We are Brian Joseph and Christopher Brown, part of a team at The Ohio State University, working on a larger overall project, the Herodotos Project, and a specific subpart that NEH has funded under the Digital Humanities rubric.

Historians working on ancient times have focused largely on noteworthy individual persons and places, and a result of that focus is that no comprehensive catalogue exists of peoples in ancient times, i.e. tribes, clans, and other groups.

We are aiming to fill that gap and to create such a catalogue and then to work with that catalogue to develop an integrated data base of the peoples mentioned in ancient classical sources, drawing on different fields of research for the data about them, fields ranging from philology and textually based work to archaeology and even genetics.

In our first phase, we are working to automate the identification of such group names in ancient sources, working first with Latin and Greek texts, utilizing the
computational methodology known as Named Entity Recognition, and it is for the development of such a system that we received our grant.

[SLIDE 2]:

What is NAMED ENTITY RECOGNITION (NER)?

Named Entity Recognition is a computational machine-learning system for recognizing important pieces of information in text, *entities* that are identified ("named"), e.g. persons, places and groups.

We began working with the existing English-based NER and translations, with the help of David Smith's work annotating the text of Herodotos for the Perseus project, with an assist from David Bamann's modification of an existing off-the-shelf English-based NER for our use, but for comprehensiveness and for accuracy we decided it is better to work in the original languages.

With NEH help we are perfecting our Latin NER, and developing an NER for Greek.

What characterizes these ancient languages is resource-sparsity of computational resources and relatively limited data sets.

Our project involves working with diverse authors and genres such as historiography, poetry, letters, natural history (e.g. Caesar, Ovid, Pliny).

This project will be useful to other scholars working in Greek, Latin and eventually other ancient languages such as Old Church Slavonic.

[SLIDE 3]:

Our solution to issues raised by NER in Latin for group names is to target the needs of the digital historian whose time is scarce and whose primary concern is accuracy. Thus, we select sentences to annotate for named entities which will be most informative to the development of and continued progress with our NER
model. The model's expected accuracy will increase as a function of how much additional annotation one feeds into it. In this way, the digital historian can control for the quality with which entities are recognized.

We have a team of student annotators currently working on Latin texts, annotating them for use as the training data for our NER system. We then test the efficacy of the system on unannotated data, and then allow the system to enter into a feedback loop, as it were, so that it learns from the tests and improves its performance.

We have reached accuracy levels of over 95% for the identification of group names, working with our Latin-based system on Latin texts.

We see this system and general approach as being of use not just to our particular goals but to digital historians more generally, especially once it is extended to languages other than Latin and once our overall project is integrated with partners like the Pelagios Project and the Perseus Project.

Thank you for listening, and thank you, NEH!
With our HERODOTOS PROJECT for ethnohistory of the ancient world, we aim to:

• catalogue peoples mentioned in ancient sources
• assemble a detailed profile of each

We are working with Latin and Greek texts in this initial phase, aiming to:

• automate the group name extraction process, drawing on Named Entity Recognition technology.

Co-PIs: Profs. Brian D. Joseph, Micha Elsner, Marie-Catherine de Marneffe (Linguistics) & Dr. Christopher Brown (Classics), with Alex Erdmann (Linguistics)

https://u.osu.edu/herodotos/
NAMED ENTITY RECOGNITION – a computational machine-learning system for recognizing entities in text that are identified ("named"), e.g. persons, places and GROUPS

Existing English-based NER allows work with translations, but it is desirable:
• to avoid possible bias of the translator
• to be able to process untranslated texts

We are developing original-language NER for texts in resource-sparse ancient languages that have limited data in diverse genres such as historiography, poetry, letters (e.g. Caesar, Ovid, Pliny)

General issues:
• How much genre adaptation required?
• Time is scarce, annotation (for training machine-learning algorithms) expensive, accuracy crucial
• Extending Latin-based system to Greek and other languages
A solution: an Efficient Active Learning strategy addressing problematic words annotation is key to training the system.

1. Identify subsets of difficult words and rank test sentences for annotation.
2. Annotate ranked sentences until predicted accuracy is sufficiently high.
3. Move newly annotated sentences to train and re-weight the model.
4. Learn updated model and tag the remaining unannotated data.

- Annotated train data
- Unannotated test data
- Predicts expected accuracy per additional sentence annotated
- Manually labeled informative data from test
- Automatically labeled remaining test data