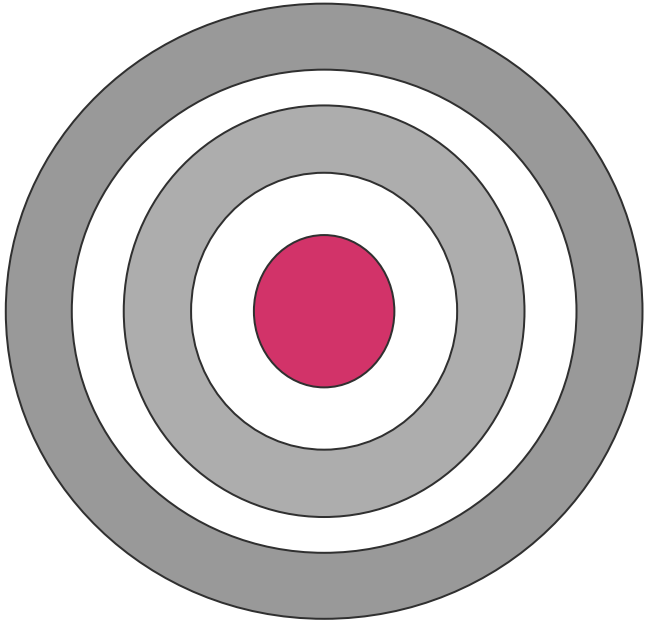
An illustrated introduction to the Classical Language Toolkit & Archive

Arts and Humanities Research Computing
Harvard University
October 26, 2016

Luke Hollis & Kyle P. Johnson
luke@archimedes.digital, kyle@kyle-p-johnson.com

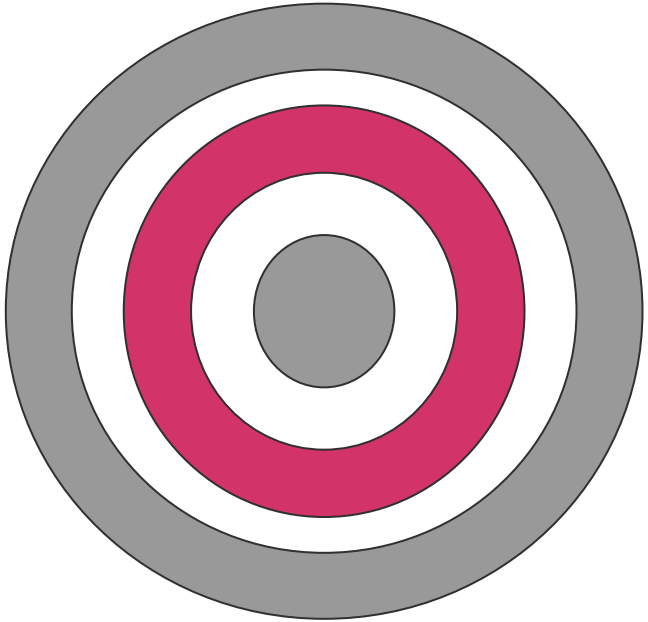
About the Toolkit and Archive

The CLTK's goals …



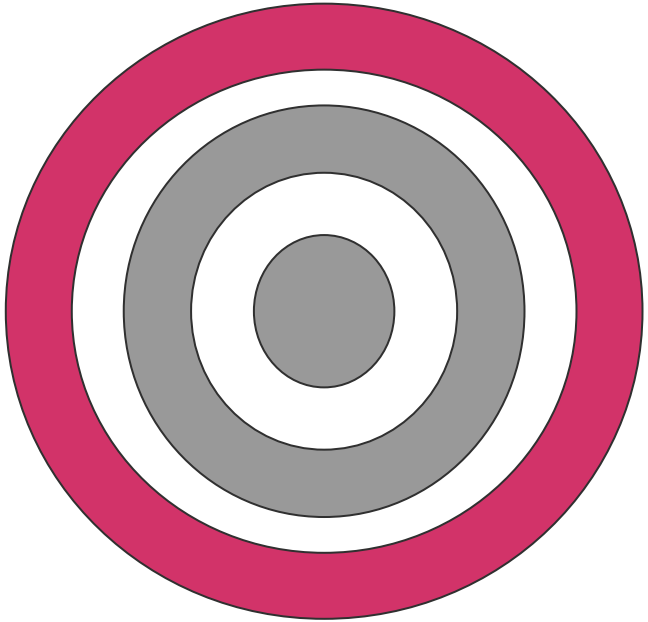
- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)

The CLTK's goals …



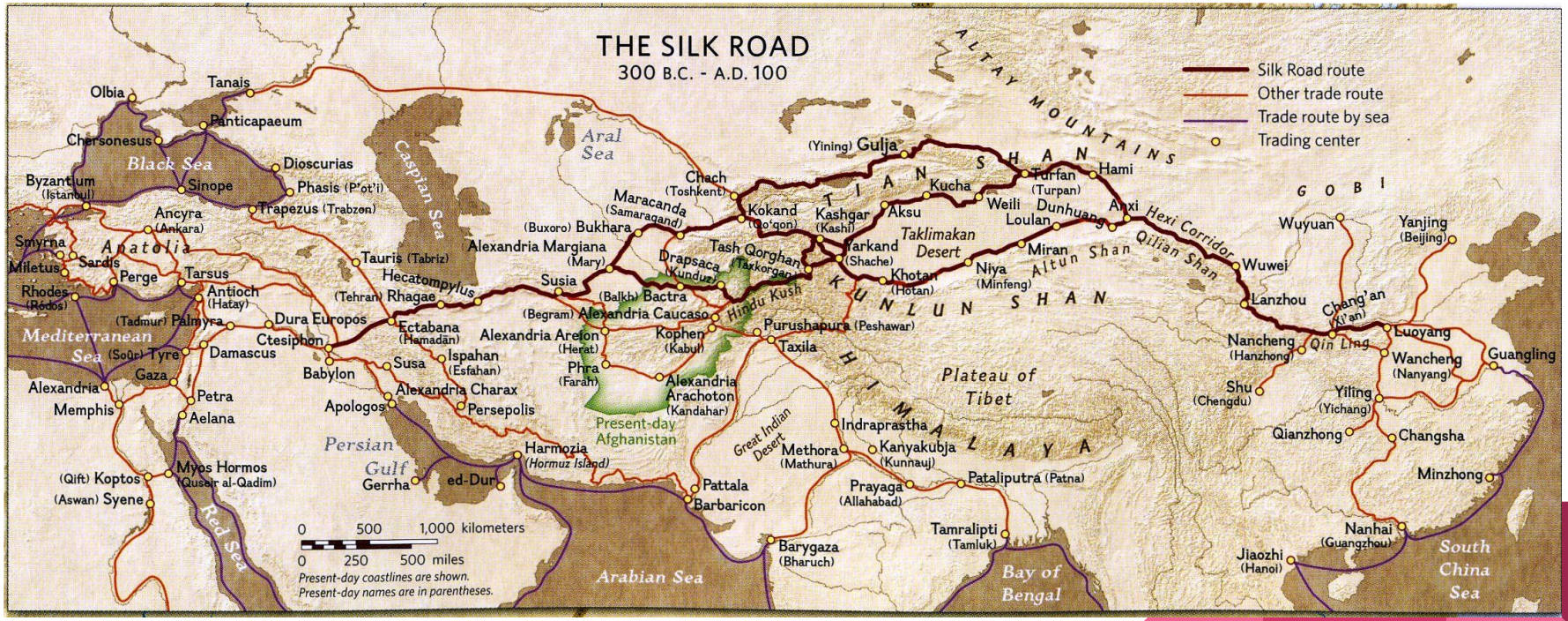
- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- Medium: Quantified Classics

The CLTK's goals …



- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- Medium: Quantified Classics
- High: Framework for an integrated study of the ancient world

… for a connected ancient world



By the numbers

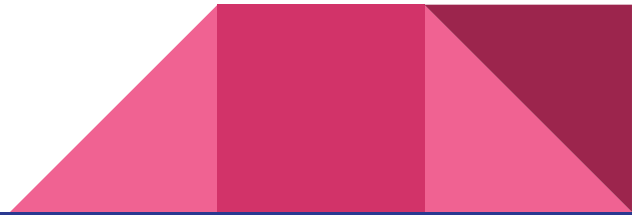
- Began 2014
- 1,660 commits
- 33 contributors
- 32 watchers, 136 stars, 98 forks
- 47 people, 18 teams
- 33 releases (with DOI for every release)
- 84% code coverage

- Full POSIX support
- 2 students, Google Summer of Code
  - Patrick Burns, PhD (ISAW)
  - Suhaib Khan (Netaji Subhas Institute of Technology, Delhi, India); mentored by Luke Hollis of Archimedes Digital)

Google Summer of Code

Some basic terms

- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing
- Package or “library”: Collection of software for a particular set of tasks
- NLTK: A prominent NLP package for the Python language
- Git: Software for distributed software development
- GitHub: A website which makes Git easy
- Jupyter (formerly IPython): “Scientific notebooks”, an easy way to share code



Design principles

Disintermediation	• Independent of academic bureaucracy • Software direct into researchers' hands
Decentralization	• Distributed by Git • Not "pet project" of one person – of many!
Transparency	• Public development on GitHub • Public, readable code
Standardization	• Scientific reproducibility • Good basic texts, but editable

Design principles

Extensibility	• Accepting of any proven NLP algorithms • 100% NLP coverage of all ancient langs
Multi–disciplinary	• Academic depts, CS, faith traditions • Intersection of industry & academe
Mutual benefit	• Full public record of all commits • Researchers develop own work
Inclusion	• Collaborative, encouraging • Free, easy communication

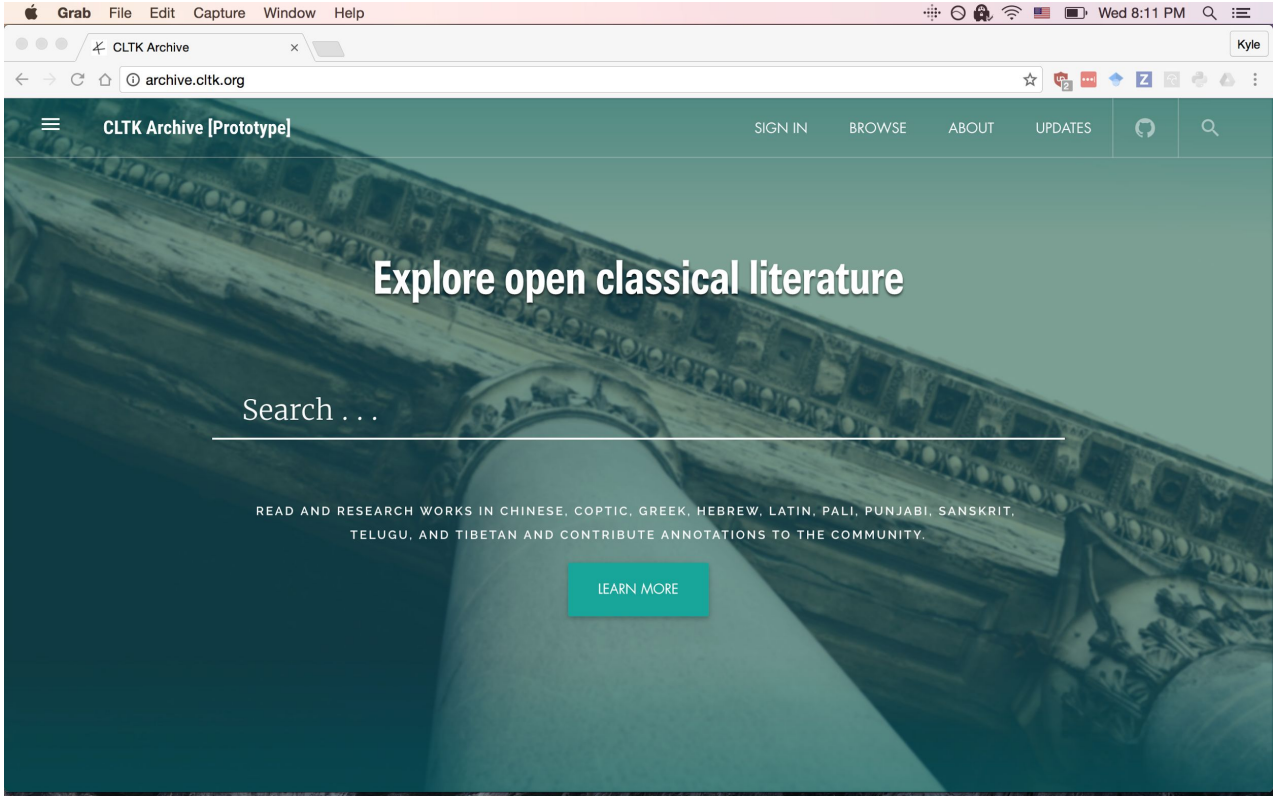
Design principles

Free & open source

- Fork, modify, merge … whatever!
- MIT licence (OK for commercial use)

CLTK Archive

- http://archive.cltk.org
- Almost to “alpha”
- Led and designed by Luke Hollis
- Node.js + React frontend
- Interactive reading, research, and educational environment
- Document annotations
- Scholarly commentaries
- Vocabulary and morphology help
- Poetry scansion
- Standard document format for adding new languages



Aeneid

Aeneidos

Arma virumque cano, Troiae qui primus ab oris

Italiam, fato profugus, Laviniaque venit

litora, multum ille et terris iactatus et alto

"Folio 45V". *The Vergilius Vaticanus*. Vatican, Biblioteca Apostolica, C
Vat. lat. 3225). Rome, Italy. Ca 400 AD.

Vergil *(Publius Vergilius Maro)*

Aeneid

Aeneidos

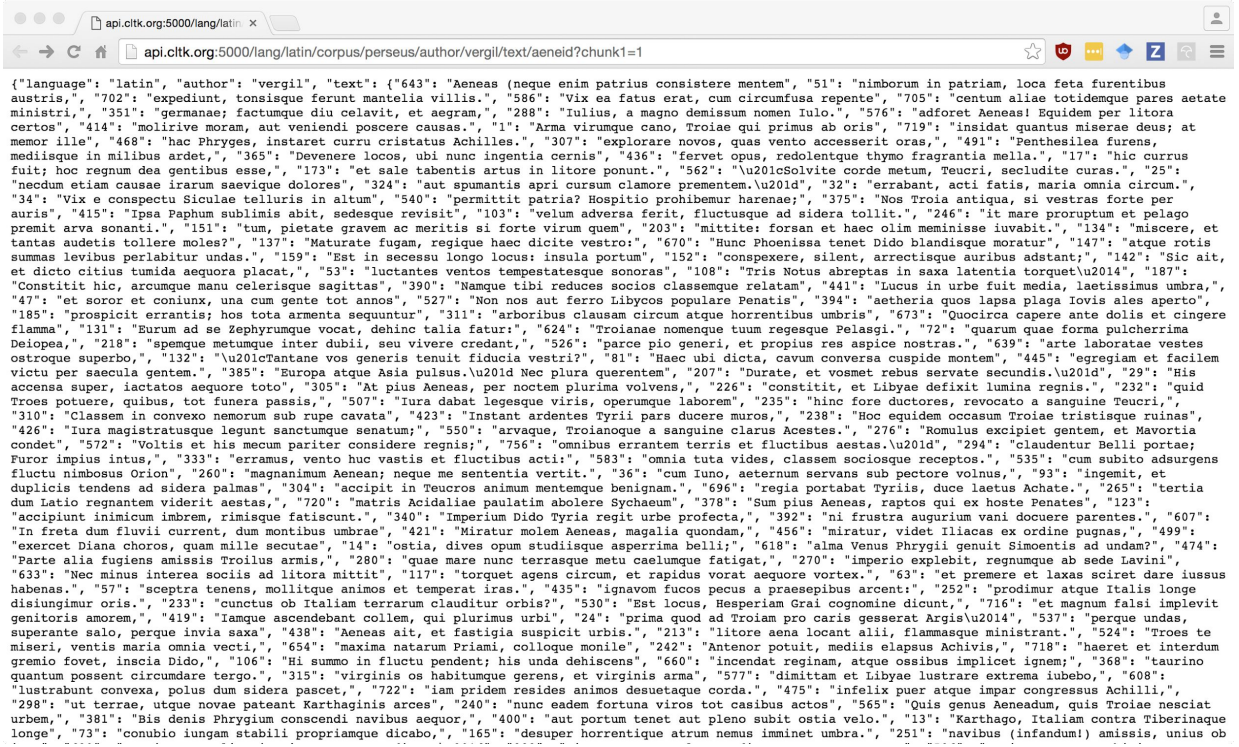
Ārmă vĭrūmqŭe cănō, Trōiāe qūi prīmŭs ăb ōrīs
Ītăl ĭām, fātō prŏfŭgūs, Lāvīnĭaqŭe vēnīt
lĭtŏră, mūltum īlle ēt tērrīs iāctātŭs ĕt āltō
vi superum saevae memorem Iunonis ob iram;
multa quoque et bello passus, dum conderet urbem,

Arms and the man I sing, who first made way,
predestined exile, from the Trojan shore
to Italy, the blest Lavinian strand.
Smitten of storms he was on land and sea
by violence of Heaven, to satisfy
stern Juno's sleepless wrath; and much in war
he suffered, seeking at the last to found
the city, and bring o'er his fathers' gods
to safe abode in Latium; whence arose
the Latin race, old Alba's reverend lords,
and from her hills wide-walled, imperial Rome.
O Muse, the causes tell! What sacrilege,
or vengeful sorrow, moved the heavenly Queen
to thrust on dangers dark and endless toil
a man whose largest honor in men's eyes
was serving Heaven? Can gods such anger feel?
In ages gone an ancient city stood—

THEODORE C. WILLIAMS

CLTK API

- http://api.cltk.org (soon)
- Logical text serving
- Post data for NLP
- RESTful design
- Python w/ the Flask framework
- Containerized with Docker, served on Google Cloud Platform with Kubernetes



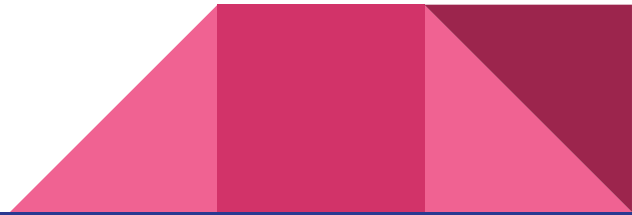
Linked Data Reading Interface for Classical Texts

Leverage metadata to create a multimedia networked reading environment

Digital narrative construction from data journalists

Linked data resources and create our own

Researchers, educators, and students tools that simplify in depth study of classical texts



Current Status

611 classical texts

317,721 nodes

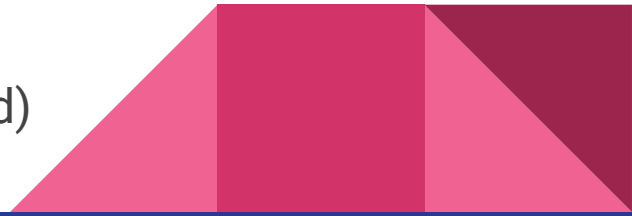
109 authors

5 languages

4,593 definitions for 14,489 word forms

19,892 entities from Wikipedia

~180,000 related passages (above significance threshold)



Reading Environment Metadata

Commentary (modern)

Translations

Definitions

Media (images / manuscripts, audio soon)

Entities (Wikipedia, Pleiades)

Related Passages (text reuse)

Scansion

Annotations

Annotation collection (soon)

Criticism (soon)

Prototype currently accessible

http://archive.cltk.org

User Story: Educator and annotation collections

Educator can

create annotations,

add annotations to a collection,

and share a link to that collection with their classroom

Metadata is not static - deployment with microservices

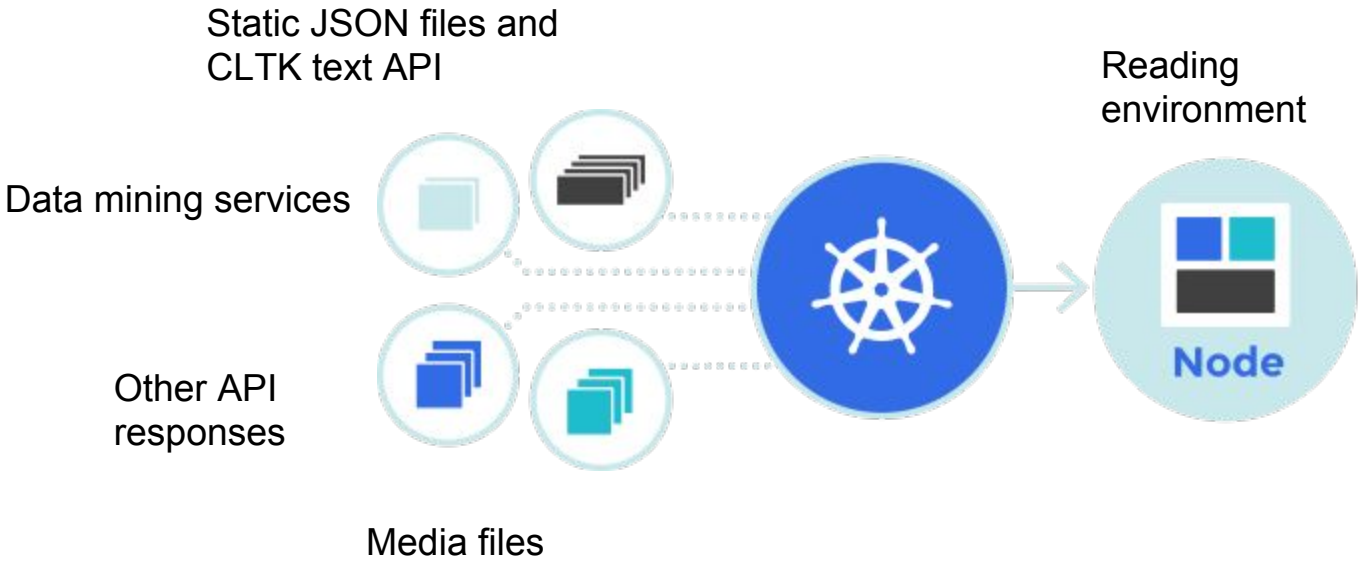
Metadata is not static - continuously evolving as corpora are added/updated

Moving away from a large application and single point of failure

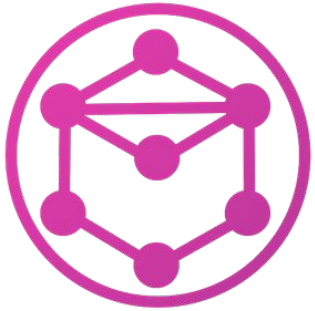
Packaging data mining scripts as services

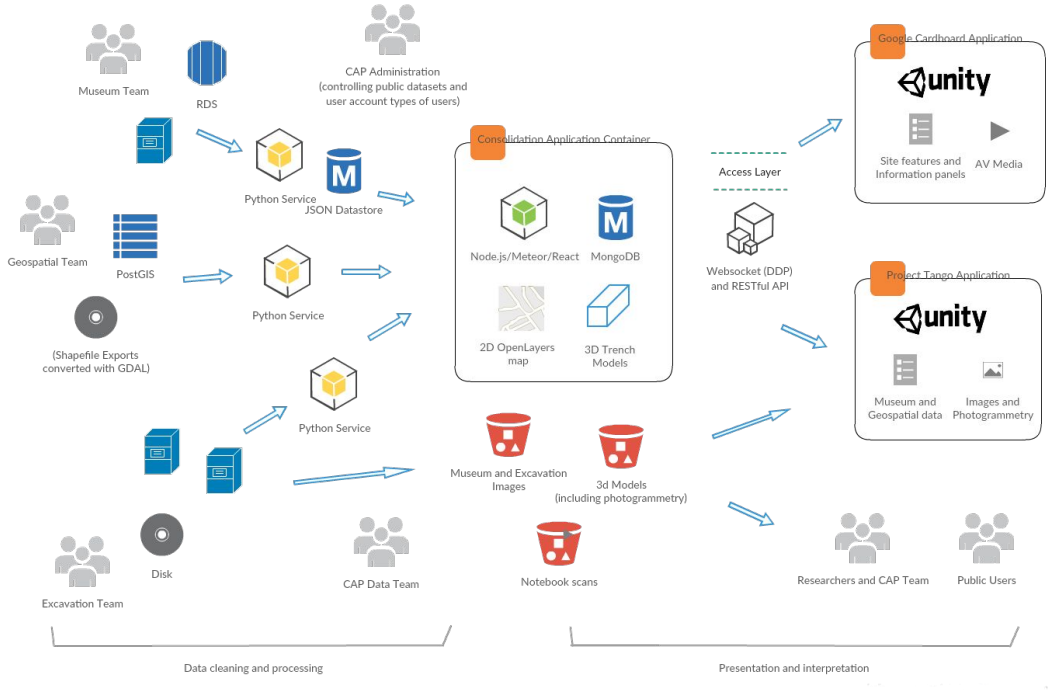
Migrating to Google Cloud Platform with Kubernetes

Metadata continually refreshing / evolving



Meteor / Apollo stack / GraphQL





Example application environment

Building content and fixing existing content

Application users:

Media / annotation creation

Exploring options for crowdsourcing correcting or adding the metadata as ‘Community Editors’

Works with any texts in our JSON format

If anyone here has documents they'd like to share in a multimedia, linked data reading environment, you can get started very quickly

```
{
   title: 'Aeneid',
   texts: {
      '1': {
         '1' : 'Arma virumque cano Troiae qui primus ab oris',
      }
   }
}
```

Workflow: Leveraging CLTK core modules for identifying and extracting metadata

Lemmatization

Named Entity Recognition

Text Reuse

Scientific method and communication

1. Make observations
2. Think of interesting questions
3. Formulate hypotheses
4. Develop testable predictions
5. Gather data to test predictions
6. Develop general theories
7. (repeat)

1. Peer review
2. Documentation
3. Reproducibility
    a. Archiving
    b. Data sharing

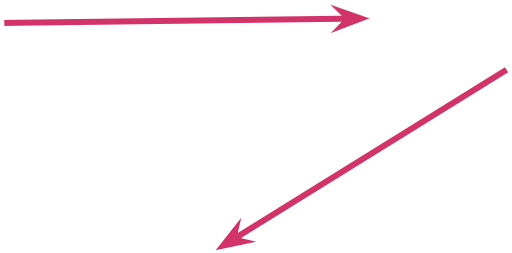
Scientific method and communication

1. Make observations
2. Think of interesting questions
3. Formulate hypotheses
4. Develop testable predictions
5. Gather data to test predictions
6. Develop general theories
7. (repeat)

1. Peer review
2. Documentation
3. Reproducibility
    a. Archiving
    b. Data sharing

Scientific method and communication

1. Make observations
2. Think of interesting questions
3. Formulate hypotheses
4. Develop testable predictions
5. Gather data to test predictions
6. Develop general theories
7. (repeat)

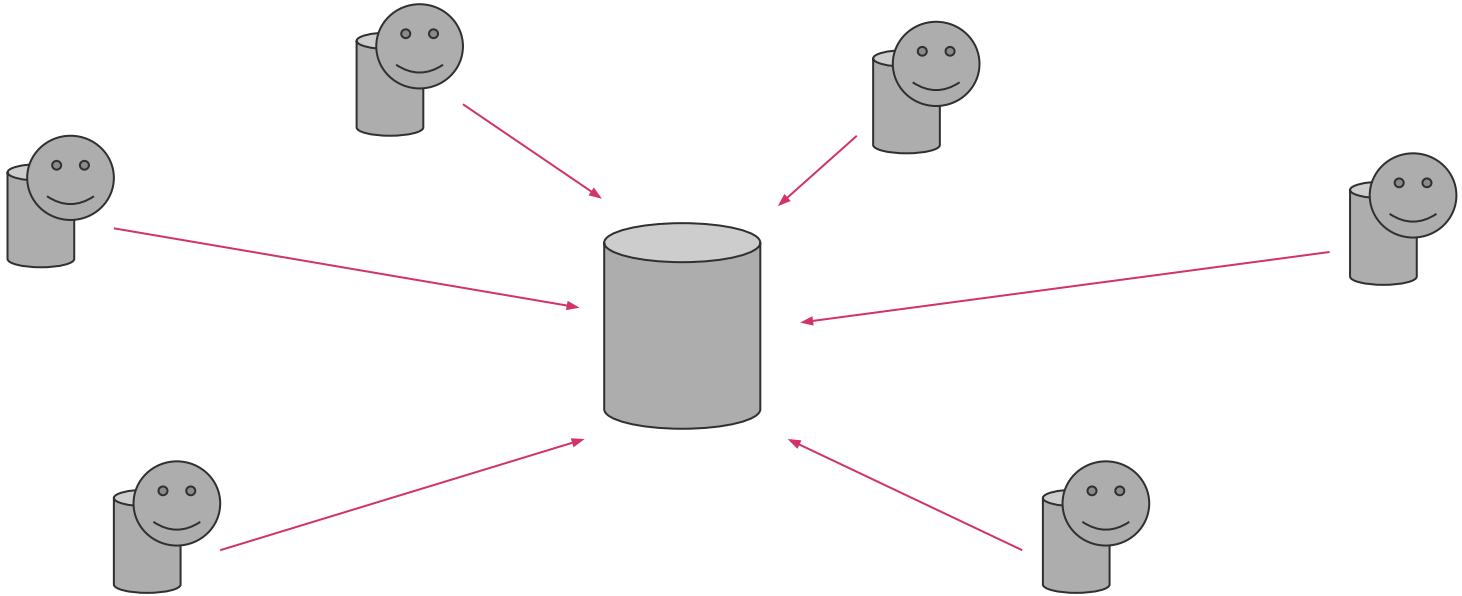


1. Peer review
2. Documentation
3. Reproducibility
   a. Archiving
   b. Data sharing

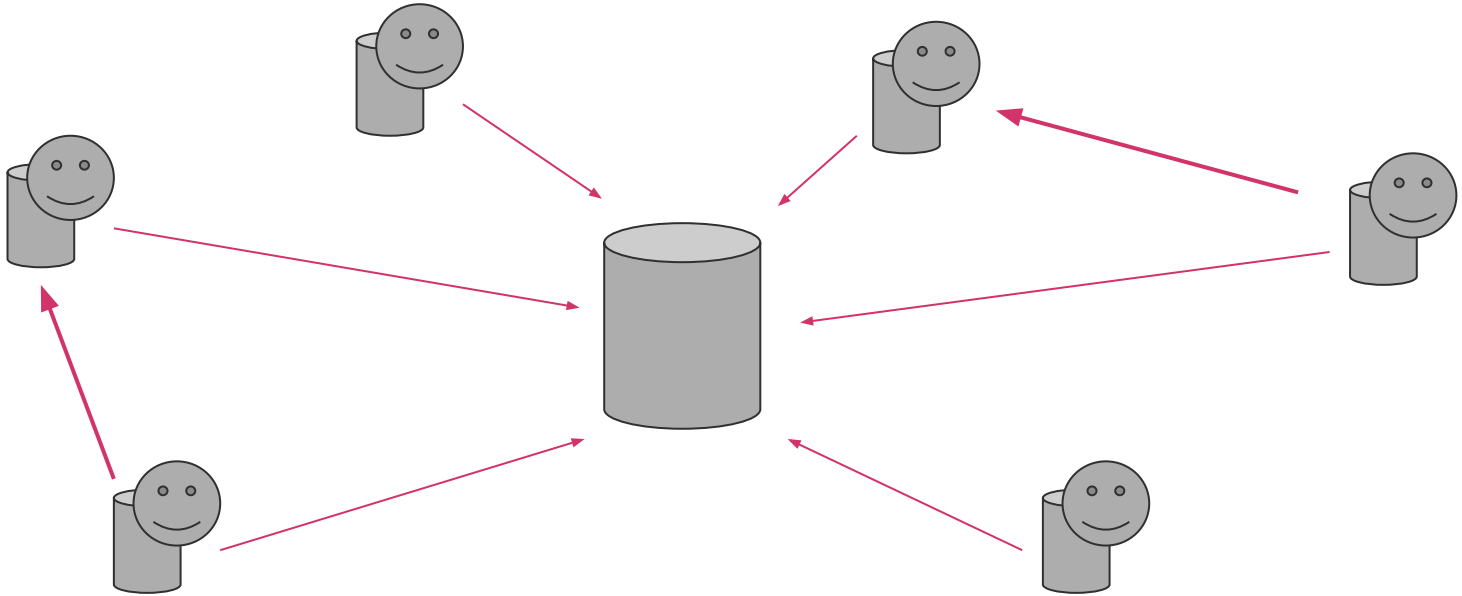
Data sets should be:

1. Versioned
2. Author-attributed
3. Auditable
4. Editable
5. Easily obtained

Technical organization: Repositories

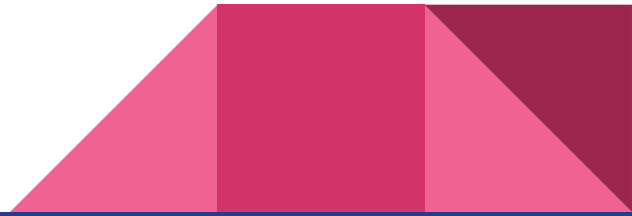


Technical organization: Repositories



Technical organization: Core vs. Data

- CLTK Core software
  - Led by programmers
  - Coordinates data processing
  - Downloads and installs data repositories
- Linguistic data repositories
  - Led by language experts
  - Plaintext corpora
  - Trained models (for machine learning)
  - Dictionaries, word lists
  - Tagged texts (for part-of-speech, dependency grammar)
- CLTK Archive and API
  - Reading environment, with NLP and research extras
  - Totally led by Luke Hollis



(Really quick) quickstart

- Make virtualenv and download core
  - `$ pyvenv venv`
  - `$ source venv/bin/activate`
  - `$ pip install cltk`
- Download and import corpora
  - `$ python`
  - `>>> from cltk.corpus.utils.importer import CorpusImporter`
  - `>>> ci = CorpusImporter('greek')`
  - `>>> ci.list_corpora`
  - `['greek_software_tlgu', 'greek_text_perseus', 'phi7', 'tlg', 'greek_proper_names_cltk', 'greek_models_cltk', 'greek_treebank_perseus', 'greek_lexica_perseus', 'greek_training_set_sentence_cltk', 'greek_word2vec_cltk']`
  - `>>> ci.import_corpus('greek_text_perseus')`

Setup for PHI and TLG corpora

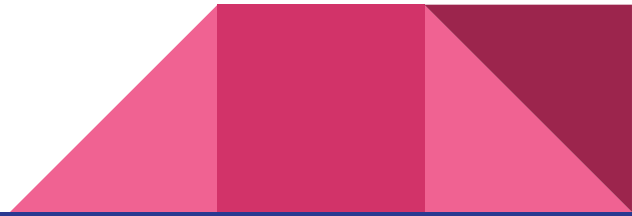
- PHI5, PHI7, and TLG_E
  - Not downloaded, but imported from local files
  - `>>> ci.import_corpus('tlg', '~/Documents/corpora/TLG_E/')`
  - Makes copy of corpus at `~/cltk_data/originals`
- Convert TLG from Beta Code into Unicode
  - `>>> from cltk.corpus.greek.tlgu import TLGU`
  - `>>> t = TLGU()`
  - `>>> t.convert_corpus(corpus='tlg')`
  - `>>> t.convert_corpus(corpus='phi5')`
  - Makes copy of corpus in `~cltk_data/greek/text/tlg` or `~/cltk_data/latin/text/phi5`

NLP for all languages

- Concordance
- Information retrieval
  - Plain and regex searching
  - Robust boolean search on the way
- n-gram: 'Ut primum nocte discussa sol’
  - bigrams: ('ut', 'primum'), ('primum', 'nocte'), ('nocte', 'discussa'), ('discussa', 'sol')
  - trigrams: ('ut', 'primum', 'nocte'), ('primum', 'nocte', 'discussa'), ('nocte', 'discussa', 'sol')
  - 5-grams: ('ut', 'primum', 'nocte', 'discussa', 'sol')
- Word frequencies
  - simple count for a word
  - complete reports for a text
- Word tokenization (via NLTK)

NLP for Greek and Latin

- Text normalization
  - j → i, v → u (Latin)
  - Beta Code conversion (for legacy Greek texts)
  - TLG and PHI corpus specific (remove formatting)
  - Unicode normalization
- Sentence tokenizer
- Lemmatizer
- Stemmer (Latin)
- Word tokenizer, for enclitics (Latin)
- Stopword filtering

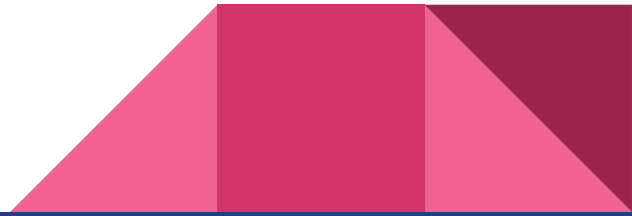


NLP for Greek and Latin (cont'd.)

- Named Entity Recognition (NER)
- Part-of-speech (POS) tagger
  - From Perseus/Alphaeus treebank
  - Great work remaining to be done, convert their codes to others (Brill, PROIEL, etc)
- Dependency grammar tagger `# In progress!`
- Prosody scanner
- Syllabifier (Greek)
- TLG and PHI5 indices
  - File to author, genre to authors, date to authors, gender to authors, etc.
- Word2Vec

Beyond Greek and Latin

- 68 repos at https://github.com/cltk
- Chinese, Coptic, Pali, Tibetan, Middle English, Telugu, Classical Hindi, Sanskrit, Hebrew, Aramaic, Old & Middle English
  - 2.5 GB (!) of Chinese Buddhist texts
  - Coptic texts (via Coptic Scriptorium)
  - Pali Tipitaka
  - Tibetan POS tagged texts and a lexicon
  - Parallel corpora – ready for statistical machine translation (hint, hint)
  - Corpus of ~50 million Hebrew words, ~20 million Aramaic (via Sefaria)
  - Entirety of Perseus/Open Philology 10s of millions of words

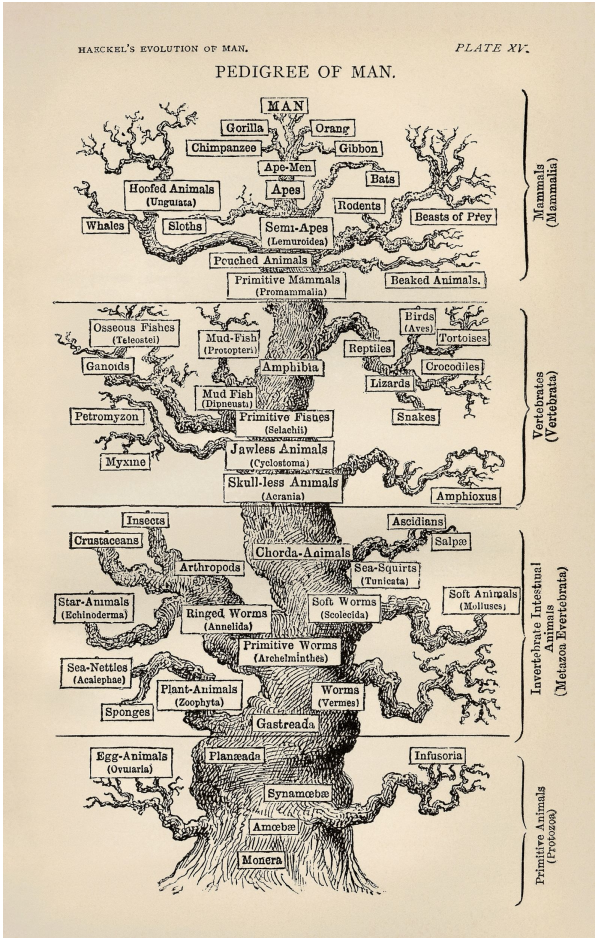


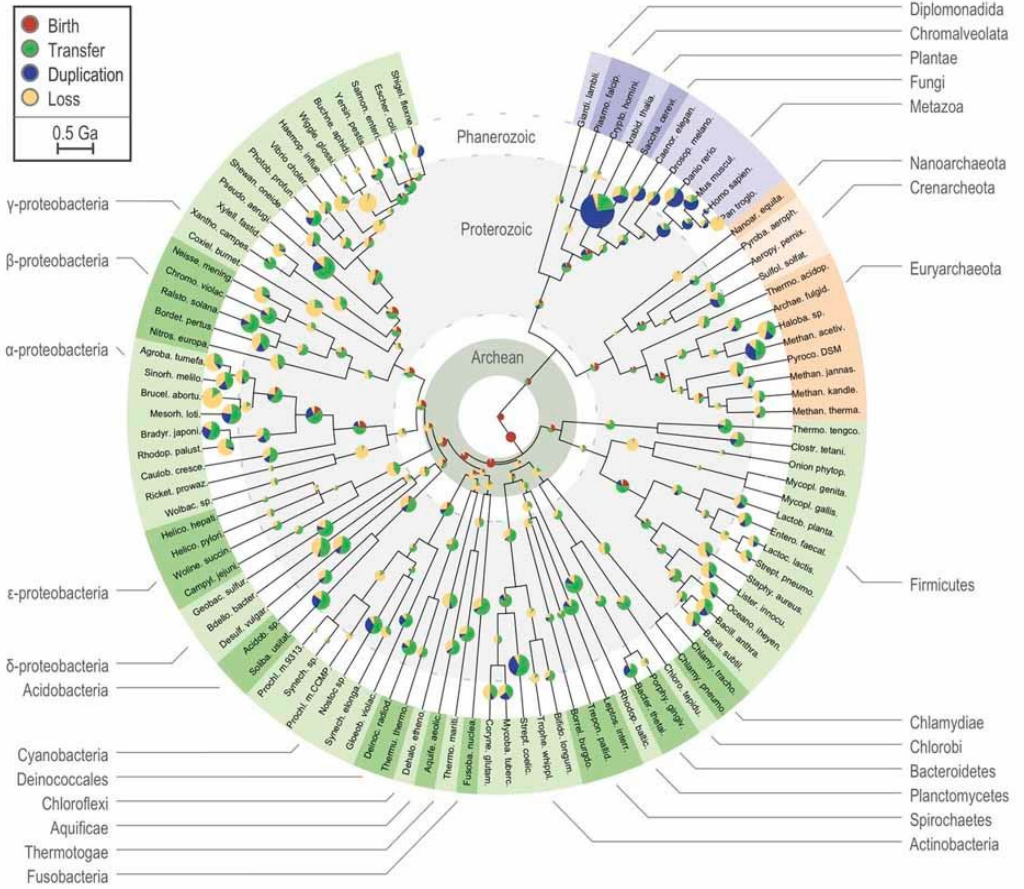
Re-classifying Greek genres

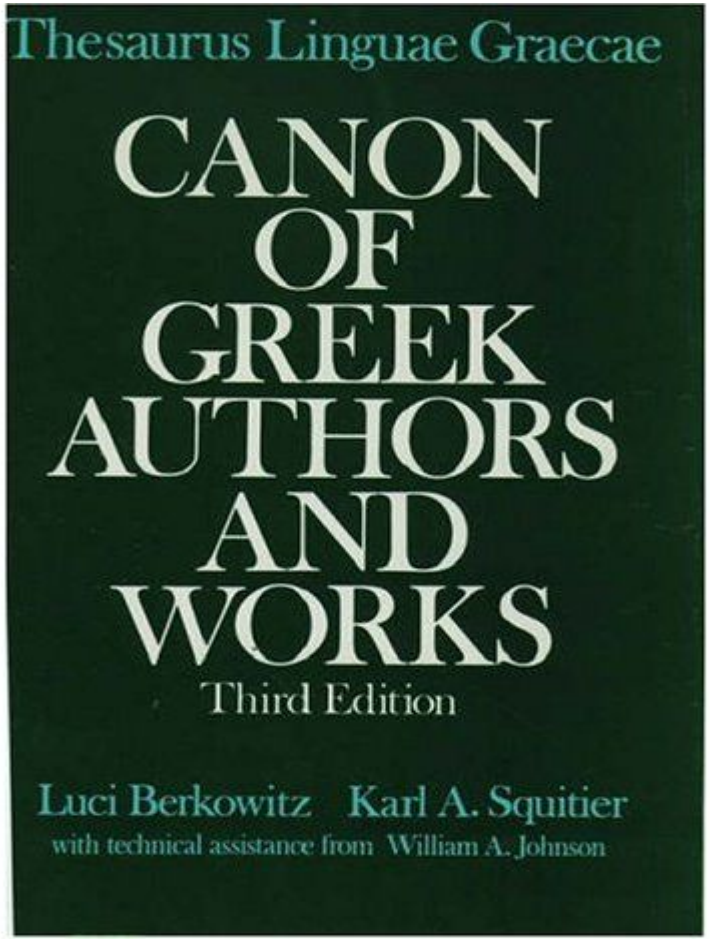
Problem

What is a genre?

Genre labels to Greek literature have slowly accumulated over the past 2,500+ years. They do not reflect the thinking of one people or method, but are the result of centuries of consensus and debate.


HAECKEL'S EVOLUTION OF MAN.
PLATE XV.
PEDIGREE OF MAN.
MAN
Gorilla
Orang
Chimpanzee
Gibbon
Ape-Men
Bats
Hoofed Animals (Unguata)
Apes
Rodents
Whales
Sloths
Semi-Apes (Lemuroides)
Beasts of Prey
Pouched Animals
Primitive Mammals (Promammalia)
Beaked Animals.
Mammals (Mammalia)
Osseous Fishes (Teleostei)
Mud-Fish (Protopteri)
Birds (Aves)
Tortoises
Ganoids
Amphibia
Reptiles
Crocodiles
Mud Fish (Dipneusta)
Lizards
Petromyzon
Primitive Fishes (Selachii)
Snakes
Jawless Animals (Cyclostoma)
Myxine
Skull-less Animals (Acrania)
Amphioxus
Vertebrates (Vertebrata)
Insects
Ascidians
Crustaceans
Chorda-Animals
Salpæ
Arthropods
Sea-Squirts (Tunicata)
Star-Animals (Echinoderma)
Ringed Worms (Annelida)
Soft Worms (Scolecida)
Soft Animals (Molluscs)
Primitive Worms (Archelminthes)
Sea-Nettles (Acalephæ)
Plant-Animals (Zoophyta)
Worms (Vermes)
Sponges
Gastræada
Invertebrate Intestinal Animals (Metazoa Enterobrata)
Egg-Animals (Ovularia)
Planæada
Infusoria
Synamœbæ
Amœbæ
Monera
Primitive Animals (Protozoa)


Birth
Transfer
Duplication
Loss
0.5 Ga
Phanerozoic
Proterozoic
Archean
Diplomonadida
Chromalveolata
Plantae
Fungi
Metazoa
Nanoarchaeota
Crenarcheota
Euryarchaeota
Firmicutes
Chlamydiae
Chlorobi
Bacteroidetes
Planctomycetes
Spirochaetes
Actinobacteria
γ-proteobacteria
β-proteobacteria
α-proteobacteria
ε-proteobacteria
δ-proteobacteria
Acidobacteria
Cyanobacteria
Deinococcales
Chloroflexi
Aquificae
Thermotogae
Fusobacteria


Thesaurus Linguae Graecae
CANON OF GREEK AUTHORS AND WORKS
Third Edition
Luci Berkowitz Karl A. Squitier
with technical assistance from William A. Johnson

?

Hypothesis

Can AI classify literature better than the scholarly tradition?

Can statistical machine learning both expose illogical elements of the traditional taxonomy and also assist in creating a new data–driven one?

Methodology

Discovery, clustering, & classification

Let’s look at what text data survives, its basic linguistic properties, how well it clusters in unsupervised ML, and how supervised models perform with various features extracted.

Basic corpus statistics

Corpus counts

- Total author files: 1,822
- Total words: 72,057,716
- Total unique words: 1,515,193

Counts per author

- Mean words per author: 39526.997257268 24
- Standard deviation of words per author: 174923.289766537 58

Avgs per author

- Mean unique words per author: 5435.820076796 489
- Standard deviation of unique words per author: 14195.290142159 112

Epithet

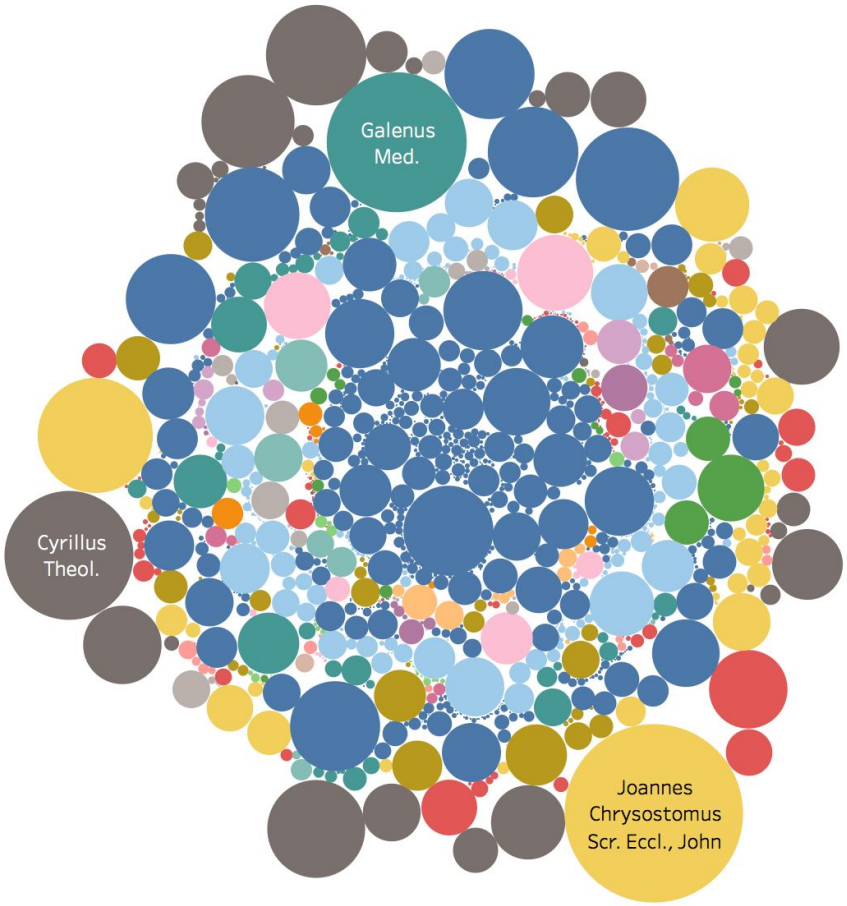
(All)

Name

(All)

Word Count All

100 4,095,369



Name. Color shows details about Epithet. Size shows sum of Word Count All. The marks are labeled by Name. The data is filtered on Action (Epithet), which keeps 54 members. The view is filtered on Epithet

Epithet

- Null
- Alchemistae
- Apologetici
- Astrologici
- Astronomici
- Atticistae
- Biographi
- Bucolici
- Chronographi
- Comici
- Doxographi
- Elegiaci
- Epici/-ae
- Epigrammatici/-ae
- Epistolographi
- Geographi
- Geometri
- Gnomici
- Gnostici
- Grammatici
- Historici/-ae
- Hymnographi
- Iambici
- Lexicographi
- Lyrici/-ae
- Mathematici
- Mechanici
- Medici
- Mimographi
- Musici
- Mythographi
- Onirocritici
- Oratores
- Paradoxographi
- Parodii
- Paroemiographi
- Periegetae
- Philologi
- Philosophici/-ae
- Poetae
- Poetae Didactici
- Poetae Medici
- Poetae Philosophi
- Polyhistorici
- Rhetorici
- Scriptores Ecclesiastici
- Scriptores Erotici
- Scriptores Fabularum
- Sophistae
- Tactici
- Theologici
- Tragici

Authors by surviving text

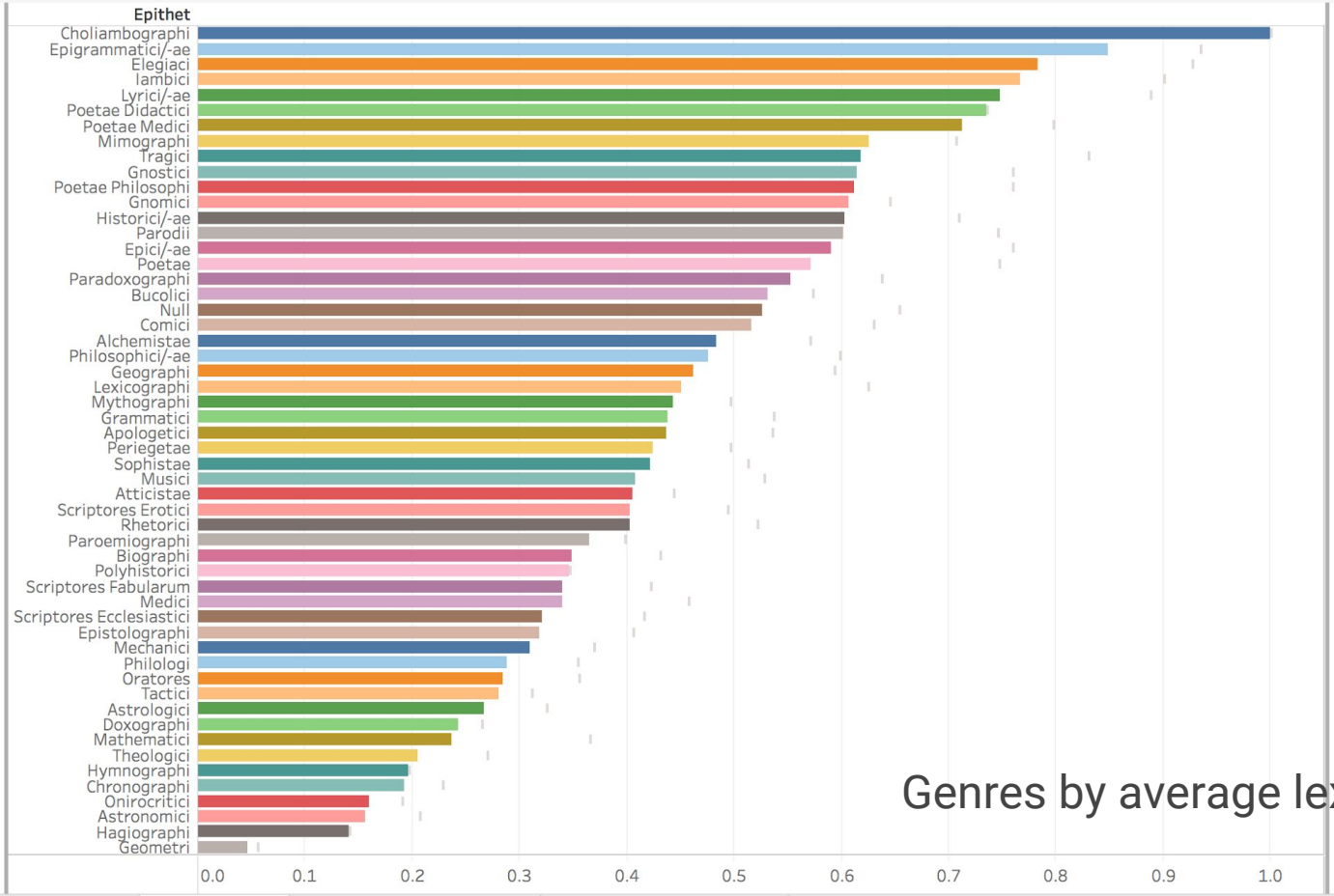
Basic corpus statistics

Lexical diversity

- Lexical diversity per author: 0.5171187962883808
- Standard deviation of lexical diversity per author: 0.2732410961564417

Tableau Reader - KPJ_EPIVIZ_10_1

Entire View



Genres by average lexical diversity

Overview-1 | Overview-2 | Epithet Summary Table | General Text Summary

108 marks 54 rows by 1 column SUM(Standard Deviation of Lexical Diversity): 28.7811

Genre stats

Overall

- Labels: 56 (including “no label”)

Highest lex. diversity

- Choliambographi, 1.0
- Epigrammatici, 0.886
- Elegiaci, 0.789
- Lyrici, 0.750
- Iambici, 0.748
- Poetae Didactici, 0.735

Lowest lex. diversity

- Geometri, 0.046
- Scriptores Rerum Naturalium, 0.151
- Onirocritici, 0.160
- Chronographi, 0.192
- Hymnographi, 0.120
- Theologici, 0.206
- Astronomici, 0.211
- Mathematici, 0.226
- Doxographi, 0.243

Genre stats

Philosophici

- Lex. diversity: 0.472702
- Mean wc/author: 51,513
- Mean unique wc/author: 5,177

Historici

- Lex. diversity: 0.598809
- Mean wc/author: 20,618
- Mean unique wc/author: 3,570

Mechanici

- Lex. diversity: 0.310574
- Mean wc/author: 24,746
- Mean unique wc/author: 3,331

Clustering

- NMF (similar to LDA)
- Topic–to–epithets
  - Under 20% diversity: Theologici, Historici/-ae, Philosophici/-ae, Scriptores Ecclesiastici
- Epithet–to–topics
  - Under 20% diversity:
    - 0: Philosophici/-ae
    - 1: Historici/-ae
    - 2: Comici
    - 3: Scriptores Ecclesiastici
    - 4: Historici/-ae
- Affinity Clustering
  - Problem: Clusters very tight

	epithet	most_common_topic	topic_diversity	total_docs	unique_topics
0	Lyrici/-ae	24	0.466667	30	14
1	Theologici	3	0.111111	36	4
2	Scriptores Fabularum	0	1	2	2
3	Atticistae	17	0.75	4	3
4	Elegiaci	24	0.5	16	8
5	Sophistae	0	0.307692	39	12
6	Medici	5	0.302326	43	13
7	Philosophici/-ae	0	0.172775	191	33
8	Polyhistorici	6	1	1	1
9	Apologetici	3	0.444444	9	4
10	Mimographi	2	0.5	2	1
11	Hymnographi	3	1	1	1
12	Biographi	3	0.666667	9	6
13	Poetae	4	0.421053	19	8
14	Gnomici	50	1	4	4
15	Mathematici	10	0.555556	9	5

Topics w/in epithets

	epithet_diversity	most_common_epithet	topic	total_epithets	unique_epithets
0	0.15	Philosophici/-ae	0	220	33
1	0.155172	Historici/-ae	1	58	9
2	0.195122	Comici	2	82	16
3	0.15	Scriptores Ecclesiastici	3	80	12
4	0.19802	Historici/-ae	4	101	20
5	0.25	Medici	5	24	6
6	0.307692	Historici/-ae	6	26	8
7	0.375	Historici/-ae	7	8	3
8	0.5	Grammatici	8	8	4
9	0.466667	Comici	9	15	7
10	0.45	Mathematici	10	20	9
11	0.545455	Historici/-ae	11	11	6
12	0.1875	Historici/-ae	12	16	3
13	0.833333	Philosophici/-ae	13	6	5

Epithets w/in topic

Most consistent topics

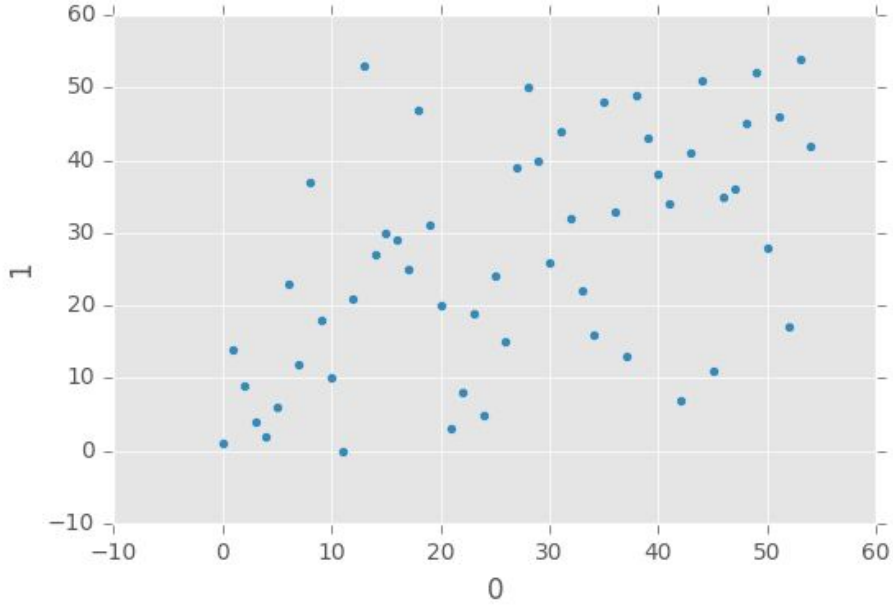
- 0: Philosophici/-ae
  - ειμι 2.85 | ου 1.01 | εχω 0.95 | τος 0.49 | ποιεω 0.47 | πα 0.37 | οιος 0.36 | τοιου 0.36 | αυτο 0.36 | μονον 0.35 | ολοξ 0.34 | δεω1 0.34 | ος 0.32 | μαλλον 0.31 | φημι 0.31 | ειπον 0.31 | παλιν 0.31 | παντα 0.3 | εαυτου 0.3 | κατ 0.28 |
- 1: Historici/-ae
  - φημι 3.03 | ειμι 0.69 | φησι 0.41 | φησιν 0.36 | λεγει 0.3 | πρωτω 0.27 | γενεσθαι 0.27 | ιστορεω 0.22 | φασι 0.21 | καθα 0.21 | προτερος 0.2 | καλεω 0.19 | διος 0.17 | λεγεται 0.17 | υστερος 0.16 | αθηνη 0.16 | ιερον 0.15 | γουν 0.15 | διο 0.14 | λεγεσθαι 0.14 |
- 2: Comici
  - ειμι 0.64 | κακοω 0.51 | ανηρ 0.39 | οραω 0.38 | ευ 0.33 | δοκεω 0.32 | κακον 0.31 | εχω 0.29 | απας 0.29 | ζαω 0.28 | ειδον 0.26 | εμεω 0.26 | αγαθος 0.25 | ποτ 0.23 | αει 0.22 | καλος 0.21 | νυ 0.21 | βιον 0.21 | οιδα 0.21 | καν 0.2 |

Most consistent topics

- 3: Scriptores Ecclesiastici
  - χριστος 0.97 | θεαομαι 0.83 | θεος 0.48 | κυριου 0.4 | ιησου 0.4 | κυριος 0.37 | θεον 0.32 | ου 0.28 | χριστον 0.28 | αγιος 0.28 | υιος 0.27 | ανθρωπος 0.22 | μα 0.22 | πνευ 0.21 | λεγει 0.2 | προφητης 0.19 | ειμι 0.18 | πνευματος 0.18 | αγιου 0.18 | σωτηρ 0.17 |
- 4: Historici/-ae
  - πολιν 0.53 | πολεως 0.5 | εκει 0.46 | πολυς 0.38 | εχω 0.38 | ανηρ 0.31 | ου 0.3 | εαυτου 0.28 | βασιλεως 0.28 | ειμι 0.28 | νος 0.27 | πολεμον 0.26 | ηδος 0.24 | αρχω 0.24 | πολει 0.23 | χωραν 0.22 | ταυτην 0.21 | εκεινου 0.21 | θεν 0.21 | απας 0.2 |

Classification

- Decision tree
- Random forest
- ADA boost
- CNN
  - Binary, Philosophy and Theology
  - acc 0.609023
  - https://github.com/kylepjohnson/cnn-te
- TODO
  - LDA-HDP



Feature extraction (tf-idf)

	ααω	αβαθης	…	αβακιον	cωμα	cωμαcιν	epithet
0	0	0		0	234	1459	Historici/-ae
1	0	0		0	23	825	Tragici
2	0	0		0	323	331	Tragici
3	0	0		0	66	417	Comici
6	0	0		0	98	2475	Historici/-ae
7	0	0		0	638	4075	Philosophici/-ae
8	0	0		0	2	2127	Sophistae
9	1	0		0	0	2074	Theologici
10	0	0		0	45	2173	Historici/-ae
11	0	0		0	3	1419	Scriptores

Feature extraction (morphology)

	verb	noun	…	participle	adjective	adverb	epithet
0	2466	5367		1445	2342	1459	Historici/-ae
1	2256	3460		2332	2089	825	Tragici
2	906	1568		736	842	331	Tragici
3	44	99		34	20	417	Comici
6	679	2001		209	309	2475	Historici/-ae
7	35	55		27	12	4075	Philosophici/-ae
8	89	85		70	36	2127	Sophistae
9	44	53		38	0	2074	Theologici
10	908	1378		798	198	2173	Historici/-ae
11	66	87		63	45	1419	Scriptores

Feature extraction (syntax)

	nsubj	det	…	amod	nmod	dobj	epithet
0	353	337		910	346	405	Historici/-ae
1	2323	4129		292	335	918	Tragici
2	233	678		383	232	248	Tragici
3	44	35		23	39	32	Comici
6	434	564		234	203	234	Historici/-ae
7	7	24		2	13	234	Philosophici/-ae
8	74	643		34	44	32	Sophistae
9	44	55		53	72	37	Theologici
10	97	10		344	234	126	Historici/-ae
11	92	322		63	27	80	Scriptores

Epithet	precision	recall	f1-score	support
Alchemistae	1	0.14	0.25	7
Comici	0.68	0.51	0.58	51
Elegiaci	0	0	0	8
Epici/-ae	0.29	0.11	0.15	19
Grammatici	0.4	0.2	0.27	20
Historici/-ae	0.66	0.75	0.7	84
Lyrici/-ae	0.29	0.17	0.21	12
Medici	0.15	0.29	0.2	7
Philosophici/-ae	0.45	0.48	0.46	46
Poetae	0	0	0	7
Rhetorici	0.38	0.25	0.3	12
Scriptores Ecclesiastici	0.44	0.31	0.36	13
Sophistae	0.33	0.25	0.29	8
Theologici	0.25	0.4	0.31	5
Tragici	0.15	0.6	0.24	15
avg / total	**0.49**	**0.45**	**0.45**	**314**

Decision tree

Random forest

Epithet	precision	recall	f1-score	support
Alchemistae	0	0	0	7
Comici	0.82	0.63	0.71	51
Elegiaci	0	0	0	8
Epici/-ae	0.4	0.11	0.17	19
Grammatici	0.75	0.15	0.25	20
Historici/-ae	0.59	0.89	0.71	84
Lyrici/-ae	0	0	0	12
Medici	1	0.29	0.44	7
Philosophici/-ae	0.47	0.74	0.58	46
Poetae	0	0	0	7
Rhetorici	1	0.17	0.29	12
Scriptores Ecclesiastici	0.5	0.38	0.43	13
Sophistae	0	0	0	8
Theologici	0	0	0	5
Tragici	0.22	0.73	0.34	15
avg / total	**0.52**	**0.53**	**0.47**	**314**

ADA Boost

Epithet	precision	recall	f1-score	support
Alchemistae	1	0.29	0.44	7
Comici	0.34	0.78	0.48	51
Elegiaci	0	0	0	8
Epici/-ae	0	0	0	19
Grammatici	0.4	0.1	0.16	20
Historici/-ae	0.6	0.58	0.59	84
Lyrici/-ae	0	0	0	12
Medici	0.17	0.14	0.15	7
Philosophici/-ae	0.3	0.54	0.39	46
Poetae	0	0	0	7
Rhetorici	0	0	0	12
Scriptores Ecclesiastici	0.5	0.31	0.38	13
Sophistae	0	0	0	8
Theologici	0.25	0.2	0.22	5
Tragici	0	0	0	15
avg / total	**0.34**	**0.39**	**0.34**	**314**

Classification results

- Overall Random Forest has slight edge over Decision tree
- 0.1 - 0.3 better in precsion and recall
- ADA Boost worst
- Interesting:
  - Alchemistae is 100% precision w/ DT, 0% RF
  - RF worse precision:: Alchemistae, Historici, Lyrici, Sophistae, Theologici
  - Medici is 15% precision w/ DT, 100% RF
  - Rhetorici is 38% precision w/ DT, 100% RF
  - Elegiaci, Poetae 0% in all

	precision	recall	f1-score
Decision tree	0.49	0.45	0.45
Random forest	0.52	0.53	0.47
ADA Boost	0.34	0.39	0.34

Preliminary observations

Corpus

- Needs aggressive cleanup: lemmatization, stopword rm, and rm short docs
- But all should be tested w/o cleanup
- “None” docs need a label

Clustering

- Traditional topics show poor consistency with lexical BOW classification
- Some noteworthy commonalities
- Need more than lexical features
- Experiment w/ K

Classification

- Lexical features perform so-so for predictive modeling
- Certain algos perform far better for subsets of data (b/c of their size or uniqueness?)
- Candidate for ensemble learning

Conclusion

Genres labels might be phantoms

Socrates on literature vs. reality: “Mimetic art is far removed from truth, and this … is the reason why it can produce everything, because it touches or lays hold of only a small part of the object and that a phantom.” (Plato, *Republic* X, 598b)

Resources

- This and other lectures
- Core software: https://github.com/cltk/cltk
  - Bug tracking: https://github.com/cltk/cltk/issues
    - Beginners' issues labeled easy
- Project repositories: https://github.com/cltk
- Docs: http://docs.cltk.org
  - Installation: http://docs.cltk.org/en/latest/installation.html
- Python + Command line basics
  - Intro to the command line: http://blog.teamtreehouse.com/introduction-to-the-mac-os-x-command-line
  - Python installation: https://www.python.org/downloads (choose 3.5)
  - Good self-paced Python lessons: http://learnpythonthehard[illegible]rg

Contribute & contact

- Classical Language Toolkit
  - Home: http://cltk.org
  - Docs: http://docs.cltk.org/en/latest
  - Source: https://github.com/cltk/cltk
  - Corpora: https://github.com/cltk
  - Import module: https://github.com/cltk/cltk/blob/master/cltk/corpus/utils/importer.py
- Contribute
  - Issue tracking: https://github.com/cltk/cltk/issues
  - Other questions: kyle@kyle-p-johnson.com

Sources

- Images
  - https://s-media-cache-ak0.pinimg.com/originals/61/b2/34/61b23411d0990a69b32a89fba0536cd3.jpg
- Git
  - GitPython: https://github.com/gitpython-developers/GitPython
  - https://en.wikipedia.org/wiki/Git_(software)
- Science
  - https://en.wikipedia.org/wiki/Scientific_method
  - https://en.wikipedia.org/wiki/Reproducibility