Classifying Latin Inscriptions of the Roman Empire: A machine-learning approach

Vojtěch Kaše[a,b], Petra Heřmánková[a] and Adéla Sobotková[a]

[a] Aarhus University, Department of History and Classical Studies, Jens Chr. Skous Vej 5, Aarhus, 8000, Denmark

[b] University of West Bohemia, Department of Philosophy, Sedláčkova 19, 30514, Plzeň, Czech Republic

Abstract

Large-scale synthetic research in ancient history is often hindered by the incompatibility of taxonomies used by different digital datasets. Using the example of enriching the Latin Inscriptions from the Roman Empire dataset (LIRE), we demonstrate that machine-learning classification models can bridge the gap between two distinct classification systems and make comparative study possible. We report on training, testing and application of a machine learning classification model using inscription categories from the Epigraphic Database Heidelberg (EDH) to label inscriptions from the Epigraphic Database Claus-Slaby (EDCS). The model is trained on a labeled set of records included in both sources (N=46,171). Several different classification algorithms and parametrizations are explored. The final model is based on Extremely Randomized Trees algorithm (ET) and employs 10,055 features, based on several attributes. The final model classifies two thirds of a test dataset with 98% accuracy and 85% of it with 95% accuracy. After model selection and evaluation, we apply the model on inscriptions covered exclusively by EDCS (N=83,482) in an attempt to adopt one consistent system of classification for all records within the LIRE dataset.

Keywords

Latin inscriptions, document classification, comparative analysis, Roman Empire

1. Introduction

A principal goal of digital scholarship is to produce new insights through the aggregation and synthesis of many sources (cf. [1] [2] [3] [4] [5] [6] [7]). Our ability to address fundamental historical questions, such as the waxing and waning of cities and civilizations, depends on our capacity to effectively reuse and integrate large evidentiary datasets [8] [9]. Data integration is a process of transforming datasets that were recorded in different ways into a single unified dataset with analytically comparable observations [10]. Achieving effective integration is often hindered by heterogeneous classification systems employed by different sources. The Epigraphic Database Heidelberg (EDH) and The

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

kase@cas.au.dk (V. Kaše); petra.hermankova@cas.au.dk (P. Heřmánková); adela@cas.au.dk (A. Sobotková)

0000-0001-8736-7152 (V. Kaše); 0000-0002-6349-0540 (P. Heřmánková); 0000-0002-4541-3963 (A. Sobotková)

CEUR Workshop Proceedings (CEUR-WS.org)

Epigraphik Datenbank Clauss-Slaby (EDCS) projects, for example, catalogue Latin inscriptions from the ancient Mediterranean, but utilize incompatible categories in their description. These discrepancies need to be systematically resolved before comparative analysis can proceed. In this paper, we reconcile inscription type categories in these two sources by training, testing, and applying machine-learning classification models.

2. Material & Methods

2.1. EDH

The Epigraphic Database Heidelberg (EDH) represents a flagship resource for the field of Latin digital epigraphy. It has been in development for 35 years, providing meticulously curated content and built with the consideration of open-research needs such as easy accessibility and reuse [11]. In 2021, the EDH dataset included over 81,000 inscriptions, offering a balanced distribution of material from the Western and Northern Roman provinces from the first century BC to the fourth century AD and facilitating quantified spatio-temporal studies of the Empire [12].

The EDH dataset is programmatically accessible via the public API [13]. Alternatively, researchers can access the data in raw EpiDoc format, an XML/TEI standard format for digital publication of inscriptions [14]. An older version of the data is stored at Zenodo and GitHub archive.

2.2. EDCS

The Epigraphik Datenbank Clauss-Slaby (EDCS) represents the most extensive digital resource for Latin epigraphy, containing over 500,000 inscriptions collected from printed publications or other digital sources. EDCS covers the Latin epigraphic production in the entire Mediterranean with the bulk of data originating from Rome, dated between the first and fourth centuries AD. Its limitations include no support for programmatic access and frequent omissions in temporal data as well as other inscription descriptors. Even after streamlining and data enrichment from available linked sources, the enriched EDCS dataset contains 29 attributes, compared to 74 attributes in the enriched EDH dataset. If a researcher can work within these constraints, EDCS offers an unparalleled spatial and temporal coverage.

2.3. LIRE: combining EDH & EDCS

The **LIRE** dataset represents an aggregate of the streamlined and enriched EDH and EDCS datasets (as published on Zenodo). The process of aggregation and filtering included several steps:

- mapping and deduplication of the records included in both sources, using the EDH-ID listed in the 'links' attribute of EDCS and linked data in the Trismegistos TexRelations API;

- filtering for records with valid geospatial data, represented by a pair of coordinates
- filtering for records that fall within the boundaries of the Roman Empire at its largest extent (under Trajan in AD 117; as delimited by the Pleiades project shapefile);
- filtering for records containing temporal information in the form of a temporal interval of creation, expressed in years and stored in attributes 'not_before' and 'not_after';
- filtering for records whose temporal interval of creation intersects with the timespan of the Roman Empire (arbitrarily set to 50 BC through AD 350).

The deduplication and filtering reduced the number of records in the aggregate substantially: the initial 500,000+ records in EDCS and 81,000+ records in EDH produced 137,305 records that had a valid date and location within the boundaries of the Roman Empire. The resulting LIRE dataset contained 49,916 inscriptions shared by the EDH and EDCS, inheriting attributes from both parent collections. In addition, there were 3,907 inscriptions recorded exclusively in EDH and 83,482 inscriptions originating solely from EDCS, containing set of attributes only from the parent dataset. After combining the valid unique records from the two sources and appending their attributes, we sought to integrate the attributes. The following attributes were shared and used consistently across the two sources, and could be combined in a straightforward way, with data originating in EDH taking precedence over EDCS whenever the attributes overlapped (cf. [12]):

- 'clean_text_interpretive_word': text of the inscription without the Leiden Conventions for editorial markup of texts;
- 'not_before': start of the chronological interval ('terminus post quem');
- 'not_after': end of the chronological interval ('terminus ante quem');
- 'geography': latitude/longitude point coordinates

.

Some of the attributes shared by both sources, however, could not be easily integrated such as the EDH attribute 'type_of_inscription_clean'. Type of inscription is an interpretive category that labels the function of an inscription following one of established domain typologies. Epigraphers classify inscription function after evaluating its content, context, and physical form during the first publication. As type definitions are broad and ambiguous, label assignment is subjective and may fluctuate through time.

The 'type_of_inscription_clean' column in EDH implements the controlled vocabularies of the EAGLE Europeana Project, a standardized list of inscription types created in 2013-2015 to tackle the vagueness of existing typologies [15] [16]. Single label per inscription and effort invested into standardisation were the main reasons we decided to use EDH as the training dataset for the present model.

EDCS stores the type of inscription information in an attribute 'status_list' together with other information extracted from the text of inscriptions, such as the social status of persons named in the text, e.g. slaves, priests, or high-ranking officials. We extracted the type values into a separate 'inscr_type' attribute. The resulting column, however, was often multi-valued and followed a different typology than EDH, relying on Latin labels and referring to non-overlapping categories. The category 'owner/artist inscription' in EDH, for example, corresponds both to 'tituli possessionis' and 'tituli fabricationis' in EDCS, hampering dataset-wide comparison. To overcome such limitations and have a consistent typology applied across the entire dataset, we reclassified the inscriptions from EDCS using the inscription type categories from EDH.

All scripts used for aggregation, filtering and enriching of the LIRE dataset are available on GitHub. The repository includes training, evaluation, selection, and application of the classification model introduced below. The final version of the dataset is also published via Zenodo [17].

2.4. Classification task

We trained a machine learning classification model using the EDH-labeled subset of inscriptions with attributes from both sources. From the 49,916 inscriptions which are shared by both the EDH and EDCS dataset, 46,171 records are properly labeled, i.e. classified using the 'type_of_inscription_clean' attribute in EDH. These were used for training of the classification model. The remaining 3,745 inscriptions were classified as 'NULL' and thus were excluded from the training.

There are 22 unique classification categories among EDH labels. Their distribution is highly imbalanced (see Table 4). The three most common categories in the training set are 'epitaph' (N=21,520) 'votive inscription' (N=11,728), and 'owner/artist inscription' (N=3,340). The three least common are 'assignation inscription' (N=15), 'calendar' (N=10), and 'adnuntiatio' (N=1).

For a preliminary model selection, we compared outcomes of several supervised machine learning algorithms commonly used for document classification [18], namely:

- Logistic Regression (LR) [19]
- Support-vector Machine (SVM)[20] [18]
- Random Forests (RF) [21] [18]
- Extremely Randomized Trees (ET)[22]

All the algorithms have been implemented using Python 3 [23] and the Scikit-learn library [24], using standardized recipes based on [25].

2.5. Features extraction and selection

For training of the models, we combined features extracted from several different EDCS attributes, namely:

- 'status_list', including information about the inscription category according to EDCS combined with other metadata ('status/titulorum distributio' in EDCS)
- 'Material', containing information about the predominant material or medium on which the inscription is found
- 'clean_text_interpretive_word', text of the inscription without the Leiden Conventions for editorial markup of texts, see [12]

The content of the three attributes was extracted independently, preprocessed, and then combined together into a 'bag-of-words' model to feed a *tfidf* vectorizer [26]. Underscore was used to treat any multi-word feature as one-word feature (e.g. 'tituli_fabricationis' instead of 'tituli fabricationis'). The text of the inscriptions in the attribute 'clean_text_interpretive_word' was treated on the level of continuous bigrams, since bigrams are suitable to capture the formulaic language of inscriptions. As a result of this preprocessing, we obtained a list consisting of the following features:

- 37 features based on unique values from the 'status_list' attribute
- 18 features based on unique values from the 'Material' attribute (e.g. 'lapis', 'opus_figlinae', 'aes')
- 100, 1,000 or 10,000 features based on a corresponding number of the most frequent bigrams from the text of the inscriptions (e.g. 'Dis_Manibus', 'vixit_annos' or 'votum_solvit')

3. Results

3.1. Classification model selection and evaluation

To evaluate performance of each model variant, we relied mainly on weighted variant of the F_1 score, referred below as $F_1(w)$. The $F_1(w)$ score is the harmonic mean of *Precision* (proportion of every observation predicted to be positive that is actually positive) and *Recall* (proportion of every positive observation that is truly positive). The weighted variant of the $F_1(w)$ score metric, which is consistently reported below, takes into account label imbalances: for each label, the average value of both metrics is weighted by the number of true instances. In some cases, we also report accuracy, which equals to the proportion of correctly classified records, what makes it very intuitive. But it has to be taken with reservation here, since it does not take into account class imbalances.

For a preliminary model evaluation, we explored different combinations of the above mentioned feature groups and trained and tested different models on a subset of 5000 labeled records (80 % for training and 20 % for testing, using *stratified k-fold cross-validation* method). For instance, using the 37 features based on the 'status_list' attribute only, LR resulted in $F_1(w)$=0.724. The performance of the model slightly improved when we added 18 features based on the 'Material' attribute ($F_1(w)$=0.734); but the model improved substantially when we included the bigrams, from $F_1(w)$=0.786

Table 1
Classification model selection (training set N=4000)

classifier	C	n_estimators	avg. $F_1(w)$
LR	1		0.808297
LR	1000		0.831937
LR	10000		0.830611
SVM	1		0.310482
SVM	1000		0.760049
SVM	10000		0.828478
RF		10	0.815809
RF		100	0.825325
RF		1000	0.826737
ET		10	0.822891
ET		100	0.831064
ET		1000	0.830801

for the 100 most frequent bigrams up to $F_1(w)$=0.832 for the 10,000 most frequent bigrams (the reported $F_1(w)$ is an average based on 5 stratified cross-fold validation tests). Thus, for the subsequent model selection we employed the features set including 10,000 bigrams.

Table 1 shows the differences in performance of the above introduced classification algorithms trained on a subset of 4,000 randomly chosen inscriptions and the above specified features set. In the case of LR and SVM, we tested different settings of the C parameter, which stands for inverse regularization strength. In the case of RF and ET, we explored several different values for the number of estimators. We see that the best results are achieved by LR (C=1000) and ET (*n_estimators*=100).

Drawing on these results, we continued with LR and ET only and trained them on the full dataset. In this setting, ET significantly outperforms LR, with $F_1(w)$=0.878 over $F_1(w)$=0.867 (the reported $F_1(w)$ is an average based on 10 stratified cross-fold validation tests). On the basis of these results, we continued with the ET model which we also saved for future reuse.

3.2. Probabilities and precision table

Each prediction of the model is accompanied by **probability** on scale 0-1, expressing a level of certainty concerning the predicted classification category. The resulting probabilities might be used to formulate thresholds under which the classification will not be accepted. In Table 2, we see that 96 % of inscriptions in the test dataset were classified with probability equal to- or higher than 0.4 and that this classification was correct in more than 90 % of cases (see the 'accuracy' score column). Further, approximately 85 % of inscriptions have been classified with probability equal to or higher than 0.6. From these more than 95 % was classified correctly. These are important observations, which might be used later on when we apply the model upon unlabeled data, where we can expect comparable ratios between threshold values, proportions of covered inscriptions,

Table 2
Classification model test results

threshold ($\geq$)	proportion	N	$F_1(w)$	accuracy
0.40	0.96	4448	0.897102	0.905800
0.45	0.91	4225	0.923481	0.929467
0.50	0.89	4123	0.935463	0.940092
0.55	0.87	4027	0.945704	0.949839
0.60	0.85	3936	0.952090	0.955539
0.65	0.83	3853	0.957309	0.960550
0.70	0.81	3755	0.962801	0.965379
0.75	0.79	3653	0.965969	0.968519
0.80	0.76	3526	0.970297	0.972490
0.85	0.70	3253	0.978090	0.979096
0.90	0.67	3082	0.979371	0.980208
0.95	0.60	2762	0.981634	0.982259

and the extent of correct classifications.

However, before we proceed to apply the model on unlabeled data, we have to evaluate the performance of the model with respect to individual categories. For that purpose, we generated a **precision table** in Figure 1, where we see the model's accuracy with respect to 10 most common categories. We see that the accuracy differs from category to category. In case of 4 categories ('epitaph', 'votive inscription', 'mile-/leaguestone', and 'defixio') the model correctly classifies 98 % or more records. For instance, it correctly classifies all 18 instances of 'defixio' in the test set. In the case of other categories, the performance is much worse: e.g. from 4 instances of 'list', 3 are incorrectly classified as 'epitaph'. The ambiguity of 'list' definition in the EAGLE vocabularies likely causes the inconsistent manual markup in EDH and the poor reclassification performance.

3.3. Classification model application

After tuning, training, and testing the model on labeled data, we proceeded to apply the model on 83,482 inscriptions which are recorded exclusively in EDCS. Further, employing the 0.6 probability threshold, we accepted the automatic labels for 82 % of inscriptions in the dataset. To estimate the proportion of correctly classified inscriptions within this subset, we generated a random sample of 100 inscriptions. The sample was labeled manually by two domain experts. The first expert was drawing on the same attributes as the automatic classifier. In this case, the manually and automatically assigned labels were in agreement in 94% of cases. Another domain expert manually labeled the data without taking into consideration the 'status_list' attribute. In this case, the agreement with the automatically assigned labels was approximately 88%. Combining this with the results from the test set, we estimate between 90 and 95 % of the automatically assigned categories to be correct (see Table 3).

As a result of applying the ET classification model, we enriched the whole LIRE dataset by adding two new attributes:

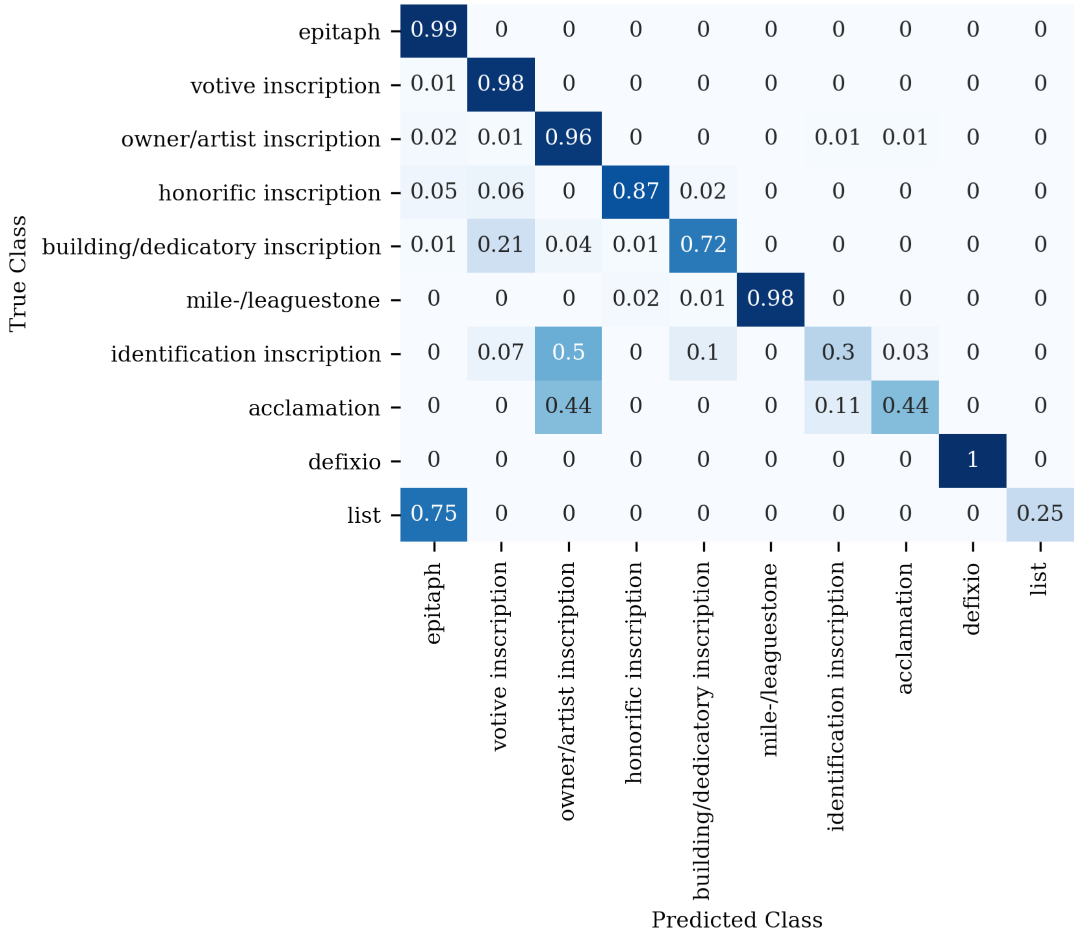

Figure 1: Precision table for 10 most common inscription categories (only inscriptions classified with probability equal to- or higher than 0.6 included.)

- 'type_of_inscription_auto', containing either the predicted label or - where available - the label from 'type_of_inscription_clean' as recorded in EDH
- 'type_of_inscription_auto_prob', expressing the probability on the scale from 0 to 1 (1 is used for datapoints where the 'type_of_inscription_clean' from EDH was used)

When we look at the LIRE dataset as a whole, we see that from the 137,305 inscriptions, 117,710 (85 %) are classified in 'type_of_inscription_auto' with probability equal to- or higher than 0.6. In the following overview of LIRE, we use this probability as a cut-off threshold, under which the automatically assigned categories are not accepted and the corresponding records are excluded from the cross-category comparison.

Table 3
Classification model main application results

threshold (≥)	proportion	N
0.40	0.96	80114
0.45	0.87	72302
0.50	0.85	70967
0.55	0.83	69502
0.60	0.82	68226
0.65	0.80	66784
0.70	0.78	65299
0.75	0.76	63790
0.80	0.74	61813
0.85	0.71	59179
0.90	0.68	56699
0.95	0.62	51400

Table 4
Classified dataset overview

inscr. type	EDH	EDCS	inscr. type	EDH	EDCS
epitaph	22902	52222	label	213	76
votive inscription	12328	3523	boundary inscription	181	233
owner/artist inscription	3851	6726	elogium	135	16
honorific inscription	3089	2511	letter	124	168
building/dedicatory inscription	2699	580	public legal inscription	119	16
mile-/leaguestone	1413	1141	seat inscription	46	9
identification inscription	1182	745	private legal inscription	38	0
acclamation	364	84	prayer	20	3
defixio	273	21	assignation inscription	16	0
list	265	32	calendar	11	3
military diploma	214	117	adnuntiatio	1	0

3.4. Cross-category comparison of the LIRE dataset

Table 4 shows the distribution of individual inscription categories across the aggregated dataset. The EDH column covers inscriptions for which the category was already available. Most of them are shared between the two datasets and used for training of the model. The EDCS column covers inscriptions exclusively from EDCS, to which the inscription type category was assigned automatically by the model. It contains only records with assignment probability equal to 0.6 or higher.

In Figure 2 we see the temporal distribution of the six most common types of inscriptions, grouped by their original source. However, since many inscriptions are dated rather vaguely by means of extensive temporal intervals (commonly on a century basis), an evaluation of temporal trends in the data is far from straightforward. To overcome this, we employ a Python implementation [27] of a Monte Carlo simulation approach to temporal uncertainty based on [28]. First, we generate 1,000 time series simulations. In each simulation, each inscription is randomly dated to a singular year within its interval of creation. Subsequently, these simulations are plotted cumulatively by means of frequency polygon lines based on 50-year-long time blocks. The width of the resulting curve, which changes from time block to time block, reflects the extent of temporal uncertainty in the underlying data. Adopting this approach, we are able to assess temporal trends in the datasets and to compare the differences in temporal distributions of individual inscription types. For instance, we see that the temporal distribution of epitaphs changes substantially if we use the EDH dataset versus the LIRE dataset, which contains a larger number of epitaphs inherited from EDCS. It reveals that the spike in the production of epitaphs in EDH (a) in the second half of the second century AD disappears once we include the epitaphs from EDCS (c). This trend is only apparent after the automatic reclassification, as the original classification systems were mutually incompatible and not yet mapped onto a single ontological system [29].

4. Discussion

Many data collections in small-science disciplines are fragmented among numerous content silos. Scholars wishing to synthesise these fragments need to ensure their analytical comparability, specifically column- and value-level consistency. Such consistency has been achieved in the past through semi-automatic mapping to relevant ontologies (see tDAR example in [10]) or through loose-coupling (see OpenContext, [30]). The machine learning applied here sits between the ontology and loose coupling approaches. With sufficiently large, representative, and well-described training dataset, an algorithm learns to make interpretive decisions like a trained epigrapher. The classifier here fully-automatically extends patterns observed in 40,000 inscription labeled records to additional 80,000 unlabelled records, creating a 'type_of_inscription_auto' attribute.

Despite the relatively high accuracy rates reported above, there is still the probability that every 20th automatically classified inscription is classified erroneously. Imperfections in the training dataset due to the ambiguity of inscriptions likely contribute. We can also look to other studies for guidance. Survey pottery specialists, for example, point to ambiguity surrounding the interpretation of type and chronology in artefacts that suffer from high wear and fragmentation [31]. Epigraphic monuments are material remains. We can expect uncertainty to be inevitable in highly fragmented and short inscriptions.

Finally, when integrating two typologies, the choice will always entail a compromise that carries with it some limitations. Inspecting the LIRE inscriptions by type across the long-term in Figure 2(c), most major trends from EDH (Figure 2a) and EDCS (Figure 2b) are preserved (e.g. dominance of epitaphs and votives) with subtle alterations to trajectories due to the combined data. Detailed labels from EDCS 'tituli fabricationis' and 'tituli possesionis' have been combined into one overarching category 'owner/artist inscription', which might be a problem if the originals have special value for the next researcher. The one category that disappears is the 'tituli christianae', utilized by EDCS

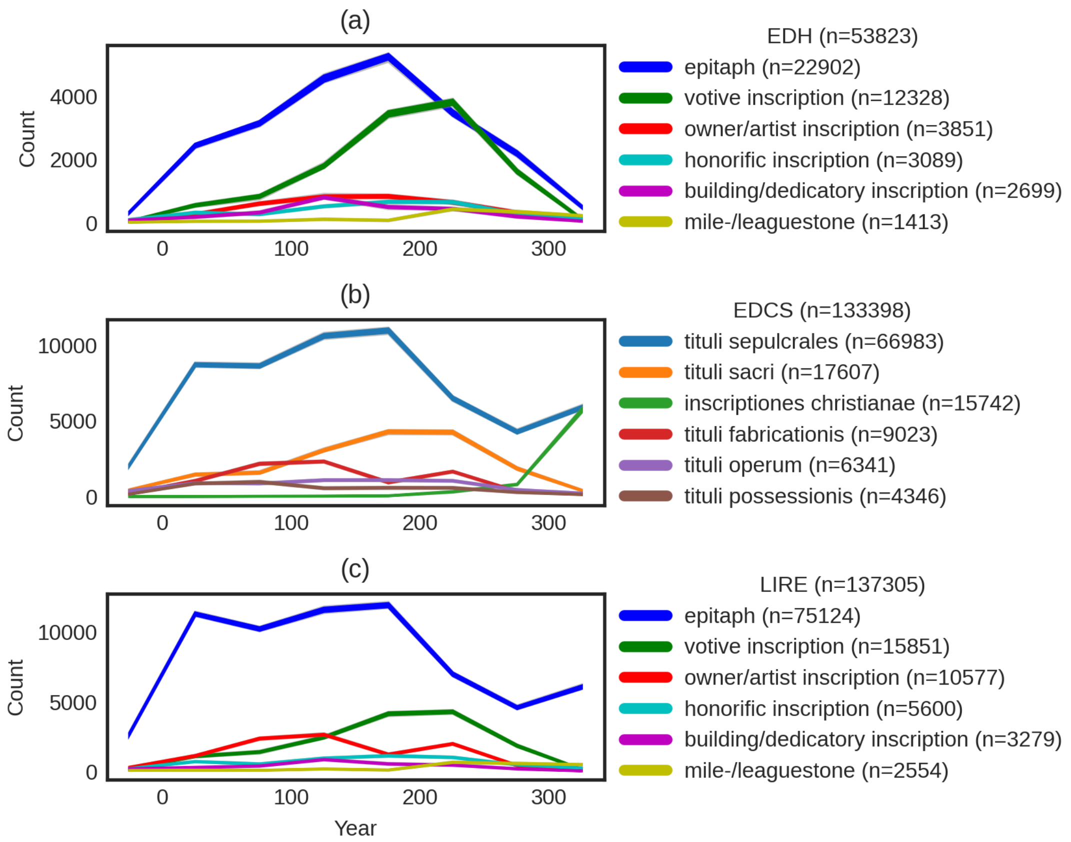

Figure 2: Temporal distribution of the six most common inscription types in EDH (a), EDCS (b), and LIRE (c). All three subplots are filtered to display only the records covered by LIRE. (a) Inscription types as labelled by EDH, one label per inscription; (b) inscription types as labelled by EDCS, where multiple labels per inscription are allowed; (c) aggregate of manually (EDH) and automatically (EDCS) classified inscriptions based on the EDH classification system.

but absent from EDH. It was relabelled as 'epitaph' in the LIRE dataset, causing a secondary rise associated with this category around AD 300. This may be a loss for scholars of Early Christianity, but represents a move towards greater consistency from a cultural label towards a functional description. While LIRE accomplishes our needs of comparability, the approach is flexible and the selection of classification can be flipped around should other scholars need it.

5. Acknowledgments

This research was funded by the Aarhus University Forskningsfond Starting grant no. AUFF-E-2018-7-22 awarded to the 'Social Complexity in the Ancient Mediterranean' (SDAM) project.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: http://www.nature.com/articles/sdata201618. doi:10.1038/sdata.2016.18.

[2] B. S. Arbuckle, S. W. Kansa, E. Kansa, D. Orton, C. Çakırlar, L. Gourichon, L. Atici, A. Galik, A. Marciniak, J. Mulville, H. Buitenhuis, D. Carruthers, B. D. Cupere, A. Demirergi, S. Frame, D. Helmer, L. Martin, J. Peters, N. Pöllath, K. Pawłowska, N. Russell, K. Twiss, D. Würtenberger, Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey, PLOS ONE 9 (2014) e99845. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0099845. doi:10.1371/journal.pone.0099845.

[3] K. Manning, S. Colledge, E. Crema, S. Shennan, A. Timpson, The Cultural Evolution of Neolithic Europe. EUROEVOL Dataset 1: Sites, Phases and Radiocarbon Data, Journal of Open Archaeology Data 5 (2016) e2. URL: http://openarchaeologydata.metajnl.com/articles/10.5334/joad.40/. doi:10.5334/joad.40.

[4] K. Smiarowski, R. Harrison, S. Brewington, M. Hicks, F. J. Feeley, Dupont-HébertCéline, B. Prehal, G. Hambrecht, J. Woollett, T. H. McGovern, Zooarchaeology of the Scandinavian Settlements in Iceland and Greenland: Diverging Pathways, in: U. Albarella, M. Rizzetto, H. Russ, K. Vickers, S. Viner-Daniels (Eds.), The Oxford Handbook of Zooarchaeology, Oxford University Press, 2017, pp. 147–163.

[5] B. J. Mills, J. J. Clark, M. A. Peeples, W. R. Haas, J. M. Roberts, J. B. Hill, D. L. Huntley, L. Borck, R. L. Breiger, A. Clauset, M. S. Shackley, Transformation of Social Networks in the Late pre-Hispanic US Southwest, Proceedings of the National Academy of Sciences 110 (2013) 5785–5790.

[6] I. McKechnie, D. Lepofsky, M. L. Moss, V. L. Butler, T. J. Orchard, G. Coupland, F. Foster, M. Caldwell, K. Lertzman, Archaeological Data Provide Alternative Hypotheses on Pacific Herring (Clupea pallasii) Distribution, Abundance, and Variability, Proceedings of the National Academy of Sciences 111 (2014) E807–E816.

[7] Forcell, The Fair Data Principles, 2021. URL: https://www.force11.org/group/fairgroup/fairprinciples.

[8] K. W. Kintigh, J. H. Altschul, M. C. Beaudry, R. D. Drennan, A. P. Kinzig, T. A. Kohler, W. F. Limp, H. D. G. Maschner, W. K. Michener, T. R. Pauketat, P. Peregrine, J. A. Sabloff, T. J. Wilkinson, H. T. Wright, M. A. Zeder, Grand Challenges

for Archaeology, Proceedings of the National Academy of Sciences 111 (2014) 879–880.
[9] J. H. Altschul, K. W. Kintigh, T. H. Klein, W. H. Doelle, K. A. Hays-Gilpin, S. A. Herr, T. A. Kohler, B. J. Mills, L. M. Montgomery, M. C. Nelson, S. G. Ortman, J. N. Parker, M. A. Peeples, J. A. Sabloff, Opinion: Fostering Synthesis in Archaeology to Advance Science and Benefit Society, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) 10999–11002.
[10] K. W. Kintigh, K. A. Spielmann, A. Brin, K. S. Candan, T. C. Clark, M. Peeples, Data Integration in the Service of Synthetic Research, Advances in Archaeological Practice 6 (2018) 30–41. URL: https://www.cambridge.org/core/product/identifier/S232637681700033X/type/journal_article. doi:10.1017/aap.2017.33.
[11] A. Hartmann, Rezension zu: Epigraphische Datenbank Heidelberg, H-Soz-Kult (2018). URL: https://www.hsozkult.de/webreview/id/rezwww-179.
[12] P. Heřmánková, V. Kaše, A. Sobotková, Inscriptions as data: digital epigraphy in macro-historical perspective, Journal of Digital History 1 (2021).
[13] F. Grieshaber, Epigraphic Database Heidelberg – Data Reuse Options (2019). URL: http://archiv.ub.uni-heidelberg.de/volltextserver/id/eprint/26599. doi:10.11588/HEIDOK.00026599, publisher: Heidelberg University Library.
[14] T. Elliott, G. Bodard, E. Mylonas, S. Stoyanova, C. Tupman, S. Vanderbilt, EpiDoc Guidelines: Ancient documents in TEI XML, 2020. URL: https://epidoc.stoa.org/gl/latest/.
[15] P. M. Liuzzo, Mapping Databases to EpiDoc, in: S. Orlandi, R. Santucci, F. Mambrini, P. M. Liuzzo (Eds.), Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference, 2017, pp. 187–199. URL: http://www.editricesapienza.it/node/7610.
[16] P. M. Liuzzo, EAGLE Continued: IDEA. The International Digital Epigraphy Association, in: A. De Santis, I. Rossi (Eds.), Crossing Experiences in Digital Epigraphy, De Gruyter Open Poland, 2018, pp. 216–230. URL: https://www.degruyter.com/document/doi/10.1515/9783110607208-018/html. doi:10.1515/9783110607208-018.
[17] V. Kaše, P. Heřmánková, A. Sobotková, LIRE, 2021. URL: https://doi.org/10.5281/zenodo.5074774. doi:10.5281/zenodo.5074774.
[18] J. Hartmann, J. Huppertz, C. Schamp, M. Heitmann, Comparing Automated Text Classification Methods, International Journal of Research in Marketing 36 (2019) 20–38. URL: https://www.sciencedirect.com/science/article/pii/S0167811618300545. doi:10.1016/j.ijresmar.2018.09.009.
[19] H. Theil, A Multinomial Extension of the Linear Logit Model, International economic review 10 (1969) 251–259.
[20] K. Crammer, Y. Singer, On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines, Journal of machine learning research 2 (2001) 265–292.
[21] L. Breiman, andom Forests, Machine learning 45 (2001) 5–32.
[22] P. Geurts, D. Ernst, L. Wehenkel, Extremely Randomized Trees, Machine learning 63 (2006) 3–42.
[23] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts

Valley, CA, 2009.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[25] C. Albon, Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning, O'Reilly, Beijing, 2018.

[26] D. Jurafsky, J. H. Martin, Vector Semantics, in: D. Jurafsky, J. H. Martin (Eds.), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition Draft), 2017, pp. 270–285.

[27] V. Kaše, tempun, 2021. URL: https://zenodo.org/record/4650989. doi:10.5281/ZENODO.4650989.

[28] E. R. Crema, K. Kobayashi, A Multi-proxy Inference of Jōmon Population Dynamics Using Bayesian Phase Models, Residential Data, and Summed Probability Distribution of 14C Dates, Journal of Archaeological Science 117 (2020) 105136. URL: https://linkinghub.elsevier.com/retrieve/pii/S0305440320300583. doi:10.1016/j.jas.2020.105136.

[29] Bodard, Cayless, Cenati, Cooley, Elliott, Evangelisti, Felicetti, Granados, Grieshaber, Gruber, Hershkowitz, Hill, Kiiskinen, Kollatz, Levivier, Liuzzo, Luciani, Mannocci, Mataix, Murano, Murphy, Mylonas, Prag, Razanajao, Stoyanova, Tsolakis, Tupman, Vagionakis, Vitale, Weise, Modeling Epigraphy with an Ontology (2021). URL: https://zenodo.org/record/4639507. doi:10.5281/ZENODO.4639507, publisher: Zenodo Version Number: 0.1.

[30] E. C. Kansa, A. Bissell, Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains, Data Science Journal 9 (2010) 42–53.

[31] A. Bevan, J. Conolly, C. Hennig, A. Johnston, A. Quercia, L. Spencer, J. Vroom, Measuring Chronological Uncertainty In Intensive Survey Finds: A Case Study From Antikythera, Greece, Archaeometry 55 (2013) 312–328.