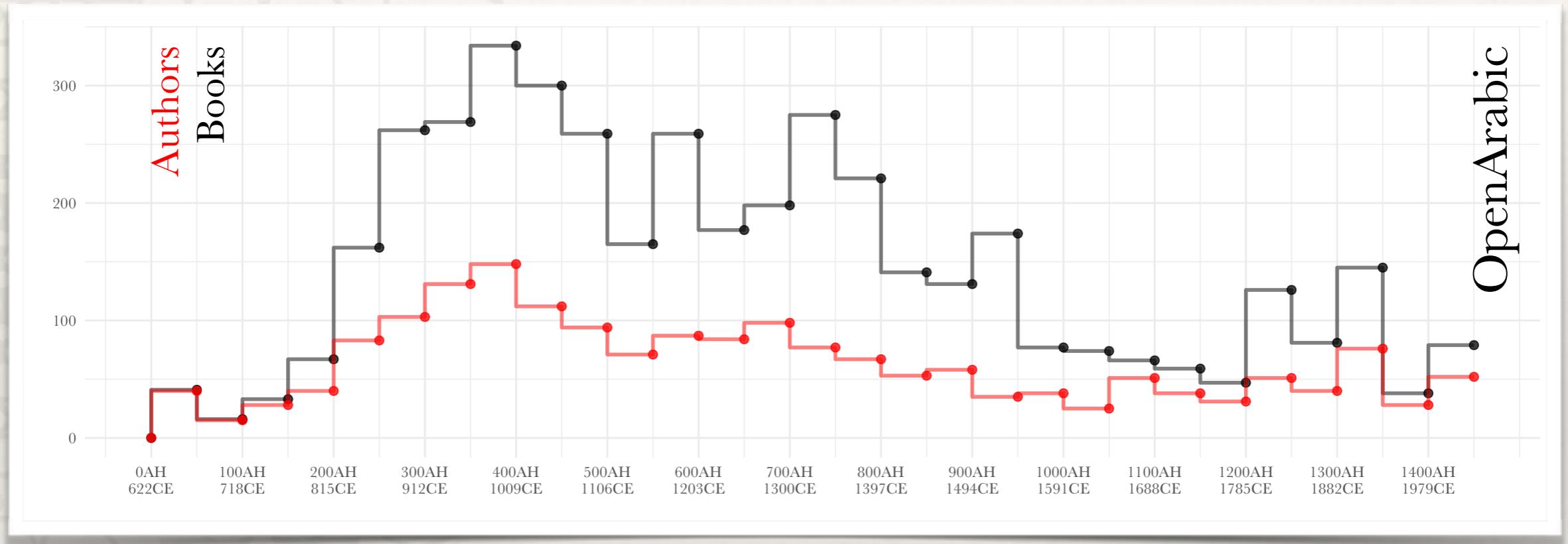
The goal of OpenArabic is to build a machine-actionable corpus of premodern texts in Arabic to encourage computational analysis of the Arabic literary tradition. Currently, most of the texts are historical in nature (chronicles, biographical collections, geographical treatises and gazetteers,

<https://github.com/OpenArabic/>

This project is maintained by [OpenArabic](#)

Hosted on GitHub Pages — Theme by [orderedlist](#)

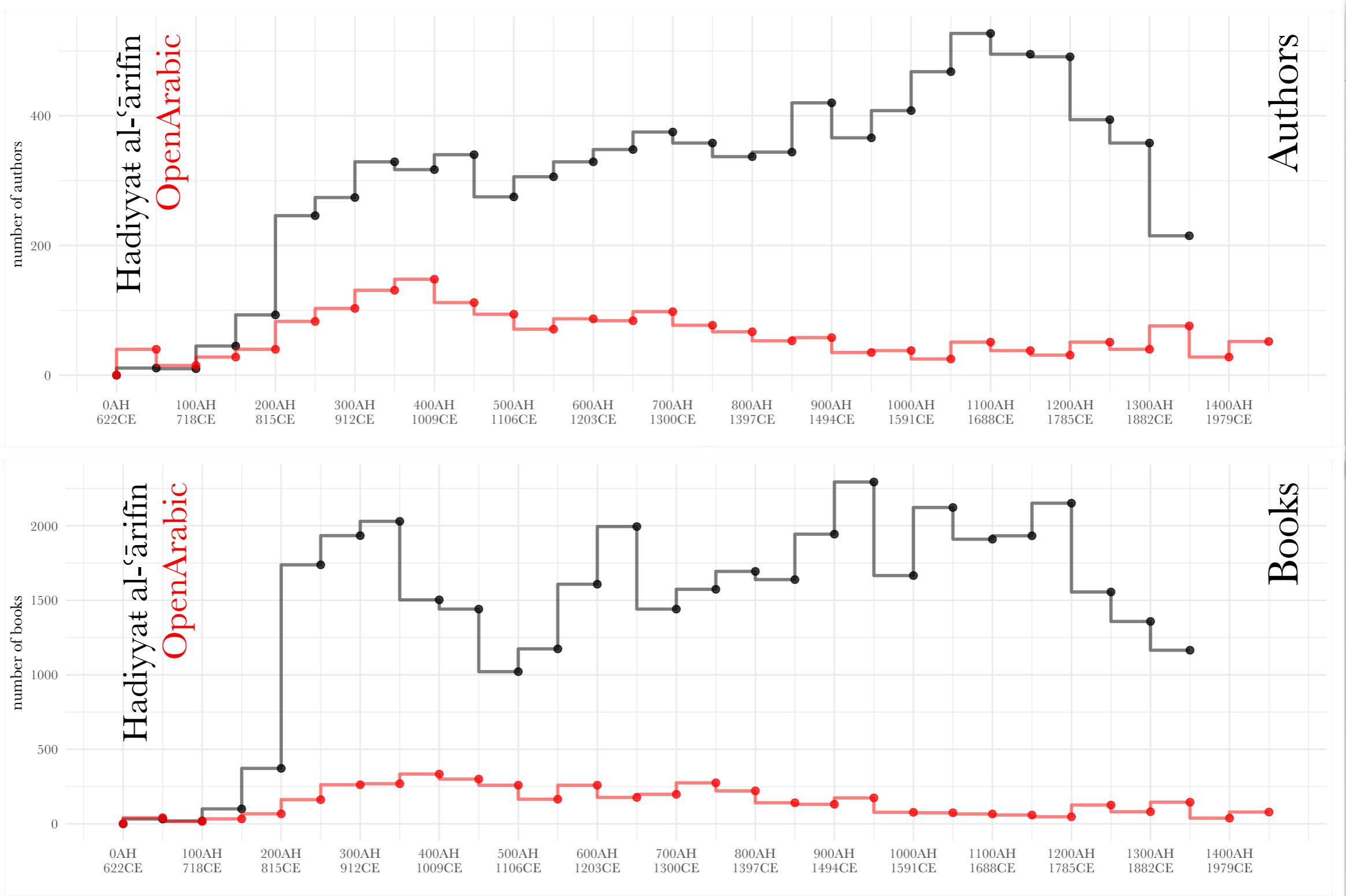
OpenArabic



**Unique Titles: 4,300
Words: 740 million
(All: 1,3 billion words)**

**Biographical Subcorpus:
Unique Titles: ~600
Words: 133 million**

OpenArabic: ~10%?



Hadiyyat al-ārifīn, a bio-bibliographical collection: 8,800 authors, 40,000 titles

OpenArabic / OpenITI: OCR

- ❖ **Kraken ibn Ocropus (a *fork* of OCropus)**

- Benjamin Kiessling, U Leipzig
- Matthew Miller, U of Maryland
- Sarah Savant, Aga Khan U—London
- Maxim Romanov, U Leipzig

Accuracy Rates in the high 90s!



<https://www.academia.edu/28923960/>



OpenArabic / OpenITI: OCR

Book*	Quality	Type	Model accuracy level			
			Size 100	Ar**	Size 200	Ar**
1 Ibn al-Athīr. <i>al-Kāmil</i>	high***	training	93.79	97.71	93.58	97.59
2 Ibn Qutayba. <i>Adab al-kātib</i>	high***	testing	82.68	95.72	80.92	94.88
3 al-Jāhiz. <i>al-Hayawān</i>	high***	testing	71.78	75.16	70.85	74.27
4 al-Ya‘qūbī. <i>al-Ta‘rīkh</i>	high***	testing	79.67	84.40	78.12	82.21
5 al-Dhahabī. <i>Ta‘rīkh al-islām</i>	low****	testing	90.68	95.95	90.37	95.78
6 Ibn al-Jawzī. <i>al-Muntazam</i>	low****	testing	93.33	98.51	92.96	98.22

* Information on the editions of these texts is supplied at the end of the report

** Performance on Arabic only (excluding punctuation and spaces)

*** 300 dpi, grayscale; scanned specifically for the purpose of testing, with ideal parameters

**** 200 dpi, black-and-white, pre-binarized; both downloaded from www.archive.org (via <http://waqfeya.org>)



OpenArabic / OpenITI: OCR

Book*	Quality	Type	Model accuracy level			
			Size 100	Ar**	Size 200	Ar**
1 Ibn al-Athīr. <i>al-Kāmil</i>	high***	testing	80.23	86.27	82.46	87.48
2 Ibn Qutayba. <i>Adab al-kātib</i>	high***	testing	80.90	91.54	82.61	93.24
3 al-Jāhīz. <i>al-Hayawān</i>	high***	training	94.86	97.59	94.82	97.41
4 al-Ya‘qūbī. <i>al-Ta‘rīkh</i>	high***	testing	90.91	95.13	91.28	94.71
5 al-Dhahabī. <i>Ta‘rīkh al-islām</i>	low****	testing	81.93	91.23	83.03	92.22
6 Ibn al-Jawzī. <i>al-Muntazam</i>	low****	testing	84.07	93.58	86.26	94.20

* Information on the editions of these texts is supplied at the end of the report

** Performance on Arabic only (excluding punctuation and spaces)

*** 300 dpi, grayscale; scanned specifically for the purpose of testing, with ideal parameters

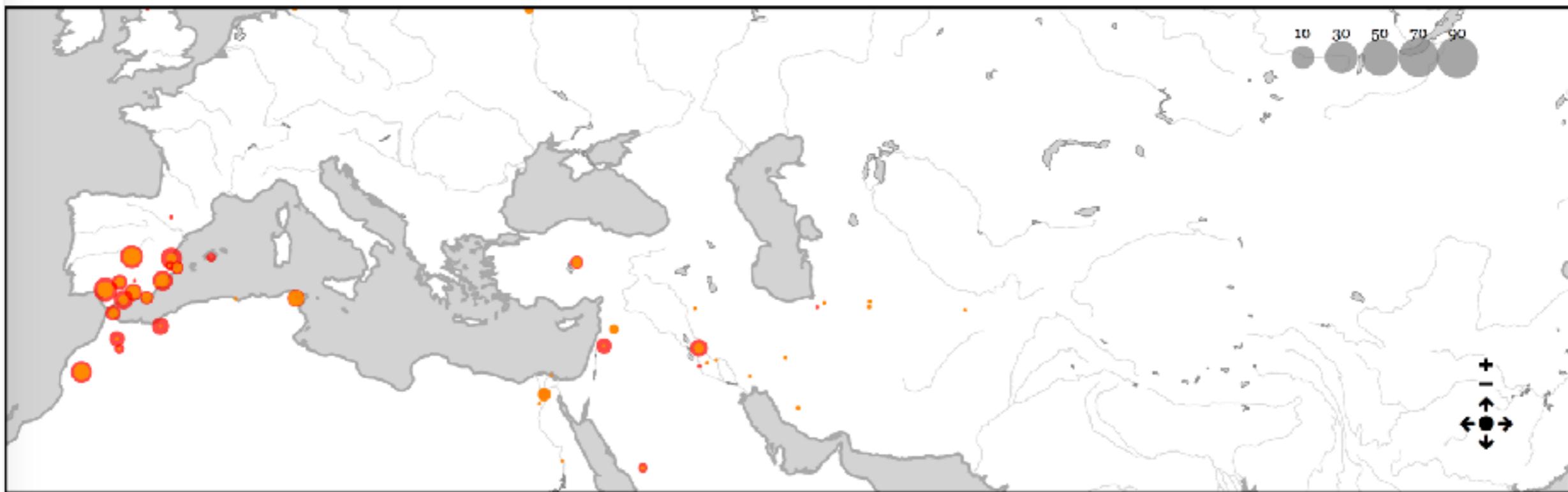
**** 200 dpi, black-and-white, pre-binarized; both downloaded from www.archive.org (via <http://waqfeya.org>)

Exploratorium for Biographical Data

Exploratorium

The page offers an exploratory insight into groups of individuals from the “History of Islām” (*Ta’rīh al-islām*) of al-Dahabī who share common ‘descriptors’ (sing. *nisbat*). By selecting a ‘descriptor’ (النسبة), you can get preliminary analyses of geographical, chronological, and network patterns.

Space



al-Turayyā Project: Gazetteer & Geospatial Model

?

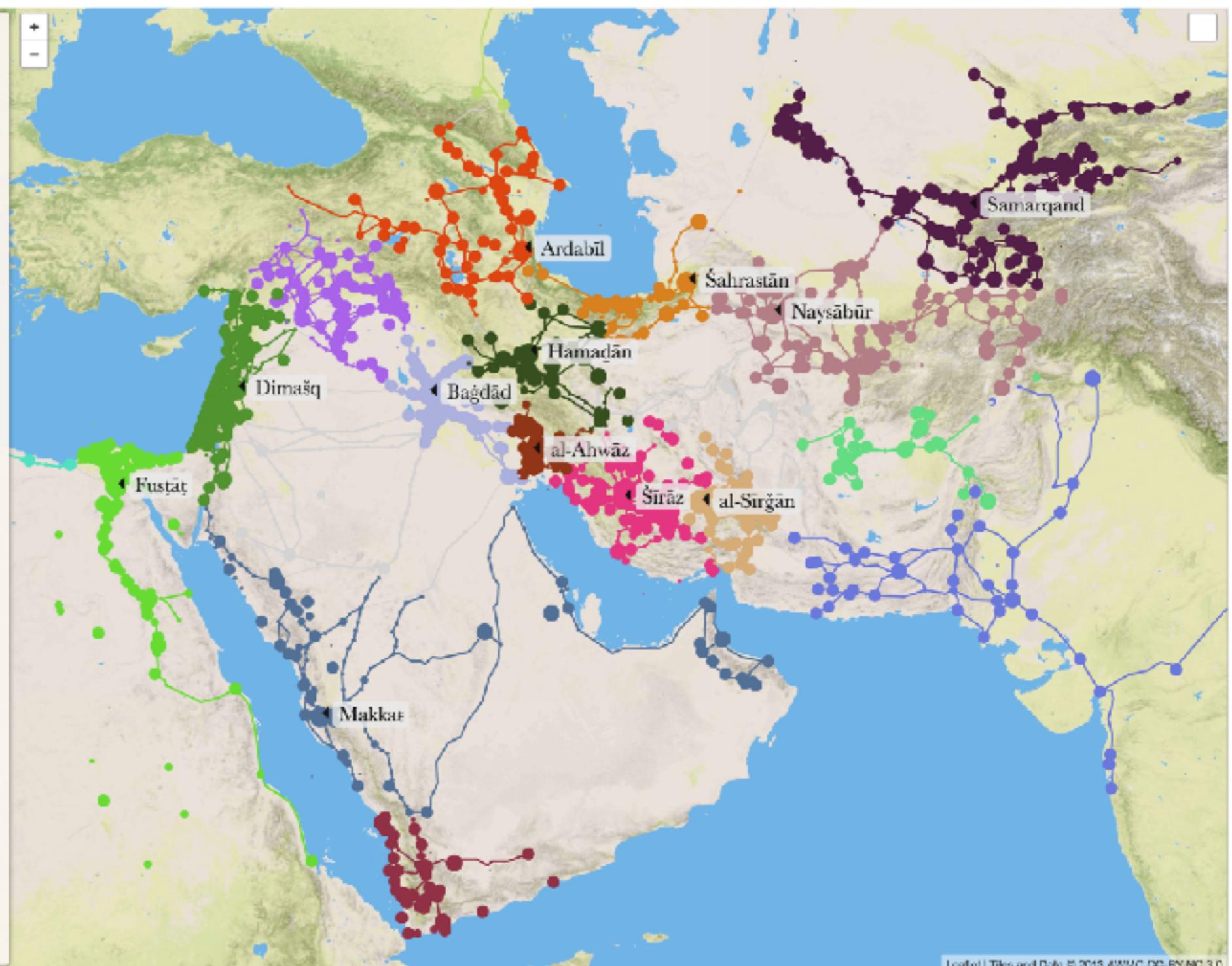
al-Turayyā Gazetteer

This is a new working version of *al-Turayyā Gazetteer* (or *al-Thurayyā Gazetteer*). Currently it includes over 2,000 toponyms and almost as many route sections georeferenced from Georgette Coerno's *Atlas du monde arabo-irano-égyptien à l'époque classique: IXe-Xe siècles* (Leiden: Brill, 1983). The functionality is still under development. You can use an earlier version of *al-Turayyā*, where you can browse the Gazetteer by clicking on any toponym marker. The popup will show the toponym both in Arabic script and transliterated. We are using a slightly modified transliteration system that facilitates conversion between fully transliterated, transliterated, and Arabic forms of toponyms. It should be easily understandable. There may be typos, because of the nature of how the data has been generated, so please, let us know if something should be corrected. The popup also offers a selection of possible sources on a toponym in question. You can check Arabic Sources: currently, al-Sam'ānī's *Kitāb al-uzūq* and Yāqūt's *Mu'jam al-bulūd*. Currently, the Gazetteer will only check for 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. Improving the precision of this lookup is on our to-do list. You can also check if there is information on a toponym in question in Brill's *Egyptological Lexicon*, *Plaids*, and Wikipedia. It can be found [here](#).

Note on Transliteration: The website uses a somewhat unconventional transliteration system, which was developed to facilitate computational analysis. Unlike more traditional transliteration schemes the current one uses one-to-one letter representation, with every Arabic letter transcribed distinctively, which allows for an automatic conversion between transliteration and the Arabic script. The overall scheme should be easily recognizable to Arabists (new letters are as follows: *t* for *كَافَةٌ*; *d* for dagger *دَالِّ*; and *d* for *دَالِّيَّةٍ*).

Credits and Acknowledgments:

Current team: Masourneh Seydi and Maxim Romanow (@ U Leipzig). **Former contributors:** 2013–2014: Cameron Jackson (class of 2014, double-major in Arabic and Computer Science, Tufts)—technical and conceptual development; 2013: Adam Tavares, programmer (@).





Thank you!