Maxim Romanov

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

Algorithmic Analysis of Premodern Arabic Biographical Collections

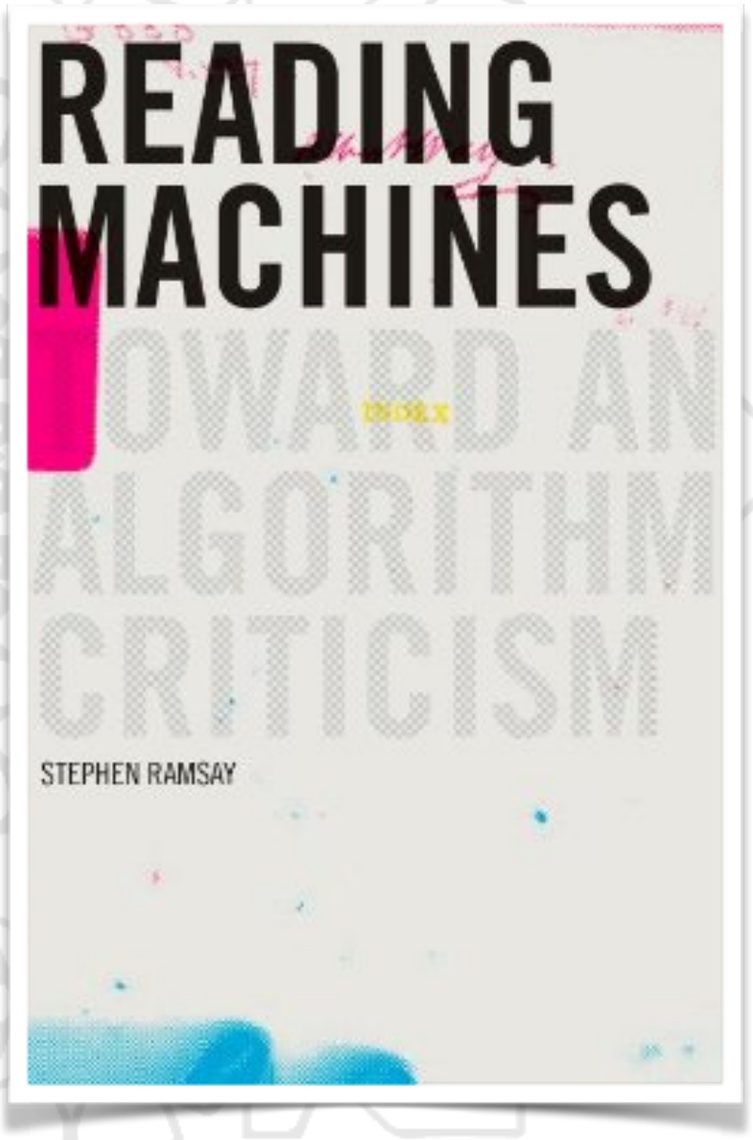
Approach, Infrastructure, Open Data

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*."

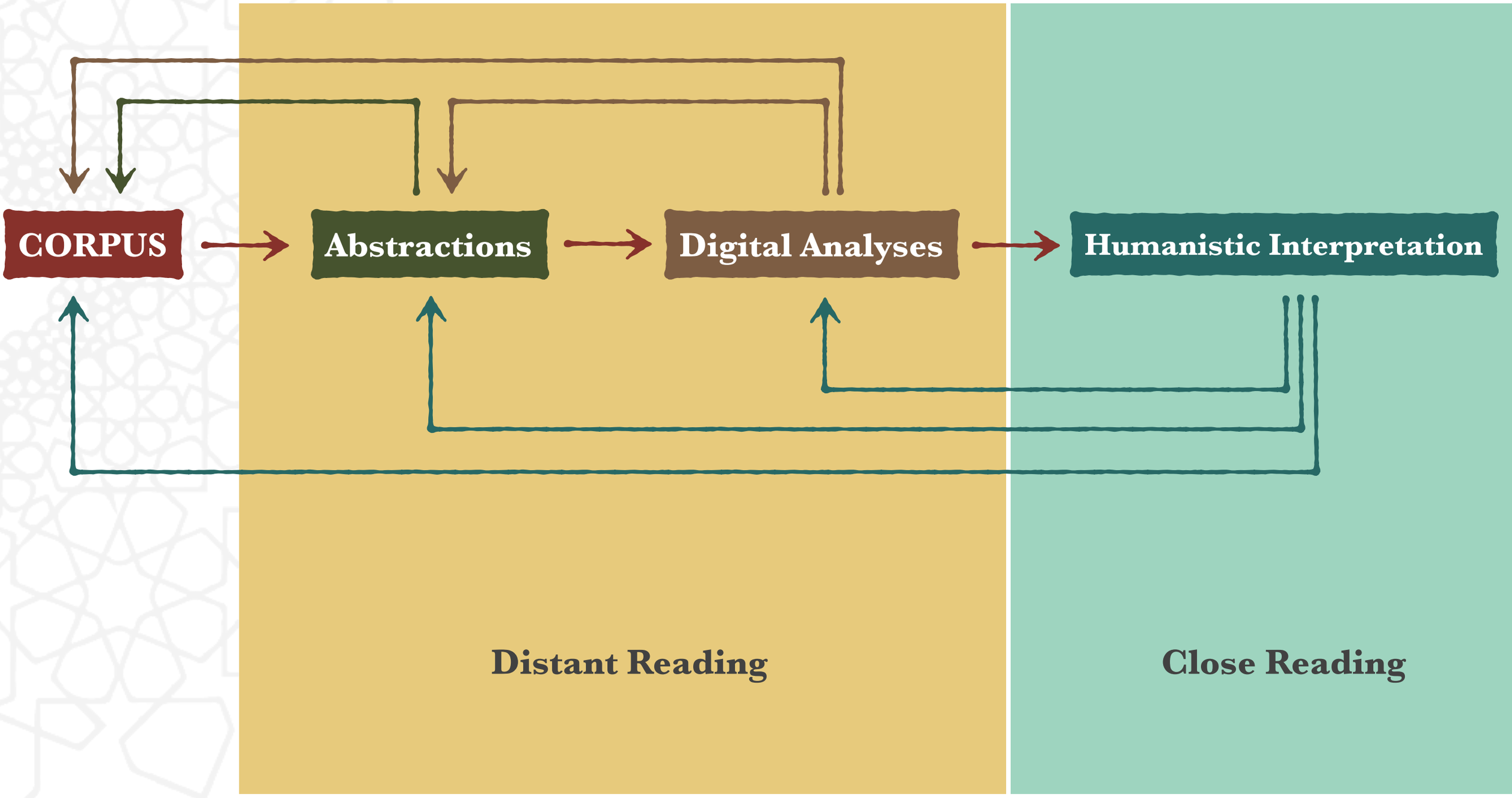
Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. 1st Edition. University of Illinois Press, 2011.

http://stephenramsay.us/

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*

```
237 # $ الهروي
238 # أبو سعيد إبراهيم بن طهمان بن شعيب من قرية باشان نزيل نيسابور
239 ~~ سافر إلى مكة ومات بها كان فقيها محدثا توفى سنة 163 ثلاث وستين
240 ~~ ومائة . صنف تفسير القرآن . سنن الفقه . كتاب العيدين . كتاب
241 ~~ المناقب .
```

```
# $ al-Harawī
# Abū Saʿīd Ibrāhīm b. Ṭahmān b. Šuʿayb, 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 ᴸF
18 # Abū Saʿīd Ibrāhīm ibn Ṭahmān ibn Šuʿayb @S01 Harawī, from the village ᴸF
19 ~~of @T01 Bāšān, a resident of @T01 Naysābūr [Nishapur]. He traveled to ᴸF
20 ~~@T01 Makkaṯ [Mecca] and died there. He was a @S01 jurist and ᴸF
21 ~~a @S01 traditionist. He died in the @YD163 year one hundred sixty three .
22 ~~He wrote : Tafsīr al-Qurʾān, Sunan al-fiqh, ᴸF
23 ~~Kitāb al-ʿīdayn, Kitāb al-manāqib. ᴸF
```

```
=====================================
id, item, category
=====================================
000006, 163, year_of_death
000006, Bāšān, toponym
000006, Naysabūr, toponym
000006, Makkaṯ, toponym
000006, Harawī, descriptive_name
000006, jurist, descriptive_name
000006, traditionist, descriptive_name
=====================================
```

Abstracting: *Enriching Data*

```
==================================
id, item, category
==================================
000006, 163, year_of_death
000006, Bāšān, toponym
000006, Naysabūr, toponym
000006, Makkaṯ, toponym
000006, Harawī, descriptive_name
000006, jurist, descriptive_name
000006, traditionist, descriptive_name
==================================
```

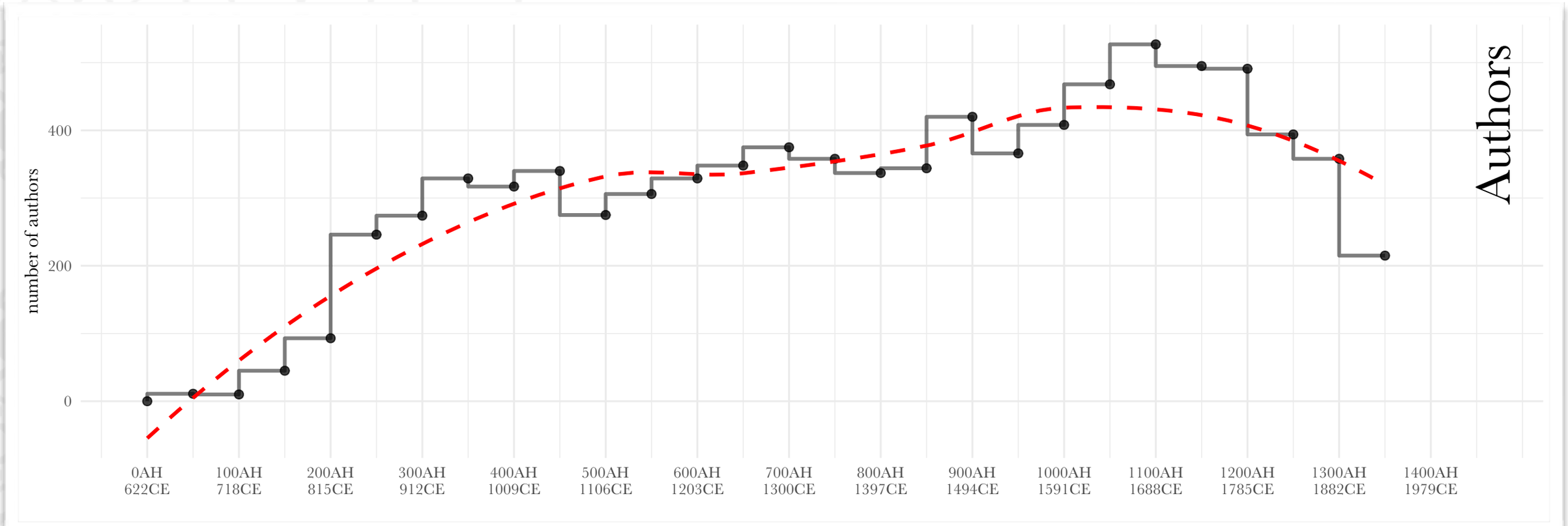
Onomastic Data Enrichment Layer (*custom scheme*)

Toponymic Data Enrichment Layer (*al-Ṯurayyā Gazetteer*)

Insight I:

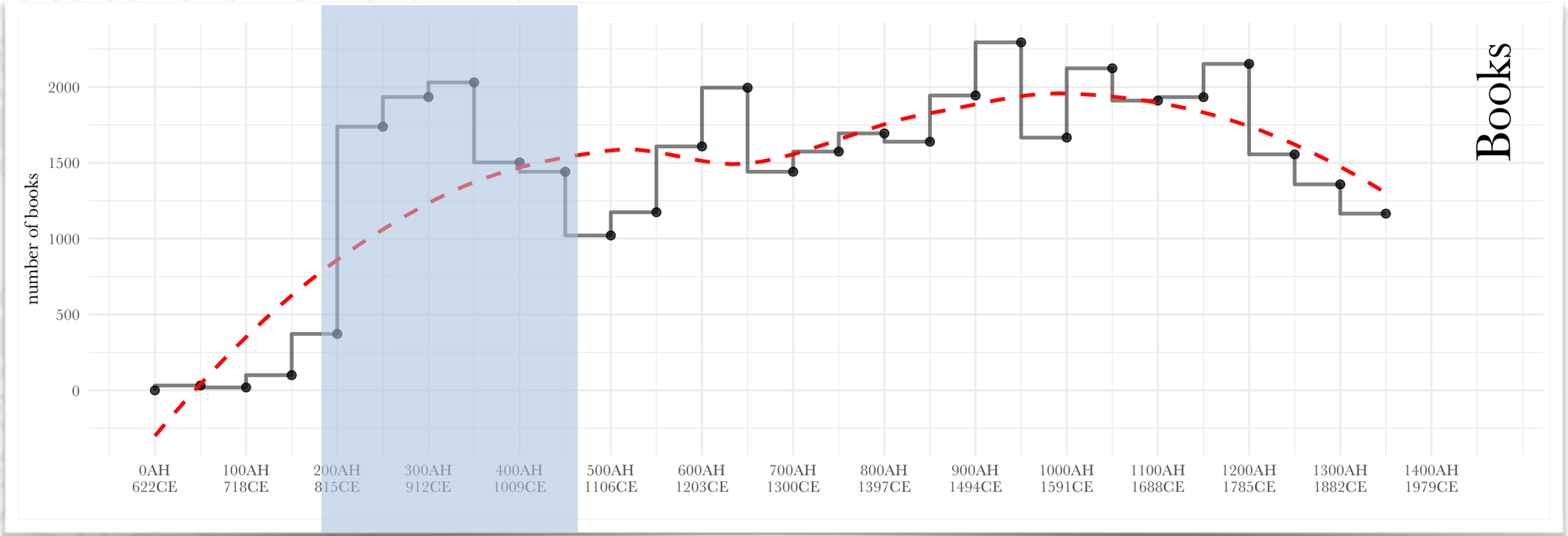
Cultural Production Over Time

Insight I: *Cultural Production Over Time*



Chronological Distribution of Authors

Insight I: *Cultural Production Over Time*

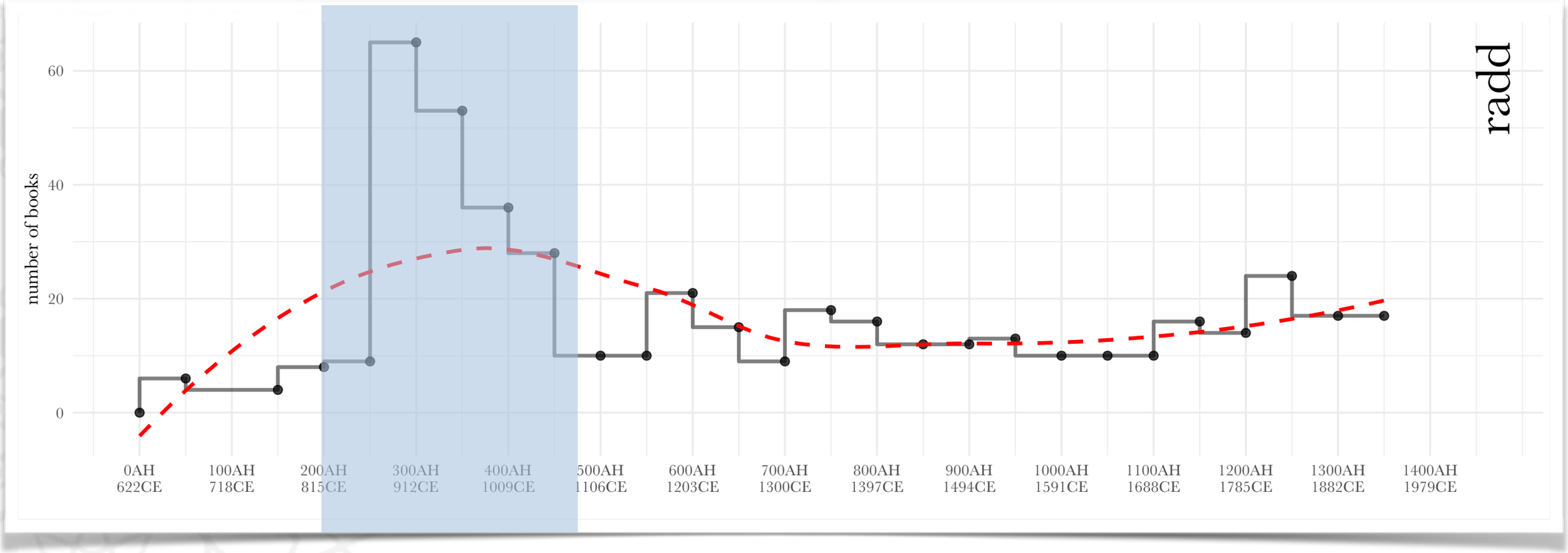


Chronological Distribution of Books

The period of:

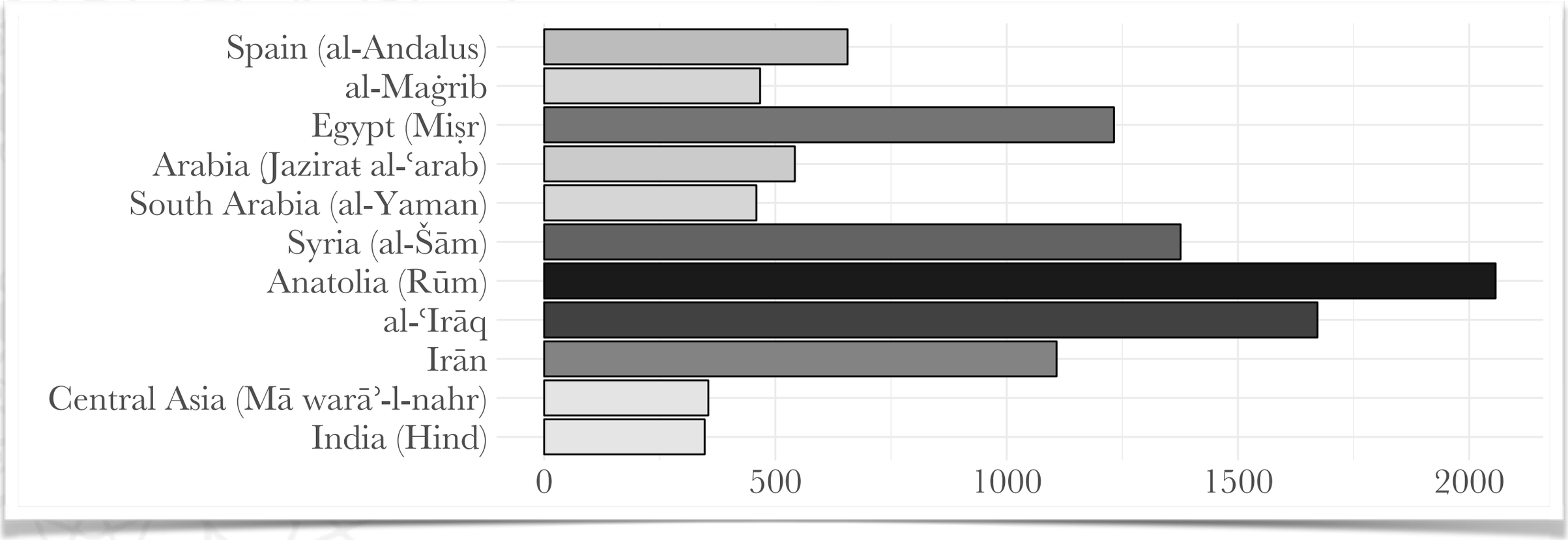
1) translations into Arabic
2) formation of the Hadith canon
3) formative period for Islam as religion more generally (radd)

Insight I: *Cultural Production Over Time*



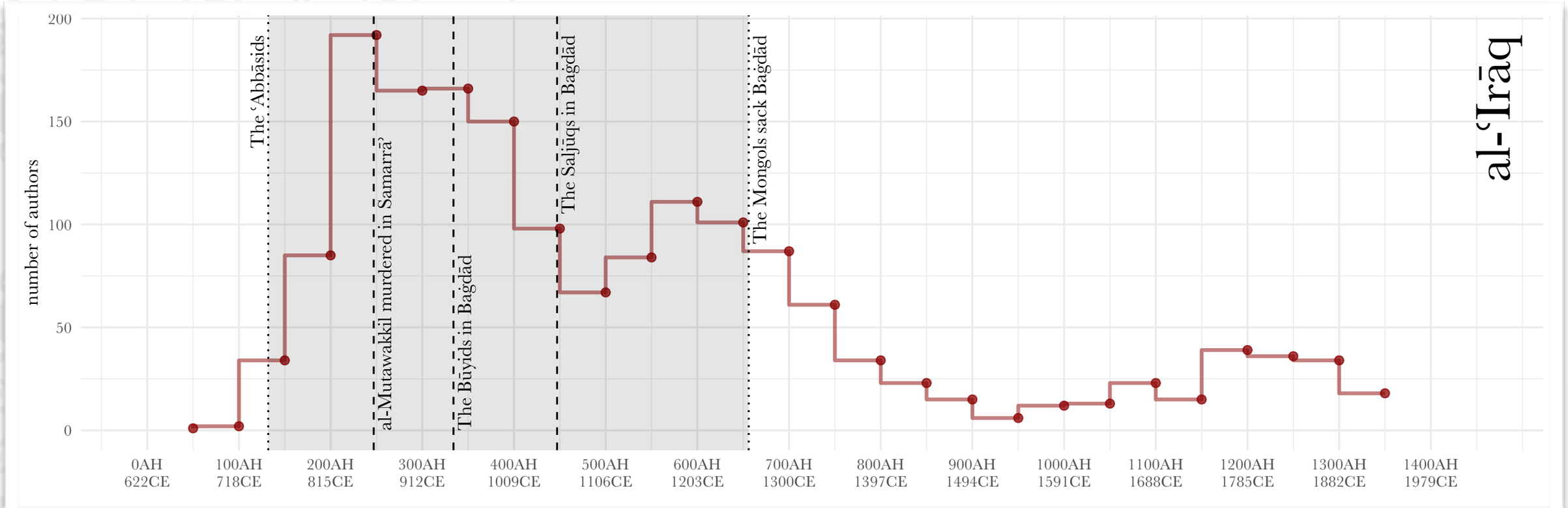
The formative period of Islam: Theological Refutations (radd)

Insight I: *Cultural Production Over Time*



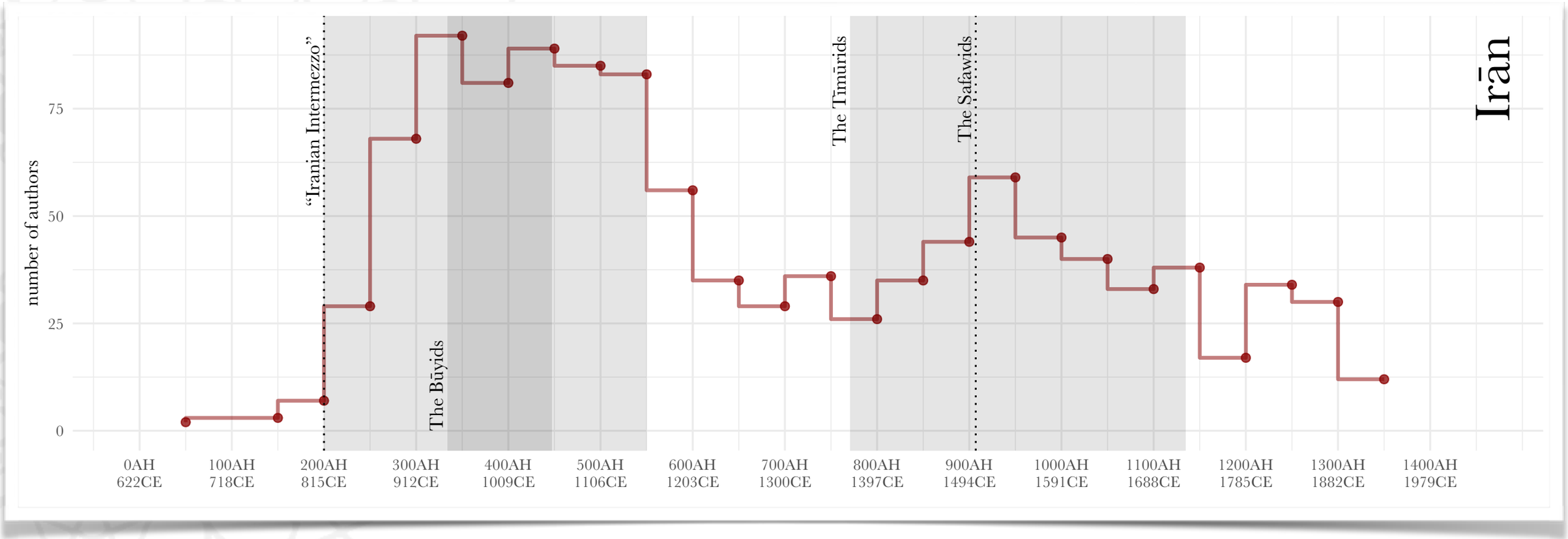
Regional Contributions

Insight I: *Cultural Production Over Time*



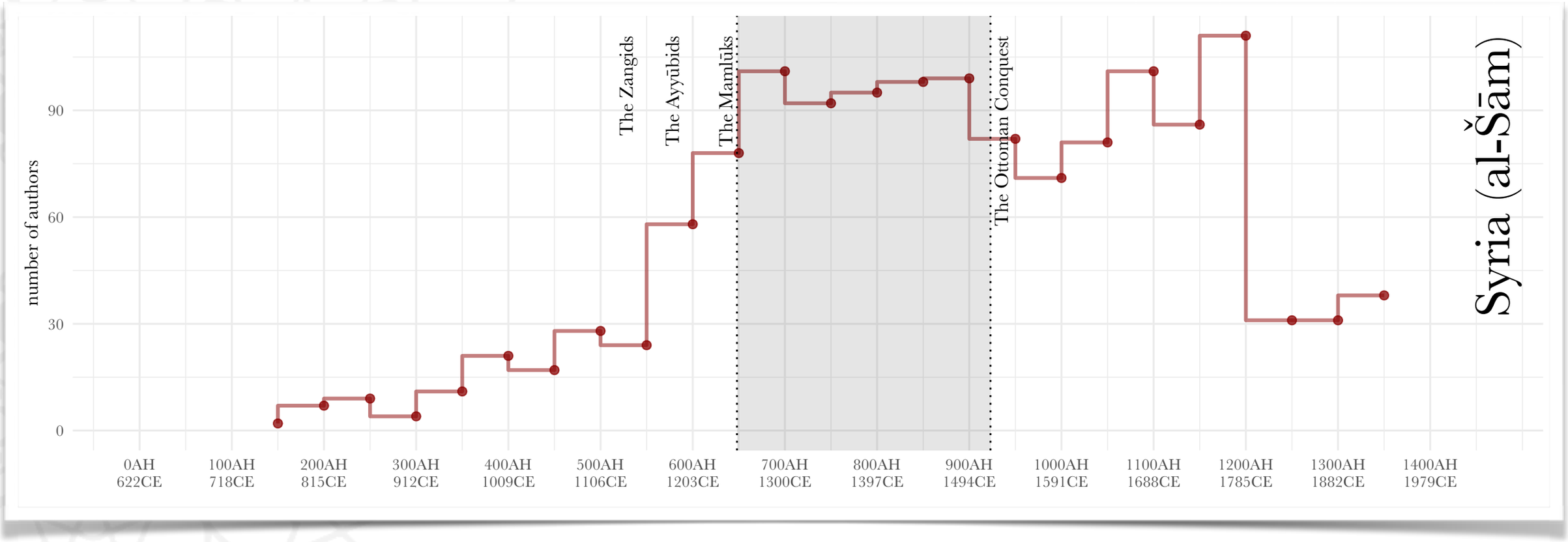
Regional Contributions

Insight I: *Cultural Production Over Time*



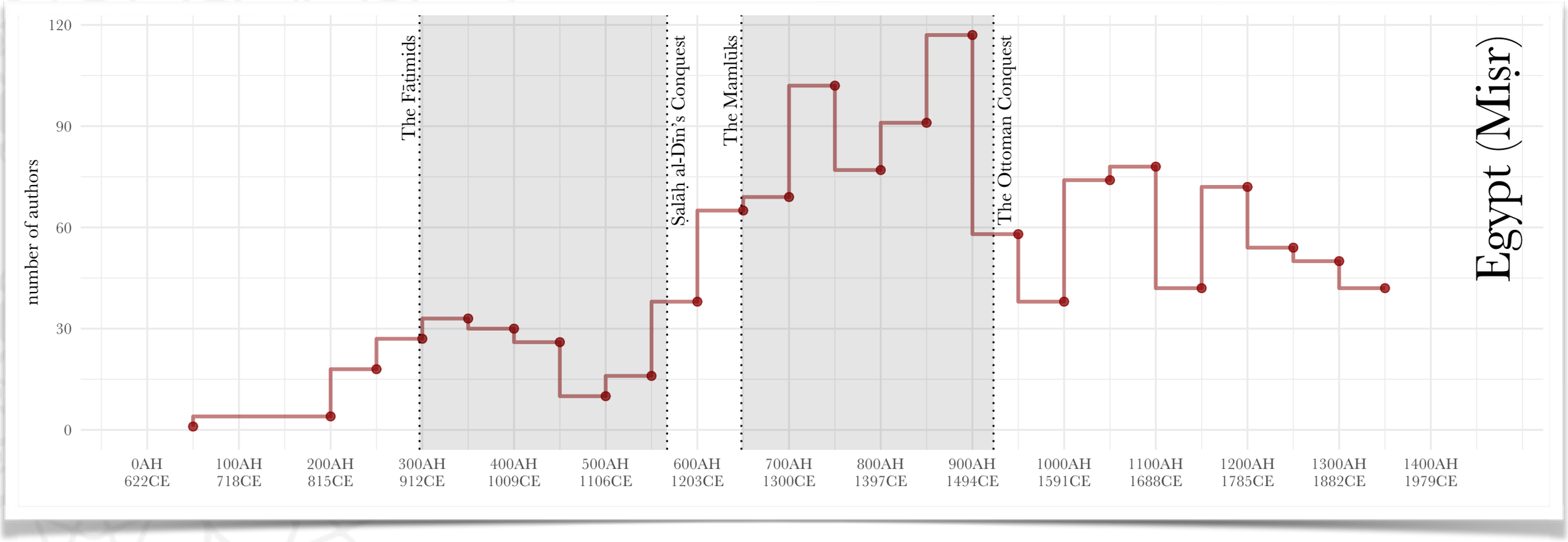
Regional Contributions

Insight I: *Cultural Production Over Time*



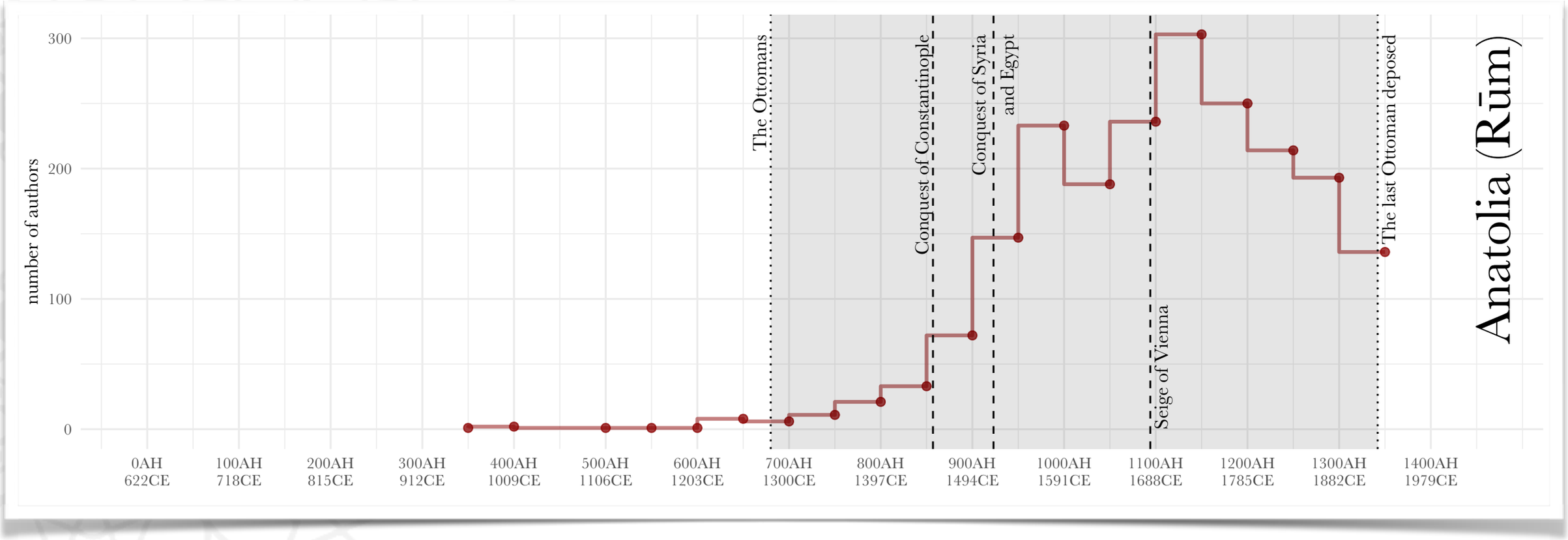
Regional Contributions

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
=====================================
```

Abstraction

Insight II: *Cultural Connections*

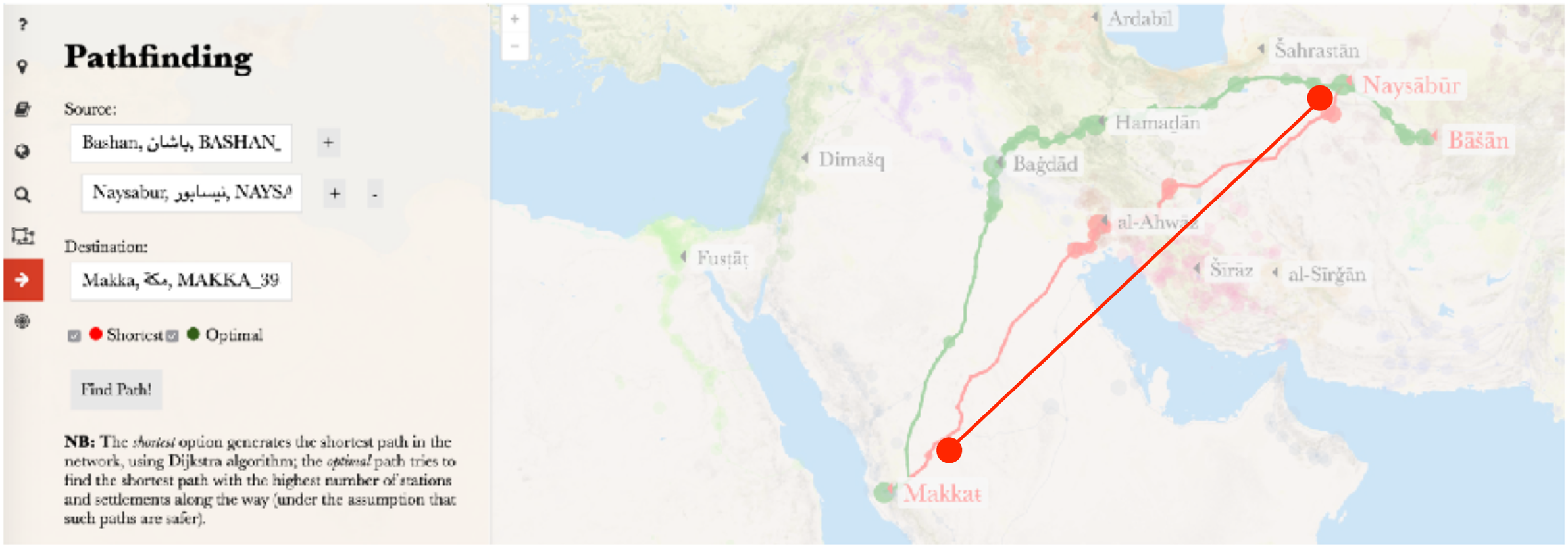
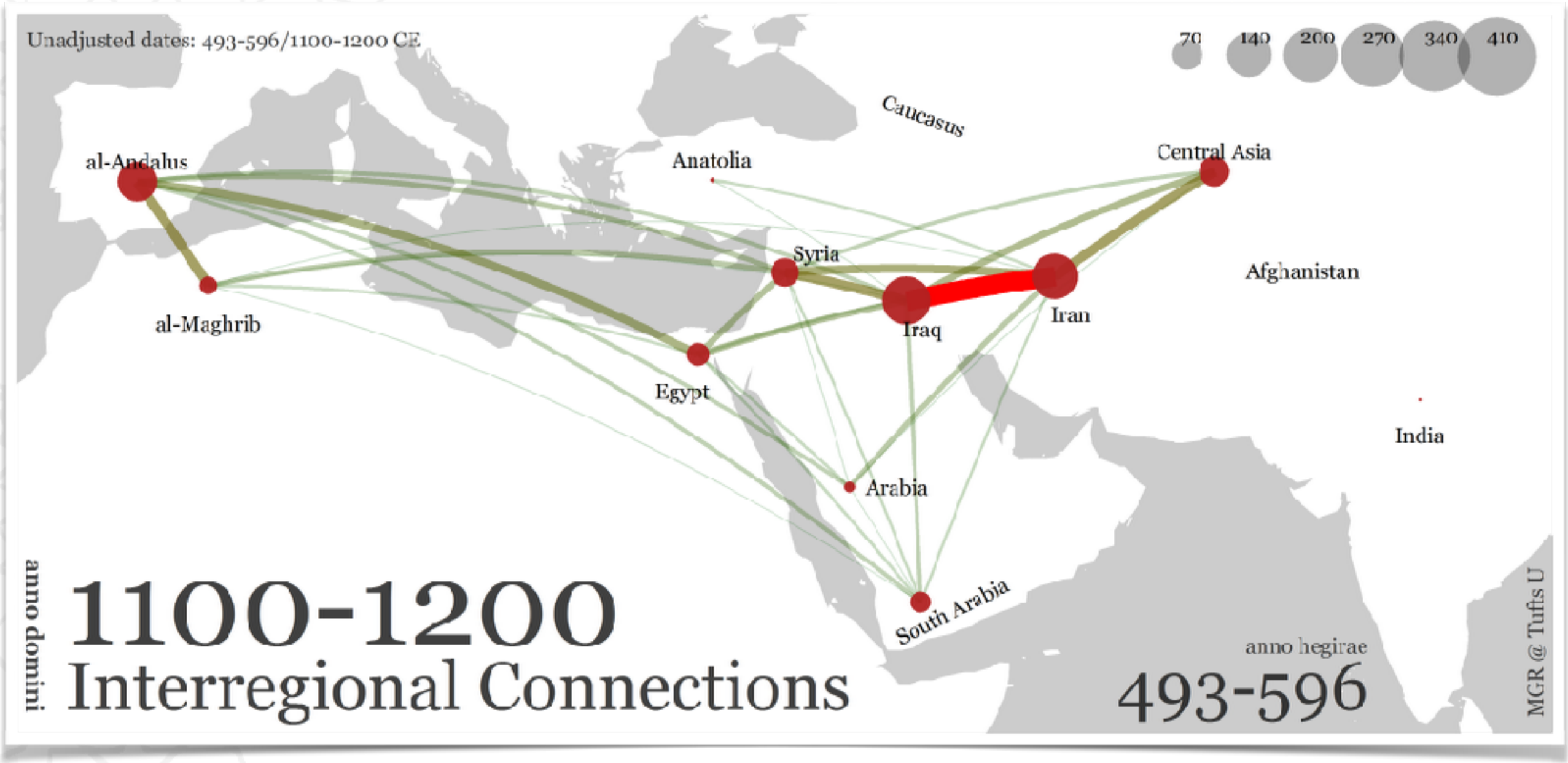


Figure 9: Geographical network of the biographee from the sample biography (using our al-Ṯurayyā Gazetteer, (`https://althurayya.github.io/`).

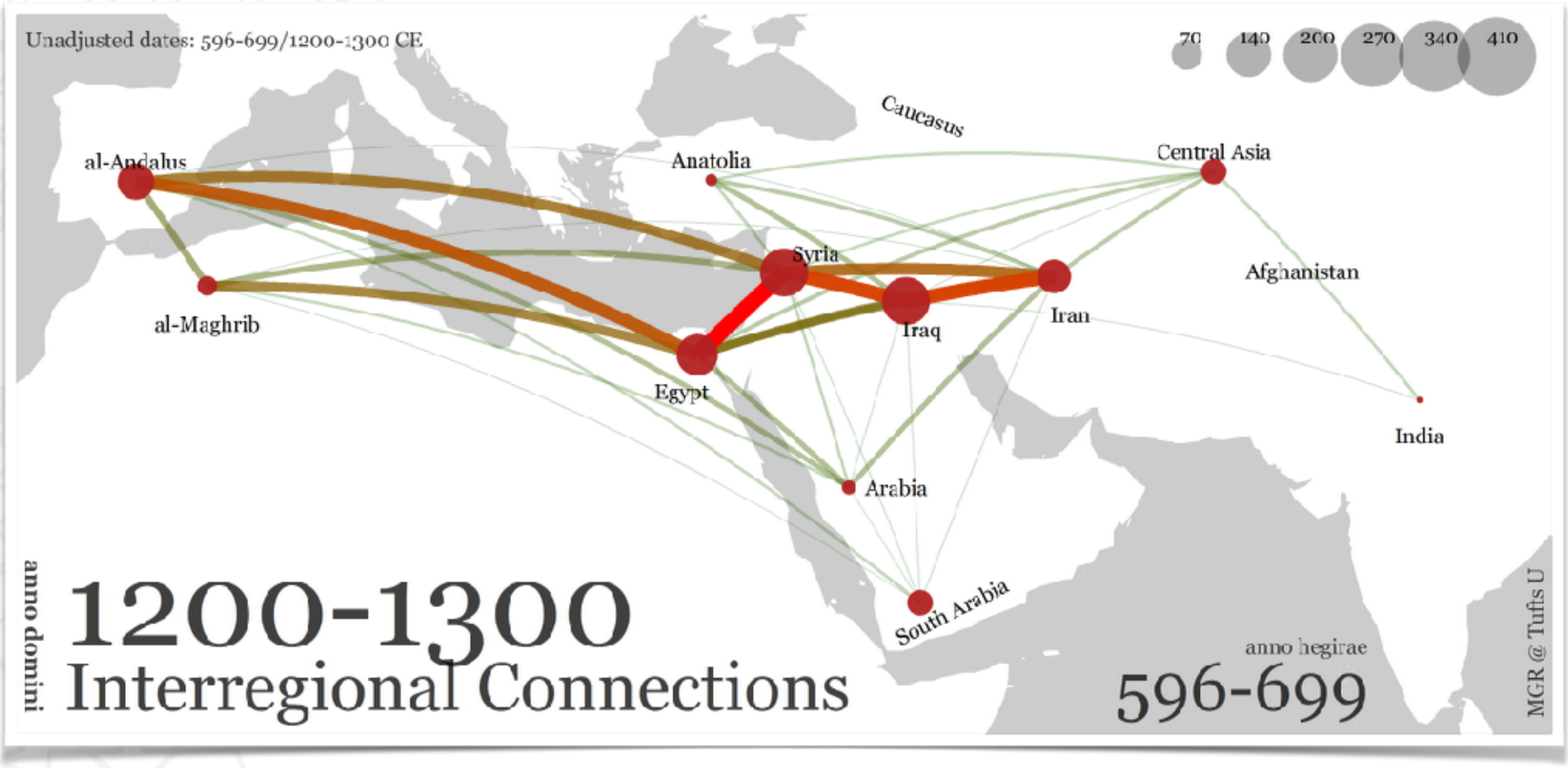
Modeling Personal Geographical Network

Insight II: *Cultural Connections*



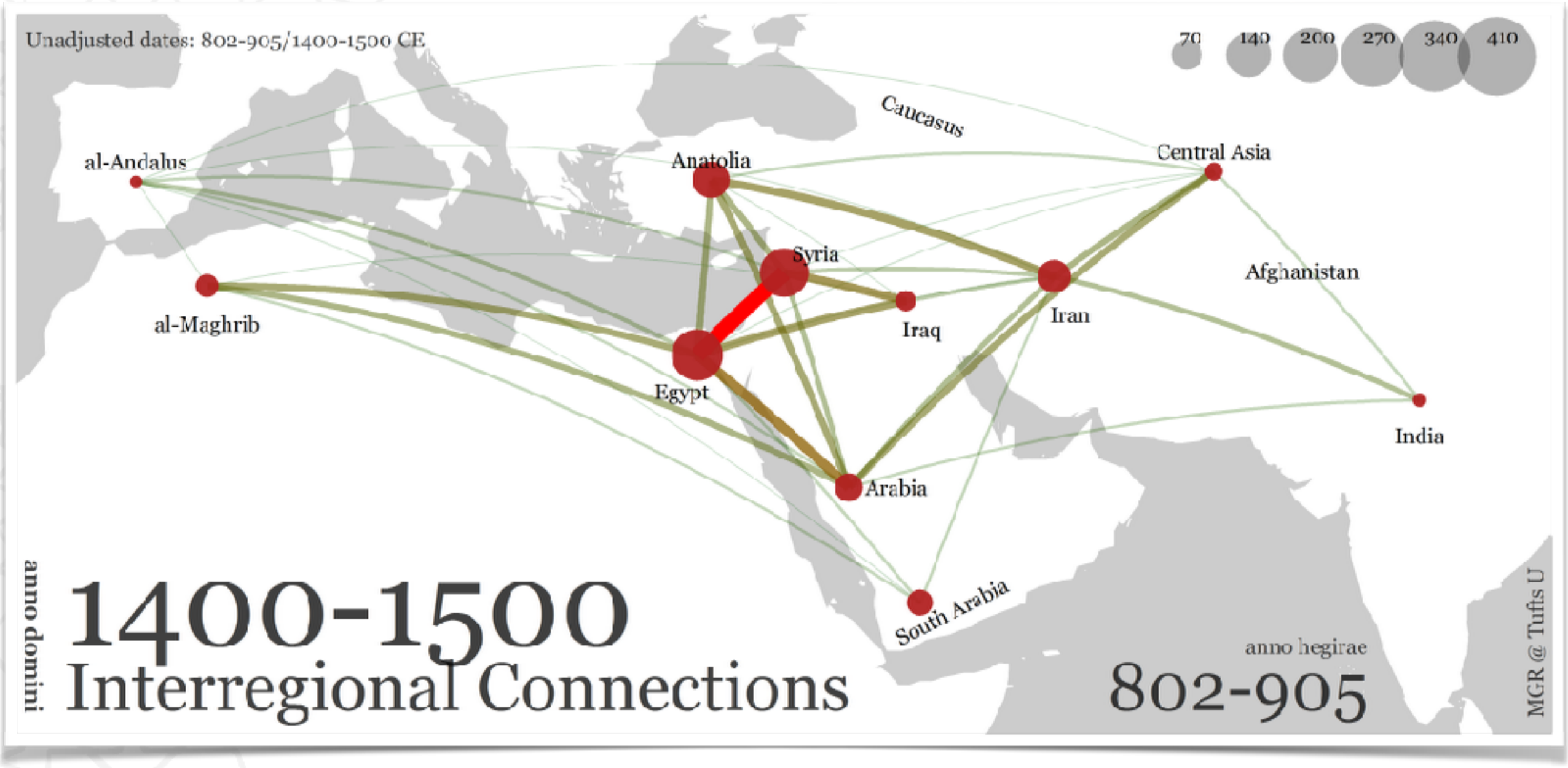
The Iraqi-Iranian Core up until the end of the 12th century CE

Insight II: *Cultural Connections*



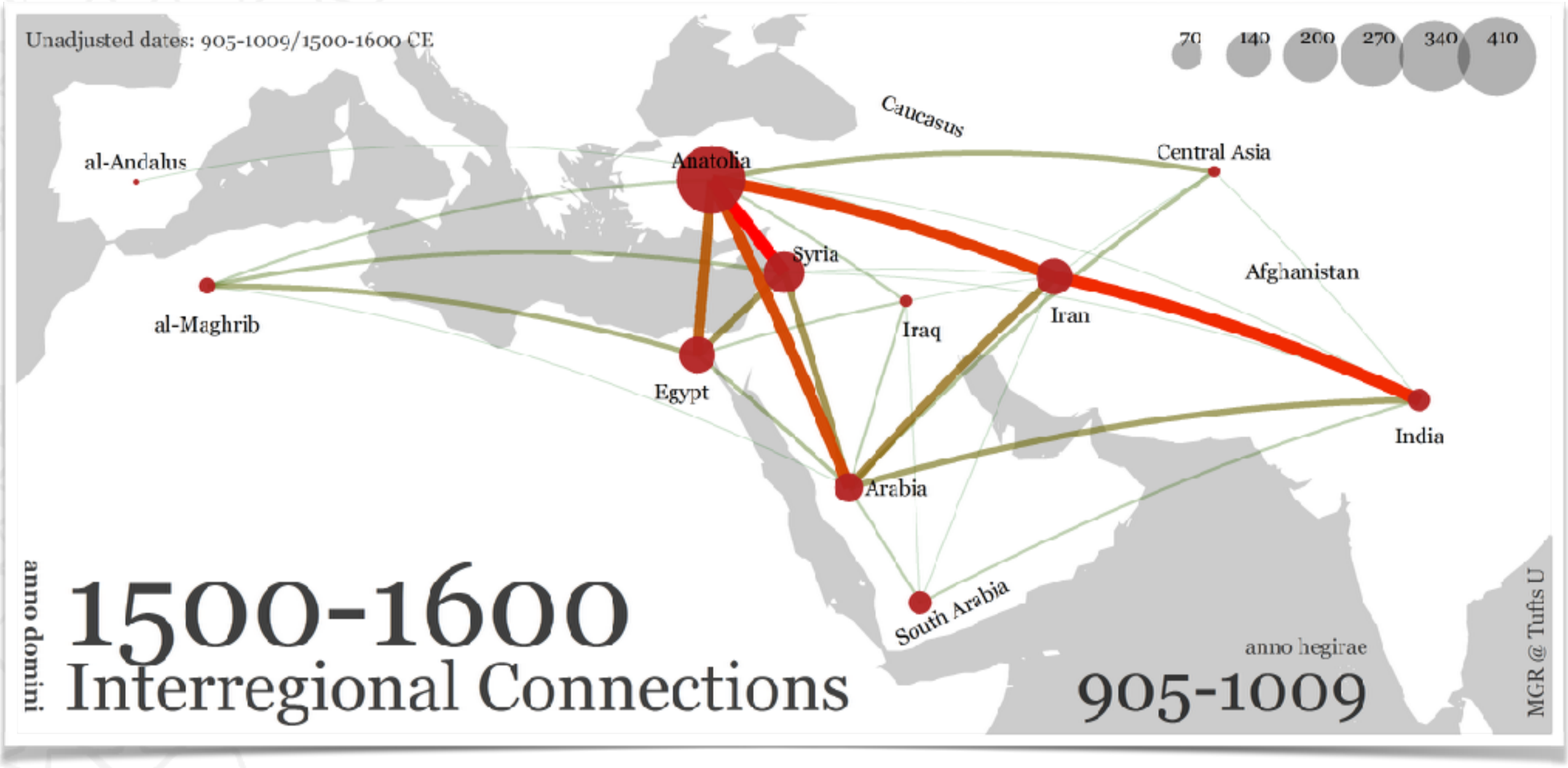
Massive migrations of the 13th century CE

Insight II: *Cultural Connections*



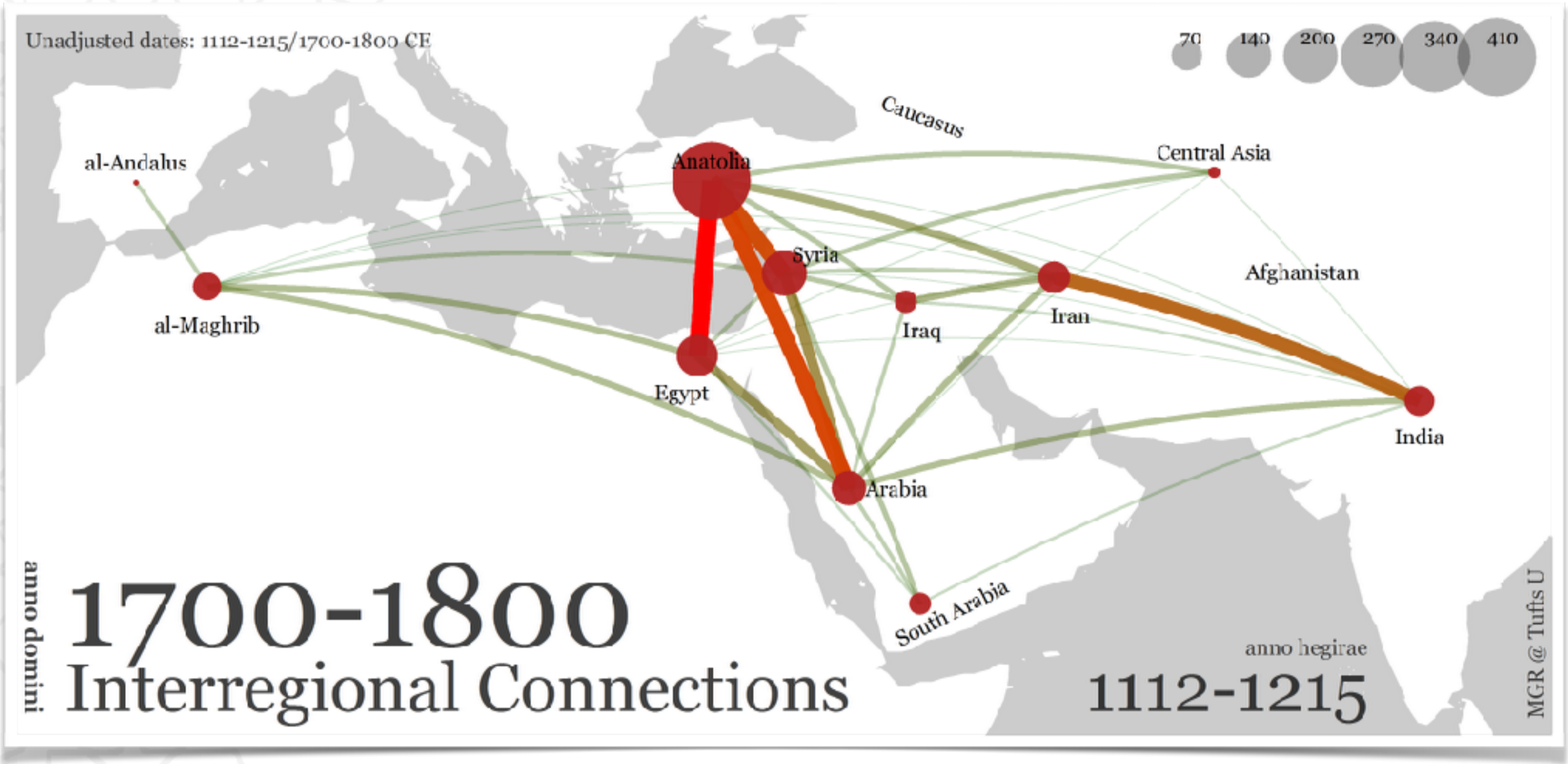
New Mamlūk Core of the 14th and 15th centuries CE

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	Download	View On
ZIP File	**TAR Ball**	**GitHub**

This project is maintained by OpenArabic

Hosted on GitHub Pages — Theme by orderedlist

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

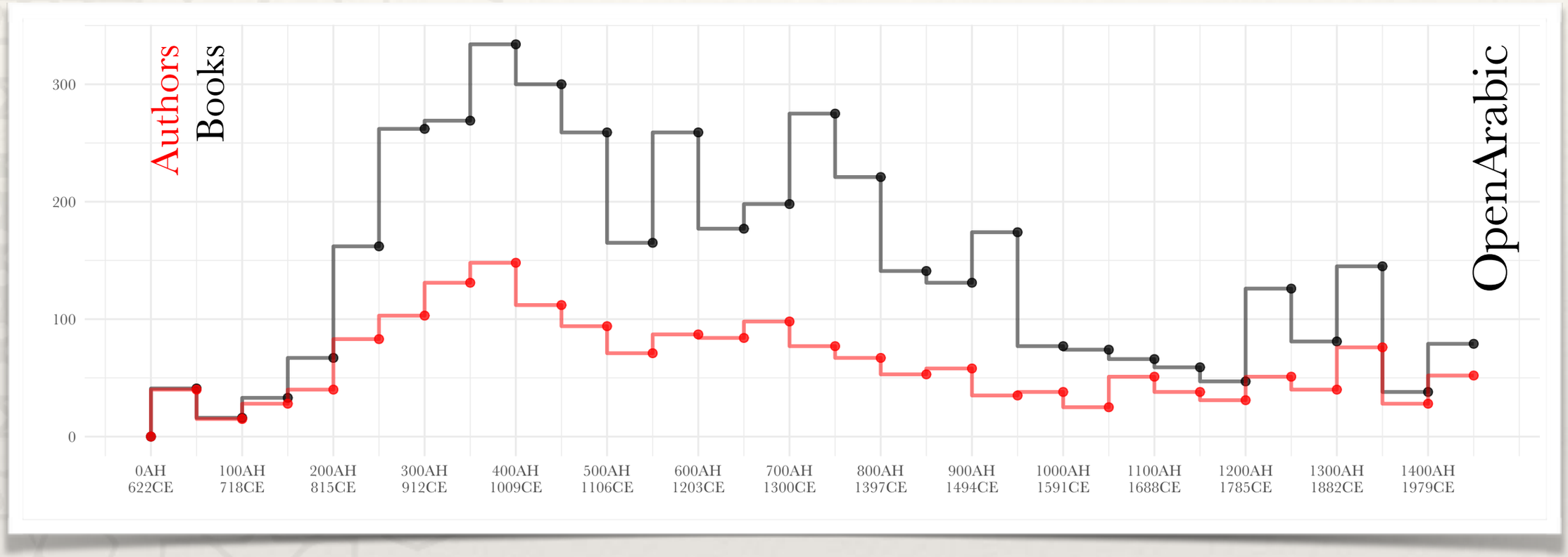
Contents

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/

OpenArabic

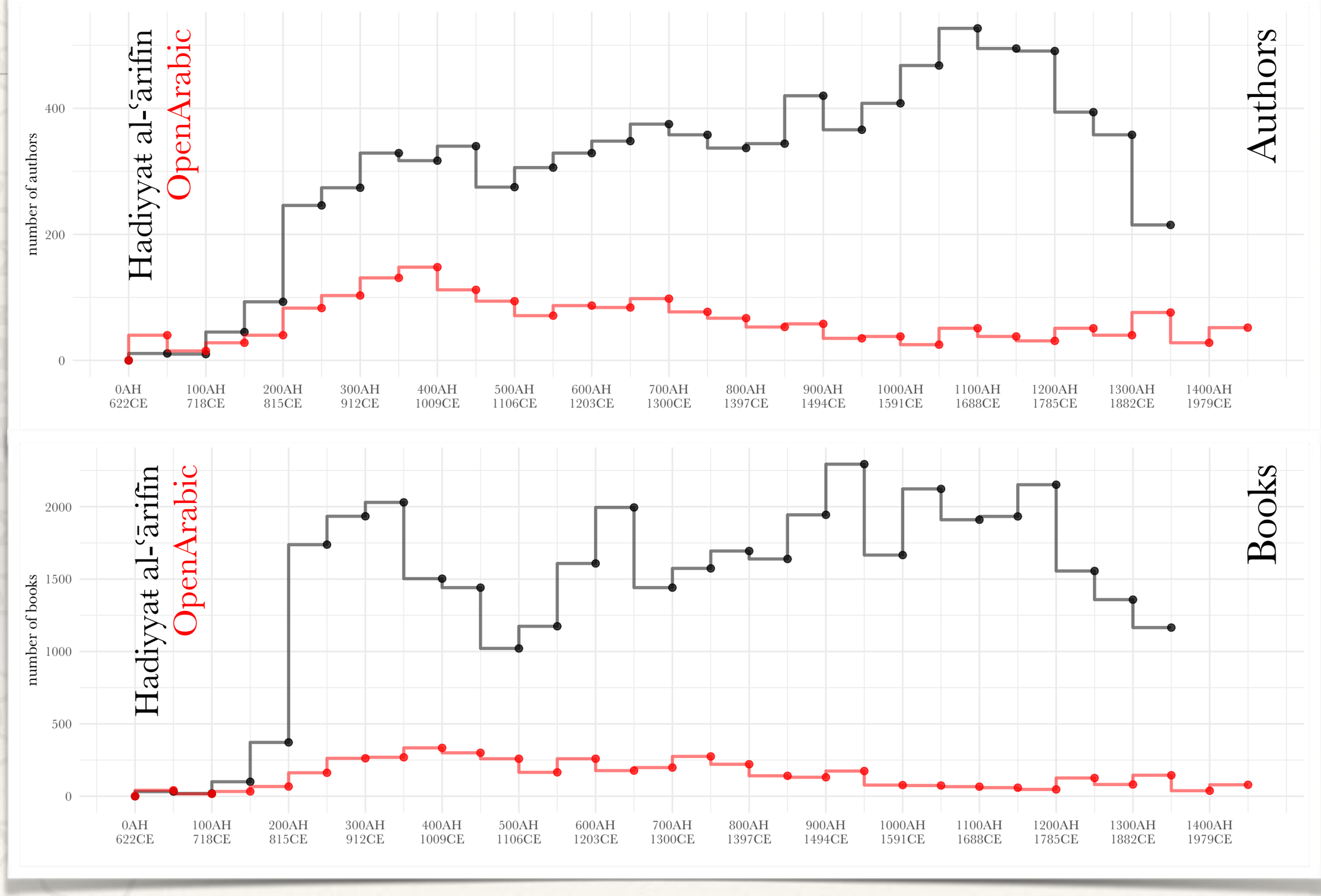


Unique Titles: 4,300
Words: 740 million
(All: 1,3 billion words)

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

https://github.com/OpenArabic/

OpenArabic: ~10%?



***Hadiyyat al-ʿārifīn*, 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

				Model accuracy level			
	Book*	**Quality**	**Type**	**Size 100**	**Ar****	**Size 200**	**Ar****
1	Ibn al-Athīr. *al-Kāmil*	*high****	***training***	93.79	97.71	93.58	97.59
2	Ibn Qutayba. *Adab al-kātib*	*high****	*testing*	82.68	95.72	80.92	94.88
3	al-Jāḥiẓ. *al-Ḥayawān*	*high****	*testing*	71.78	75.16	70.85	74.27
4	al-Yaʿqūbī. *al-Taʾrīkh*	*high****	*testing*	79.67	84.40	78.12	82.21
5	al-Dhahabī. *Taʾrīkh al-islām*	*low*****	*testing*	90.68	95.95	90.37	95.78
6	Ibn al-Jawzī. *al-Muntaẓam*	*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)*

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

OpenArabic / OpenITI: OCR

				Model accuracy level			
	Book*	**Quality**	**Type**	**Size 100**	**Ar****	**Size 200**	**Ar****
1	Ibn al-Athīr. *al-Kāmil*	*high****	*testing*	80.23	86.27	82.46	87.48
2	Ibn Qutayba. *Adab al-kātib*	*high****	*testing*	80.90	91.54	82.61	93.24
3	al-Jāḥiẓ. *al-Ḥayawān*	*high****	***training***	94.86	97.59	94.82	97.41
4	al-Yaʿqūbī. *al-Taʾrīkh*	*high****	*testing*	90.91	95.13	91.28	94.71
5	al-Dhahabī. *Taʾrīkh al-islām*	*low*****	*testing*	81.93	91.23	83.03	92.22
6	Ibn al-Jawzī. *al-Muntaẓam*	*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)*

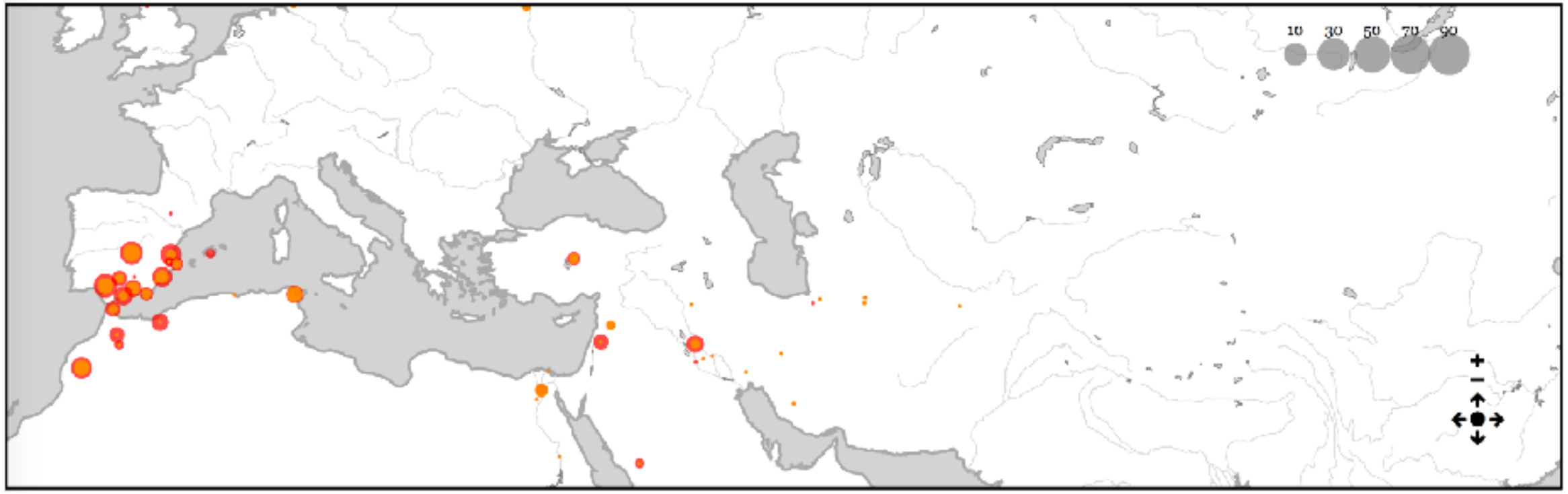
https://www.academia.edu/28923960/

Exploratorium for Biographical Data

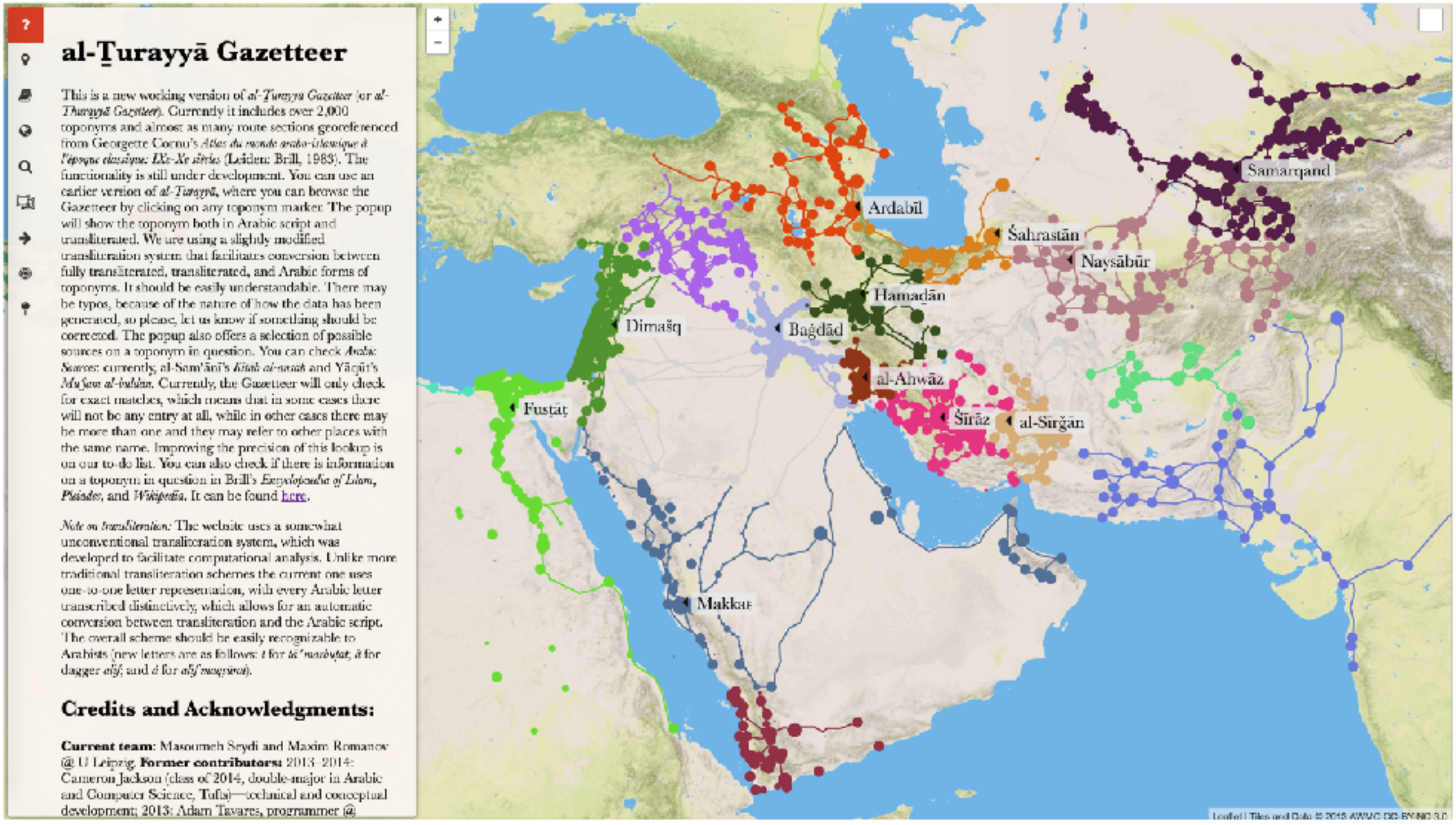
Exploratorium

The page offers an exploratory insight into groups of individuals from the "History of Islām" (*Taʾrīḫ al-islām*) of al-Ḏahabī who share common 'descriptors' (sing. *nisbat*). By selecting a 'descriptor' (الإمام), you can get preliminary analyses of geographical, chronological, and network patterns.

Space



al-Ṭurayyā Project: Gazetteer & Geospatial Model



Thank you!