Lorentz Workshop:

**Processing Ancient Text Corpora**

17-21 February 2020

Part A: Corpora Primer

# Syriac texts

*Wido van Peursen*

*Eep Talstra Centre for Bible and Computer*

www.etcbc.nl /etcbc @etcbc_vu company/etcbc/

# Workshop Overview

1. Corpora
   A. Primer
   B. In-depth discussion
2. Data structure and text model
3. Corpus analysis
4. Making knowledge accessible

ETCBC

# Part A: Corpora Primer

“In the primer (Part A) various researchers will present a snapshot of their corpora, addressing questions such as: What challenges do scholars meet when dealing with the various scripts, languages and text types? How does text model and data structure interact with the scholarly traditions and conventions in the specific disciplines. How do we deal with multiple witnesses of the same texts?”

ETCBC

# Why study Syriac texts?

Syriac is the language of a long literary tradition including poems, ancient Bible translations, apocryphal stories, historiography, Bible commentaries, philosophical tractates, scientific works. It is the language of a neglected branch of early Christianity, and one of the most important languages to study the religious and cultural context in which Islam emerged. Syriac translations served as a bridge between Greeks and Arabs, and Syriac is still the liturgical language of many Christians in the Middle East and, increasingly, in the diaspora.

# Why study Syriac texts?

**The Aramaic languages, in which Syriac takes pride of place, cover a period from the 10th cent. BC up to the present, and hence is one of the most well documented languages that allows for longitudinal linguistic analysis over three millennia.**

# Annotated corpora

# Annotated corpora

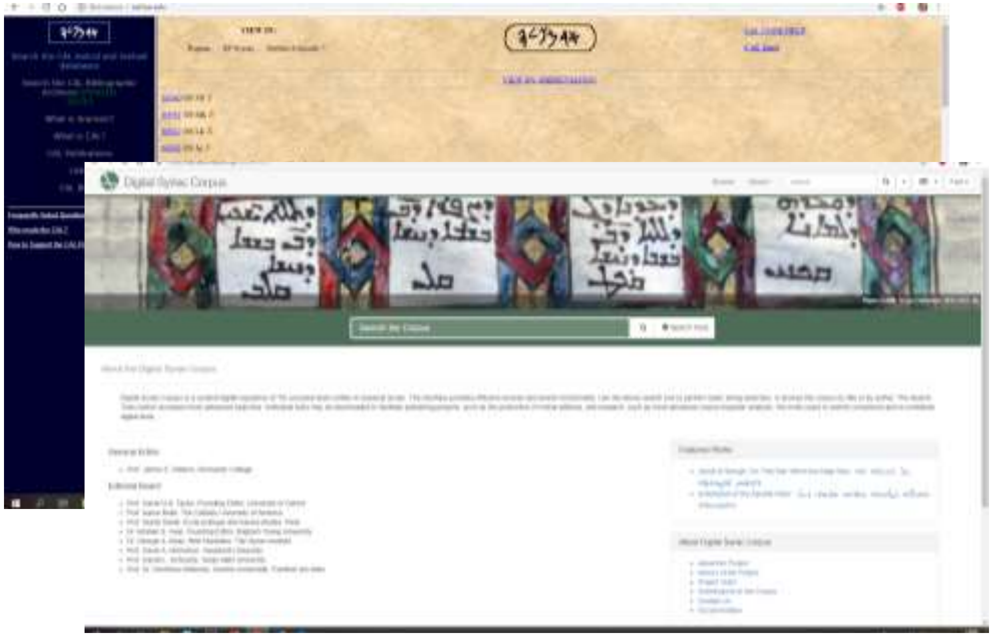
- Comprehensive Aramaic Lexicon (CAL)
- Digital Syriac Corpus
- Text-Fabric (TF)

# Annotated corpora

- Comprehensive Aramaic Lexicon (CAL)
- Digital Syriac Corpus
- Text-Fabric (TF)

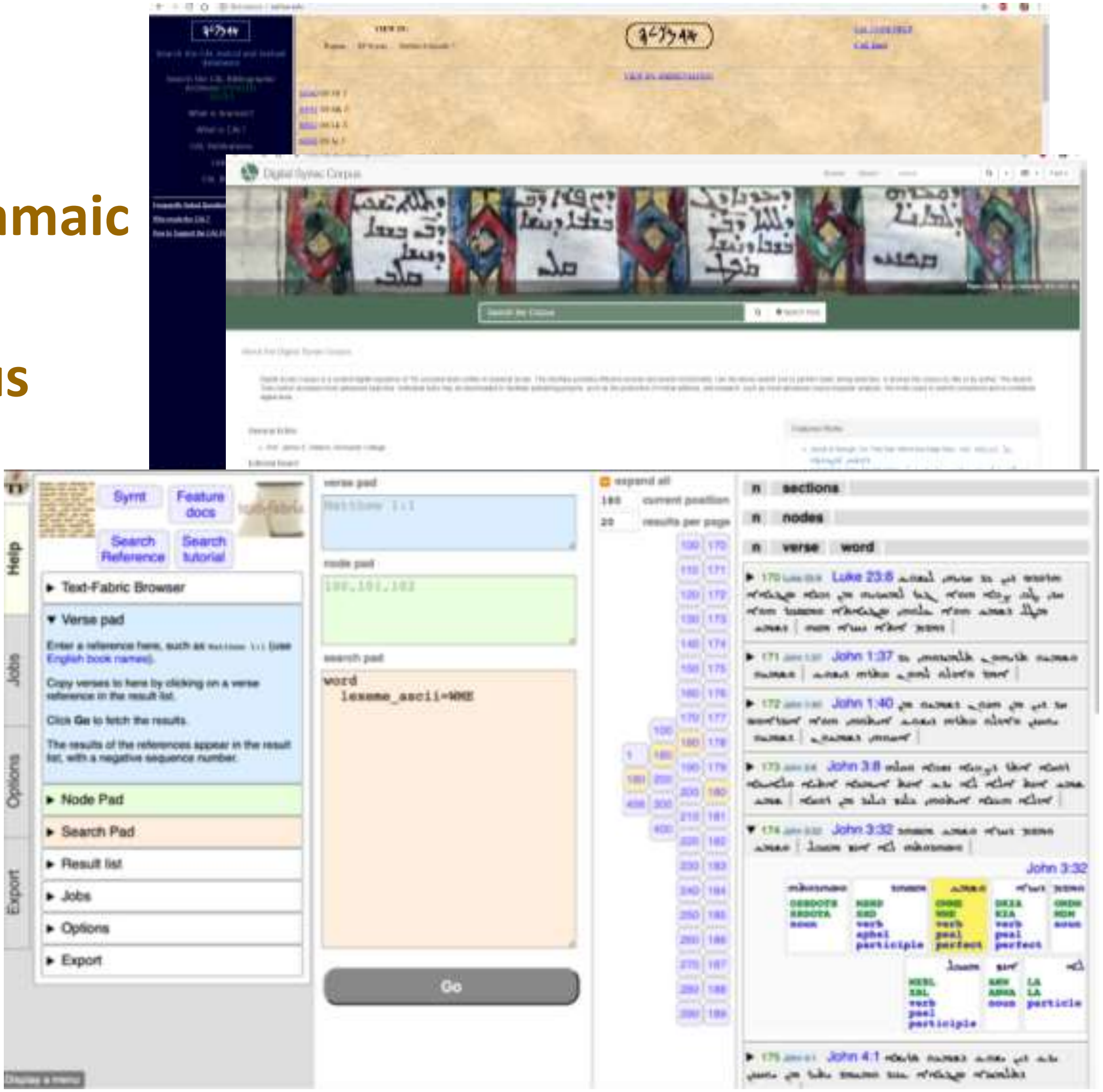
# Annotated corpora

- Comprehensive Aramaic Lexicon (CAL)
- Digital Syriac Corpus
- Text-Fabric (TF)



# Annotated corpora

- Comprehensive Aramaic Lexicon (CAL)
- Digital Syriac Corpus
- Text-Fabric (TF)



Enter a reference here, such as Matthew 1:1 (use English book names).

Copy verses to here by clicking on a verse reference in the result list.

Click **Go** to fetch the results.

The results of the references appear in the result list, with a negative sequence number.

```
word
  lexeme_ascii=HWE
```

Luke 23:8

John 1:37

John 1:40

John 3:8

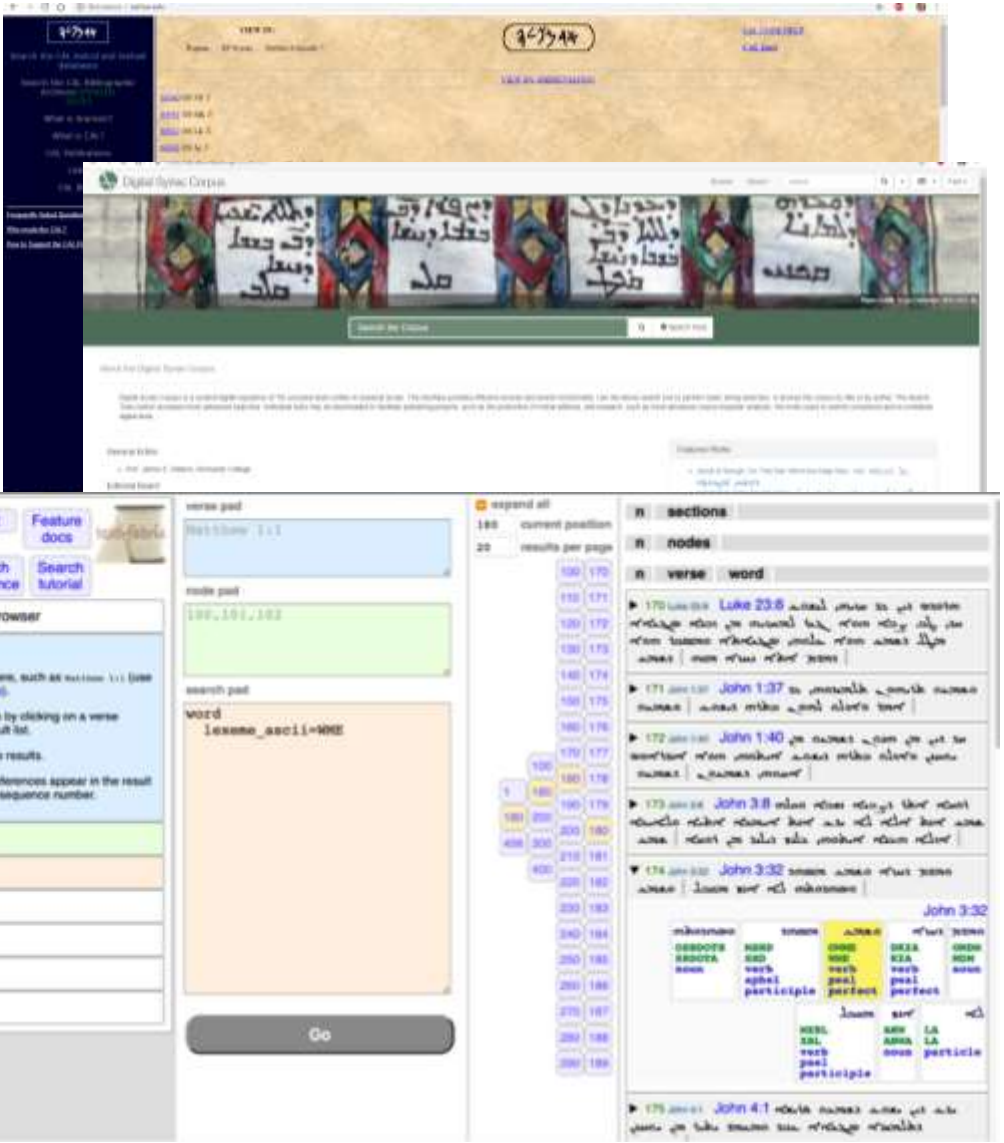
John 3:32

John 4:1

ETCBC

# Annotated corpora

- Comprehensive Aramaic Lexicon (CAL)
- Digital Syriac Corpus
- Text-Fabric (TF)



Digital Syriac Corpus

Syrnt

Feature docs

Search Reference

Search tutorial

Help

Jobs

Options

Export

▸ Text-Fabric Browser

▾ Verse pad

Enter a reference here, such as Matthew 1:1 (use English book names).

Copy verses to here by clicking on a verse reference in the result list.

Click Go to fetch the results.

The results of the references appear in the result list, with a negative sequence number.

▸ Node Pad

▸ Search Pad

▸ Result list

▸ Jobs

▸ Options

▸ Export

verse pad

Matthew 1:1

node pad

100,101,102

search pad

```
word
  lexeme_ascii=MLK
```

Go

expand all

100 current position

20 results per page

n sections

n nodes

n verse word

▸ 170 Luke 23:8

▸ 171 John 1:37

▸ 172 John 1:40

▸ 173 John 3:8

▾ 174 John 3:32

John 3:32

verb

peal

perfect

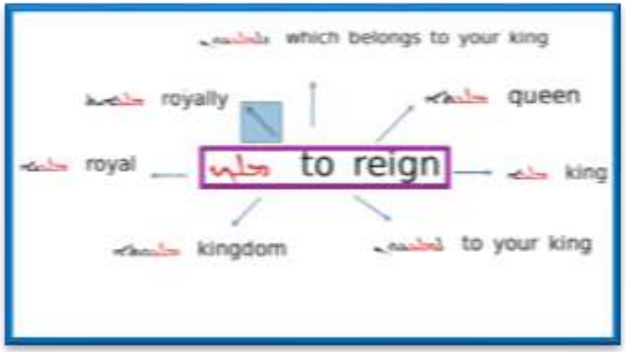
participle

noun

particle

▸ 175 John 4:1

- Parsing!



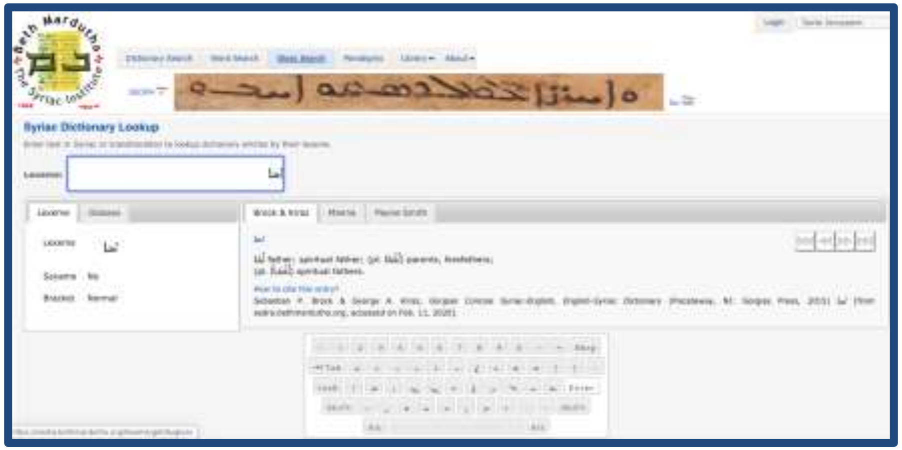
# Structured data

# Structured data

- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data

# Structured data

- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data



Syriac Dictionary Lookup

Lexeme: ܐܒܐ

Lexeme | Glosses

Lexeme ܐܒܐ

Scheme No

Bracket Normal

Brock & Kiraz | Moshe | Payne Smith

ܐܒܐ

ܐܒܐ father; spiritual father; (pl. ܐܒܗܐ) parents, forefathers;
(pl. ܐܒܗ̈ܬܐ) spiritual fathers.

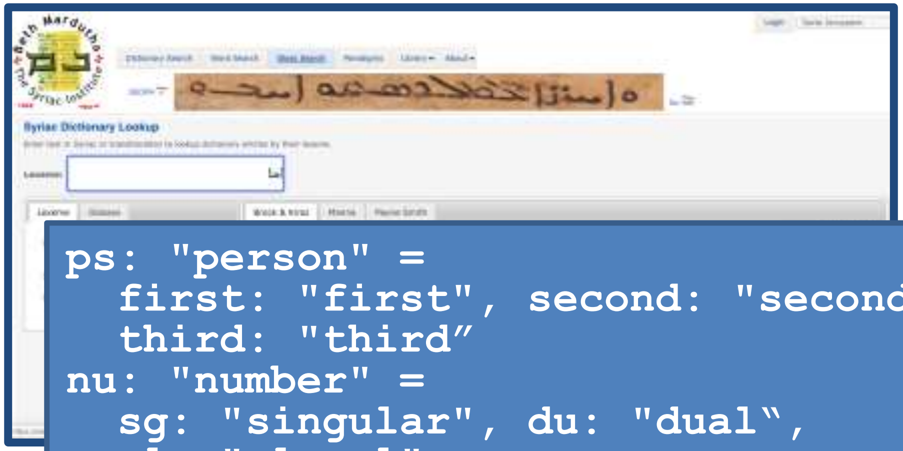
How to cite this entry?

Sebastian P. Brock & George A. Kiraz. Gorgias Concise Syriac-English, English-Syriac Dictionary (Piscataway, NJ: Gorgias Press, 2015) ܐܒܐ (from sedra.bethmardutho.org, accessed on Feb. 11, 2020).

ETCBC

# Structured data

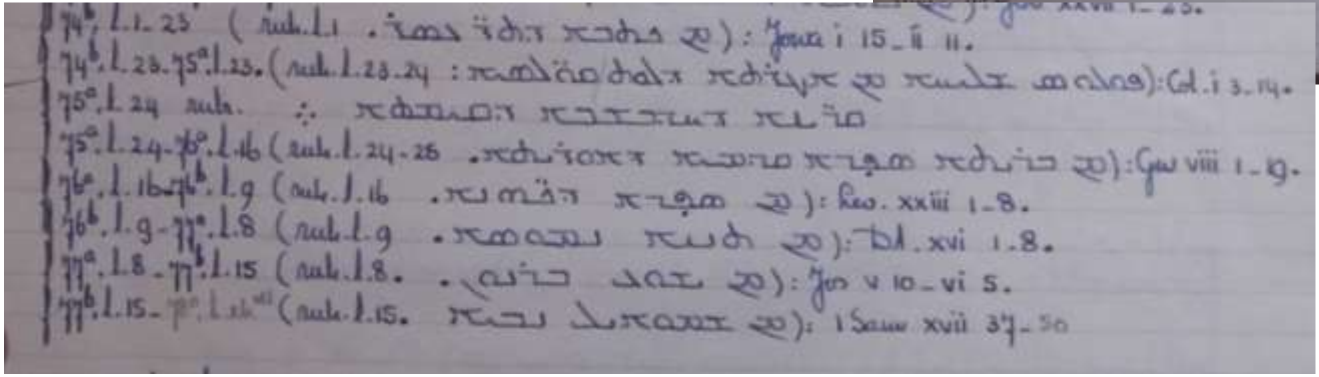
- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data



```
ps: "person" =
   first: "first", second: "second",
   third: "third”
nu: "number" =
   sg: "singular", du: "dual‶,
   pl: "plural",
   unknown: "unknown”
gn: "gender" =
   f: "feminine", m: "masculine"
```

# Structured data

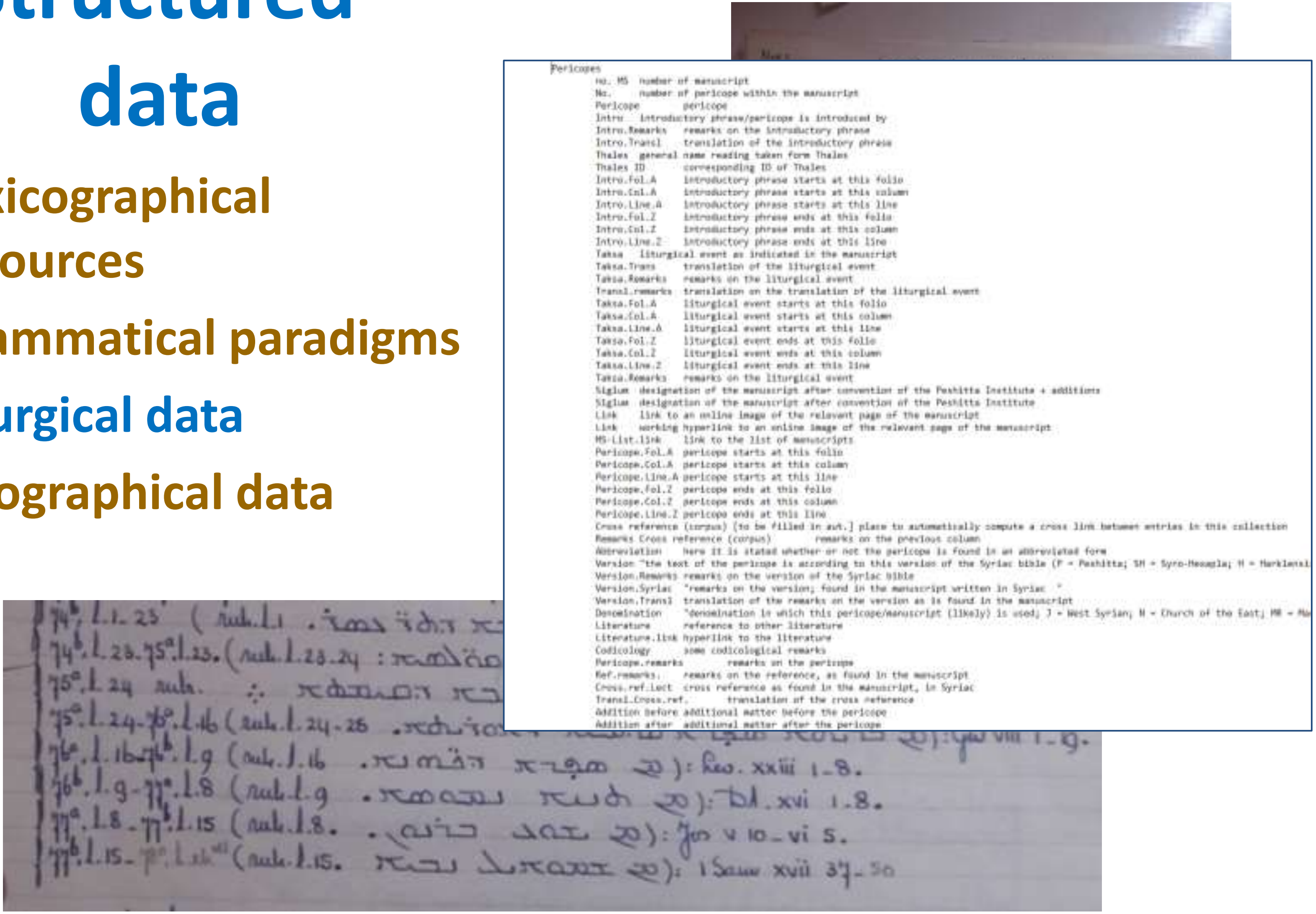
- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data



74b. l.1-23 (rub. l.1 . ܡܢ ܟܬܒܐ ܕܝܫܘܥ ܒܪ ܢܘܢ): Josua i 15 - ii 11.
74b. l.23 - 75a. l.23. (rub. l.23.24 : [illegible]): Col. i 3-14.
75a. l.24 rub. ∴ [illegible]
75a. l.24 - 76a. l.16 (rub. l.24-26 . [illegible]): Gen viii 1-19.
76a. l.16 - 76b. l.9 (rub. l.16 . [illegible]): Reg. xxiii 1-8.
76b. l.9 - 77a. l.8 (rub. l.9 . [illegible]): Jd. xvi 1-8.
77a. l.8 - 77b. l.15 (rub. l.8. . [illegible]): Jos v 10 - vi 5.
77b. l.15 - 78a. l.[illegible] (rub. l.15. [illegible]): 1 Sam xvii 37-50

# Structured data

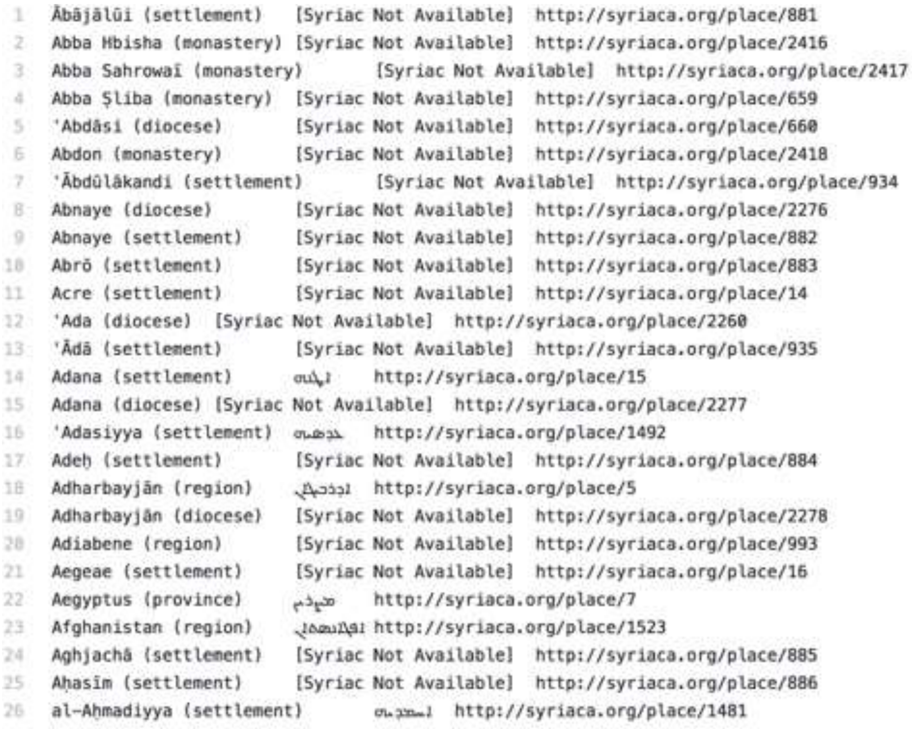
- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data



```
Pericopes
        no. MS  number of manuscript
        No.     number of pericope within the manuscript
        Pericope        pericope
        Intro   introductory phrase/pericope is introduced by
        Intro.Remarks   remarks on the introductory phrase
        Intro.Transl    translation of the introductory phrase
        Thales  general name reading taken form Thales
        Thales ID       corresponding ID of Thales
        Intro.Fol.A     introductory phrase starts at this folio
        Intro.Col.A     introductory phrase starts at this column
        Intro.Line.A    introductory phrase starts at this line
        Intro.Fol.Z     introductory phrase ends at this folio
        Intro.Col.Z     introductory phrase ends at this column
        Intro.Line.Z    introductory phrase ends at this line
        Taksa   liturgical event as indicated in the manuscript
        Taksa.Trans     translation of the liturgical event
        Taksa.Remarks   remarks on the liturgical event
        Transl.remarks  translation on the translation of the liturgical event
        Taksa.Fol.A     liturgical event starts at this folio
        Taksa.Col.A     liturgical event starts at this column
        Taksa.Line.A    liturgical event starts at this line
        Taksa.Fol.Z     liturgical event ends at this folio
        Taksa.Col.Z     liturgical event ends at this column
        Taksa.Line.Z    liturgical event ends at this line
        Taksa.Remarks   remarks on the liturgical event
        Siglum  designation of the manuscript after convention of the Peshitta Institute + additions
        Siglum  designation of the manuscript after convention of the Peshitta Institute
        Link    link to an online image of the relevant page of the manuscript
        Link    working hyperlink to an online image of the relevant page of the manuscript
        MS-List.link    link to the list of manuscripts
        Pericope.Fol.A  pericope starts at this folio
        Pericope.Col.A  pericope starts at this column
        Pericope.Line.A pericope starts at this line
        Pericope.Fol.Z  pericope ends at this folio
        Pericope.Col.Z  pericope ends at this column
        Pericope.Line.Z pericope ends at this line
        Cross reference (corpus) [to be filled in aut.] place to automatically compute a cross link between entries in this collection
        Remarks Cross reference (corpus)        remarks on the previous column
        Abbreviation    here it is stated whether or not the pericope is found in an abbreviated form
        Version "the text of the pericope is according to this version of the Syriac bible (P = Peshitta; SH = Syro-Hexapla; H = Harklensi
        Version.Remarks remarks on the version of the Syriac bible
        Version.Syriac  "remarks on the version; found in the manuscript written in Syriac  "
        Version.Transl  translation of the remarks on the version as is found in the manuscript
        Denomination    "denomination in which this pericope/manuscript (likely) is used; J = West Syrian; N = Church of the East; MR = Ma
        Literature      reference to other literature
        Literature.link hyperlink to the literature
        Codicology      some codicological remarks
        Pericope.remarks        remarks on the pericope
        Ref.remarks.    remarks on the reference, as found in the manuscript
        Cross.ref.Lect  cross reference as found in the manuscript, in Syriac
        Transl.Cross.ref.       translation of the cross reference
        Addition before additional matter before the pericope
        Addition after  additional matter after the pericope
```

# Structured data

- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data

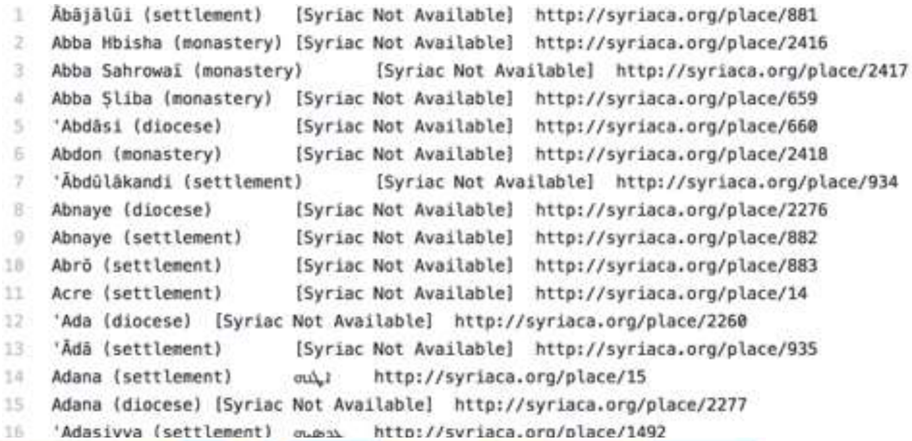


```
Ābājālūi (settlement)	[Syriac Not Available]	http://syriaca.org/place/881
Abba Hbisha (monastery)	[Syriac Not Available]	http://syriaca.org/place/2416
Abba Sahrowaī (monastery)	[Syriac Not Available]	http://syriaca.org/place/2417
Abba Ṣliba (monastery)	[Syriac Not Available]	http://syriaca.org/place/659
'Abdāsi (diocese)	[Syriac Not Available]	http://syriaca.org/place/660
Abdon (monastery)	[Syriac Not Available]	http://syriaca.org/place/2418
'Ābdūlākandi (settlement)	[Syriac Not Available]	http://syriaca.org/place/934
Abnaye (diocese)	[Syriac Not Available]	http://syriaca.org/place/2276
Abnaye (settlement)	[Syriac Not Available]	http://syriaca.org/place/882
Abrō (settlement)	[Syriac Not Available]	http://syriaca.org/place/883
Acre (settlement)	[Syriac Not Available]	http://syriaca.org/place/14
'Ada (diocese)	[Syriac Not Available]	http://syriaca.org/place/2260
'Ādā (settlement)	[Syriac Not Available]	http://syriaca.org/place/935
Adana (settlement)	ܐܕܢܐ	http://syriaca.org/place/15
Adana (diocese)	[Syriac Not Available]	http://syriaca.org/place/2277
'Adasiyya (settlement)	ܥܕܣܝܐ	http://syriaca.org/place/1492
Adeḥ (settlement)	[Syriac Not Available]	http://syriaca.org/place/884
Adharbayjān (region)	ܐܕܪܒܝܓܢ	http://syriaca.org/place/5
Adharbayjān (diocese)	[Syriac Not Available]	http://syriaca.org/place/2278
Adiabene (region)	[Syriac Not Available]	http://syriaca.org/place/993
Aegeae (settlement)	[Syriac Not Available]	http://syriaca.org/place/16
Aegyptus (province)	ܡܨܪܝܢ	http://syriaca.org/place/7
Afghanistan (region)	ܐܦܓܢܣܛܢ	http://syriaca.org/place/1523
Aghjachā (settlement)	[Syriac Not Available]	http://syriaca.org/place/885
Aḥasīm (settlement)	[Syriac Not Available]	http://syriaca.org/place/886
al-Aḥmadiyya (settlement)	ܐܚܡܕܝܐ	http://syriaca.org/place/1481
```

ETCBC

# Structured data

- **Lexicographical resources**
- **Grammatical paradigms**
- **Liturgical data**
- **Geographical data**

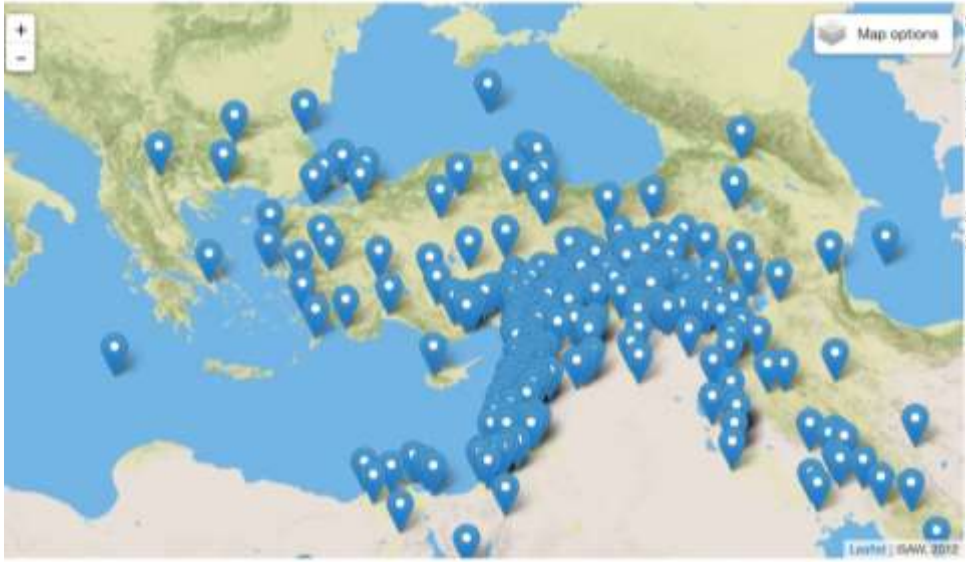


```
Ābājālûi (settlement)   [Syriac Not Available]  http://syriaca.org/place/881
Abba Hbisha (monastery) [Syriac Not Available]  http://syriaca.org/place/2416
Abba Sahrowaī (monastery)       [Syriac Not Available]  http://syriaca.org/place/2417
Abba Ṣliba (monastery)  [Syriac Not Available]  http://syriaca.org/place/659
'Abdāsi (diocese)       [Syriac Not Available]  http://syriaca.org/place/660
Abdon (monastery)       [Syriac Not Available]  http://syriaca.org/place/2418
'Ābdûlākandi (settlement)       [Syriac Not Available]  http://syriaca.org/place/934
Abnaye (diocese)        [Syriac Not Available]  http://syriaca.org/place/2276
Abnaye (settlement)     [Syriac Not Available]  http://syriaca.org/place/882
Abrō (settlement)       [Syriac Not Available]  http://syriaca.org/place/883
Acre (settlement)       [Syriac Not Available]  http://syriaca.org/place/14
'Ada (diocese)  [Syriac Not Available]  http://syriaca.org/place/2260
'Ādā (settlement)       [Syriac Not Available]  http://syriaca.org/place/935
Adana (settlement)      ܐܕܢܐ    http://syriaca.org/place/15
Adana (diocese) [Syriac Not Available]  http://syriaca.org/place/2277
'Adasiyya (settlement)  [illegible]     http://syriaca.org/place/1492
e/884

e/2278
e/993
e/16



e/885
e/886
```



ETCBC

# Structured data

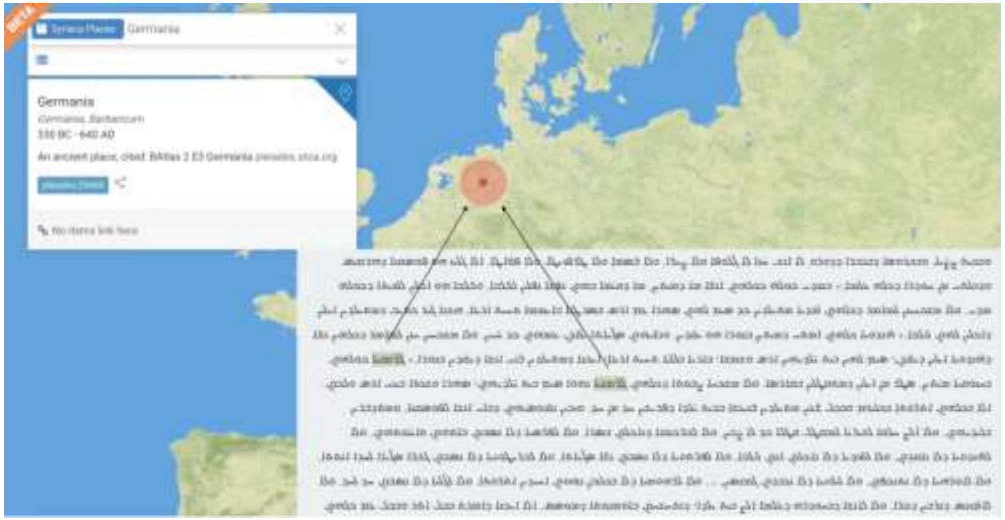
- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data

- Combining textual data and structured data

# Structured data

- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data

- Combining textual data and structured data



Germania

*Germania, Barbaricum*

330 BC - 640 AD

An ancient place, cited: BAtlas 2 E3 Germania pleiades.stoa.org

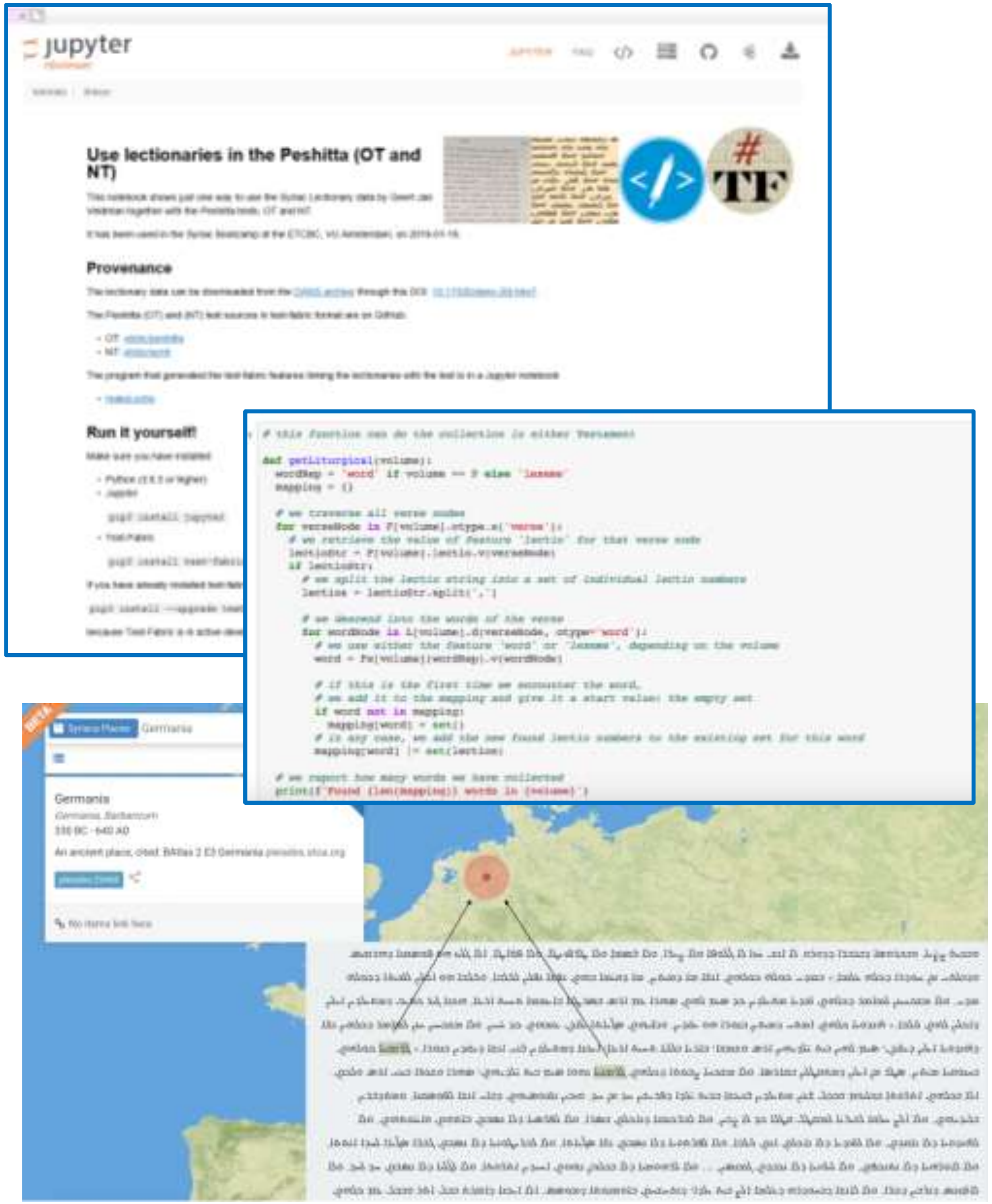
ETCBC

# Structured data

- Lexicographical resources
- Grammatical paradigms
- Liturgical data
- Geographical data

↓

- Combining textual data and structured data



jupyter

## Use lectionaries in the Peshitta (OT and NT)

This notebook shows just one way to use the Syriac Lectionary data by Geert Jan Veldman together with the Peshitta texts, OT and NT.

It has been used in the Syriac Bootcamp at the ETCBC, VU Amsterdam, on 2019-01-16.

### Provenance

The lectionary data can be downloaded from the DANS archive through this DOI: 10.17026/dans-26t-hhv7

The Peshitta (OT) and (NT) text sources in text-fabric format are on GitHub:

- OT: etcbc/peshitta
- NT: etcbc/syrnt

The program that generated the text-fabric features linking the lectionaries with the text is in a Jupyter notebook:

- makeLectio

### Run it yourself!

Make sure you have installed

- Python (3.6.3 or higher)
- Jupyter

```
pip3 install jupyter
```

- Text-Fabric

```
pip3 install text-fabric
```

If you have already installed text-fabric before

```
pip3 install --upgrade text-fabric
```

because Text-Fabric is in active development

```
# this function can do the collection in either Testament

def getLiturgical(volume):
  wordRep = 'word' if volume == 'NT' else 'lexeme'
  mapping = {}

  # we traverse all verse nodes
  for verseNode in F[volume].otype.s('verse'):
    # we retrieve the value of feature 'lectio' for that verse node
    lectioStr = F[volume].lectio.v(verseNode)
    if lectioStr:
      # we split the lectio string into a set of individual lectio numbers
      lectios = lectioStr.split(',')

      # we descend into the words of the verse
      for wordNode in L[volume].d(verseNode, otype='word'):
        # we use either the feature 'word' or 'lexeme', depending on the volume
        word = Fs[volume](wordRep).v(wordNode)

        # if this is the first time we encounter the word,
        # we add it to the mapping and give it a start value: the empty set
        if word not in mapping:
          mapping[word] = set()
        # in any case, we add the new found lectio numbers to the existing set for this word
        mapping[word] |= set(lectios)

  # we report how many words we have collected
  print(f'Found {len(mapping)} words in {volume}')
```

Germania

Germania, Barbaricum

330 BC - 640 AD

An ancient place, cited: BAtlas 2 E3 Germania pleiades.stoa.org

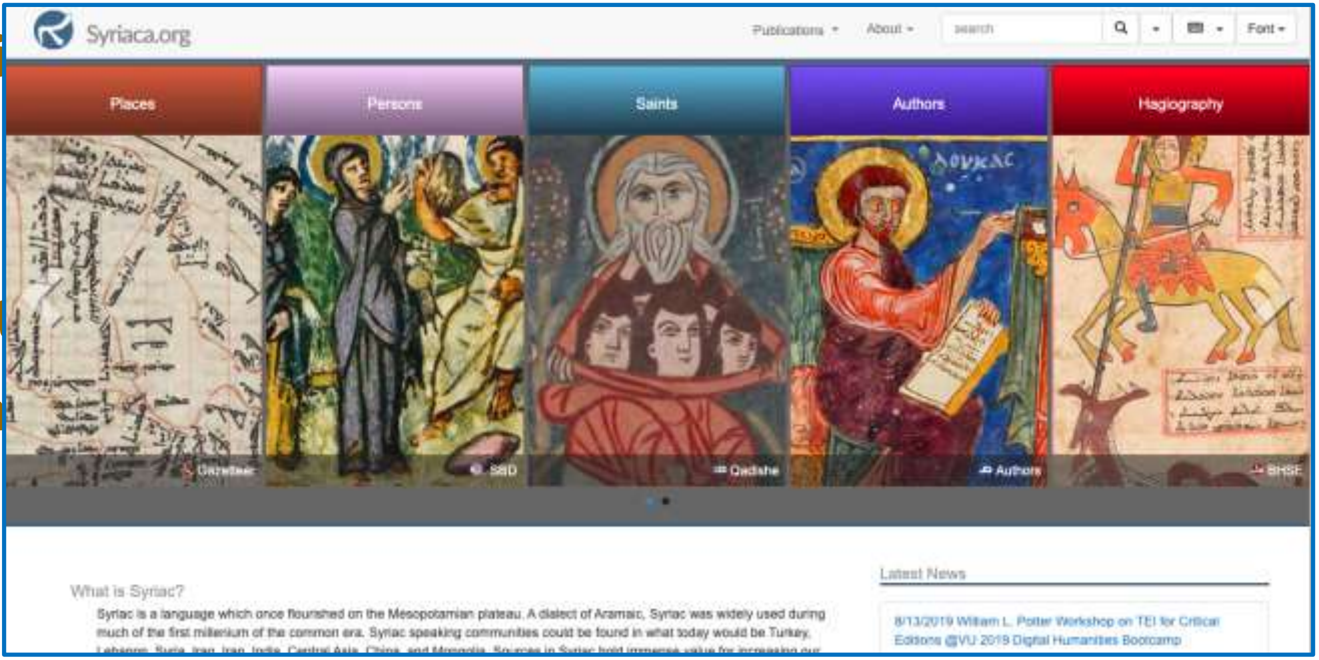
No items link here

ETCBC

# Structured data

- Lexicographi resources
- Grammatica
- Liturgical dat
- Geographica

- Combining textual data and structured data
- Linked data



What is Syriac?

Syriac is a language which once flourished on the Mesopotamian plateau. A dialect of Aramaic, Syriac was widely used during much of the first millenium of the common era. Syriac speaking communities could be found in what today would be Turkey, Lebanon, Syria, Iraq, Iran, India, Central Asia, China, and Mongolia. Sources in Syriac hold immense value for increasing our

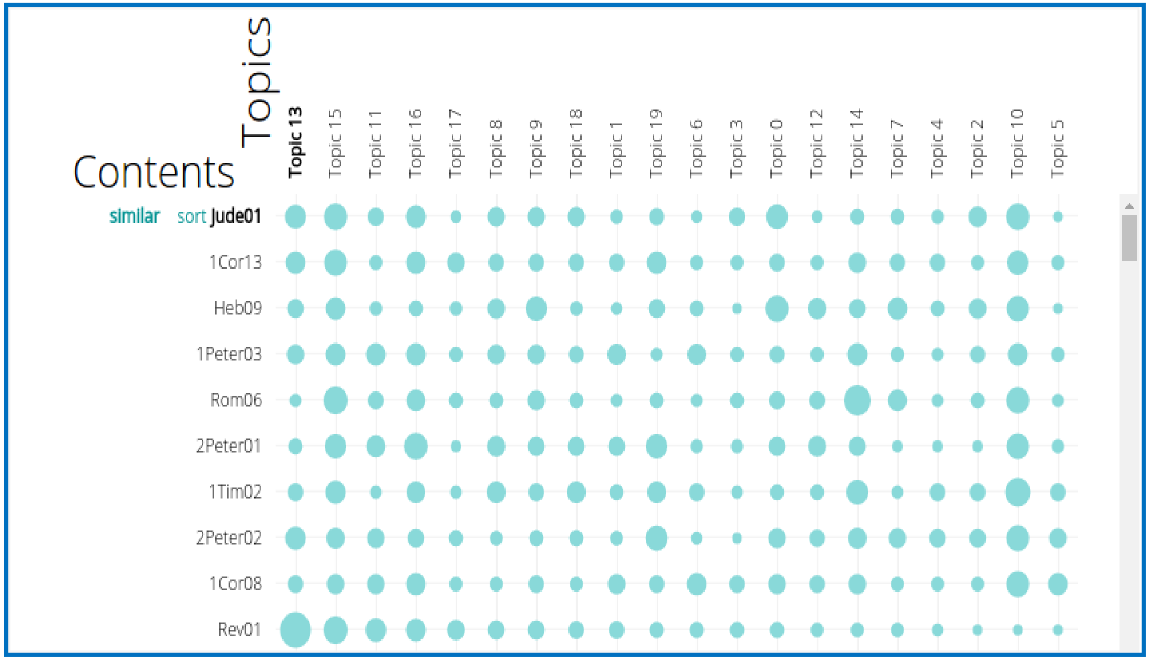
8/13/2019 William L. Potter Workshop on TEI for Critical Editions @VU 2019 Digital Humanities Bootcamp

ETCBC

# Text analytics

# Text analytics

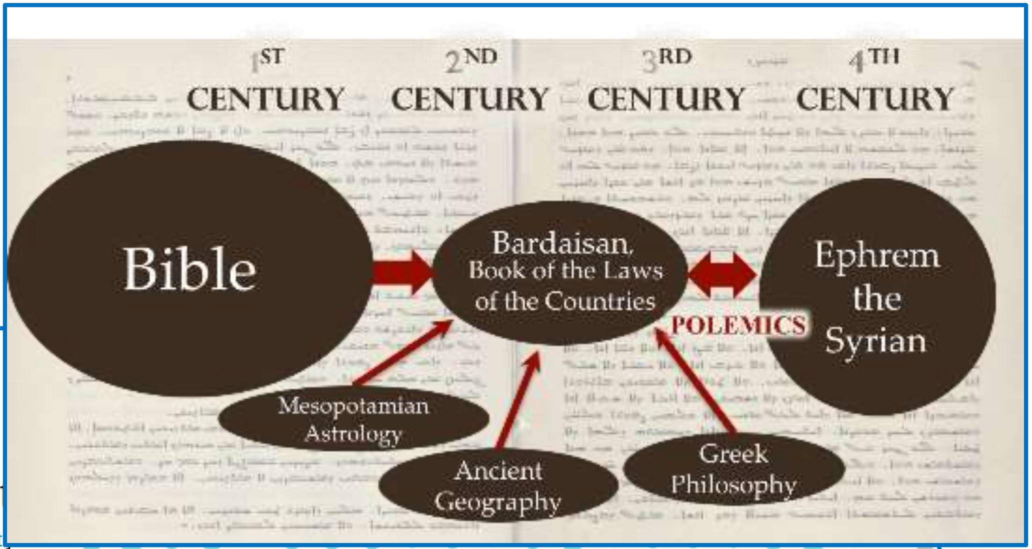
- Topic modelling
- Text clustering
- Etc, etc.

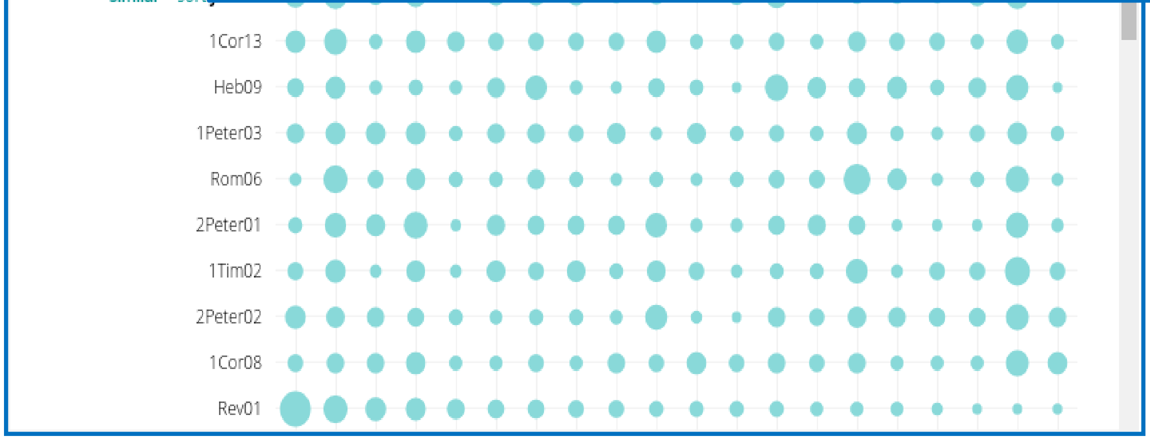


ETCBC

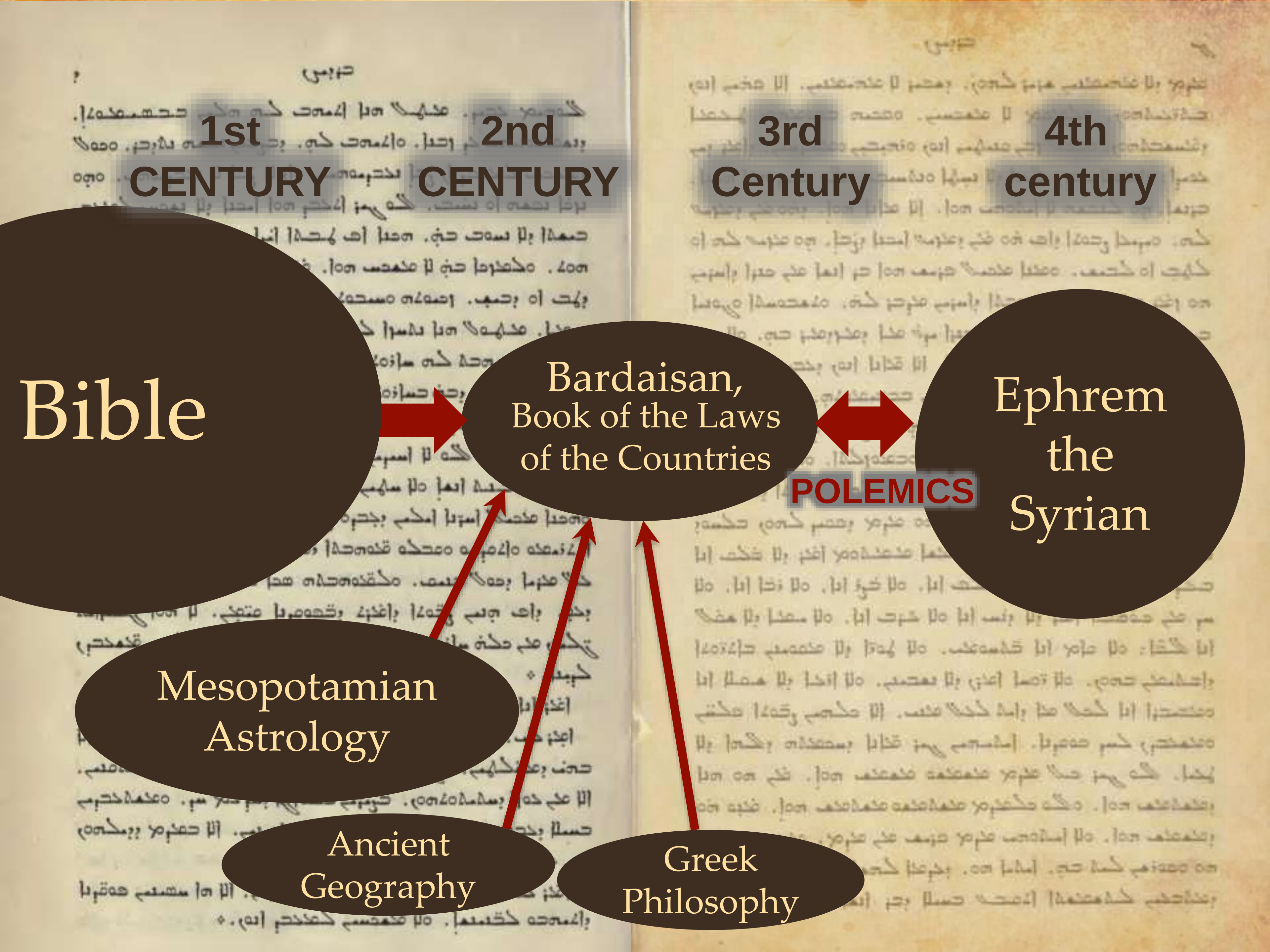
# Text analytics

- Topic modelling
- Text comparison
- Etc, etc.





ETCBC



1st CENTURY

2nd CENTURY

3rd Century

4th century

Bible

Bardaisan,
Book of the Laws
of the Countries

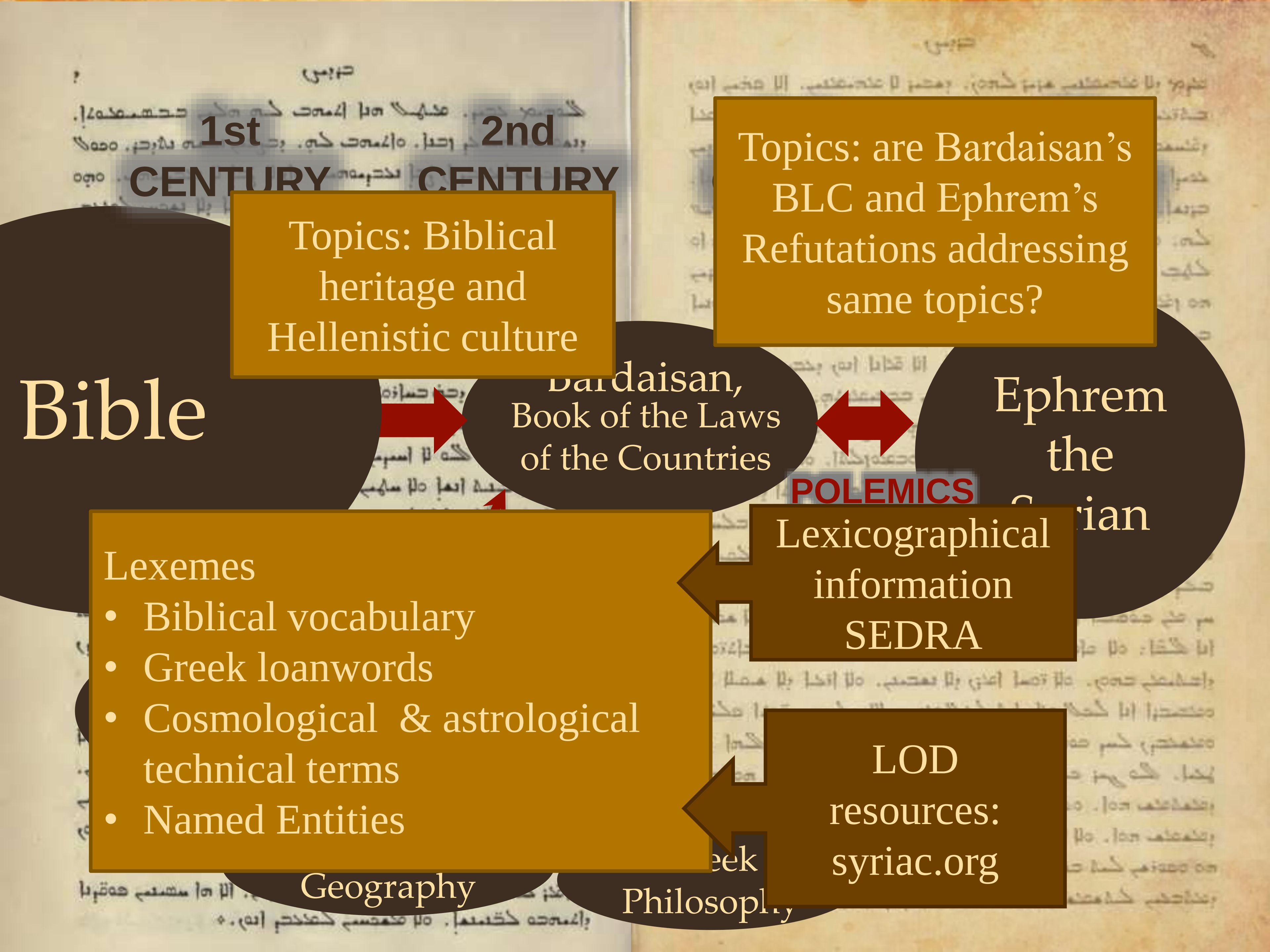
Ephrem
the
Syrian

POLEMICS

Mesopotamian
Astrology

Ancient
Geography

Greek
Philosophy



1st CENTURY

2nd CENTURY

Topics: Biblical heritage and Hellenistic culture

Bible

Bardaisan, Book of the Laws of the Countries

Topics: are Bardaisan's BLC and Ephrem's Refutations addressing same topics?

Ephrem the Syrian

POLEMICS

Lexicographical information SEDRA

Lexemes

- Biblical vocabulary
- Greek loanwords
- Cosmological  & astrological technical terms
- Named Entities

LOD resources: syriac.org

Geography

Greek Philosophy

# Desiderata

- Annotated corpora
  - Syriac
  - Other forms of Aramaic, esp. Neo-Aramaic
- Analytical tools
- Human resources / Project funding