

White Paper for Grant HK-50161-14

Berkeley Prosopography Services: Implementing the Toolkit

Principal Investigators: Niek Veldhuis, Laurie Pearce, University of California, Berkeley

Date Due: December 31, 2018

Date Submitted: February 9, 2019

Table of Contents:

1. Project Description	3
2. Accomplishments	3
3. Participants and Collaborating Organizations	5
4. Changes/Problems	8
5. Special Reporting Requirements (N/A)	14
6. Project Outcomes	14
7. Technical Products	19
8. Supplementary Materials	21

1. Project Description[†]

Berkeley Prosopography Services (BPS) streamlines prosopographical research by offering researchers a customizable out-of-the-box toolkit and environment for their work. The toolkit includes a user-customizable **probabilistic disambiguator**, a program that determines the likelihood that two or more instances of the same name refer to the same person; a **Social Network Analysis engine** that computes, utilizing well-established SNA metrics, the mathematical measures that define the social network; and a **graph visualizer** that automatically generates interactive visual representations of the social networks reflected in the data set.

Prosopography, the identification and disambiguation of name instances and the study of the interactions and life histories of the people involved, is at the core of many humanities research agendas. BPS offers a set of tools that facilitate various steps in this process and an environment in which researchers may dynamically interrogate different models of the same data set. The corpus-agnostic architecture of BPS assures that its disambiguation tools are available to researchers working in diverse languages and periods of history.

During the grant period, BPS developed preprocessing support to facilitate input of existing digitized data sets into BPS, refined core functionality in response to user requests, and integrated software modules to ensure a seamless flow of data from one research stage to the next. We continued our program of community engagement, on campus, in the digital humanities community, and in humanities research.

2. Accomplishments

The goal of this project was to extend the functionality of Berkeley Prosopography Services (BPS), an interactive toolkit for analyzing and visualizing datasets, and to expand its accessibility and utility to researchers working with data across diverse disciplines. During this implementation phase of the project, activities focused on building on the project's existing software base, as well as its sound conceptual and architectural structure, and targeted specific areas of technical development and increased user functionality:

- extending preprocessing support to facilitate input of existing digitized data sets into BPS:

User input had already identified ease of inputting digitized data sets into BPS as a priority for a positive user experience. In response, we developed the TEI4BPS converter tool that automates the conversion of users' database files (in CSV format) to the TEI the BPS disambiguation engine ingests (the code is available at <https://github.com/berkeleyprosopography/TEI4BPSConverter>). This change reflects (a) the robustness of the BPS architecture, which allows for modifications in response to user requests without triggering need for major code revision, and (b) the evolution of the dialogue between the BPS technical team and domain specialists. Our original intention to have BPS directly ingest TEI was challenged both by users who identified the need to generate TEI tagged text added a major,

[†] Numerals in square brackets (e.g., [1], etc.) refer to items in the list of products, for which a table of contents is provided on p. 21

unwelcome time commitment to their workflow, and in the instability of XML IDs produced by the [Oracc](#) TEI generator. To be sure, we gratefully acknowledge the support of the Oracc technical team in creating a mechanism to generate TEI from texts in Oracc subprojects. Over the course of our discussions and testing with Oracc corpora, we experienced the reality that well-architected projects, using best engineering practices and open-source code does not guarantee interoperability or ease of communication between such projects, even (or especially) when the project and/or toolkit is designed to support growth and exploration of a research project. In Oracc, as a researcher develops his/her corpus, the process of rebuilding renders mutable the XML IDs for personal names. As BPS is predicated on probabilistic disambiguation of name instances, secure identifiers are crucial to anchor name instances in specific texts as researchers add to the data set and/or modify the parameters used in disambiguation. The development of a tool to convert database content to BPS compliant TEI, which we term TEI4BPS, eliminated a bottleneck in the UX and stabilized the form of the content provided to the BPS disambiguation engine.

- refining core functionality in response to user requests and integrating software modules to ensure a seamless flow of data from one research stage to the next

End-to-end integration of the BPS toolkit components was achieved. Users can: (a) upload corpora; (b) using the role matrix, identify (in)compatible roles for similar name instances; (c) submit these inputs to the disambiguation engine; and (d) generate, in real time, a social network visualization of the computed graph. Functionality of the following features has yet to be implemented: sharing or freezing of workspaces; recording assertions; export of the graph.

- continuing our program of community engagement, on campus, in the digital humanities community, and in humanities research

We are proud of our active community engagement on campus. Events and programs in which the PIs and technical lead participated are listed in §6.2. Here, we note the impact that the involvement of the co-PIs has had on pedagogy in the Cuneiform Studies program and in the Department of Near Eastern Studies as a whole. Both Veldhuis and Pearce have integrated digital humanities into their teaching and mentoring. Veldhuis' undergraduate researchers worked on the Social Network Analysis of a corpus of late third millennium Sumerian administrative texts from Puzriš-Dagan (modern Drehem). Both Veldhuis and Pearce have successfully recruited and mentored undergraduate researchers through the campus [URAP](#) (Undergraduate Research Apprentice Program). Pearce's undergraduate researchers have been more directly involved in BPS-related activities, developing and maintaining the development corpus (HBTIN), and generating additional content associated with the presentation of that corpus. The graduate students in Cuneiform Studies have all received training in digital humanities research tools and methods; through their task-specific involvements with BPS, Eduardo Escobar and Gil Breger have integrated social network analysis to their own research agendas. Breger and Jason Moser have participated in the collaboration with the Department of Ancient History at the Ludwig-Maximilians-Universität München, as noted below.

An unexpected and exciting outcome of the project has been the development of research opportunities and projects between BPS team members and researchers at other universities, notably the Ludwig-Maximilians-Universität München. Our teams have developed a research agenda that includes re-examination and documentation of the heuristics employed in a set of established, well-respected and widely used studies (S. Parpola, *Letters from Assyrian Scholars* [LAS]) of the scholars associated with the court of the Neo-Assyrian empire (c. 900-600 BCE; northern Mesopotamia). Whereas BPS supports documentation of all assertions, the assumptions and heuristics underlying the disambiguations in the LAS corpus are poorly, if at all, articulated. Members of the BPS team have embarked on research aimed at redefining the disambiguation of individual scholars, and, with the help of digital tools unavailable to the LAS project, recomputing the astronomical and chronological data that contributed to its conclusions. The BPS and LMU teams are committed to promoting full documentation of assumptions, heuristics, and computational methods to establish reproducibility and replicability in humanities research.

3. Participants and Collaborating Organizations

3.1 The BPS team

3.1.1 The core-team: co-PIs Niek Veldhuis, Laurie Pearce, both of the Department of Near Eastern Studies, UC Berkeley, and technical lead Patrick Schmitz, Associate Director, Research IT, UC Berkeley

3.1.2 Additional staff members and affiliates

3.1.2.1 Staff

- Davide Semenzin, MS. Software engineer, Internet Archive. Semenzin’s affiliation with BPS began with his unpaid internship with the project during Spring 2012, when he worked with Pearce and Schmitz in developing some of the early articulations of the conceptions and frameworks of the scholarly workflows as part of his Master’s thesis for the Department of Business Informatics at the Utrecht University (NL). With the award of the NEH grant, we invited Semenzin to join the staff as a core software developer. As an independent contractor, he worked under Schmitz and was responsible for developing the simplified input model—,preprocessing CSV files into the standard XML format our services take as input (including coordination with student interns for the design and some of the implementation); for refactoring the visualization code and upgrading the core rendering libraries to be more performant; for refactoring the deployment stack using “dev-ops” tools so that BPS can be easily and consistently set up by new developers and be easily deployed by the maintenance team; for helping to integrate the core services and the visualization tools; and for numerous smaller functionality improvements and bug fixes.
- Terri-lynn Tanaka, PhD. Visiting Scholar, Department of Near Eastern Studies, UC Berkeley. Project-Policy Analyst. January 4, 2016—Sept. 30, 2018. Tanaka worked with Pearce and Schmitz documenting project workshops, developing project documentation, and identifying development and improvement needs in the user interface and

experience. Together with Pearce and Escobar, she reviewed project workflow and contributed to the development of an improved UI.

- Eduardo Escobar, PhD Candidate, Department of Near Eastern Studies, UC Berkeley. Spring 2017. Escobar worked with Pearce and Tanaka to develop a revised workflow and UI for BPS.
- Joanne Tan, UC Berkeley undergraduate intern. March–September 2018. Tan worked under Pearce and UC Berkeley Department of Near Eastern Studies graduate student Gil Breger in developing research that evolved as a result in connection with our collaboration with colleagues at Ludwig-Maximilians-Universität München to explore using BPS in researching social networks of scholars and astronomers in the corpus of letters between scholars and the court in the Neo-Assyrian empire (c. 900–600 BCE).

3.2.1.2 Affiliates

- Gil Breger, PhD Candidate, Department of Near Eastern Studies, UC Berkeley. Spring-Summer 2018. Breger was a member of the team of the LMU-UCB collaboration, working to develop documentation of probabilistic assumptions about astronomical events utilized in the dating of a previously published corpus and analysis of letters between scholars and kings of the Neo-Assyrian empire.
- Caroline Cheung, PhD Candidate, Ancient History and Mediterranean Archaeology, UC Berkeley. Fall 2016. Cheung worked with Pearce and Semenzin in testing and refining the CSV-to-TEI script; she reviewed the output for accuracy, in consultation with research partner Langellotti, whose corpus was used for development of the script.
- Jason Moser, PhD Candidate, Department of Near Eastern Studies, UC Berkeley. Spring-Summer 2018. Moser was a member of the team of the LMU-UCB collaboration, working to develop NLP methods of assessing the paleographic criteria used to distinguish scribal hands (and by extension, to disambiguate individual scribes) documented in the previously published corpus of letters between scholars and kings of the Neo-Assyrian empire.
- Gautami Sharma, Undergraduate, UC Berkeley. Summer-Fall 2016. Sharma worked with Cheung, Pearce and research partner Langellotti to develop the TEI4BPS script. Sharma had taken basic programming courses in the preceding semesters and this was an opportunity for the BPS team to integrate student researchers into the development of this research toolkit.

3.2 Collaborating Projects, Workshop Participants, Potential Collaborators

3.2.1 [The Center for the Tebtunis Papyri](#) [14]

Under the direction of Professor Todd Hickey, students and post-doctoral fellows associated with the Tebtunis Center collaborated with BPS. Michaela Langellotti (Postdoctoral Fellow, The Center for the Tebtunis Papyri, Bancroft Library, UC Berkeley. AY 2015–2016) worked with Pearce, Sharma, and Cheung to identify features in the Grapheion Archive texts for input into the TEI4BPS conversion protocol and to articulate rules that would apply in the disambiguation of name instances in the registers of transactions in the corpus. New features introduced into the TEI4BPS converter protocol included fields that allowed researchers to identify the chronological sequence of multiple events within a record and to add professional titles as person attributes as distinct from roles associated with the name-instance. Langellotti's corpus

was the first for which BPS produced a full, real-time visualization of the persons in the 125 transactions in a summary register of financial events in the Grapheion archive.

3.2.2 [The Perseids Project](#)

In our Social Science Matrix workshop, Bridget Almas, Frederik Baumgardt, and Marie-Claire Beaulieu, members of the Perseids team, focused on the pedagogic applications of BPS, identifying its on-the-fly graph visualization engine as a valuable contribution to the digital humanities pedagogy exercises that Beaulieu (Tufts University) assigns to her students: HTML-tagging the social and biological relationships between entities named in William Smith's *Dictionary of Greek and Roman Biography and Mythology*. They found that students ignore the directionality of gender-specific relationships, an error not apparent until the SNA graph viz is produced by the instructor, who then must engage in data cleaning. With the addition of an easy-to-use, real-time graph visualizer, students would be more likely to review the process and product of their work and to learn through critical review of their own or classmates' work. The BPS mission includes support for pedagogy, and the Perseids application demonstrates that there are potential pedagogic uses BPS did not anticipate in its original conception.

3.2.3 Contributors and Potential Collaborators

The following individuals are or have been in conversation with the BPS team about the applicability of BPS to their research. The scope and diversity of projects across archives from a variety of geographic/linguistic/temporal frameworks reinforces the perceived value of the probabilistic disambiguation model underlying BPS.

- Cuneiform:
 - David Danzig (PhD candidate, ISAW/NYU, NY [2017]): a corpus of >1000 texts documenting ethnic groups in late first millennium BCE Mesopotamia.
 - Prof. John Nielsen, Prof. Tachun Lin (Bradley University, IL [2016]): Neo-Babylonian (726-589 BCE) economic texts. Matrix workshop participants; see §3.3.1, below.
 - Paulina Pikulska (PhD candidate, Warsaw, PL [2018]): A Social Network of Scribes in Neo-Babylonian Sippar.
 - Yoko Watai (Post-doctoral Fellow, JP [2016]): A Network of Women in the Economic Records of the Neo-Babylonian Archives.
- Egyptology:
 - Esther deGroot (PhD candidate, Leuven, BE [2015]): Prosopography of Workers in the Village of Deir el-Medina, Egypt.
 - Gaye Wilson (Independent scholar, Sydney, AU [2014]): Officials in the Fifth-Dynasty Sun Temples of Egypt.
- Chronology:
 - Prof. Adam Rabinowitz (University of Texas, Austin; [perio.do](#) [2016]): Rabinowitz participated remotely in our Social Science Matrix workshop, contributing to discussions of the use of absolute dates and date ranges as a feature in disambiguating common namesakes.
- Islamic Intellectual History:
 - Prof. Jamill Ragep, Prof. Sally Ragep (McGill University, Montreal, CA [2016]); Islamic Sciences Manuscripts Initiative (<https://ismi.mpiwg-berlin.mpg.de/>). The researchers

expressed interest in the BPS workspace model which allows documentation of the rule sets implemented in the probabilistic disambiguation as a methodology for justification of scholarly conclusions and the sharing of workflows with collaborators or other independent researchers.

3.3 Collaborating organizations

The following UC Berkeley units/organizations collaborated with BPS, providing community for digital researchers in the social sciences and humanities, as well as support, both material and in kind.

3.3.1 [Social Science Matrix](#) (Matrix) – UC Berkeley [5-6, 12, 15]

BPS participated in the initial cohort of Matrix-sponsored Prospecting Seminars ([Exploring Reusable Tools for Prosopography and Historical Social Networks](#), 2014) and Research Seminars ([Developing Tools and Collaborations in Prosopographical and Historical Social Network Research Environments](#), 2015-2016). Pearce and Schmitz served as co-PIs in both seminars. This program was developed to identify and support research agendas that integrated social science research/ers and methods across multiple disciplines, including the humanities. Our Prospecting Seminar meetings (which included remote participation from a number of partners) helped refine use-case scenarios and user requirements for researchers in a variety of disciplines. The monthly meetings of the Research Seminar (2105-2016) focused on developing a community of BPS users. For the final, on-campus workshop, BPS had achieved end-to-end integration, and all researchers were able to input and visualize networks in their specific research corpus. The material support from Matrix—in particular, the dedicated hi-tech meeting space—nurtured a community of researchers who, over the course of an academic year, developed more nuanced understanding of the problems and requirements of their own, as well as their colleagues', prosopographical research.

3.3.2 [DH@Berkeley](#) UC Berkeley DH-Mellon Capacity Building Grant

A Digital Humanities at Berkeley grant, funded through the Mellon Foundation Grant for Capacity Building and Integration in the Digital Humanities, supported two student interns who were responsible for coding a CSV-to-TEI4BPS conversion script and for developing the articulation of the conceptual frameworks in which researchers adapt their existing data into the BPS formats.

- Gautami Sharma, undergraduate intern, UC Berkeley. AY 2015-2016. Under Pearce's direction, Sharma worked with research partner Langellotti to determine the disambiguation rules appropriate for Langellotti's corpus (the Grapheion archive) and to assist in data cleaning and organization for input into the TEI4BPS conversion protocol.
- Carolyn Cheung, PhD Candidate, Program in Ancient History and Mediterranean Archaeology, UC Berkeley. AY 2015-2016. Working with Pearce and Semenzin, Cheung wrote the basic code for the TEI4BPS conversion protocol.

The frameworks they developed are discussed in §4.2.1.1 below.

4. Changes/Problems

BPS made no significant changes to the planned functioning and digital structure of the toolkit. However, in the course of interactions with research partners, the need to clarify

the conceptual framework of the fundamental input unit and modify the method of data input became clear. The change is described below. The greatest obstacle was in the area of personnel, described in §4.3 below.

4.1 Conceptual Framework

The BPS team identified the components of the basic prosopographical unit and labeled the entity an **NRAD** (Name in a Role in an Activity in a Document). We achieved consensus with research partners that all prosopographical research depends on the disambiguation of namesakes, and that in all instances, similar steps are taken (if in different orders) across disciplines and corpora:

1. Researchers first encounter each name instance within a document as a unique occurrence. The researcher applies his/her expert knowledge of the probability that other name instances that share attributes of the original instance could be collapsed into instances of individuals. To translate this process (which our collaborators widely acknowledged as a nearly intuitive process of “how prosopographers work”) into a digital methodology requires that each personal name in a text be treated as a discrete datum, identified with a unique identifier.

2. Researchers articulate **General Rules** expressing the number of name elements that constitute each name instance (e.g., three elements in the expression “Anu-uballiṭ son of Nidintu-Anu son of Anu-bēlšunu”), as well as make reasonable assumptions about the length of an individual’s professional (or business) career and the length in years between generations. Researchers acknowledge both: (1) that they may modify these parameters as their understandings of a corpus develops and that (2) difference researchers may assign different values to those parameters, thus affecting the disambiguation of namesakes within a corpus.

3. Researchers consider the likelihood that two or more name instances within a text refer to the same individual (**Intra-document rules**). Factors affecting this likelihood include the degree to which each name instance represents a most complete expression of a name formula and an evaluation that two different roles within a single text could be performed by the same individual (as represented by two instances of the same personal name): for example, could a name instance “John son of Bob”, who is a seller in a text represent the same person as a name instance “John son of Bob”, who is a witness to that sale? The likelihood of two individuals holding multiple roles in a document can be expressed in a **role-matrix**.

4. Researchers also consider the likelihood that two or more name instances within a text refer to the same name instances in other texts (**Inter-document rules**). Although the same name components and structures are considered, the probabilities that two (or more) name instances across texts may differ significantly from the probabilities within a single text. Of less significance at the inter-document level is the role-matrix, as the possibility of a single individual assuming two different roles in two (or more) texts is great.

We discovered that researchers have so internalized these processes that they may have difficulty in articulating the steps they take and the features they use in disambiguating namesakes in their specific corpora; for most researchers articulation of the conceptual

framework as well as the rules they use in disambiguation is an elusive process. In all cases, as researchers came to understand the conceptual framework behind BPS, a logistical problem emerged as it became clear to the researchers and to the BPS team that a substantial amount of data-transformation would be necessary to support uptake of a researcher's data into the BPS disambiguation engine. Examples are discussed in Technical Challenges, §4.2.1 below.

4.2 Technical Challenges

4.2.1 Modification of data input and upload methods

We originally expected that BPS users would upload data as TEI tagged with the primary elements relevant to prosopographical research: name, role of the individual in the recorded activity, biological relationships to other names in a name formula, date, activity, and document. We further planned that BPS would develop or adapt existing Natural Language Processing tools to recognize and tag personal names and other potentially important text elements in TEI, in order to simplify the task of manually marking up digitized data. In conversations with potential users, we encountered resistance to producing digital text editions that, in most cases did not already exist, and that would require pre-processing by the corpus experts before prosopographical analysis could begin. Researchers indicated that their existing databases contained the necessary information, and that this was a more efficient use of time and labor than producing digital text editions from scratch and mastering TEI and mark-up tools required for BPS input.

4.2.1.1 Unstable XML IDs as factor in developing TEI4BPS conversion protocol

The demonstrator/development text corpus of BPS is an archive of legal texts from the Hellenistic period (the time subsequent to Alexander the Great's conquest of Mesopotamia). Pearce has prepared digital editions of the texts for the cuneiform consortial project *Open Richly Annotated Cuneiform Corpus* (Oracc; oracc.org). Texts in the Oracc corpus adhere to common standards for transliteration of the Akkadian language, and project directors lemmatize corpus texts for the goal of building an Oracc-wide lexicon. The Oracc team provided a conversion script to transform RTF into TEI, and BPS ingested this TEI tagged for the NRAD elements. This is a clear and strong example of the co-ordination of research needs and the reuse of the product developed for one project for another. Unfortunately, the XML IDs Oracc generated changed on every corpus rebuild; the researcher attempting to identify name instances as s/he is building the text corpus would find this instability untenable. This prompted creation of a tool or protocol to convert researcher's database content (output as CSV format) to the TEI format requisite for input into the BPS structure (TEI4BPS). The scope of the conversion protocol project was sufficiently defined that it was a good opportunity to engage student researchers. Cheung and Sharma were hired to assess the needs and develop the code for the conversion tool; Semenzin supervised and validated the code. The converter is linked at: <https://github.com/berkeleyprosopography/TEI4BPSConverter>. It was important to ensure that conversion involved minimal manipulation of the researcher's existing database to include the requisite fields and to present them in compliance with specific typographic conventions and vocabulary items for each field (At the present, the limitations reflect current user needs; as other terms, e.g. for role,

activity, etc., are identified by other users, the conversion script will be edited to support them as well. See discussion under Roles, below.). These fields are:

- **Text_ID:** This is a unique identifier for each text in a corpus. Publication sigla, digital catalogue numbers, or other markers commonly recognized by discipline experts may all serve as this identifier.
- **Date:** This field indicates the date on which the text was produced and/or recorded, or on which the recorded activity occurred. Users are provided links to tools that convert dates preserved in local calendric systems to standard Julian dates.
- **Activity:** This field designates the type of transaction the individual text records, including, but not limited to sale, lease, quitclaim, and donation. At this time, the conversion script recognizes only the limited number of activities commonly encountered in the cuneiform corpus; the script will be expanded to accommodate additional terms as future users identify the kinds of activity recorded in other corpora.
- **Activity_sequence:** At this time, the BPS parser assumes that each document contains only one activity; however, this is not the case in all corpora. The addition of this field allow corpus experts to identify multiple activities in documents such as a summary ledger.
- **Activity_attribute:** To simplify the identification of activities, input for the activity field is restricted to a vocabulary of manageable size. However, we recognize that refinements of the activity field may facilitate disambiguation of participants by corpus experts. For example, it is conceivable that individuals who participated in the sale of houses did not participate in the sale of prebends. The activity attribute field supports identification of the object of the transaction (e.g.: object:house) and can, in the future, be adapted to extend the attributes encountered in a variety of humanistic disciplines. The activity_attribute is something of a work-around for the rendering complexity issue (see §4.2.1.3). It allows users to preserve some of the fine-grained info, while simplifying the set of roles used in the graph.
- **Role:** Each name instance in a text is assumed to have a role. At the current time, the converter recognizes only a limited number of terms. The inventory of accepted roles can be expanded in the future as users identify additional roles assumed by each name instance in a text.
- **Role_attribute:** Researchers might feel the need to further refine the role played by a name instance in an activity. As the development corpus (HBTIN) had no requirements for this field, there is no list of terms that the conversion script recognizes. This will be developed in response to the needs of future BPS users.
- **Person_name:** Each personal name in a document is recorded as a person name. In the case of texts written in non-alphabetic scripts, the user enters a normalized version of the name. At this time, collapsing of orthographic variants of the same name is not supported. However, the user may wish to enter such information into the Person_attribute field so that it is available when this feature is developed.
- **Person_sequence:** In some corpora, the order in which person names are presented can be a useful factor in disambiguating between multiple instances of a namesake, as in the sequence of witnesses in the list appended to many legal documents.
- **Person_attribute:** In some corpora, individuals bear professional titles that may

contribute to the disambiguation of namesakes. These titles or attributes (life-roles) are not of significance in the individual's participation in the activity instance, but may point to social markers that could distinguish two individuals.

- **Relation:** Each name instance can be qualified by his/her relationship to another individual in the name formula, or perhaps to another individual in the text. These relationships, explicitly stated in the text, are marked in this field as: "familial term:index number of the related person,". e.g. brother:1114, where the Person_name has an index number 1113, and the text specifically labels him as the brother of the individual indexed at 1114.

4.2.1.2 Common data cleaning/transformation issues:

Researchers commonly use databases to collect and record prosopographical data; however most databases that were not originally developed for use with BPS create a number of challenges for BPS users.

- a. Data fields in many researchers' databases group information from multiple texts: Many researchers record the sigla of all texts in which instances of a specific personal name in a corpus appear. Even when the researcher is convinced of the validity and accuracy of the collapsing of the multiple name instances into a single individual bearing a particular name, such streamlining introduces complications for common social network analysis tasks, such as determining co-occurrences between actors across texts in a particular archival context.
- b. Inconsistent transcription of names and orthographies: Human researchers can better tolerate variation and inconsistency in transcription of foreign names than can the computers utilized in common prosopographical tasks. This problem is particularly compounded when names originating in languages which do/did not employ the Latin alphabet constitute the data set. The problem arises out of the lack of consistent conventions, even within specific disciplines, for the rendering of sounds and characters that lie beyond the phonological and orthographic repertoire of languages that are natively represented by the Latin script. It is not so much a Unicode issue but rather a lack of scholarly consensus about how to represent some linguistic and orthographic features in the Latin script. As BPS considers each name instance (and thus each transcription) to be a unique instance that may or may not be collapsed into an individual, consistent transcription practices are imperative; almost all researchers acknowledge that this is an area of improvement to address in their own research flows.
The problem of orthography transcription (normalization) is related to, but stands apart from the issue of multiple orthographies used in writing a single name (including those used to represent a specific individual). Both Oracc and BPS tolerate multiple orthographies, as variants for a specific name and as a disambiguation criterion, respectively. However, as most researchers work with normalizations of name orthographies, orthographic variation was not included in the CSV-to-TEI conversion script.
- c. The role matrix poses conceptual and logistical challenges, as the disambiguation engine accepts all terms a researcher might use to label roles individuals held in activities. Without limiting the number of terms used in the matrix, this matrix expands quickly, increasing the computational demands for rendering the SNA graph visualization. We encourage researchers to develop a consistent and limited inventory of terminology that

capture the fundamental roles demonstrated in their respective text corpora. As the BPS architecture is corpus agnostic, there is no constraint that all corpora employ a single vocabulary for roles. The converter tool supports both role and role attribute, mitigating the impact of simplifying the roles in this way.

- d. Identification and labeling of the activity in each text similarly requires a controlled vocabulary. The variants that a human researcher tolerates as references to a single activity frequently have to be edited for consistency and scope.

4.2.1.3 SNA Graph Visualization Rendering

A number of factors affect the rendering of the SNA graph visualization:

- a. Computational resources on user computers (and available in the browser environment): Because BPS is a web-service, and the SNA graph is rendered on the client browser, the computing power of the user's machine affects the time required to generate a graph following completion of the disambiguation process. Relatively modern laptops and workstations provide a responsive user experience. However, some users have much older laptops or workstations, lacking graphics accelerators and/or with very limited RAM; users on these platforms experienced slower rendering and a less responsive user experience.
- b. Complexity of the resulting graph when there are many distinct roles: For corpora with a broad range of finely distinguished roles, the mix of these roles in activities expands combinatorically, resulting in many (duplicate) arcs between a given pair of nodes; across the entire graph of activity, this can greatly increase the graph complexity, and the associated requirements for layout computation and rendering. We discussed simplifying the characterization of roles with corpus owners in the near term, to reduce the layout and rendering overhead (See, for example, §4.2.1.2.c). We also discussed (future) support for simple taxonomies of roles, allowing the graph to be simplified at large scale, while supporting finer-grained distinctions when considering a local focus.
- c. Graph viz libraries: The graph rendering libraries we originally chose to use did not scale well when rendering in a browser context. We updated the libraries, and implemented several optimizations to reduce rendering time and to improve the interactive user experience. We have designed several further optimizations to the total data flow from the database to the visualization/rendering engine so that very large graphs perform responsively without sacrificing the greater detail and functionality that is useful when exploring smaller graphs.
- d. Display of SNA metrics: At the present time, the graph viz supports a modest display of the most basic SNA metrics. These may not be needed by most researchers, and they are likely to be moved off the main rendering page in future iterations of the graph viz presentation.

4.3 Personnel

Personnel was the greatest challenge throughout the course of the project. BPS remains committed to engaging student researchers for the purposes of pedagogy and to provide opportunities for students to participate in developing digital humanities projects and the digital humanities presence on the Berkeley campus. BPS took its first steps in development as UC Berkeley was developing campus awareness and infrastructure for the digital humanities in research and pedagogy. We were hopeful that student hires would be easy to implement and prove to be mutually productive and rewarding for the

students and the BPS agenda. However, it remained difficult to identify students with sufficient technical ability and a commitment to developing a humanities research agenda over participation in projects with obvious connections to the more lucrative and professionally rewarding tech and STEM sectors. Recent Masters degree graduates of the School of Information were promising candidates, but insurance requirements for independent contractors presented a financial obstacle to several prospective hires. Undergraduate prospects who initially expressed interest in the project were lured away by internships in the for-profit sector.

5. Special Reporting Requirements (N/A)

6. Project Outcomes

6.1 Technical components

The bulk of the technical work was focused on four areas: (1) extension and refinement of the core BPS model and associated functionality; (2) updating the visualization support and integrating this with the core services; (3) developing the corpus importer/converter; and (4) developing “DevOps” scripts for deployment of the BPS software to a server. All of this work was done following contemporary development good practices, including the use of an open software repository ([github: https://github.com/berkeleyprosopography/](https://github.com/berkeleyprosopography/)), where the products are freely available and will be supported beyond the current grant phase. Of the six technical goals laid out in the project work plan (§4.5.1-2 of the grant application), we accomplished two. Although we did other work, we missed some of these goals in large part because we lost our core software development capacity, which we were unsuccessful in replacing.

6.1.1 Extension and refinement of the core BPS model and associated functionality

We completed extensive rework of the PersonCollapser logic to address problems with the distribution of weight to reasonable candidates, deferring normalization until all candidates have been considered. This entailed associated cleanup of various classes supporting the disambiguation, refactoring to make classes more maintainable, etc.

We addressed a number of bugs and issues related to the probabilistic model, and to stability of the system under load from many users, which were surfaced through broader use and testing. We also fixed various problems in the provisioning scripts that build the BPS services in a virtual machine (VM).

We began the work to persist the data models for Person and Clan objects that are created in the disambiguation process. This will allow BPS to scale and support a range of filtering queries across many workspaces, as users explore the social networks in their corpora. This work was interrupted when we lost our primary developer, but will be completed in ongoing development.

We added support to persist and serialize all user settings for the workspace; these control the functioning of the disambiguation rules for a given workspace. The added functionality includes database support as well as new web service calls to expose the user settings as a RESTful payload. We added APIs to support update of the rule weights from the browser client. We used these new web APIs to rewrite the

UI logic to dynamically synthesize the user settings page, based upon the rules configured for the workspace. We developed logic to synthesize the role-matrix UI from the roles discovered in the current corpus, and to rebuild this when the corpus is updated. This surfaced a need to either simplify the roles (working with the corpus owners), or to model a simple taxonomy of roles that will simplify the role-matrix UI for users, when there are many different roles in the corpus. In the current phase, we worked with corpus owners to reduce the granularity of their activity roles, and have discussed various approaches to providing role-grouping, or a simple taxonomy editor/importer to balance fine-grained distinctions and the need to make general rules about larger classes of roles.

We addressed a number of feature requests that emerged in testing with our community, especially as we had a variety of corpus owners working in the system. For example, we added a feature allowing users to clear the corpora out of the workspace environment (so they could start over with a workspace, or to import a new version of the corpus). We added related functionality to preclude editing of corpora that the current user does not own (even if she has corpus manager rights in the system). In addition to new features, we addressed various issues discovered in testing and use of the system.

Finally, we conducted a thorough code-review of the core services, and cleaned up some issues and documentation gaps that were revealed in this review.

Various issues and many feature requests remain for future development phases; these are maintained in our JIRA issue tracking database, which will continue to be supported beyond the current grant phase.

6.1.2 Updating visualization support and integrating this with core services

We integrated a new visualization library to address gaps in functionality, and performance in the original libraries. We updated out-of-date libraries, to ensure stability and maintainability of the software. We then completed the integration of the full data flow from corpus input into the core services, to disambiguation and construction of a graph model, to delivery via web services to the client, to rendering within the context of the web interface. This included refactoring of the existing layout model to maximize the screen real-estate available for graph rendering, while maintaining a pleasing experience with tabular and text-centric pages.

We designed and built specifications for an extended protocol between the client renderer and the web services, to optimization of the graph computation and ultimate rendering in the client, and to support level-of-detail and user-controlled filtering. We did not complete all of this functionality, and will address the remaining work in ongoing development.

6.1.3 Developing the TEI4BPS corpus importer/converter

The impetus and resulting functionality for this tool is discussed in §4.2 above. The tool itself was developed to run either as a command line script, or as a simple web application (parallel to the primary BPS services). We may consider how to better integrate the user experience of this tool with the rest of BPS, based upon further user evaluation. The code is available on our github repository: <https://github.com/berkeleyprosopography/TEI4BPSConverter>.

6.1.4 Developing DevOps scripts for BPS deployment

From the beginning, we developed the software with standard tools for building and packaging the core software components, including Maven and Ant to compile and package the software for the web services, and to package the HTML, CSS, PHP, etc. for the web application. This automated build process provided consistency of day-to-day development.

However, as we brought different developers onto the project, and as we deployed test environments for new versions of the BPS web application, it became clear that we needed to automate the process of creating and deploying the full software environment (or “stack”), from the operating system to the core libraries and platforms for the web server. The associated problems became more acute when one of the virtual machines upon which the BPS services were hosted experienced a hardware failure, and we had to rebuild that server from backups.

We had originally architected the software stack to be platform independent, however we encountered various and subtle inconsistencies across the various development and deployment environments (including Linux, Windows, and MacOS operating systems). We resolved to standardize on Linux as a deployment operating system, leveraging VirtualBox to support an embedded (virtualized) environment on Windows and MacOS development platforms. We then created base “box” patterns and registered them on a public repository of VirtualBox VM images. Working from this base operating system image, we created configuration and deployment scripts using the Ansible toolset, which captures all the steps for installing and configuring software such as the Apache web server, the MySQL database engine, the Tomcat web service platform, etc.

Many of the steps to set up the server had been described in prose installation notes, but these required a skilled developer or admin to follow the steps, and were subject to misinterpretation and error at many points. The work to develop the automated deployment scripts took some time and effort, but resolved a number of issues that had plagued our development team, and ensured that the BPS application can be easily, reliably, and consistently installed on any developer’s machine, as well as on our public hosting environment.

On a related note, we devoted some effort to “harden” the web services environment in order to provide greater stability in the open web environment. In particular, a security review of the server platform was carried out, to identify and address weaknesses and attack vectors. This work included cleaning up sensitive data from the existing deployment; updating network configuration (such as more restrictive firewall rules, IP blacklisting); hardening the web stack, by disabling unnecessary default modules, protecting from distributed denial of service attacks by setting timeouts, hiding admin URLs; and performing other industry-standard security checklist items.

6.2 Community engagement, on campus, in the digital humanities community and in humanities research[‡]

6.2.1 Campus engagement:

The BPS project and its core team members contributed to a burgeoning digital humanities initiative and environment on the UC Berkeley campus. Although BPS is

[‡] Numerals enclosed in square brackets, e.g. [5], refer to items in the Supplement.

corpus-agnostic (its architecture accepts correctly formatted data from any discipline), its co-PIs are scholars of cuneiform, a script used to write Sumerian and Akkadian, two of the world's oldest documented languages. The PI's participation in campus digital initiatives has led to campus-wide recognition of the Department of Near Eastern Studies as an innovative and leading partner in campus DH.

6.2.1.1. 2015 DH Faire: Poster Session and Roundtable. Pearce was invited to be a member of the panel discussing "The Landscape of Berkeley DH." Her comments related to the poster included in the poster session convened at the Social Science Matrix venue on campus. [7-10]

6.2.1.2. 2015 DH@Berkeley Summer Institute: This was the inaugural year of the DHBSI, a [partnership between the Dean of Arts and Humanities and Research IT](#). Both co-PIs participated in the institute: Veldhuis in the Computational Text Analysis course thread, and Pearce, teamed with NES PhD student, and part-time BPS team member Eduardo Escobar, in the Data Workflows and Network Analysis thread. Pearce and Escobar worked with a data set drawn from the BPS demonstrator corpus [HBTIN](#) in preparation for a jointly-authored article, "Bricoleurs in Babylonia: The Scribes of Enūma Anu Enlil." In *The Scaffolding of Our Thoughts: Essays on Assyriology and the History of Science in Honor of Francesca Rochberg*, 264-287. Eds. C.J. Crisostomo, E. Escobar, T. Tanaka, N. Veldhuis; Leiden: Brill, 2018. As they refined their analysis of this corpus using Gephi, they contributed feedback to the BPS team. Escobar expanded on this experience by participating in the 2016 [SMART \(Student Mentoring and Research Teams\)](#) program, mentoring an undergraduate history major (with no prior knowledge of cuneiform) in developing visual networks of Babylonian knowledge. [16]

6.2.1.3. 2017: School of Information Friday Seminar: Professor Michael Buckland (emeritus) of the School of Information invited Pearce and Schmitz to report on BPS development to this monthly, informal gathering of members of the School of Information faculty and graduate students. This was the second time they presented to the Friday Seminar (first in 2013, prior to the start of the current grant period). [20]

6.2.1.4. 2017 Data Science Connector Course presentation: Together, Escobar and Pearce presented "Social Networks and Ancient Texts" to Professor Dennis Feehan's Data Science Connector Course on Social Networks. Students in Data Science Connector Courses apply foundational principles and skills learned in Data 8: Foundations of Data Science to domain specific or cross-disciplinary content. The importance of this presentation lay in making students aware that the computation of social networks and the implementation of visualization techniques that represent those outcomes are increasingly employed in humanities research; that the content was based in texts from the world's earliest writing system drove home the utility of the methods in a wide variety of disciplines. [19]

6.2.1.5. 2018 Research IT Reading Group: At the invitation of Patrick Schmitz, associate director of Research IT and BPS technical lead, and of Steven Masover, Research IT Architect, Pearce presented to the Reading Group on the topic of [Computational Assyriology and Reproducible Research](#). One of the features of BPS is the workspace, in which researchers conduct their analysis and formulate assertions about the corpus and disambiguations. Built into the architecture of the workspace, although not yet implemented, is the feature of recording the data set and assumptions applied in the disambiguation process. This feature was included from the inception of

BPS as the team recognized the significance of documenting the research process, much as one does (or should do) in traditional print publications. As humanities disciplines embrace the tools and methods of social science, they are challenged to document ever more precisely the heuristics and workflows of their research. This presentation built on the structures in BPS that support such documentation and introduced the members of the Research IT Reading Group, which includes librarians and other collections managers, to a new project in which teams of Assyriologists and specialists in information and technology services at the University of California, Berkeley and the Ludwig-Maximilians-Universität München are working to reconstruct and document heuristics Assyriologists apply in their research and to demonstrate the reproducibility of their conclusions. [21]

6.2.2 Digital Humanities Community and DH research

6.2.2.1 2014 DH-CASE II: Collaborative Annotations in Shared Environments: Metadata, Tools, and Techniques in the Digital Humanities, co-located with ACHM DocEng 2014. Schmitz, Pearce, and Quinn Dombrowski (Research Applications Developer in the Research IT group at UC Berkeley) organized this refereed workshop, which focused on tools and environments that support annotation activities. Participants presented outcomes of their research and innovations, shared challenges and differing approaches, and identified best practices and novel approaches that may have potential for wider application or influence. Papers accepted for the workshop addressed the intersection of theory, design and implementation, emphasizing a “big-picture” view of architectural, modeling, and integration approaches in digital humanities. A major concern in all talks was data and tool reuse. These interactions with colleagues from UC Berkeley, UCLA, and the Leibniz Institute of European History (Mainz, Germany) afforded the BPS team an opportunity to engage with digital humanities researchers working on a variety of project and to develop lines of communication and co-operation. [1-3]

6.2.2.2 As reported in Fall 2017, our collaboration with partners at the Ludwig-Maximilians-Universität München resulted in implementation of an extension of the probabilistic reasoning model at the heart of BPS through an exploration of the evaluation of heuristics employed in early studies of the SAA corpus (1970s, by Simo Parpola) to assign cuneiform letters to dossiers grouped by their writers, whose identities were disambiguated on the basis of three features: (a) distinctive paleographic features, deemed “scribal hands”; (2) mentions of officials whose names were used to label years and thus could serve as chronological indicators in the letters’ absence of absolute dates; (3) mentions of astronomical phenomenon that could be linked to securely dateable events. We engage in the digital analysis of these features not to challenge the accuracy of our colleague’s work, which remains highly regarded and foundational to the discipline, but rather to establish a specific instance of examining the reproducibility of results achieved through analog heuristics applied by humanities researchers against those obtained through the use of recognized digital tools and computational techniques. This is a direct extension of the conceptual framework of probabilistic modeling that supports the BPS architecture.

Funds awarded to our colleagues at the Ludwig-Maximilians-Universität München by the [UC Berkeley-Ludwig Maximilians Universität](#) München Research in the Humanities

program supported the travel of BPS team members Veldhuis, Pearce, Schmitz, along with UC Berkeley digital humanities postdoctoral fellow Adam Anderson, graduate students Gil Breger and Jason Moser (Near Eastern Studies), and undergraduate Joanne Tan (Astronomy) to Munich in July 2018 for a four-day workshop. The goal of this workshop was to integrate the work of the Berkeley and Munich researchers. The Berkeley team had focused on evaluation of the heuristics employed in (a) identifying astronomical events on the basis of textual references and (b) the reliability of paleographic and orthographic variation as disambiguation criteria for scribal identities; the Munich team has been developing a methodology to digitally assess references to specific events and named, datable personages and harmonize them with generally accepted chronological frameworks for Assyrian historiography.

7. Technical Products

7.1 Technical products

The primary code repository for the BPS web application software is publicly available at <https://github.com/berkeleyprosopography/bps>. This includes the two main code packages for the web-service software, as well as the web-content and web user-experience support. The repository also includes some test corpora, and the support configuration files for the automated deployment tools.

A secondary code repository includes the software and basic documentation for a conversion utility that simplifies corpus input by converting comma separated values (CSV) files into the TEI (xml) files that the BPS services ingest. This repository is available at <https://github.com/berkeleyprosopography/TEI4BPSConverter>. A third related repository provides a web service wrapper for the conversion utility. This repository is available at <https://github.com/berkeleyprosopography/convert-service>.

7.2 Project sites

7.2.1 Project site and UI: <http://dev.berkeleyprosopography.org>

7.2.2 bps: berkeley prosopography services, a WordPress site for project blog and documentation <http://berkeleyprosop.digitalhumanities.berkeley.edu/>.

7.3 Publications during the grant period: (page numbers refer to the supplemental materials)

2014

1. Humanist-centric tools for ‘Big Data’: Berkeley Prosopography Services (with Patrick Schmitz). Pp. 179-188 in Proceedings of ACM DocEng 2014. doi>[10.1145/2644866.2644870](https://doi.org/10.1145/2644866.2644870). 22
2. Social Networks from History. Chuck Kapelke. [Social Science Matrix website](http://socialsciencematrix.org). 7/28/2014. 50

2015

3. [Historical Texts, Modern Tools: Berkeley Prosopography Services](http://libro.digital.ub.edu). Blog Post, Libro Digital: Universitat Oberta de Catalunya. 5/19/2015. 75
4. [Prosopography: Toward a Toolkit](http://socialsciencematrix.org). Social Science Matrix Article. 2015. 77

5. [Analyzing Social Networks and Semantic Networks in Assyriology](#). Eduardo Escobar. DH@Berkeley Blog Post. 11/24/2015. 80

2016

6. [Berkeley Prosopography Services and the Tebtunis Papyri](#). DH@Berkeley Blog Post. Patrick Schmitz. 2/1/2016. 83
7. [Semantic Network Analysis and Cuneiform Intellectual History](#). Research IT Blog. Quinn Dombrowski. 4/22/2016. 89
8. [Berkeley Prosopography Services Successfully Demonstrates New Toolkit](#). Research IT Blog. Quinn Dombrowski. 5/11/16; cross-posted to [DH@Berkeley](#) Blog 91; 94

8. Supplementary Materials (pdf)

Over the course of the grant, BPS team members presented the progress of the project in a variety of academic and DH fora. The supplementary materials document the scope of those presentations.

2014

1. Humanist-centric tools for ‘Big Data’: Berkeley Prosopography Services (with Patrick Schmitz). Pp. 179-188 in Proceedings of ACM DocEng 2014. doi>[10.1145/2644866.2644870](https://doi.org/10.1145/2644866.2644870). 22
2. DH-CASE II Workshop Description in Proceedings (2014) 31
3. BPS Paper to be presented at DocEng 2014. Steve Masover. [Research IT Blog Post](#). 7/2/2014. 33
4. Presentation to Bay Area DH Community meet-up: slide deck. Laurie Pearce. 7/8/2014. 35
5. Social Networks from History. Chuck Kapelke. [Social Science Matrix website](#). 7/28/2014. 50
6. Social Network Visualizations: use case with HBTIN data. Social Science Matrix Workshop slide deck. Laurie Pearce, Patrick Schmitz, Davide Semenzin. 11/14/2014. 54

2015

7. The Landscape of Berkeley DH: panel discussion. Laurie Pearce presented. Poster from DH Faire. 4/8/2015. 63
8. Abstract of Laurie Pearce presentation to DH Faire panel discussion. 4/8/2015. 64
9. Historical Texts, Modern Tools: Berkeley Prosopography Services. Slide deck for Laurie Pearce’s presentation to DH Faire panel discussion. 4/8/2015. 65
10. Historical Texts, Modern Tools: Berkeley Prosopography Services. Poster for DH Faire. 4/8/2015. 74
11. [Historical Texts, Modern Tools: Berkeley Prosopography Services](#). Blog Post, Libro Digital: Universitat Oberta de Catalunya. 5/19/2015. 75
12. [Prosopography: Toward a Toolkit](#). Social Science Matrix Article. 2015. 77
13. [Analyzing Social Networks and Semantic Networks in Assyriology](#). Eduardo Escobar. DH@Berkeley Blog Post. 11/24/2015. 80

2016

14. [Berkeley Prosopography Services and the Tebtunis Papyri](#). DH@Berkeley Blog Post. Patrick Schmitz. 2/1/2016. 83
15. [Social Science Matrix Sponsors Research Seminar on Berkeley Prosopography Services](#). Research IT Blog. Patrick Schmitz. 2/16/16 86
16. Visualizing Networks of Knowledge in ancient Babylonia: SNA, Semantics, and Pedagogy. Eduardo Escobar and Laurie Pearce. DH Faire Poster. 2016. 88
17. [Semantic Network Analysis and Cuneiform Intellectual History](#). Research IT Blog. Quinn Dombrowski. 4/22/2016. 89
18. [Berkeley Prosopography Services Successfully Demonstrates New Toolkit](#). Research IT Blog. Quinn Dombrowski. 5/11/16; cross-posted to [DH@Berkeley](#) Blog 91; 94
19. Social Networks and Ancient Texts. Presentation to UC Berkeley Data Science Connector Class. 11/21/2016. slide deck. Laurie Pearce and Eduardo Escobar 96

2017

20. Berkeley Prosopography Services (BPS): a toolkit supporting humanities research. UCB I-School Friday Afternoon Seminar. Slide deck. Laurie Pearce and Patrick Schmitz. 04/14/2017. 121

2018

21. Computational Assyriology and Reproducible Research. Research IT Reading Group. 10/4/2018. pre-circulated slide deck. Laurie Pearce 143

Humanist-centric tools for “Big Data”: Berkeley Prosopography Services

Patrick Schmitz
IST/Research IT
U.C. Berkeley
Berkeley, CA
pschmitz@berkeley.edu

Dr. Laurie Pearce
Dept. of Near Eastern Studies
U.C. Berkeley
Berkeley, CA
lpearce@berkeley.edu

Abstract: In this paper, we describe Berkeley Prosopography Services (BPS), a new set of tools for prosopography - the identification of individuals and study of their interactions - in support of humanities research. Prosopography is an example of “Big Data” in the humanities, characterized not by the size of the datasets, but by the way that computational and data-driven methods can transform scholarly workflows. BPS is based upon re-usable infrastructure, supporting generalized web services for corpus management, social network analysis, and visualization. The BPS disambiguation model is a formal implementation of the traditional heuristics used by humanists, and supports plug-in rules for adaptation to a wide range of domain corpora. A workspace model supports exploratory research and collaboration. We contrast the BPS model of configurable heuristic rules to other approaches for automated text analysis, and explain how our model facilitates interpretation by humanist researchers. We describe the significance of the BPS assertion model in which researchers assert conclusions or possibilities, allowing them to override automated inference, to explore ideas in *what-if* scenarios, and to formally publish and subscribe-to asserted annotations among colleagues, and/or with students. We present an initial evaluation of researchers’ experience using the tools to study corpora of cuneiform tablets, and describe plans to expand the application of the tools to a broader range of corpora.

Keywords— *Annotation, Assertions, Big Data, Cyberinfrastructure, Digital Humanities, Prosopography, Social Network Analysis, Web-services.*

I. INTRODUCTION

A long tradition of data-driven research in the physical sciences has made for a relatively straight-forward transition to the techniques of data science and so-called “Big Data.” However, in much of the humanities, traditional scholarship has focused on close examination of texts and the interpretation of associated corpora. While linguistic and other analytic tools have shifted some research more toward data science, the

transition to the requisite/accompanying tools has not been a natural one for many humanist scholars.

In both the sciences and the humanities, the promise of “Big Data” is that new insights may be garnered through the analysis of datasets that are significantly larger than those that could reasonably be considered before, that new questions may be asked and answered, and that new connections may be found across corpora and domains. In the sciences, the transformative aspect is often understood in terms of the huge increase in data gathering, storage, and processing techniques, and so “Big Data” is often associated with measures of scale like petabytes or even exabytes of data. However, applying these same measures and notions of scale to the humanities misses the point that the transformative aspect is simply in a shift from a scholarly workflow focused on close reading and interpretation of texts, to one in which analytic tools enable the consideration of larger corpora and the associated relationships among entities, linguistic patterns, etc. Moreover, while numerical and analytic tools comport with the tradition of research in the sciences, many such tools do not fit well with the conceptual models and scholarly workflows of research in the humanities.

Against this background, we describe Berkeley Prosopography Services (BPS): an interactive tool-kit for analyzing and visualizing prosopographical datasets, available to researchers working in diverse disciplines and operating on a range of data that derives from a variety of text sources and formats. BPS innovates by providing software tools to perform association and computation tasks for name disambiguation, long done by hand, by adding a new model for curation and collaboration, and by connecting Social Network Analysis (SNA) tools and visualizations that reveal patterns and key features in a social network. BPS is built as an automation of the same conceptual models that humanist researchers have used in their analyses, and avoids (or de-emphasizes) abstract mathematical models that are in wider use in machine-learning or purely statistical tools. The BPS tools are developed using current enterprise software best practices, which provide a reusable, scalable and more sustainable software base than is commonly implemented by digital humanities research tools. BPS provides novel productivity tools, visualization tools, and workspace support for exploration and collaboration.

A. Finding “Big Data” in <500K Clay Tablets

The total number of cuneiform texts cataloged in museums, libraries, and private collections throughout the world is well under 500,000 (cdli.ucla.edu). The number is stunningly large and small. These clay and stone documents, composed in one of the world’s oldest writing systems, provide the first documentation of three and one-half millennia of human activity—economic, religious, intellectual, and scientific. It is remarkable both that so many have survived and, at the same time, frustrating that the uneven distribution of their number over such a long time span means that the study of various times and places often depend on fragmentary data (Figure 1). Nonetheless, scholars of the ancient Near East, lands and empires that flourished in modern Iraq, parts of Syria, Turkey, and Iran, ask of the available resources questions recognizable from many other humanities disciplines, and are increasingly turning to methodologies and tools from the digital and social science worlds to frame and answer their research agendas.

An exploration of the remarkable embrace of digital tools by practitioners of this recondite specialty deserves more attention than might be expected, particularly in light of the size of text corpora with which researchers are concerned—from several hundred to ±10,000—and the corresponding (small) digital footprint of such data as presented in a standard off-the-shelf database. Nonetheless, the data and the questions researchers are asking of it suggest that even “Small Data” faces challenges in modeling authority and data review, areas of concern in “Big Data”.

BPS is unique as a tool-kit for prosopography: it emulates the workflow—including, and notably, the *uncertainty*—of real scholars grappling with data analysis. While the small cuneiform corpus of some 500 texts from the city of Uruk in the 4th-3rd centuries BCE would seem to consign it irrevocably to the realm of “Small Data”, that same size served BPS well as a development corpus as it attempted to capture and replicate in digital tools the human researcher’s analytic process(es) of disambiguation of multiple instances of name-

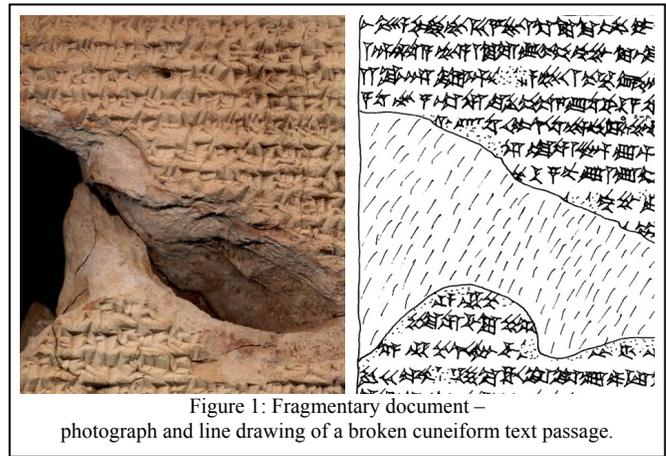


Figure 1: Fragmentary document – photograph and line drawing of a broken cuneiform text passage.

sakes into discrete name instances, even as it allowed for a process of modeling authority, scholarly debate and the *communis opinio*.

The central problem in prosopography, and thus, in the social network analysis that draws upon it, is the disambiguation of the many individuals who share the same names. Culturally specific naming practices may simplify or complicate the process: kings of England named Henry are numbered VI, VII, VIII, just as men in non-royal families might be called John Sr., John Jr., John III, etc. Disambiguation is made more difficult when, for example, individuals in alternating generations are named for their grandfathers (a practice called papponymy), with the result that a document may record the participation of Anu-uballit, son of Nidintu-Anu, son of Anu-uballit, son of Nidintu-Anu.

The specialist in any corpus knows that clues that inhere in the data and the context in which they appear facilitate disambiguation. Were the specialist asked how he distinguishes between close namesakes, he might claim “intuition”, his expertise so ingrained that it obscures the sequential process considering attributes and the likelihood that any one or a

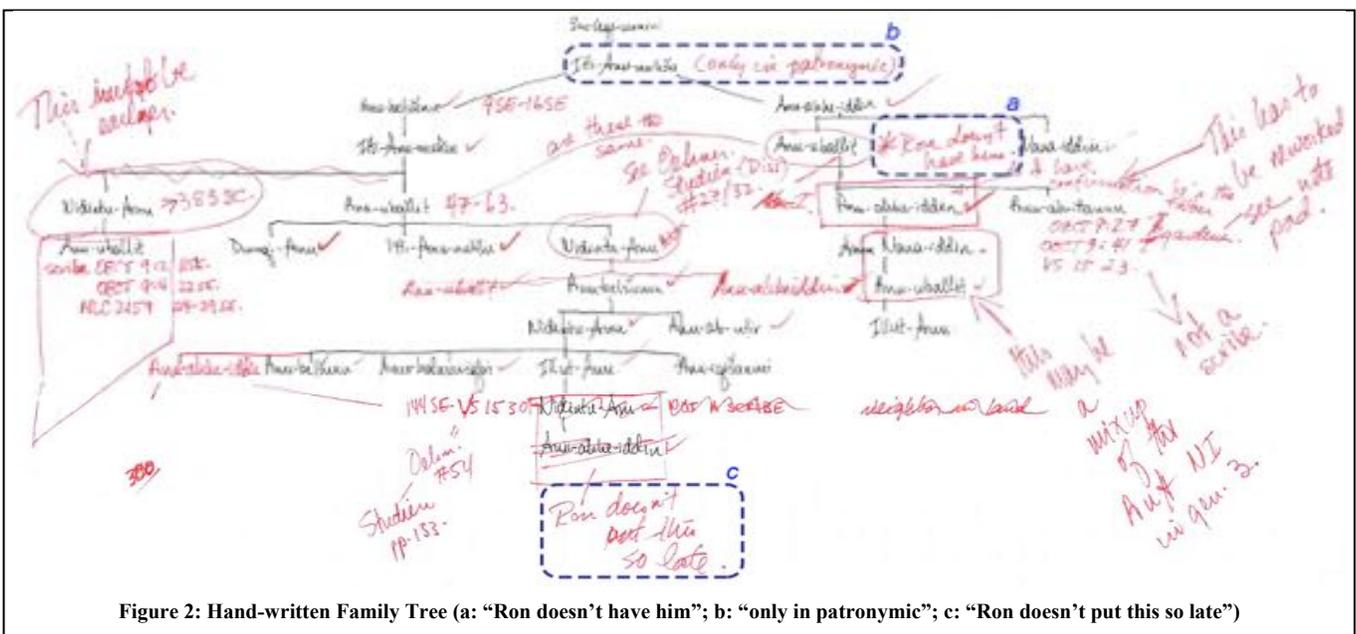


Figure 2: Hand-written Family Tree (a: “Ron doesn’t have him”; b: “only in patronymic”; c: “Ron doesn’t put this so late”)

collection of those features applies to one, but not the other, of the namesakes. For example, dates in documents that record two transactions occurring seventy-five years apart would, in some social contexts, make it highly unlikely that Anu-uballit(1) is the same individual as Anu-uballit(2). Other attributes do not provide such clear criteria, and different scholars might variously assess the utility of those measures (and/or meta-data) in disambiguation, effectively assigning them different weights in their evaluation, for example: how likely is a buyer of real-estate also to participate in slave sales? Differences in outcomes are chalked up to scholarly debate, which may range on the continuum from friendly to flame. Scholars committed to the exploration of ideas may annotate the differences and note how their procedures and assumptions led to variant outcomes. Examples include: (a) the inclusion or exclusion of data from different researchers' data sets, (b) identification of few attestations, appearing in limited contexts, of data across the corpus, (c) discrepancies reflecting lacking or corrupted metadata. These are highlighted in Figure 2.

When these differences are small, few in number, and occur in a small number of texts, dissenters or the merely curious can easily retrace the analytic process that brought assertions and facts together to a result. But when a researcher studies a corpus of 10,000 economic records from temples, and tracks the price of the sale of a liter of barley on specified festival days in several months over a thirty year period in three cult centers, in the hopes of determining the economic prowess of similarly-named members of families of cult officiants, he faces a much greater challenge. The research quickly becomes a complex interplay of layers of data (prices, quantities, days, years and location) associated with relevant individuals who have been disentangled, with varying degrees of likelihood, from multiple namesakes. The researcher's colleagues may simply choose to accept all of his assertions on the basis of the researcher's reputation and standing in the field. Alternatively, they might dismiss the same results produced at the hand of a relative unknown, perhaps a recent PhD (even if a student of a legendary professor) or by an interested dilettante.

Common notions of scientific exploration include "reproducible results", which carries with it the corollary that the data and research method can be tracked and implemented repeatedly. In order to approach that standard in research fields such as prosopography, where degrees of uncertainty factor into the methodology, tracking of the changes and the researchers who introduced them, are a necessary component of responsible practice. A feature of the BPS architecture designed to support probabilistic assertions is the creation of workspaces for individual experimentation, recording of parameters applied in any particular research question, and authority tracking. In this way, BPS engages with one of the questions "Big Data" faces, even as the number of bytes any BPS researcher may process is small.

B. Workflows in prosopography

Prosopography is at the core of many humanities research agendas. It enables researchers to identify in text corpora the unique individuals who populated a documented social milieu.

Disambiguation, the initial task in prosopographical research, is fundamentally a probabilistic endeavor. Initially, the prosopographer considers all namesakes in each text and throughout the corpus as references to discrete individuals. As the impact or validity of factors (metadata) — such as lifespan, co-occurrences with other individuals, mutually exclusive roles within a transaction-context — are taken into account, multiple name instances may be collapsed or maintained as different individuals. The process may be complicated by the state of preservation, where broken sources (Figure 1) may be reconstructed on the basis of specialized knowledge; different researchers may advocate for variant reconstructions. The implications of variant disambiguations carry great importance for the generation of the social networks which articulate the links between individuals, and for the graph visualizations which facilitate their exploration by multiple investigators. It is common, also, for different researchers to assign different weights to parameters resulting in potentially and dramatically differing disambiguations. The uncertainty inherent in all prosopography fosters discussion and, in the most collegial of research environments, promotes advancements in the field of study. Prosopography transforms probabilistic assertions about data into narrative and provides foundations for additional discovery.

C. Traditional curation and annotation

A common theme in "Big Data" workflows is the amount of time and effort spent cleaning and normalizing datasets in preparation for analysis. In many areas of humanities, these activities directly correspond to *curation*. Martin Mueller [10] describes processes to Complete, Correct, and Connect. Texts must be completed to address content that is missing or illegible due to damage or problems with digital capture. Corrections address OCR errors, incorrect markup produced in transcription, etc. Connection makes texts machine actionable by adding markup and relations on entities, features, structure, etc. to support subsequent analysis. Especially in the humanities, however, these curation activities are often the subject of discourse and debate. While many annotation models include support for provenance (authorship) of annotations, curation models rarely incorporate this aspect, and annotation models generally do not actually change the underlying text (so that the changes are reflected in further processing workflows).

In addition, despite the emergence of digital curation models (e.g., [9]) that will support "access, use and reuse of digital materials throughout their lifecycle," the workflows to perform these curation activities are still largely human-based; people create the metadata, and commonly assume that it will be interpreted by other people. Aside from the problems of scaling this to larger corpora, the unstructured or semi-structured nature of the resulting documents, curatorial metadata and even many annotations are often ill-suited for the algorithmic processing that underlies many digital humanities tools.

A final challenge is one not uncommon across digital humanities, especially as tools from other "Big Data" applications are adapted to humanities corpora and domains: where analytic tools are applied to these new corpora, many of

the underlying algorithms are based upon mathematics that is unfamiliar and opaque to most humanists. For certain tools (e.g., algorithms that support SNA visualization) this not an issue, as the researcher will interpret the results and draw her conclusions referencing the corpus directly. However, in many other cases (e.g., supporting name disambiguation), humanists have expressed reluctance to “trust” an algorithm that they cannot (at least conceptually) understand. Inasmuch as the tools *support* rather than *replace* scholarly workflows, researchers’ willingness to integrate the tools into their reasoning process is essential to adoption and ultimate utility.

D. BPS approach to infrastructure

The BPS project originated in the context of a larger exploration of cyberinfrastructure in support of humanities research¹. As such, the requirements included both functional needs of the domain researchers, as well as software architecture and design requirements that the solution be generalized, re-usable across domains, scalable, and sustainable. These technical requirements led to a number of basic design decisions for the infrastructure, and approaches to the implementation of specific functional needs. The core application logic is implemented as a set of RESTful [7] web services, following the principles of Service- (and Resource-) Oriented Architecture (SOA/ROA). Communication between services is loosely coupled, and generally uses standard or abstract XML (or json) payload formats. This facilitates the re-use of individual components or services in other projects. It also makes it easier to replace individual pieces of the underlying infrastructure, since the details of the implementation are abstracted behind a service API. For example, the initial implementation took a lightweight approach to Identity and Access Management (IAM – a.k.a. authentication and authorization), and to corpus content management. To scale up these components, the IAM implementation could be tied into campus IAM infrastructure, and the content management could be tied into common repository services, without requiring changes to the rest of the application.

BPS streamlines prosopography and SNA by offering an integrated and customizable out-of-the-box digital analysis tool-kit and work environment. The tool-kit includes: (1) corpus services to parse TEI [14] and build an internal model of name-citations, etc. in documents, (2) a probabilistic disambiguator that determines the likelihood that two or more name instances refer to the same person, (3) support for assertions that parameterize various tools, and that let a user confirm or override disambiguation results, (4) SNA services that compute features and aspects of the resulting social network, (5) a visualization module that renders interactive visual representations of the networks based on data drawn from the text sources, and (6) workspace support that allows users to manage various corpora and experiment with *what-if* scenarios for each corpus.

BPS leverages an assertion model that reflects the tradition of discourse among researchers. The BPS productivity tools support the essential disambiguation steps, and also allow the

researcher’s individual conclusions or conjectures to be modeled explicitly as assertions. These assertions can be published to collaborators, reviewed and accepted (or rejected) by these research peers, maintaining the original provenance.

In order to support a wide range of domains, the disambiguation engine is based upon a plug-in model for abstract *rules* that influence the disambiguation process. Some rules can be used across many corpora (e.g., if they are based upon document dates), where others may be more domain specific (e.g., in a legal transaction, a named witness cannot be the same person as a named principal, even if the names are the same; a rule can express this constraint). The rules are applied in an automated framework that closely follows the researchers’ mental model of the process they traditionally (if laboriously) applied by hand. This ensures that algorithmic results can reasonably be understood and vetted by the humanist researchers. Thus, rather than imposing a foreign paradigm from an engineering or mathematical domain, BPS incorporates the humanists’ scholarly workflow into a computational model that can scale up to much larger corpora than were traditionally manageable by hand.

In the following sections we describe work related to aspects of BPS, present details of the BPS models and implementation, and relate our work to the larger context of “Big Data” and the humanities.

II. RELATED WORK

A. Related projects in prosopography

BPS is unique among existing digital prosopography projects in its corpus-agnostic architecture that ensures reusability of technical components to solve comparable problems across corpora, in its modeling of probabilistic assertions for disambiguation, and in its workspace environments that support and encourage individual and collaborative exploration, authority tracking and reputation building.

Most existing digital prosopography projects superimpose elegant interfaces over relational databases (e.g., PASE², CCEd³, and PBW⁴). These models all present a single editorial view produced by humans, and have no tools for disambiguation, no support for assertions or exploratory research, and limited or no support for SNA and visualization.

In addition to support for online queries in both English and Chinese, the Chinese Biographical Database⁵ (CBDB) [8] is one of a very limited number of projects that make their databases accessible via download, in an effort to share digital resources. While downloads support offline visualization of graphs, they do not provide an integrated online environment for research and inquiry. The project has created a tool to help extract information from source documents (the “RegEx

² <http://pase.ac.uk/>

³ <http://theclergydatabase.org.uk/>

⁴ <http://pbw.kcl.ac.uk/>

⁵ <http://isites.harvard.edu/icb/icb.do?keyword=k16229>

¹ <http://www.projectbamboo.org/>

Machine”), but lacks a more general disambiguation engine, and has no support for assertions or exploratory workspaces.

Mapping the Republic of Letters⁶ provides, under a single banner, analytic and visualization tools to a group of projects grounded in corpora that differ in content. In essence, however, each project remains a standalone presentation of the research results of a single research group, and there is no possibility of actively engaging with the data. There is no support for disambiguation, assertions, workspaces, etc.

Mining Social Structures from Genealogical Data⁷ (MISS) includes support for disambiguation, leveraging probabilistic machine learning (ML). However, ML approaches are better suited to very large data sources with extensive common patterns upon which a model can be trained; common corpora in humanities domains often number in the hundreds of documents, and so are not well suited to ML approaches. In addition, when the model makes a recommended disambiguation choice, researchers have expressed a desire to see an explanation for the suggestion. An ML model trained to maximize a function over a vector expression of features will be hard for most humanists to understand, whereas the BPS heuristics represent the same mental model they have used in the past – the researchers describe the rules that they used by hand, and BPS implements these formally as algorithms. There is often a temptation to apply tools commonly used in “Big Data” applications across other domains, however the needs of researchers in established scholarly workflows contra-indicate ML tools for the central problem of name disambiguation in prosopography.

In [6], a method for deriving social networks from English novels is described, using Natural Language Processing (NLP) to recognize named persons, and when persons are depicted in conversation. This approach has merit for those corpora for which good language models exist and for corpora in which conversations are depicted. For many of the ancient corpora in the BPS applications, no language model exists (nor can one easily be generated given the nature of the corpus), and so many NLP tools are not appropriate. In addition, in many corpora the documents are not narratives with conversation, so the approach (while interesting) cannot be applied.

Explorations in novel visualization for prosopography are described in [12]; however these are not yet integrated into a larger framework for research. Initial experiments with SNA visualization of related corpora [16] yielded a dense, difficult-to-interpret “thicket”; filtering and other query tools are needed for effective prosopographical visualization.

Booth [2] points to the impact of varieties of unstable data, a term seemingly applied to attributes associated with an individual in the BPS model, on the creation of digital prosopographies. The call to use XML markup (BESS) to identify “basic facts” in the construction of biographies and prosopographies resonates with identification and integration

of attributes (role, status, active life-span) into the BPS assertion model.

III. THE BPS MODEL AND TOOLS

A brief user story clarifies the role of the main functional components in BPS for prosopography research. The disambiguator performs an activity central to all prosopographical research, regardless of discipline or corpus content: it determines which of multiple name-instances can refer to a single person (discussed in I.A., above). Typically, researchers use criteria of date, provenance, profession or title. Since the data that are relevant and available differ from one text corpus to the next, some disambiguation rules are highly specific, not only to a discipline, but to a particular corpus. Although certain cases may be very clear, in many cases there may remain at least some doubt or margin of error. In order to feed such data into a family tree or into an SNA computation, a decision (even a provisional one) is needed, and may change as the result of the introduction of new data or different valuation of the criteria. BPS provides the researcher with the possibility of generating and exploring *what-if* scenarios, and with easy means to redraw graph visualizations to explore the consequences of these changes.

Although SNA is widely used in the social sciences, its penetration in the humanities and in historical studies is relatively low, in spite of early success in diverse implementations: studies of the history of early Islamic intellectuals [1] and the development of the anti-Persian movement among the Neo-Babylonian elite [15] illustrate the utility of SNA for understanding social dynamics. As BPS’s integrated disambiguator and visualization service replicate processes the humanist already employs, the tool-kit supports the iterative computational aspect of prosopography and SNA.

A. BPS Architecture

The architecture is divided into three major areas that correspond to the processing steps⁸:

1. Text Preprocessing
2. Disambiguation and Social Network Analysis
3. Presentation, Visualization, and Reporting.

In Text Preprocessing, a corpus is converted from some native format to TEI (possibly transliterating from, e.g., cuneiform, to a Unicode representation of Akkadian). The TEI markup includes elements denoting the individual documents, activities within each document, and persons that have roles in those activities. This markup may be generated by hand or by some semi-automated processes to recognize names, filiation, roles and activities (in any case, most of this happens external to the BPS system). In many “Big Data” and text analysis applications, NLP tools are an essential part of the processing workflows. However, there are fundamental challenges with applying traditional NLP tools to many corpora like cuneiform texts. For most ancient languages, no language models exist

⁶ <http://republicofletters.stanford.edu>

⁷ <http://swarmlab.unimaas.nl>

⁸ Diagram available at: <https://wikihub.berkeley.edu/x/BoPCB>

that enable the immediate application of typical NLP tools. Although the total number of texts under consideration may be large, the individual texts are spread across large geographic regions and considerable time spans (several millennia), making it very difficult to train a reliable language model. Nevertheless, we are exploring the application of simple NLP tools to automatically enrich metadata (markup) in the TEI (e.g., entailment phrases that indicate the role a named person has in an activity). Planned work in the next phase of the project will include adding services to support a broader range of corpora formats as input (e.g., direct from an existing database).

In Disambiguation and Social Network Analysis, TEI is ingested and parsed by corpus services, and a native data model is built internally. The workspace services share this model, and leverage authentication and authorization components to support login and access controls on corpus and workspace resources. The disambiguation engine incorporates rules that may be generic or may be corpus-specific, and associates the name citations in each document with actual persons depicted in the texts. It includes support for assertions that researchers make to confirm or reject the possibilities suggested by the engine. Finally, GraphML (a standard XML format) is passed to the SNA services to compute significant features of the social networks.

The Presentation, Visualization, and Reporting area presents results from various core model and analysis components, including the declared data model in each corpus (names, activities, etc.), assertions that the researcher has made or imported from others, family tree visualizations, as well as interactive network graphs for exploration and understanding.

The following sections describe the major areas of functionality in BPS. The assertions model underlies several areas, but is described in the primary context of making assertions about disambiguation.

B. Corpus input and management

A set of web services provides basic corpus services, including TEI parsing of a corpus file and conversion to the internal model of documents, activities, and name citations with an associated role.

RESTful APIs provide access to the basic document metadata, activities, roles, and names present in the corpus. We chose to implement the core functionality as REST services both to facilitate re-use of individual components, but as well to enable other applications to more easily integrate with BPS data and services. This has become a common architectural pattern in many applications, even if it is not all that common in digital humanities tools. The RESTful APIs support query and filter options, e.g., to list names that occur in a given role, and/or with specified features (e.g., a given value for gender). The associated web-application functionality provides a UI for these corpus services, as well as for uploading new corpus files, etc. All property values used in the UI, e.g., the list of activities and the list of roles, are derived dynamically from the

corpus; the BPS UI need not be rewritten or customized to support different corpora or domains.

C. Disambiguation support

A central task in prosopography is to associate each name instance to some real-world person. All name instances in a corpus, both within a single document (intra-document) and in documents across the corpus (inter-document), provide evidence for disambiguation. The algorithmic model is based upon the heuristics that researchers have long used, and so is familiar to users. To begin, a unique person is posited for each name instance in each document. Then, the model attempts to *collapse* persons into one another, so that the persons posited for name instances that refer to a given real-world person are collapsed into a single person in the model as well. It does this according to user-configured rules that operate on various features (properties) of each original person.

Filiation (declaration of parents and ancestors) is a primary feature used by the model. Additional features include the *activity* in which each associated name instance is cited, the *roles* that the citation had in the activity, the date of the respective activities, etc. In response to interviews with researchers, the disambiguation engine is being expanded to support a more generic model allowing an extensible set of features (e.g., life-role or offices held, titles, etc.).

The rules operate on these features and can have one of three functions:

1. **Shift** rules shift weight from one person to another.
2. **Boost** rules magnify the effect of applied shift rules.
3. **Discount** rules reduce the effect of applied shift rules.

A rule that produces a conclusive match between two person/name instances may shift 100% of the weight from one to the other. A rule that is only *likely* but not certain, may shift less weight. Name-matching rules are generally modeled as *shift* rules. Rules that provide additional evidence for a match are modeled as a *boost*, and tend to leverage features like location or activity. Rules that provide evidence of counter-indication are modeled as a *discount*; examples include date rules that consider the typical life-span and span of activity, along with the dates of respective activities (if two activities are 30 years apart, there is less likelihood that two person/name instances refer to the same real-world person, and so even if a name matches, a discount reduces the effect of the collapse). The date discount rule is implemented as a modified Gaussian that accommodates a user parameter for the expected duration of an individual active life span.

Rules may apply only to person/names within a document (intra-document rules), or to persons across the corpus (inter-document rules). Many rules can operate in either model, but function slightly differently in the two contexts.

The end result of applying the rules is a set of probabilities (i.e., a probability distribution) for each name-instance, for the

set of real-word persons to which that name instance may correspond (low weight probabilities can be filtered out to remove unlikely matches, and simplify the results).

The model closely follows the approach researchers have traditionally applied “by hand”: within each document, the person posited for each name is considered against each other person (with at least a minimally matching name). Researchers consider the roles, filiations, and other features associated to each name-citation, and come to at least a tentative conclusion about whether the two could be the same person. Once all the intra-document collapsing has been considered, researchers look across the corpus to correlate persons among different documents. Again, features such as filiation and roles are considered, but additional rules come into play like the dates of the documents (two persons mentioned 5 years apart are more likely to be the same than persons mentioned 40 years apart), the place in which respective activities took place, etc. BPS formalizes these research workflows as the described classes of rules operating on a probabilistic mathematical representation. The collapse algorithm also has an initial pass over each document to apply intra-document rules, and then a subsequent pass at the corpus level to apply inter-document rules. As each rule is applied, the calculated weight-shifts are normalized to produce a consistent probabilistic distribution; an optional threshold can be specified to filter very low probability matches (after which the distribution is re-normalized).

Each rule is implemented as a plug-in to a standard API, and each can be configured separately as it applies to a given corpus. The API includes methods for the computation of weight shift (or boost, or discount, respectively, for each class of rule). These methods just take two person instances and associated metadata for the respective documents, and return a floating point value for the shift/boost/discount. A second set of methods in the API supports the generation of a configuration UI with which researchers can parameterize the rules. Each rule returns a simple string that describes the impact of the rule (“what it does”, in plain language). Each rule also returns any named parameters that can be set (commonly, a confidence weight, but other values can also be described). The generated UI allows a researcher to specify her confidence in a given rule (i.e., the scaling weight in the interval 0 to 1 that is applied to the computed result of the rule; the UI often presents a simplified set of values like “Never: 0%”, “Conservative: 30%”, “Aggressive: 70%”, and “Always: 100%”). Each researcher can configure her confidence in each rule that is configured for her corpus, and thereby individually control how the heuristic proceeds. Changing these values in a different workspace (described below, III.F.) allows researchers to explore *what-if* scenarios.

BPS provides base classes that support common functionality (e.g., serialization, a confidence parameter) for each class of rule. BPS also provides some generic rules (e.g., a discount based upon the dates of two name-citations) that can be re-used without modification in many domains. A more complex but still-generic rule leverages a half-matrix of all the

role-names specified in the corpus (discovered when the corpus is parsed). The UI for this is a table of parameters allowing the user to specify which pairs of roles are compatible (or not).

Additional effort is of course required to support the plug-in infrastructure, and the abstraction that supports automated generation of a configuration UI. However, the infrastructure makes BPS much more easily adapted to other domains (where different rules must be defined). The broader applicability and re-use makes BPS a more sustainable project, especially given the relatively small communities in each individual domain.

1) Assertion support

Additional infrastructure supports an assertion model by which researchers can confirm or discard results of the automated disambiguation engine. Assertions are also used internally to model the users’ chosen values for rule parameters, etc. More details of the BPS assertion model are described in [13], but it is worth noting that BPS assertions connect “Big Data” algorithms that are necessarily more generic and abstract to “Small Data” workflows in which researchers draw their own specific conclusions in a given instance. In addition, by providing provenance tracking in a model that can be shared among researchers, assertions support reproducible research in a manner that is relatively uncommon in the humanities.

An important infrastructure challenge arises for corpora in which the content of individual documents is changing (as curators correct errors, etc.) – maintaining the references for assertions requires a robust linking model. We are refining the serialization and internal support for BPS assertions, and plan to extend the Hypothes.is fuzzy reference model described in [5] to provide a more robust model.

D. SNA support

BPS includes a set of RESTful web services that provide common algorithms from graph theory and social network analysis, including clustering, statistical analysis, and calculation of network distances, flows, and ranking measures (centrality, PageRank, HITS, etc.). The services are implemented as a thin abstraction layer over the Java Universal Network/Graph (JUNG) Framework [11], which implements the actual methods. The BPS services abstract the graph model, and handle translation to and from GraphML [4] for service payloads. We define a graph-context that maps user-level queries and filters on the underlying corpus, persons, features, etc., to database queries to select the nodes of interest, and then produces GraphML for the SNA algorithm services.

The abstractions (e.g., GraphML as a payload format) were chosen to make the BPS SNA services reusable as core cyber-infrastructure across a range of applications. The services provide common functionality as is, and can also be more tightly integrated with other applications by providing a graph context implementation that is specific to the application data model.

E. Visualization support

The BPS web application includes a range of visualization supports – lists, tables and reports with query filters and graph visualization support that lets users view family trees, social networks, clusters – that let users explore the corpus and the prosopographical analysis. The visualization support is particularly important to most humanist researchers, who prefer narrative to computation or numerical representations. The challenge has been to provide a seamless model for querying and filtering the corpus and/or persons in both simple report views and in the visualization; naïve application of SNA and visualization to large datasets is often hard to use. BPS includes various filters and queries that narrow in on specific questions and produce more useful visualization.

BPS defines script libraries to manage the data from the services (especially the SNA services), but the visualization support is built upon the D3 script library [3].

F. Workspaces in BPS

The Digital Humanities research environment exposes the functionality of the probabilistic model and assertions in workspaces. Each user can have one or more workspaces for his projects, and can import one or more corpora into each workspace for analysis; each workspace has an independent set of model parameters and curatorial assertions. After setting parameters for disambiguation rules and making specific assertions, a researcher can see the effects on the resulting SNA visualizations. Support for *freezing* a workspace and bookmarking specific visualizations means that results and views of a model can be shared and cited in publication. The current implementation, which supports this workspace model, has been evaluated with current users, who embrace the semblance of the digital workflow to that carried out with the familiar pen-and-paper one. User requests for future expansion include implementation of support for publishing an experimental result as a workspace and shared workspaces that enable collaboration among colleagues. These move humanities research into the realm of “Big Data” by inviting both speculative and collaborative investigation and encouraging the study of larger and more complex data sets. Pedagogical use-cases enable students to follow along as a faculty member works through the process of prosopographical research in a corpus, learn the process, and track the judgments of the experienced user, all while being trained in best digital humanities practices.

IV. EVALUATION WITH USERS

The merits of data-driven science, social science research, and text-based humanities derive in large measure from the approbation of colleagues qualified to assess the validity of data utilized as well as the methodological underpinnings of the research program. As humanists approach larger text corpora and ask wider-ranging questions of that data, they face increasing computationally complexity, regardless of the “Small Data” number of bytes they process. Tools that replicate workflows with which they are already familiar are transformative not because of the way they work, but because of the ways they enable the humanist to expand his enquiry.

In early stage workshops, humanists expressed strong support for the BPS model and tools, suggesting that they would result in “aha!” moments, characterized by new research directions and outcomes not previously possible. The assertions-based model and the workspaces in which different parameterizations could find expression and evaluation that would be the catalyst to further discourse and dialogue.

Users confirmed that the disambiguation and assertion models reflect actual workflows, and affirmed the value of assigning attribution provenance and the creation of workspace environments for individual and collaborative research. Users explored the implications of the flexible BPS data model, which allowed them to filter on features of persons and activities in ways that had not been practical before. As they did so, they discovered that BPS allowed them to ask new questions about their corpora. The BPS team and workshop participants compiled and prioritized a list of desired features and tools for, e.g., performing date conversions between different calendars, adding new import models, and supporting multiple workspaces. A subsequent webinar included a real-time demonstration of the implementation of the SNA graph visualization. Response was good, and users endorsed the approach presented in the proof of concept demonstration.

Future development phases will expand and refine core functionality, and will also consider whether and how BPS is changing scholarly workflows.

V. CONCLUSIONS AND FUTURE WORK

“Big Data” in the humanities is not about the number of bytes, but rather about the nature of the work and accompanying workflows. BPS is a “Big Data” humanities tool-kit that changes workflows by expanding the capacity to ask new questions, to handle larger data sets and to discover features that were previously impractical to identify. Extending the research model means that humanists could: discover the relation of mercantile activity between multiple families in differing locations, assess individuals’ wealth accumulation against patterns of seasonal variation in agricultural productivity, and establish the agency of women as principals and witnesses in a social network of economic transactions — questions previously deemed not possible or practical to answer.

“Big Data” applies to humanities research, even when a corpus contains <500K items. Humanists rejoice when archaeological excavations or cataloging projects in libraries turn up masses of new documentation. But at the same time, they may be reluctant or ill-equipped to incorporate the data and metadata from 20,000 new documents into a standing research project. Changes in the size of data-sets on the order of 10-20% (regardless of the absolute number of data items) can best be transformative when tools incorporate the new data easily into an existing workflow.

BPS enables the implementation of “Big Data” in humanities research by automating humanist workflows.

Tools at home in the disciplines of computer science, natural language processing, and statistics are not part of the historical workflow in this kind of disambiguation and prosopographical analysis, in spite of the fact that the process is probabilistic and computational. A thoughtful, well-engineered tool-kit brings to a field of humanities research tools that serve existing needs and facilitates the answering of new and expanding questions.

A tool-kit such as BPS also facilitates and promotes communication and collaboration in the research process. In the humanities, debate and authority is traditionally expressed and established on static pages of learned journals. The workspace environment of BPS affords researchers the opportunity to engage immediately and intimately in digital “conversations”, and to track, through their own, shared, or others’ BPS workspaces.

In the next phase of work we will be expanding upon the assertions model, generalizing some of the model for features used in disambiguation, and adding additional workspace support. We plan to conduct a series of formal evaluations to track and quantify the computations implemented by the humanists and correlate those with the self-reported impact that the tools have on the framing of humanities scholarship, with particular focus on the contribution of BPS toward extending the complexity of research agendas and opening viable humanities approaches to the realm of “Big Data”.

VI. ACKNOWLEDGEMENTS

The early development of BPS was supported in part by a Digital Humanities Start-Up Grant from the National Endowment for the Humanities. We would also like to acknowledge the Oracc project⁹ which provides infrastructure support for cuneiform corpus transliteration, and contributed supporting technology for TEI markup of the corpora. Finally, we would like to acknowledge the contributions of Davide Semenzin, Utrecht University ICS, who implemented a number of key features as part of his Master’s final project.

REFERENCES

[1] Ahmed, A. 2011. *The Religious Elite of the Early Islamic Hijāz : Five Prosopographical Case Studies*. Occasional Publications of the Oxford Unit for Prosopographical Research 14. Oxford: Unit for Prosopographical Research Linacre College University of Oxford.

[2] Booth, A. 2013. *Brief Overview of Curating Lives: Museums, Archives, Online Sites, Autobiography, Biography, and Life Writing* session, MLA Commons, Jan 5 2013. Available at: <http://commons.mla.org/docs/a-brief-synopsis-of-curating-lives-mla-paper-alison-booth/>

[3] Bostock, M., et al. 2011. *D3: Data-Driven Documents*, IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011.

[4] Brandes, U., et al. 2000. *The GraphML file format*.

[5] Csillag, K. 2013. *Fuzzy anchoring* (blog post, April 22, 2013). Available at: <http://hypothes.is/blog/fuzzy-anchoring>.

[6] Elson, D., Dames, N., and McKeown, K. 2010. *Extracting Social Networks from Literary Fiction*, Proc. of 48th Annual Meeting of the Association for Computational Linguistics, pages 138–147, Uppsala, Sweden, 11-16 July 2010

[7] Fielding, R., and Taylor, R. 2002. *Principled design of the modern Web architecture*. ACM Trans. Internet Technol. 2,2 (May 2002) DOI=10.1145/514183.514185

[8] Gerritsen, A. 2008. *Prosopography and its Potential for Middle Period Research*, Journal of Song-Yuan Studies Volume 38, 2008 pp. 161-201.

[9] Higgins, S. 2011. *Digital Curation: The Emergence of a New Discipline*, International Journal of Digital Curation, 2011, Vol. 6, No. 2, pp. 78-88, doi:10.2218/ijdc.v6i2.191. <http://ijdc.net/index.php/ijdc/article/view/184>

[10] Mueller, M. 2011. *Collaboratively Curating Early Modern English Texts*. Contributed essay to Project Bamboo wiki. <https://wikihub.berkeley.edu/x/QAdRB>. Accessed June 2013.

[11] O’Madadhain, J., et al. 2005. *Analysis and visualization of network data using JUNG*. Journal of Statistical Software 10.2 (2005): 1-35.

[12] Pasin, M. 2012. *Exploring Prosopographical Resources Through Novel Tools and Visualizations: a Preliminary Investigation*, Digital Humanities 2012, Hamburg.

[13] Schmitz, P., and Pearce, L., 2013. *Berkeley Prosopography Services: Ancient Families, Modern Tools*, DH-Case 2013 (workshop), ACM Document Engineering 2013, Florence, Italy

[14] TEI P5: *Guidelines for Electronic Text Encoding and Interchange*, Ch. 13 Names, Dates, People, and Places, V 2.3.0. Available at: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html> (accessed June 2013).

[15] Waerzeggers, C. 2003-2004. *The Babylonian Revolts Against Xerxes and the ‘End of Archives’*, Archiv für Orientforschung 50: 150–173.

[16] Waerzeggers, C. 2013. *Social Network Analysis of Cuneiform Archives: A New Approach*. Proc. of the Second START Conference in Vienna (17-19th July 2008) Too Much Data? Generalizations and Model-building in Ancient Economic History on the Basis of Large Corpora of documentary Evidence , edited by H. D. Baker and Michael Jursa.

⁹ <http://oracc.org>

DH-CASE II: Collaborative Annotations in Shared Environments: metadata, tools and techniques in the Digital Humanities

Patrick Schmitz
OCIO/Research IT
UC Berkeley
Berkeley, CA
pschmitz@berkeley.edu

Laurie Pearce
Dept. of Near Eastern Studies
UC Berkeley
Berkeley, CA
lpearce@berkeley.edu

Quinn Dombrowski
OCIO/Research IT
UC Berkeley
Berkeley, CA
quinnd@berkeley.edu

ABSTRACT

The DH-CASE II Workshop, held in conjunction with ACM Document Engineering 2014, focuses on the tools and environments that support annotation, broadly defined, including modeling, authoring, analysis, publication and sharing. Participants explored shared challenges and differing approaches, seeking to identify emerging best practices, as well as those approaches that may have potential for wider application or influence.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *Linguistic processing*, H.3.5 [Information Storage and Retrieval]: Online Information Services–*Web-based services*, D.2.13 [Software Engineering]: Reusable software.

General Terms

Design, Experimentation, Human Factors.

Keywords

Annotation, Metadata, Cyberinfrastructure, Digital Humanities.

1. INTRODUCTION

Digital Humanities is rapidly becoming a central part of humanities research, drawing upon tools and approaches from Computer Science, Information Organization, and Document Engineering to address the challenges of analyzing and annotating the growing number and range of corpora that support humanist scholarship.

From cuneiform tablets, ancient scrolls, and papyri, to contemporary letters, books, and manuscripts, corpora of interest to humanities scholars span the world's cultures and historic range. More and more documents are being transliterated, digitized, and made available for study with digital tools. Scholarship ranges from translation to interpretation, from syntactic analysis to multi-corpus synthesis of patterns and ideas. Underlying much of humanities scholarship is the activity of annotation. Annotation of the "aboutness" of documents and entities ranges from linguistic markup, to structural and semantic relations, to subjective commentary; annotation of "activity"

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference'10, Month 1–2, 2010, City, State, Country.

Copyright 2010 ACM 1-58113-000-0/00/0010 ...\$15.00.

around documents and entities includes scholarly workflows, analytic processes, and patterns of influence among a community of scholars. Sharable annotations and collaborative environments support scholarly discourse, facilitating traditional practices and enabling new ones.

2. WORKSHOP CONTRIBUTIONS

Contributions were solicited related to the intersection of theory, design, and implementation, emphasizing a "big-picture" view of architectural, modeling and integration approaches in digital humanities. Submissions were encouraged that discuss data and tool reuse, and that explore what the most successful levels are for reusing the products of a digital humanities project (complete systems? APIs? plugins/modules? data models?). We noted that submissions discussing an individual project should focus on these larger questions, rather than primarily reporting on the project's activities.

The workshop was a forum in which to consider the connections and influences between Digital Humanities annotation tools and environments, and the tools and models used in other domains, that may provide new approaches to the challenges we face. It was also a locus for the discussion of emerging standards and practices such as OAC (Open Annotation Collaboration) and Linked Open Data in Libraries, Archives, and Museums (LODLAM).

We received nine submissions, of which we accepted four for inclusion in the proceedings. The remaining authors were encouraged to attend and participate in the workshop discussion. We received additional requests to participate from members of the community who did not submit papers, and we welcomed their participation.

DH-CASE II was a full day workshop. Authors of accepted papers were given time to present their research/project, followed by a discussion of emerging themes, best practices, and the potential for integration or collaboration. We summarized the workshop and discussion to the broader DocEng community during the conference.

Proceedings will be published via the ACM International Conference Proceedings Series. Workshop organizers will produce and submit a paper to "Digital Humanities Quarterly" (www.digitalhumanities.org/dhq/) summarizing topics that arise in the workshop. We may consider a special issue on workshop topics or key findings to an appropriate journal.

3. PROGRAM COMMITTEE

We would like to thank the program committee members for their help reviewing submissions, and ensuring that the proceedings were of high quality. The PC members were:

Antoine Isaac, Vrije Universiteit Amsterdam
Cerstin Mahlow, University of Stuttgart
Christof Schoch, University of Würzburg
Claus Huitfeld, University of Bergen
Corey Harper, New York University
Elisabeth Burr, University of Leipzig
Fabio Vitelli, University of Bologna
Francesca Tomasi, University of Bologna
Jacco van Ossenbruggen, CWI, Amsterdam
Jody Perkins, Miami University of Ohio
John Bradley, King's College London
Lisa Spiro, Rice University
Michael Piotrowski, Leibniz Institute of European History

Paolo Ciccicarese, Harvard University
Paul Spence, King's College London
Ryan Shaw, University of North Carolina
Silvio Peroni, University of Bologna, Italy

4. ACKNOWLEDGMENTS

Our thanks to Research IT at UC Berkeley for sponsoring the workshop, and to the ACM ICPS for publishing support.

See also the workshop website, at:

<http://research-it.berkeley.edu/dhcase2014>



Research IT (RIT) provides research computing technologies, consulting and community for the Berkeley campus. Our goal is to advance research through IT innovation.

PROGRAMS	SERVICES	PARTNERSHIPS	PROJECTS	NEWS	ABOUT
----------	----------	--------------	----------	------	-------

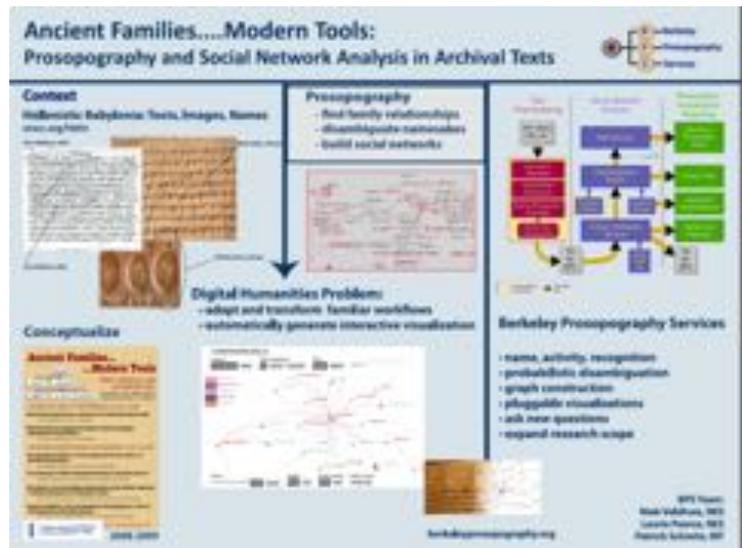


Berkeley Prosopography Services paper to be presented at DocEng 2014

Submitted by [Steve Masover](#) on July 2, 2014

A conference paper describing [Berkeley Prosopography Services](#) (BPS) has been accepted for presentation at this year's ACM Symposium on Document Engineering ([DocEng 2014](#)).

In *Humanist-centric tools for Big Data: Berkeley Prosopography Services*, authors Patrick Schmitz of Research IT, and Laurie Pearce of UC Berkeley's Department of Near Eastern Studies frame BPS as "a new set of tools for prosopography - the identification of individuals and study of their interactions - in support of humanities research," and note that "prosopography is an example of Big Data in the humanities, characterized not by the size of the datasets, but by the way that computational and data-driven methods can transform scholarly workflows."



Technically, Schmitz and Pearce explain that:

BPS is based upon re-usable infrastructure, supporting generalized web services for corpus management, social network analysis, and visualization. The BPS disambiguation model is a formal implementation of the traditional heuristics used by humanists, and supports plug-in rules for adaptation to a wide range of domain corpora. A workspace model supports exploratory research and collaboration.

Using the tools that constitute BPS, "researchers assert conclusions or possibilities, allowing them to override automated inference, to explore ideas in what-if scenarios, and to formally publish and subscribe-to asserted annotations among colleagues, and/or with students." The paper describes researchers' experience using BPS in the study of corpora of cuneiform tablets, as well as plans to apply to the tools to other types of textual corpora.

The authors will present their paper at the DocEng meeting in Fort Collins, Colorado, to be held September 16 - 19.

Tags: [humanities](#) [BAM/PFA](#) [Art](#) [image](#) [Information Services and Technology](#)

Program: [Digital Humanities](#)

Service: [Digital Humanities](#)

Project: [Berkeley Prosopography Services](#)

Partnership: [Arts & Humanities Division](#)

Technology @ Berkeley

Email us: research-it@berkeley.edu



[@research_it_ucb](#)

What is Prosopography?

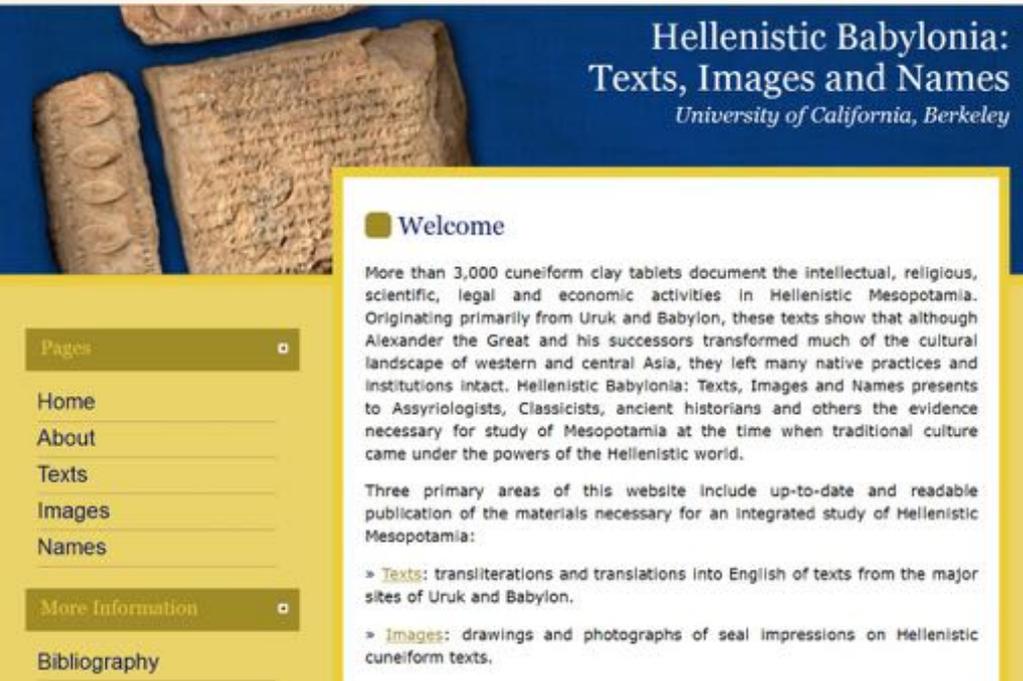
- Identifying people referenced in corpora: onomasticon
- Building genealogies: family lineages
- Recovering relationships: social networks

Dependencies:

- Scope and condition of media & data
- Disambiguating namesakes
- Finding family relations
- Recognizing activities and roles
- Controlling chronological framework



Project Content: Hellenistic Uruk



**Hellenistic Babylonia:
Texts, Images and Names**
University of California, Berkeley

Welcome

More than 3,000 cuneiform clay tablets document the intellectual, religious, scientific, legal and economic activities in Hellenistic Mesopotamia. Originating primarily from Uruk and Babylon, these texts show that although Alexander the Great and his successors transformed much of the cultural landscape of western and central Asia, they left many native practices and institutions intact. Hellenistic Babylonia: Texts, Images and Names presents to Assyriologists, Classicists, ancient historians and others the evidence necessary for study of Mesopotamia at the time when traditional culture came under the powers of the Hellenistic world.

Three primary areas of this website include up-to-date and readable publication of the materials necessary for an integrated study of Hellenistic Mesopotamia:

- > **Texts:** transliterations and translations into English of texts from the major sites of Uruk and Babylon.
- > **Images:** drawings and photographs of seal impressions on Hellenistic cuneiform texts.

Pages

- Home
- About
- Texts
- Images
- Names

More Information

Bibliography



530

legal texts

8-20

name citations/text

3

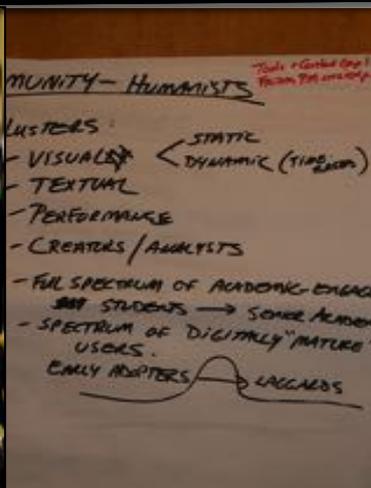
individuals/citation

10,000

name instances



“Talk the Talk”: DH & the world’s oldest writing system in conversation



Oracc Activity data mining affiliation
 ATF Corpus HTML TEI
 DIRT papyrology NLP
 Digital Research Tools XSLT SOA
 dirtdirectory.org Lemmatization



decipherment

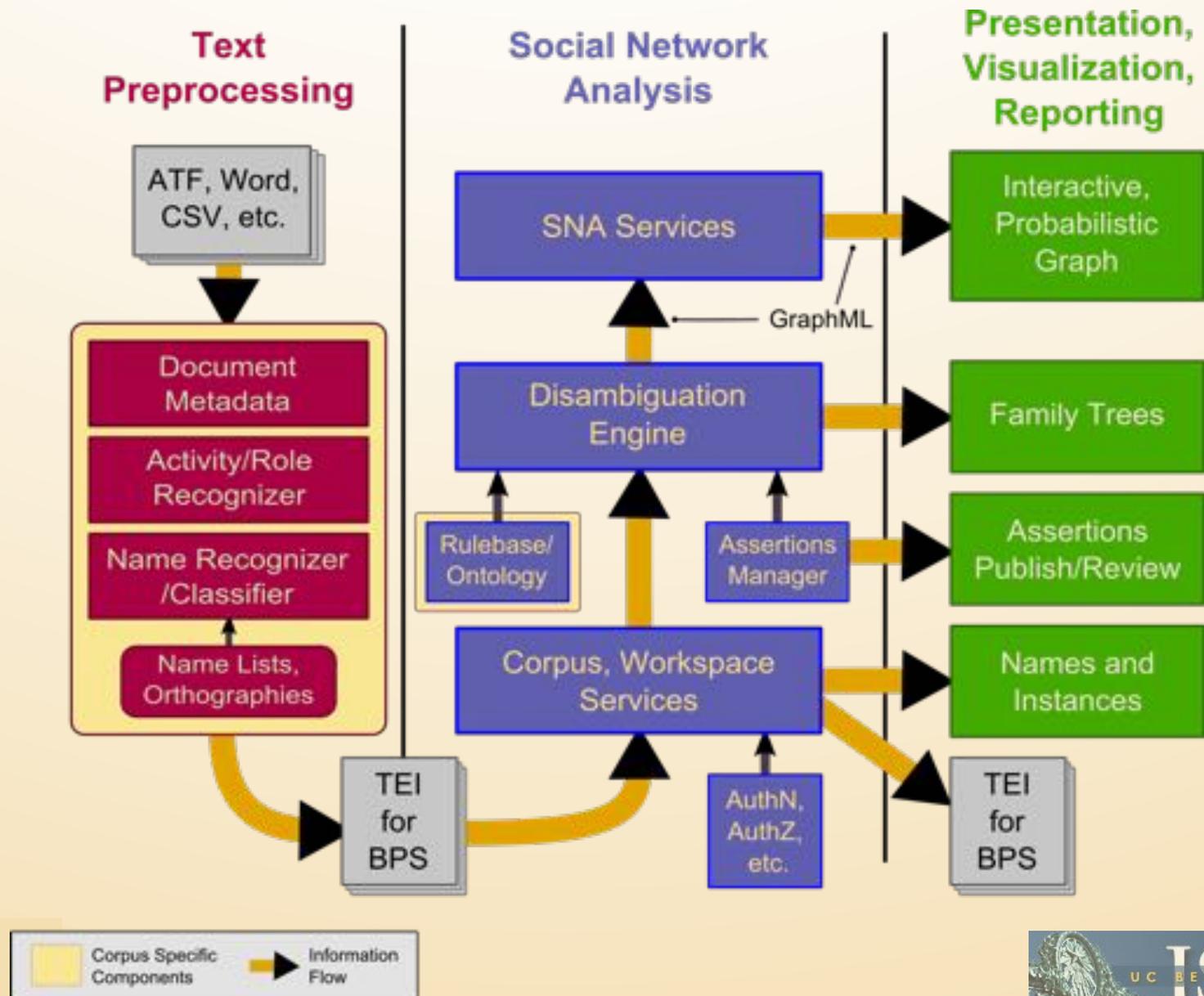


BPS - The Grand Vision

- A generalized, extensible tool-box
- Communities build recognizers for names, activities, etc.
 - Domain-specific rule-bases for various languages, corpora
- Social Network Analysis (SNA) features
- Graph viz. tools for families, social networks
- Support **humanities-research workflows**: *what-if*, uncertainty, and disagreement



BPS System Architecture



High-level Processing Model

1. Import TEI for corpus, build model:
2. Corpus has Documents, each of which has:
 - One or more *Activities*, each of which has:
 - One or more *Name* citations, in *Roles*
 - Name-Role-Activity-Document → *nrad* is base unit
3. (Clone into workspace, set params)
4. Collapse Persons using disambiguation rules
 1. Apply locally within a document, normalize
 2. Apply globally across the corpus, normalize



Disambiguation Rules

- Classes of rules, normalized in context
 - Shifts, Boosts, and Discounts
 - Name heuristics, General feature rules (e.g., place), Date heuristics/constraints
 - Role matrices
- Rules are configurable/pluggable/extensible
- Rules expose user-facing aspects (meta-data)
 - For parameterization UI, allowing researchers to control impact of rules



BPS innovations

- Assertions, not factoids
 - Probabilistic model
 - Support uncertainty
- Network analysis
- Workspaces
 - Support hypotheses
 - Build community
 - Track authority



Assertions

- Assertions integrate directly into model
 - Override disambiguation results
 - Control model (rule) parameters
- Assertions encapsulate judgment by user
- Assertions are sharable
 - Publish-from/Consume-into workspaces
- Assertions expose user's approach and meta-data
 - Natural language description of effect
 - Include provenance (who originally published)



What is a probabilistic model?

- Posit a Person for each citation in a document
 - Each name cited *might be* a unique person (but isn't really)
- Citations refer to one of several real persons
 - Express each possible link as probability or weight (0-1)
 - Shift weight around with heuristic rules.
- Persons relate to one another thru documents (roles, activities, family links)
 - Express person-person links as probability or weight (0-1), based upon above weighted links to citations in docs
- Yields a graph with weighted edges/links
- Users can filter or focus to simplify the graph



Berkeley Prosopography Services

Logged in as Laurie | [Help](#) | [Admin](#) | [Sign Out](#)

Home Corpora Workspace

[Documents](#) [People](#) [Clans](#) [Settings](#) [Admin](#)

Set Model Parameters

Background on the model

The BPS analyzer will try to disambiguate among citations using the same name(s). To do this, it will basically model a new citation-person for each name it finds in a document (including fathers, grandfathers, etc. that are mentioned as qualifiers to the named actors). Then, it will attempt to collapse some of those citation-persons to get to the set of actual (real world) persons mentioned in all the corpus documents. Each citation-person is compared to other citation-persons, and a set of rules is applied to determine how likely it is that the two citations are the same person. The analyzer proceeds in two steps: first it considers all the citation-persons within each single document (intra-document), and then it considers the citation-persons across the entire corpus (inter-document).

When comparing two citation-persons, the analyzer will first require that there is no conflicting information about the two citation-persons - e.g., if they have different declared fathers, they will be considered as distinct, and will not be collapsed. The rules below allow you to configure whether specific rules must be considered to be distinct, and to control how strong the likelihood that two persons with partial matching name information are the same real world person.

General settings:

Number of qualifications (father/grandfather/ancestor/clan) in addition to forename required to consider a name citation "fully qualified"	<input type="text" value="2"/>
Assumed typical length of active business life (years)	<input type="text" value="25"/>
Assumed typical separation of generations (years)	<input type="text" value="15"/>

Step 1: Intra-document rules:

Rule Steps 1A, 1B, and 1C collapse citations within a single document.

Step 1A: Consider equally qualified names

Collapse equal, fully qualified citations (e.g., "PN _a , son-of PN _b , in-clan CN _c " and "PN _a , son-of PN _b , in-clan CN _c ")	<input type="text" value="Always: 100%"/>
Collapse equal, partly qualified citations (e.g., "PN _a , son-of PN _b " and "PN _a , son-of PN _b ")	<input type="text" value="Conservative: 30%"/>
Collapse equal, unqualified citations (e.g., "PN _a " and "PN _a ")	<input type="text" value="Aggressive: 75%"/>

Step 1B: Consider compatible, but not equally qualified names

Collapse partly qualified citations with compatible, fully qualified citations (e.g., "PN _a , son-of PN _b " and "PN _a , son-of PN _b , in-clan CN _c ")	<input type="text" value="Conservative: 30%"/>
Collapse unqualified citations with compatible, more qualified citations (e.g., "PN _a " and "PN _a , son-of PN _b , in-clan CN _c ", OR, "PN _a " and "PN _a , son-of PN _b ")	<input type="text" value="Aggressive: 75%"/>



For more information:

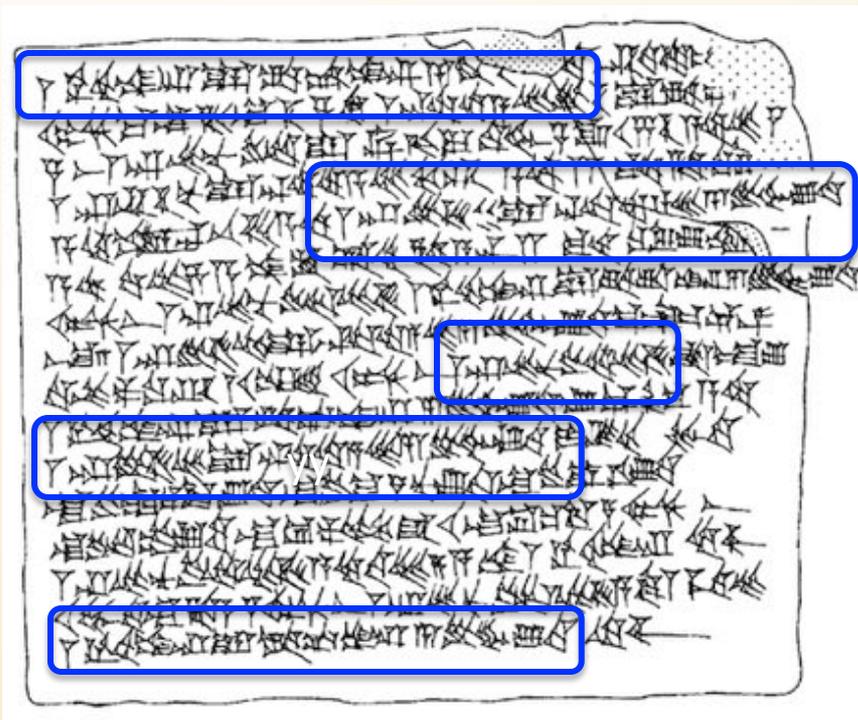
<http://www.berkeleyprosopography.org>

- HBTIN project home:
 - <http://oracc.museum.upenn.edu/hbtin/>
- Project wiki
 - <https://wikihub.berkeley.edu/display/istds/Berkeley+Prosopography+Services+Wiki+Home>
- Code:
 - <http://code.google.com/p/berkeley-prosopography-services>
- Contact us:
 - Laurie Pearce (lpearce@berkeley.edu)
 - Patrick Schmitz (pschmitz@berkeley.edu)
 - Niek Veldhuis (veldhuis@berkeley.edu)



Data Mining in Uruk Legal Texts

- **Boilerplate text**
 - repetitive patterns
 - attributes
 - many names!



- **Onomastic data**

- **standard naming pattern:**
A / son of B / son of C // descendant of D
- **papponymy:** name child for (male) ancestor





Research Streams

Events

Initiatives

Services & Support

Affiliated Centers

About

[Home](#) › [Research Streams](#) › [Matrix](#) › **Research Streams**

<https://matrix.berkeley.edu/research/social-networks-history>

Matrix

Social Networks from History

Associated Channels

[Matrix](#)

[Identities](#)

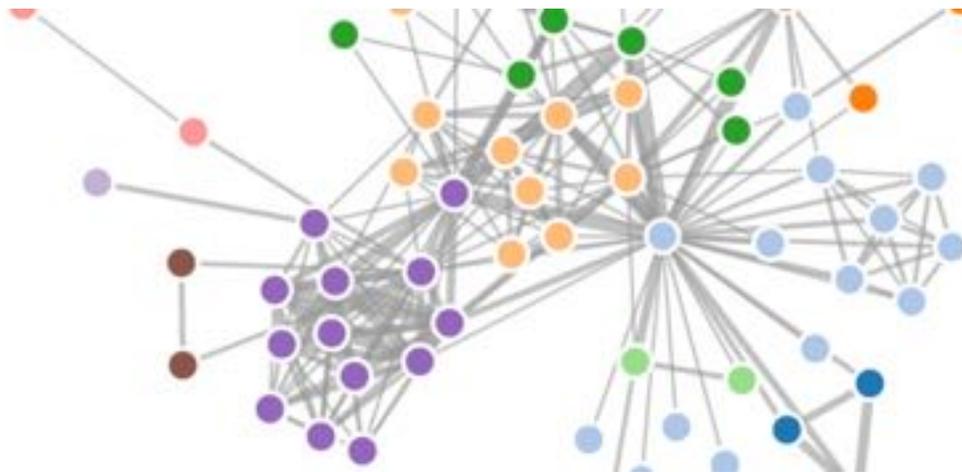
[Technology](#)

[Complexity](#)

July 28, 2014 by *Chuck Kapelke*

SHARE

A new tool developed by a team at UC Berkeley can help build “prosopographies,” social networks based on names, affiliations, and other historical data.



Social scientists working to study modern society have an array of data sources available, from government censuses to digital social networks. But for researchers trying to study people who lived, say, 2500 years BFB (before Facebook), trying to figure out who was friends (or business associates, of family members, or political allies) with whom can pose significant challenges.

Enter “prosopography,” the practice of gaining insights into individuals based on attributes of their families, business associates, or other affiliates, based on information preserved in historic documentation. Prosopography has long been an important tool for the study of all kinds of past societies; the name stems from the Greek *prosopoeia*, or “face created,” suggesting how this methodology enables researchers to “put a face on” individuals about whom little is known based on information about their connections to other people.

For Laurie Pearce, lecturer in Near Eastern Studies at UC Berkeley, prosopography has been a crucial resource for understanding the relationships among elite individuals in ancient Mesopotamia. Pearce’s research entails using approximately 700 clay tablets written in cuneiform script to tease out the lives of people who lived in southern Iraq during the fourth century BC; many of these texts detail real estate transactions or small transfers of income. “It’s a challenge, because people were usually named after their fathers or grandfathers, so they had names like, ‘Joe Son-of-Fred Son-of-Joe Son-of-Fred,’” she explains. “Straightening out who lived where and when, and what they did, is a non-trivial problem.”

Pearce has been working together with Patrick Schmitz, Berkeley’s Associate Director for Research IT and Strategy, to establish [Berkeley Prosopography Services](#), or BPS, which has created an XML-based tool to help map out relationships and discern

individuals from each other. “The texts that I work on seemed like a good demonstrator corpus to explore a set of tools for a common research problem: that is, identifying individuals mentioned in text of any sort and trying to establish the social networks in which those individuals appear,” Pearce explains.

The BPS tool applies methods from the fields of natural language processing (NLP) and social network analysis (SNA) to extract the names and basic familial relationships of people mentioned in texts, and weigh the probabilities that individuals were connected to each other. The resulting graph model can be used to produce reports and visualizations ranging from simple name lists and family trees to interactive models. “There is a lot of uncertainty in these historical records,” Schmitz says. “The tool can represent that uncertainty explicitly. We can say, it is 70 percent certain that this name corresponds to this person, but it could also be one of these three people.”

Once you know who was working with whom—for example, who was in the same location at the same time—you can see how ideas might have spread and become more accepted in the community.

To further refine this tool, Pearce and Schmitz are working with the UC Berkeley Social Science Matrix to coordinate a seminar focused on exploring how historical social networks might be used by researchers working in different domains. “Prosopography has traditionally been used in the humanities, but it became clear to us that social scientists and some natural scientists are facing similar challenges of disambiguation and understanding relationships,” Pearce explains. “In the Matrix seminar, we’d like to explore how our approach and tools might benefit them. The point is not just to get together people from Near Eastern Studies who want to do this; it’s to say, what can we

learn from other disciplines? We are specifically searching for people who want to understand how they would use a tool like this.”

As an example of a broader application, BPS has potential to help staff members from Berkeley’s Museum of Vertebrate Zoology, who are sifting through 100+ year-old records of the activities of collectors and researchers. “They’re trying to understand how those people related to one another, and how they collaborated,” Schmitz says. “They are facing issues ... that make the identification of some of the people involved non-obvious. There’s so much data in all the thousands of records, it takes an effort to come to grips with, what were the patterns of collaboration and influence among these people?”

The power of prosopography, Schmitz explains, is that it can use these data sets to generate insights into bigger questions. “Researchers can use the network analysis to say, what are all the patterns of interaction? Who are the subgroups? Who had the most influence?” he says. “Once you know who was working with whom—for example, who was in the same location at the same time—you can see how ideas might have spread and become more accepted in the community. It has implications far beyond just mapping out a tree; it’s looking at the construction of knowledge within institutions.”

Berkeley Prosopography Services was recently awarded an NEH Digital Humanities Implementation Grant that will support this work for a two-year period. Schmitz will serve as the technical lead, and Niek Veldhuis (professor in Near Eastern Studies) will join Pearce as co-principal investigator (PI).

For more information, see the home page of [Berkeley Prosopography Service](#), or read a [research paper](#) by Schmitz and Pearce.

Article Type

Research Highlights

Comments

SHARE

Social Network Visualizations: use case with HBTIN data

Social Science Matrix Workshop #3

November 14, 2014

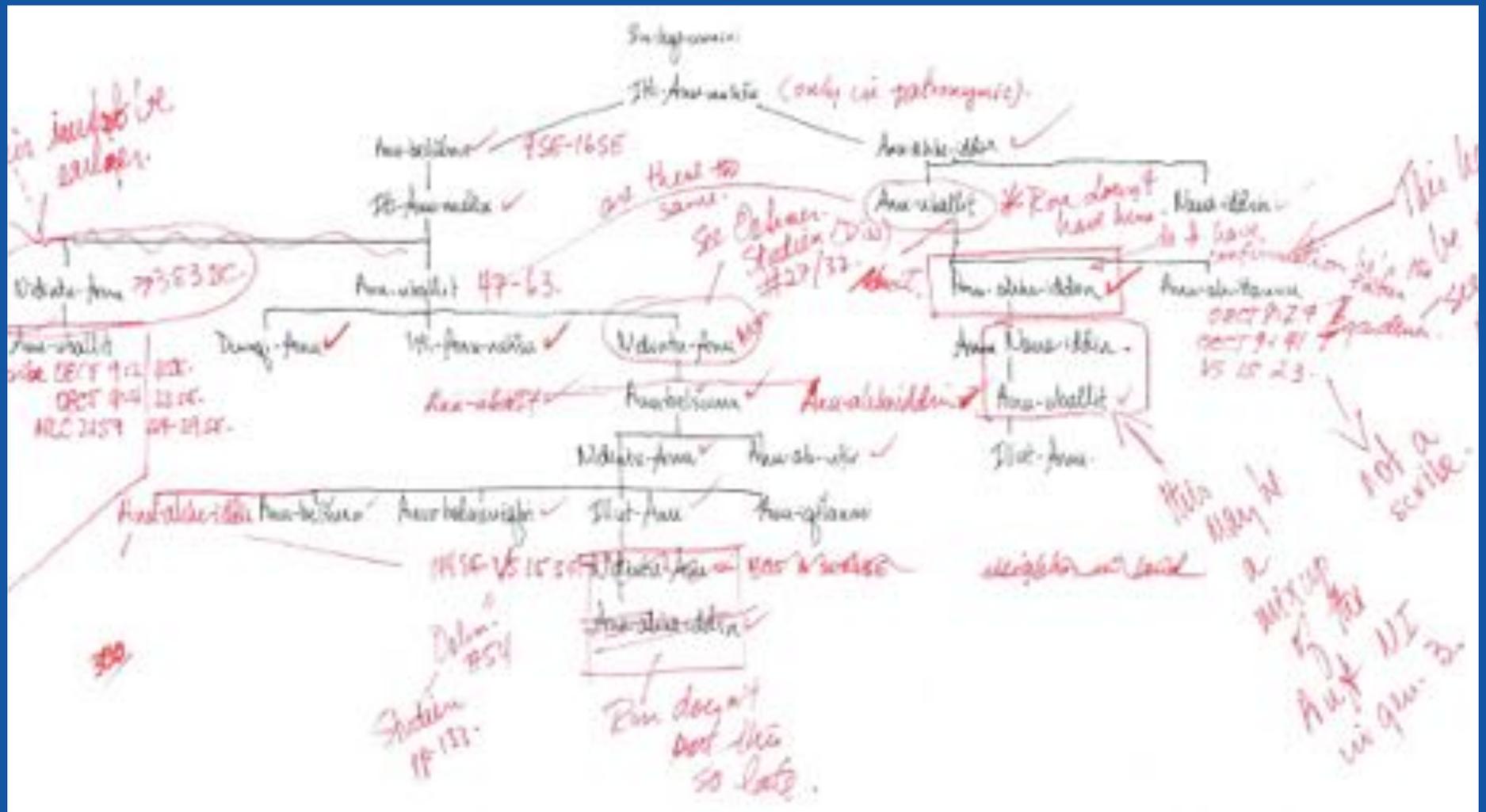
Laurie Pearce, Patrick Schmitz, Davide Semenzin

Picture perfect data?



Text	Individual Index	PN	FN	GFather
BRM 2 24		Anu-iksur	Kidin-Anu	Dannat-Belti
BRM 2 37		Anu-iksur	Zeriya	Anu-iksur
BRM 2 44		Anu-iksur	Kidin-Anu	Anu-iksur
CM 12, 06	Anu-iks,ur[02]	Anu-iksur	Anu-ahhe-iddin[07]	(Anu-ah-ittannu[11])
MLC 2124A		Anu-iksur	Kidin-Anu	Dannat-Belti
MLC 2172		Anu-iksur	Zeriya	Ina-qibit-Anu
NCBT 1963		Anu-iksur	Labasi	Mukin-apli
OECT 9 63		Anu-iksur	Ina-qibit-Anu	Anu-ah-iddin
OECT 9 63		Anu-iksur	Ina-qibit-Anu	Anu-ah-iddin
TCL 13 243	Anu-iks,ur[02]** may equal individual in CM 12, 06	Anu-iksur	Anu-ahhe-iddin[07]	Anu-ah-ittannu
VAS 15 34		Anu-iksur	Kidin-Anu	Anu-iksur
VAS 15 48	Anu-iks,ur[02]** may equal	Anu-iksur	Anu-ahhe-iddin[07]	(Anu-ah-/MU/-[nu?])
OECT 9 42		Anu-iqisannu	Liblut	Nana-iddin
BIMes 24 45		Anu-iqisannu	Anu-ittannu	Nana-iddin

Broken orthographies complicate disambiguation.
Constraints of tabular presentation of data.



Can a diagram show “what if”?

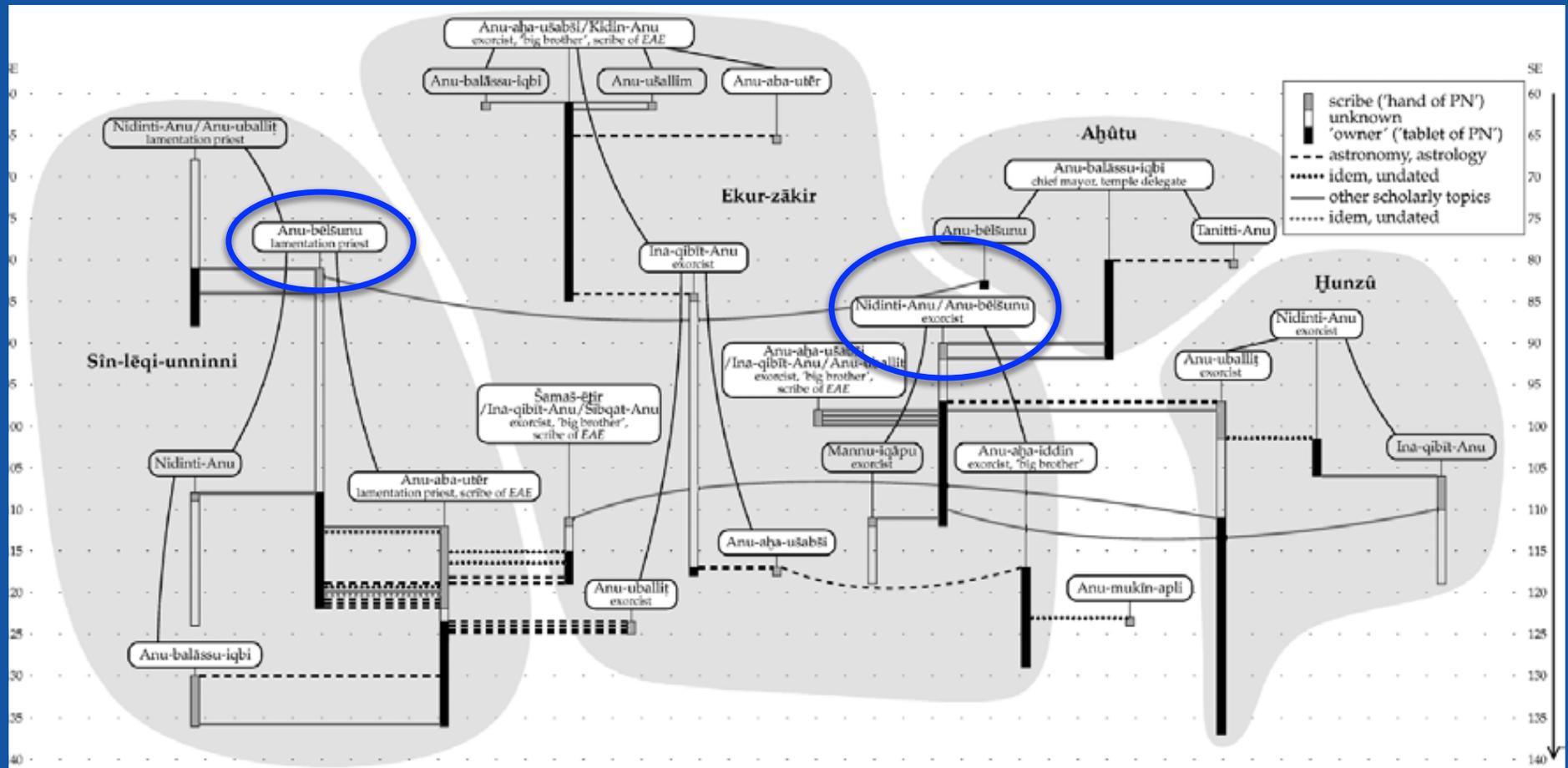


Figure B. Printed diagram of network relations

Ekur-zākir

Ina-qibit-Anu
exorcist

Anu-balassu-igbi
chief mayor, temple delegate

Anu-bēšunu

Nidinti-Anu / Anu-bēšunu
exorcist

Anu-aha-ušabši
/ Ina-qibit-Anu / Anu-uballit
exorcist, 'big brother',
scribe of EAE

Šamaš-ēbir
/ Ina-qibit-Anu / Sibqat-Anu
exorcist, 'big brother',
scribe of EAE

Anu-aba-utēr
ritual priest, scribe of EAE

Mannu-ikāpu
exorcist

Anu-aha-iddin
exorcist, 'big brother'

Anu-aha-ušabši

Anu-uballit

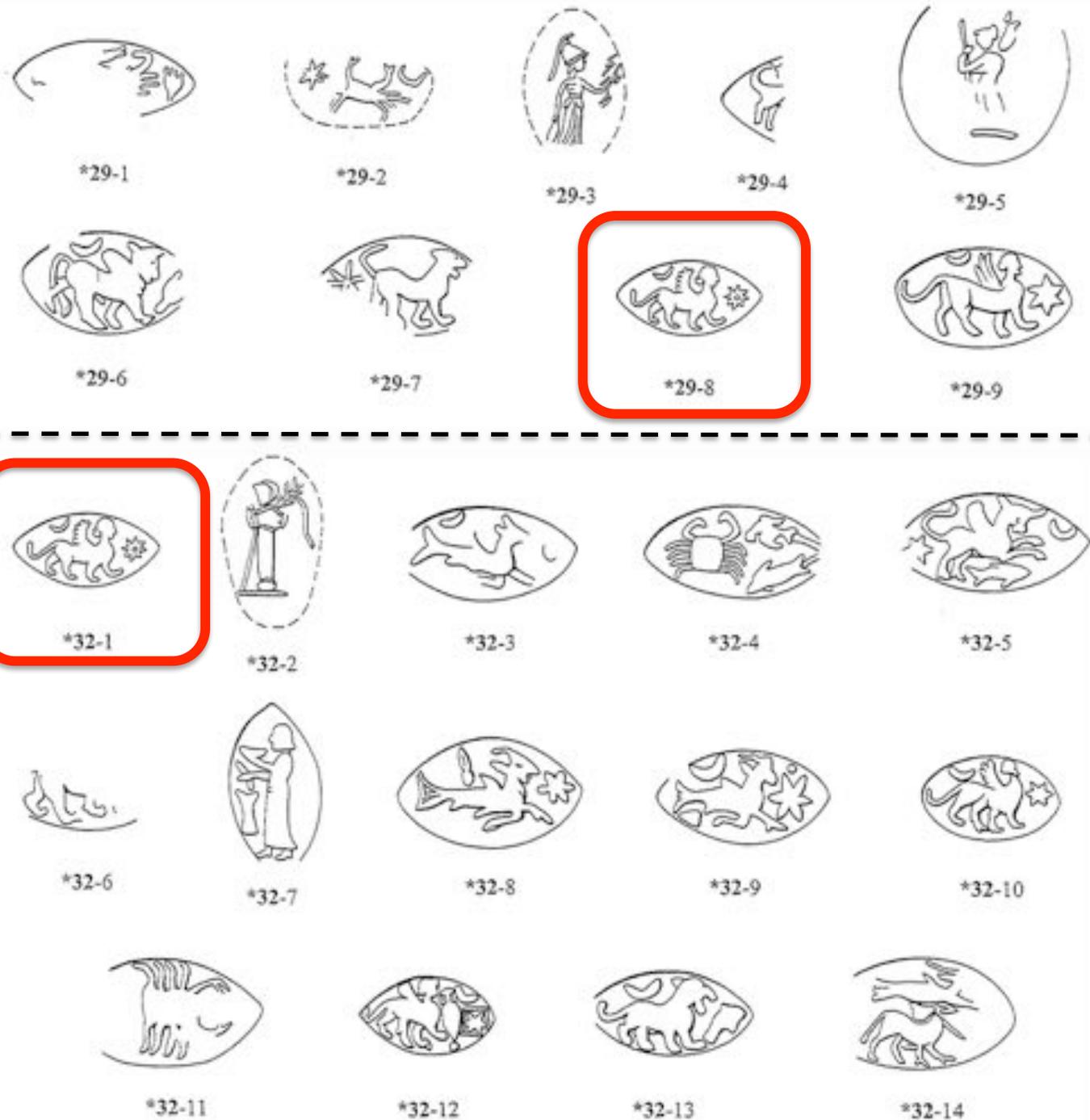
Anu-muktu-anli

Will the real Anu-aha-ušabši / Ina-qibit-Anu, exorcist, please stand up?

OECT 9, 12
P342349



Can a network of images support a network of text / text ids?



[http://berkeleyprosopography.org/
docs/UXIdeasSpr2014](http://berkeleyprosopography.org/docs/UXIdeasSpr2014)

- [http://berkeleyprosopography.org/docs/
UXIdeasSpr2014#Figure0](http://berkeleyprosopography.org/docs/UXIdeasSpr2014#Figure0)

lemmatized text

```
#alt: use math
@tablet
# note: = Ashm 1923-732; Del Monte 2000, 194-196 (partial edition)
@obverse
1.      IM HA.LA sza2 {m}BAD4#-{d#}GASZAN# u {m}ni-din-tu4--{d}60 DUMU#-MESZ# sza2 {n}lib#-lut,# A
{m}lu-usz#-[t#m]-nar--{d}ISZKUR
#lem: {uppi|letter|N; zitti|share|N; {a|of|DET; Dannat-Beltu|PN; u|and|CNJ; Nidintu-Anu|PN; m#ri|son|N; {a|of|DET;
Liblut|PN; m#r|descendant|N; Lu#tammar-Adad|i|LN
2.      ina# ki-szub-ba#a' sza2 bi-rit#-szu2#-nu# sza2 KI#(+t13) E2# {d#}ISZKUR# sza2# qe2#-reb# UNUG{ki}
#lem: ina|in|PRP; ki#ubb#|waste field|N; {a|of|DET; bi-riti#unu|between|PRP; {a|of|DET; er#seti|city quarter|N; bit|house|N;
Adad|i|DN; {a|of|DET; qe-#reb|interior|N; Uruk|i|SN
@date
3.      MU 28-KAM2 {m}si#-lu#-ku# u {n}at-te-i-ku#-su# LUGAL-MESZ ina hu-ud lib3#-bi#-szu2-nu
#lem: {an|at|year|N; n; X; u|and|CNJ; X; {arr#|king|N; ina|in|PRP; h#d|happiness|N; libbi#unu|heart|N
4.      KI a-ha-a-mesz a-di u4-nu s,a-a-tu2 i-zu-zu-'
#lem: itti|with|PRP; ah#ni#|one another|AV; adi|as far as|PRP; #u|day|N; {#tu|distant|AJ; izuz#|divide|V
# ruling
5.      E2 ki-szub-ba-a' KI(+t13) E2 {d}ISZKUR sza2 qe2-reb UNUG{ki#} 86 KUSZ3#
#lem: bit|house|N; ki#ubb#|waste field|N; er#seti|city quarter|N; bit|house|N; Adad|i|DN; {a|of|DET; qe-#reb|interior|N;
Uruk|i|SN; n; am#ati|unit|N
6.      US2# AN(+u2) {tu15}MAR.TU <<DA>> 85 KUSZ3 US2 a-na szu-u2-x# x#
#lem: {iddi|length|N; eli|upper|AJ; amurru|west|N; n; am#ati|unit|N; {iddi|length|N; ana|to|PRP; u; NENNI|PN
7.      kie-szube-ba-a' HA.LA sza2 {m}{d}60--SZESZ-MESZE-MU# DUMU# [sza2] {m#}{d#}60#-SZESZE-MU(+nu) u
{m}szib-qat2-{d#}60# DUMU# sza2 {m}ana#-GAL#--{d}60
#lem: ki#ubb#|waste field|N; zitti|share|N; {a|of|DET; Anu-ah#e-iddin|PN; m#ri|son|N; {a|of|DET; Anu-ah-ittannu|PN;
u|and|CNJ; Sibqat-Anu|PN; m#ri|son|N; {a|of|DET; Ana-rabut-Anu|PN
8.      85 KUSZ3 US2 KI#(+u2) {tu15#}KUR.RAW DA mesz-hat sza2-ni-tu4 HA.LA#
#lem: n; am#ati|unit|N; {iddi|length|N; {apl#|lower|AJ; {ad#|east|N; t#h|adjacent to|PRP; me#hat|measurement|N;
```

DH Faire 2015

Berkeley
UNIVERSITY OF CALIFORNIA

A Panel
Discussion on

Tuesday, APRIL 7, 2015

Digitally Supported Research and Pedagogy

Edmund **CAMPION**

CNMAT & Music

Andrew **GARRETT**

Linguistics

Alex **TARR**

Graduate Student, Geogra

Mila **OIVA**

Visiting Scholar, ISEEEES Dept.

9:30-11:00 AM 180 Doe Library

Hosted by the Library

Moderated by

Mary Elings, The Bancroft Library

A Panel
Discussion on

Wednesday, APRIL 8, 2015

The Landscape of Berkeley DH

Elizabeth **HONIG**

Art History

Laurie **PEARCE**

Near Eastern Studies

Francesco **SPAGNOLO**

The Magnes Collection of Jewish
Art and Life

3:10-6:00 PM Social Science Matrix,
8th Floor Barrows Hall (east entrance)

Moderated by

Cathryn Carson, History Department

**Poster presentation of current digital humanities
projects.** Please RSVP for the poster session and reception.



Keynote address by Professor

Zephyr Frank, Department of History,
Founder and Director of the Spatial History Project
at Stanford University

This event is co-sponsored by Computing and the Practice of History, the History Department, Digital Humanities at Berkeley (a collaboration between Research IT and the Dean of Arts and Humanities), Digital Humanities Fellows, the D-Lab, Social Science Matrix, the Townsend Center for the Humanities, and the Library.

DH Faire blurb
for April 8 2015

Historical Texts, Modern Tools: Berkeley Prosopography Services

Berkeley Prosopography Services (BPS) streamlines prosopographical research by offering researchers a customizable out-of-the-box tool-kit and environment to support the recreation of social, economic and intellectual environments preserved in texts. Prosopography, the identification of individuals in texts and the determination of their relationships to others, is foundational to understanding questions such as “In what contexts do ancient Babylonian scholars and business-people in extended families interact, and what can this tell us about the intersection of spheres of activity?”, or “What is the impact of the dissemination of Berkeley zoologist Joseph Grinnell’s ground-breaking models of annotation and research?” Conceived as a solution to real research problems, BPS provides humanities researchers a powerful digital environment that emulates familiar and comfortable processes of interacting with data, and presents opportunities to explore familiar data in new ways, and, in turn, develop new avenues of research.

The tool-kit includes a user-customizable **probabilistic disambiguator**, a program that determines the likelihood that two or more instances of the same name refer to the same person, a **Social Network Analysis engine**, that computes, utilizing well-established SNA metrics, the mathematical measures that define the social network, and a **graph visualizer** that automatically generates interactive visual representations of the social networks reflected in the data set. A collaboration between NES faculty, Professor Niek Veldhuis and Dr. Laurie Pearce, and Associate Director of Research IT, Patrick Schmitz, BPS demonstrates the potential and power of extensible, re-usable digital tools to extend humanities research.



Historical Texts, Modern Tools: Berkeley Prosopography Services

Digital Humanities Faire DH@Berkeley

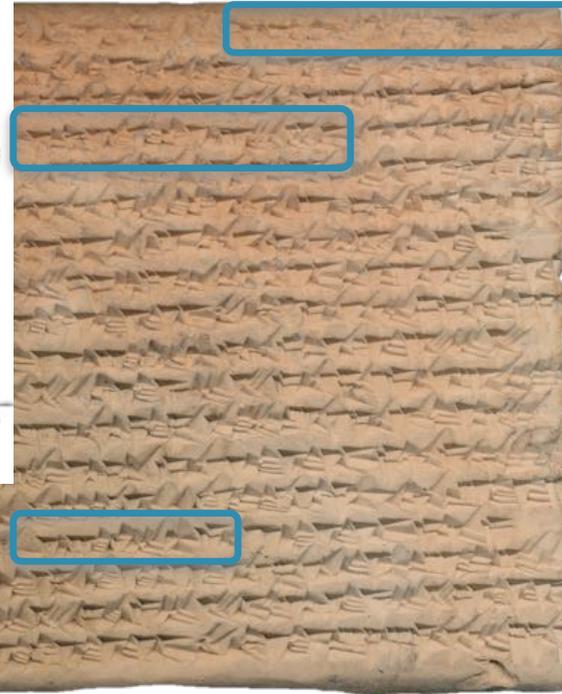
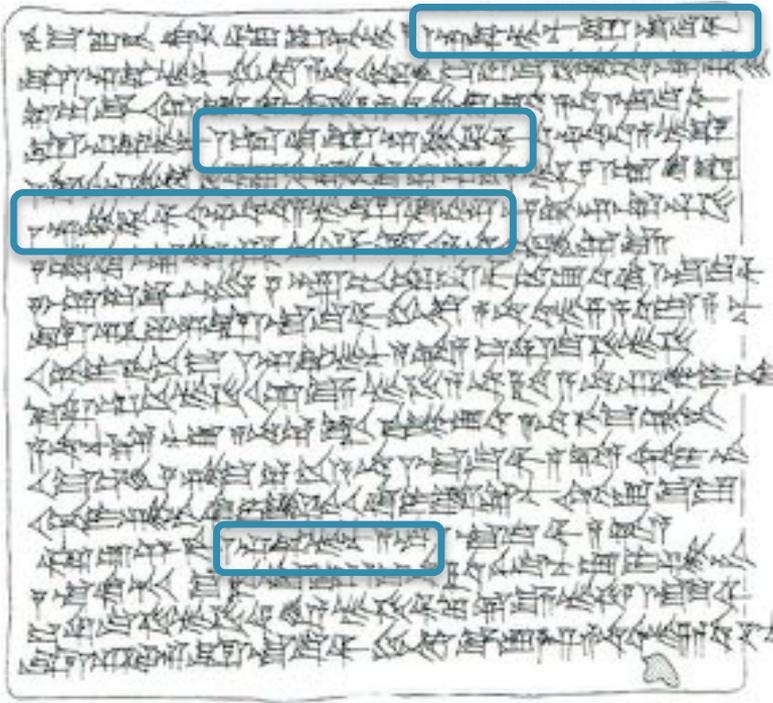
April 8, 2015

Laurie Pearce, Near Eastern Studies

Niek Veldhuis, Near Eastern Studies

Patrick Schmitz, Research IT

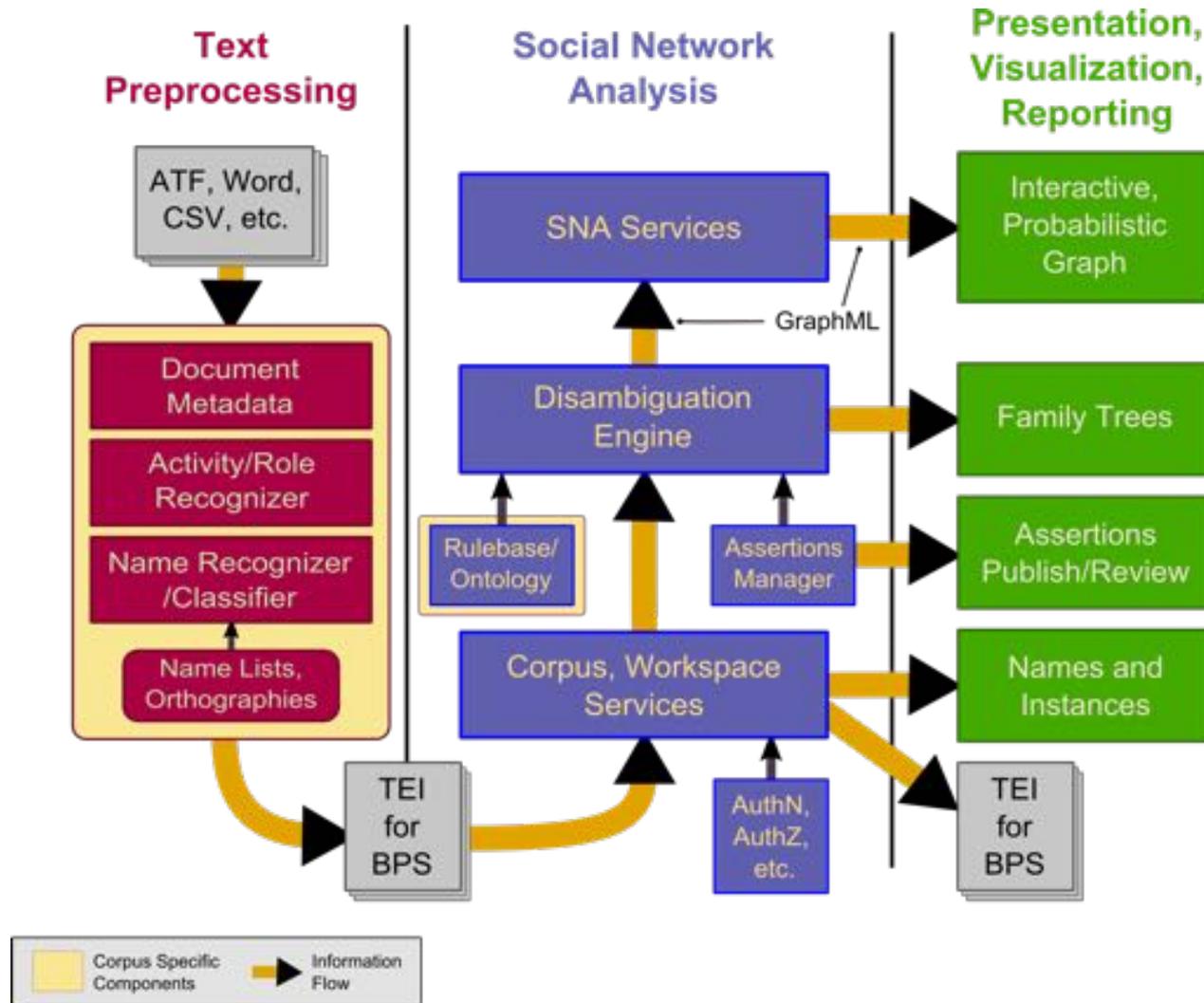
+ Prosopography



**Naming patterns:
Papponymy**

Anu-bēlšunu / Nidintu-Anu / Anu-bēlšunu // Ekur-zakir

+ BPS Architecture



+ Disambiguation engine

Berkeley Prosopography Services Logged in as Laurie | Help | Admin | Sign Out

Home Corpora Workspace

Documents People Clans Settings Admin

Set Model Parameters

Background on the model

The BPS analyzer will try to disambiguate among citations using the same name(s). To do this, it will basically model a new citation-person for each name it finds in a document (including fathers, grandfathers, etc. that are mentioned as qualifiers to the named actors). Then, it will attempt to collapse some of these citation-persons to get to the set of actual (real world) persons mentioned in all the corpus documents. Each citation-person is compared to other citation-persons, and a set of rules is applied to determine how likely it is that the two citations are the same person. The analyzer proceeds in two steps: first it considers all the citation-persons within each single document (intra-document), and then it considers the citation-persons across the entire corpus (inter-document).

When comparing two citation-persons, the analyzer will first require that there is no conflicting information about the two citation-persons - e.g., if they have different declared fathers, they will be considered as distinct, and will not be collapsed. The rules below allow you to configure whether specific roles must be considered to be distinct, and to control how strong the likelihood that two persons with partial matching name information are the same real world person.

General settings:

Number of qualifications (father/grandfather/ancestor/clan) in addition to forename required to consider a name citation "fully qualified"	<input type="text" value="2"/>
Assumed typical length of active business life (years)	<input type="text" value="25"/>
Assumed typical separation of generations (years)	<input type="text" value="15"/>

Step 1: Intra-document rules:

Rule Steps 1A, 1B, and 1C collapse citations within a single document.

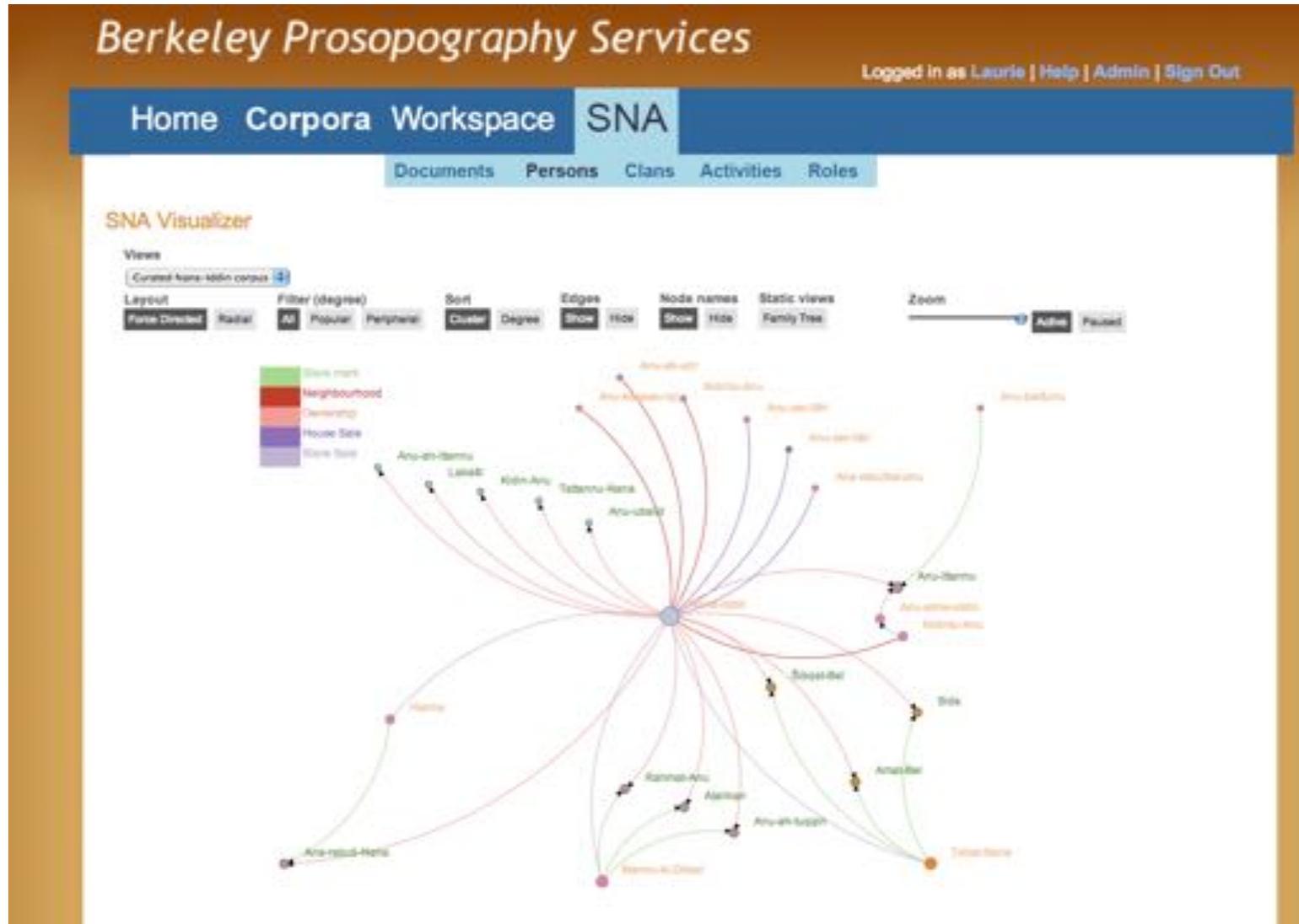
Step 1A: Consider equally qualified names

Collapse equal, fully qualified citations (e.g., "PN _a son-of PN _b , in-clan CN _c " and "PN _a son-of PN _b , in-clan CN _c ")	<input type="text" value="Always: 100%"/>
Collapse equal, partly qualified citations (e.g., "PN _a son-of PN _b " and "PN _a son-of PN _b ")	<input type="text" value="Conservative: 30%"/>
Collapse equal, unqualified citations (e.g., "PN _a " and "PN _a ")	<input type="text" value="Aggressive: 75%"/>

Step 1B: Consider compatible, but not equally qualified names

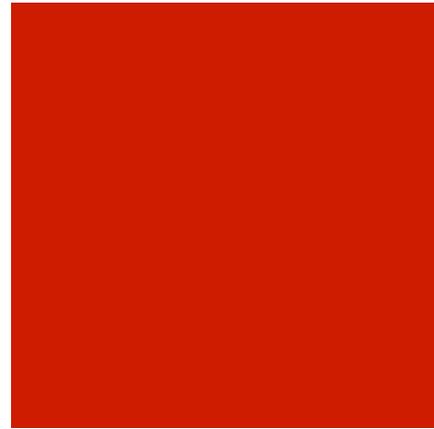
Collapse partly qualified citations with compatible, fully qualified citations (e.g., "PN _a son-of PN _b " and "PN _a son-of PN _b , in-clan CN _c ")	<input type="text" value="Conservative: 30%"/>
Collapse unqualified citations with compatible, more qualified citations (e.g., "PN _a " and "PN _a son-of PN _b , in-clan CN _c ", OR, "PN _a " and "PN _a son-of PN _b ")	<input type="text" value="Aggressive: 75%"/>

+ SNA: graph visualization





- “Talk the talk”



BPS & the DH Landscape @Berkeley



- “Talk the talk”
- Think like a data scientist

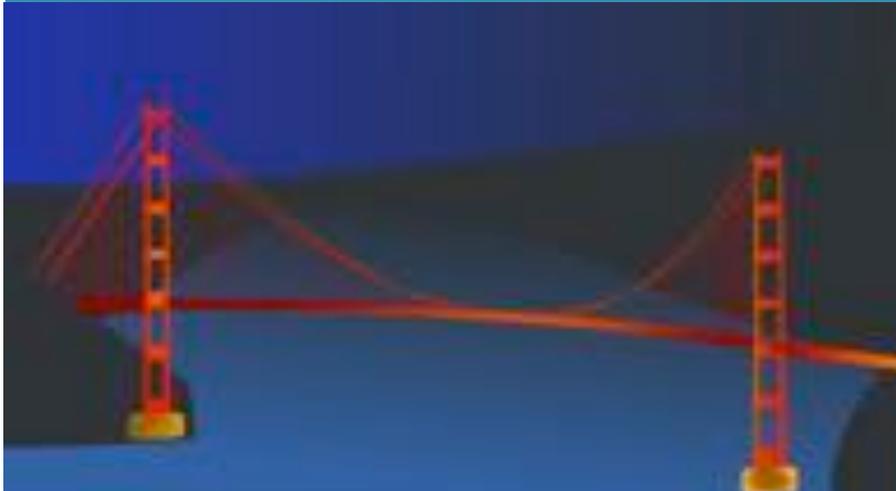
You don't have to
be a data scientist
to think like one



BPS & the DH Landscape @Berkeley



- “Talk the talk”
- Think like a data scientist
- Build bridges across domains and projects

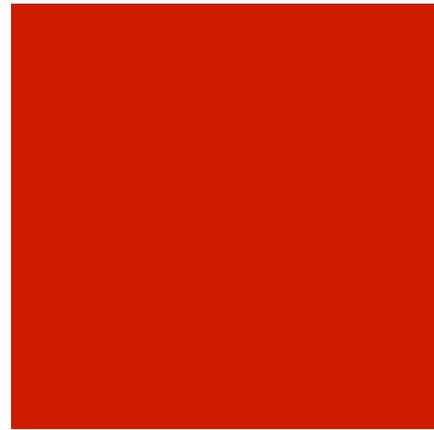


Standards for Networking Ancient Prosopographies

Data and Relations in
Greco-roman Names



A COLLABORATIVE EDITING PLATFORM FOR SOURCE DOCUMENTS IN CLASSICS



BPS & the DH Landscape @Berkeley





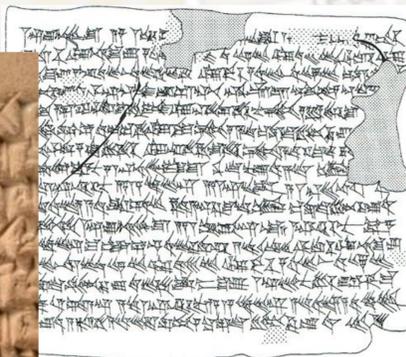
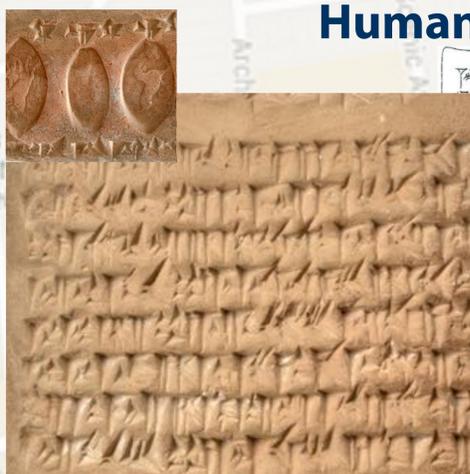
Historical Texts, Modern Tools: Berkeley Prosopography Services

berkeleyprosopography.org

Digital Context

- 3 components:
 - * text-preprocessing
 - * social network analysis
 - * visualizations
- service oriented architecture
- corpus-agnostic
- customizable, pluggable rules

Humanities Context



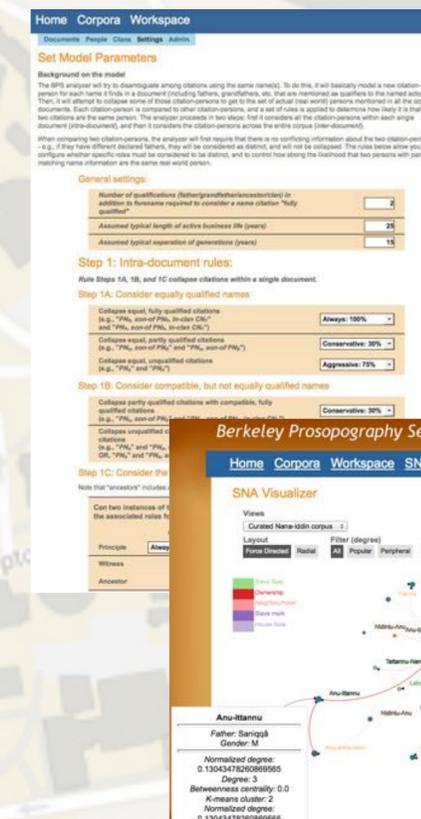
Hellenistic Babylonia: Texts, Images, Names
oracc.org/hbtin



Center for Tebtunis Papyri
tebtunis.berkeley.edu

BPS and DH Landscape

- adapt & transform familiar workflows
- support new research directions
- encourage asking new questions
- build new collaborations



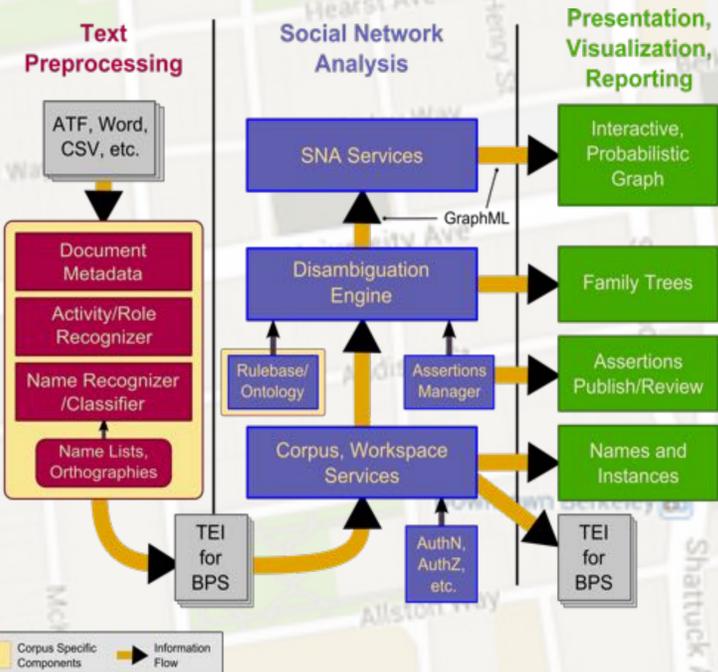
Archival texts are the foundation of social and economic history. In spite of differences in media and in times and places of composition, all archival texts preserve details that help identify individuals taking part in activities at specific times and places, and who act in relationships with persons who may or may not be family.

Prosopographers mine archival texts for:

- personal names
- family & social relationships
- dates and activities
- toponyms

Complex naming patterns, limited onomastic inventory, and damage to original sources challenge experienced and novice researchers' efforts to disambiguate namesakes: repeated, not always identical, names. Databases help with organizing data, but domain expertise is necessary to assess the value of each datum in what is often an imprecise process that leads to a declaration that 2+ name instances identify a single person: the expert uses criteria appropriate to the corpus — e.g., "how likely are name instances in texts separated by 45 years to refer to the same individual?" Recovery of new evidence, reassessment of old evidence, and differences of scholarly opinion complicate what is essentially a process of probabilistic reasoning.

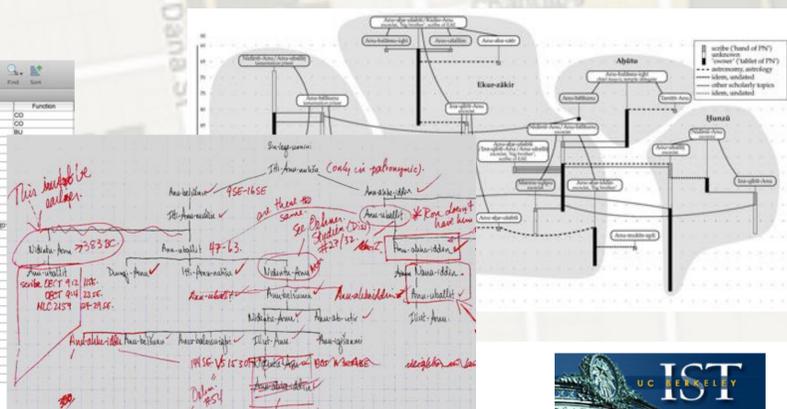
Researchers internalize the rules specific to disambiguating namesakes in their corpora; they become so proficient that the process seems intuitive. While they sense the degree to which variability in the parameters of inquiry may support investigation of "what if" scenarios in the data, they may be reluctant to explore alternative research directions because of the complexity of manual disambiguation. Subsequent researchers may avoid new interpretations because data and outcomes appear in traditional print publications, and would essentially require full re-working of the analytic process.



To serve the needs of humanities researchers, BPS:

- accepts TEI presentation of text or data-base content
- uses domain-expert defined disambiguation rules
- hides the math in computing SNA metrics
- tracks assertions of identifications
- produces GraphML for network visualizations
- creates interactive family trees
- generates interactive visualizations
- operates in individual workspaces
- encourages exploration and collaboration

Year	Text	Person Name	Primary Name	Function
1892 2 03
1892 2 03
1892 2 04
1892 2 05
1892 2 06
1892 2 07
1892 2 08
1892 2 09
1892 2 10
1892 2 11
1892 2 12
1892 2 13
1892 2 14
1892 2 15
1892 2 16
1892 2 17
1892 2 18
1892 2 19
1892 2 20
1892 2 21
1892 2 22
1892 2 23
1892 2 24
1892 2 25
1892 2 26
1892 2 27
1892 2 28
1892 2 29
1892 2 30



Partners and Collaborators

The BPS team:
Laurie Pearce, NES
Nick Veldhuis, NES
Patrick Schmitz, RIT

Standards for Networking Ancient Prosopographies
Data and Relations in Greco-Roman Names

PERSEIDS
A COLLABORATIVE EDITING PLATFORM FOR SOURCE DOCUMENTS IN CLASSICS

COPTIC SCRIPTORIUM

TRISMEGISTOS

UC BERKELEY SOCIAL SCIENCE MATRIX

Department of Near Eastern Studies

IST UC BERKELEY

FONDS FRANCE-BERKELEY FRANCE BERKELEY FUND

NATIONAL ENDOWMENT FOR THE HUMANITIES

Berkeley Digital Humanities

This post grows out of the roundtable session of the [DH Faire](#) held on April 8, 2015 at the University of California, Berkeley, that provided an opportunity for researchers of established and emerging digital humanities projects to share their work, to expose their ideas and methods to the Berkeley campus community. The charge to the roundtable participants was either to locate the project on the [digital humanities \(DH\) landscape at Berkeley](#), or to present our current research project(s).

Berkeley Prosopography Services ([BPS](#)) is a customizable, out-of-the-box toolkit and environment that supports prosopographical research. It is designed to solve a problem --- a complex research methodology --- and is not a problem or algorithm in search of a problem to which it may be applied. Prosopography is the process of discovering, through references to personal names, familial relationships, professional designations, as well as other attributes, pictures of the social, economic, intellectual activities and connections that link them. As the foundational task of prosopography is the collecting of name instances and attributes, it is hardly surprising that prosopographers were early adopters of digital tools, especially databases, which facilitated sorting, searching, and storage.

BPS is innovative as it was conceived as a reusable and generalizable toolkit, rather than as a “one-off” to serve a single, specific project or question. The [customizable probabilistic disambiguator](#), a program that determines the likelihood that two or more instances of the same name refer to the same person, processes a set of pluggable rules that each domain expert develops; the rules articulate the sequence of steps a domain expert uses to collapse multiple instances of names into the individuals that populate the corpus. Once individuals are distinguished out of multiple name instances, they are entered into the social network analysis engine that uses well-established SNA mathematical metrics that define the social network, and a graph visualizer automatically generates [interactive visual representations](#) of the social networks reflected in the data set.

In the construction of BPS, we have learned that a digital humanities project depends on more than a single conversation between humanities researcher and IT professional. On-going development of a shared vocabulary between the domain and technical components and a self-reflective ethnography leads to a strong foundation for confronting problems and problem-solving for both parties. Even this confirmed humanities researcher has learned to think more like a data scientist – not learning to code or engineer software – integrating process, abstraction, and the potential for shaping new directions into a research agenda.

Conceived as a solution to real research problems, BPS provides humanities researchers a powerful digital environment that emulates familiar and comfortable processes of interacting with data, and presents opportunities to explore familiar data in new ways, and, in turn, develop new avenues of research. This is consistent with the best in humanities research, an organic process that generates new and innovative avenues of investigation building on previous work. Well thought out digital scholarship can and should be comfortable; it should reflect and support research flows with which an investigator is familiar and should not limit or define the investigative process. Our lead researchers, Laurie Pearce, a lecturer in the Department of Near Eastern Studies, and Patrick Schmitz, Associate Director of

Research IT, welcome the support of the broadening Berkeley DH landscape that is providing new vistas for humanities research and for growing interactions between the humanities and social sciences.



Menu



Berkeley

Research Streams

Events

Initiatives

Services & Support

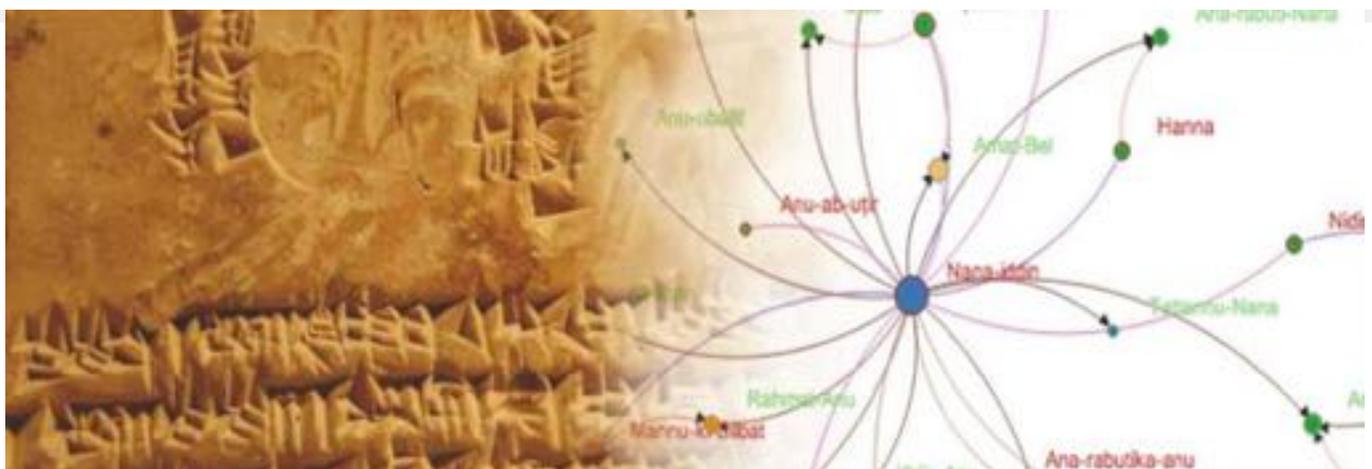
Affiliated Centers

About

<https://matrix.berkeley.edu/initiatives/matrix-seminars/prosopography-toward-toolkit>

Home > Initiatives > Initiatives

Prosopography: Toward a Toolkit



About

In Fall 2015, Social Science Matrix will be launching a year-long research seminar focused on “Developing Tools and Collaborations in Prosopographical and Historical Social Network Research Environments”.

While its name might be a mouthful, this seminar—a continuation of a [2014 Matrix prospecting seminar](#)—has a goal to develop “research toolkit” that will help faculty, staff, and post-doctoral researchers use sophisticated data modeling tools for the purposes of *prosopography*, the art of researching relationships recorded in historical documentation and using data to identify social patterns and relationships.

Based upon probabilistic models, prosopography can help determine the likelihood that two or more instances of the same name refer to the same person, and can graph social networks based upon a data set. Prosopography has long been an important tool for the study of all kinds of past societies; the word stems from the Greek *prosopoeia*, or “face created,” suggesting how this methodology enables researchers to “put a face on” individuals about whom little is known, based on their connections with other people.

Led by Patrick Schmitz, Associate Director for Research IT and Strategy at UC Berkeley, and Laurie Pearce, Lecturer in Near Eastern Studies, [Berkeley Prosopography Services](#) (BPS) helps research institutions both on campus and beyond with a customizable, out-of-the-box toolkit for “reconstruction of social contexts,” modeling the relationships between people based on available data.

“Prosopography is the practice of identifying individuals mentioned in texts and setting them in their social contexts: families, social/professional groups, etc.,” the seminar’s organizers explain in their proposal. “Prosopography is encountered in many humanities research agendas where a fundamental task is the extraction and identification of persons from records across all areas of human endeavor.”

The seminar will explore common approaches that can be employed across domains to identify reusable components of the BPS toolkit, and elicit user-driven needs for future expansion of BPS. Among the participants are the Department of Near Eastern Studies and the Social Networks and Archival Context Project, based at the UC Berkeley School of Information, as well as external partners, such as the Perseus Project, at Tufts University; the Coalition for Networked Information; the University of Pacific's Coptic Scriptorium; and the UC Berkeley Center for Tebtunis Papyrii.

Image Credit: [Berkeley Prosopography Services](#)

Contact(s)

Patrick Schmitz

Associate Director, IST-Research and Content Technologies

pschmitz@berkeley.edu

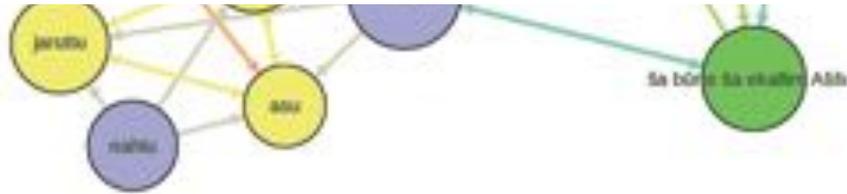
Laurie Pearce

Lecturer

Near Eastern Studies

lpearce@berkeley.edu

Introduction”,
taught by Scott
Weingart, Digital
Humanities



Specialist at Carnegie Mellon University. Though Escobar received a solid foundation in network analysis and theory, he advises other graduate students to follow up and continue taking workshops and training during the semester. “You start rethinking the basics [as you begin applying tools to your own project],” Escobar shared.

Throughout the summer, Escobar collaborated with his advisor, DH Fellow Laurie Pearce, Lecturer in Near Eastern Studies and co-director of BPS, to combine their research interests in one network graph of intellectuals in Hellenistic Uruk (ancient Iraq, 4th-3rd centuries BCE). Whereas Pearce investigates these individuals in legal contexts, Escobar is interested in their work as astronomers. Though prosopographical work is performed in a variety of fields, the problem of constructing biographies is particularly prominent when studying ancient societies, due to the relative scarcity of available texts. While these biographies may be incomplete, they are enriched by understanding the contexts in which these figures move about: what kinds of texts did they write? In what records do they appear? Who writes about them? Who were they holding transactions or contracts with? Escobar’s research on Babylonian astronomers is enhanced by understanding how they also participated in society as legal scholars, authors of literary works, and owners of land. By combining legacy data sets that often separate social and economic history from the intellectual history of the sciences, Pearce and Eduardo hope to create a more complete picture of elite families in ancient Uruk.

At the inaugural [Digital Humanities at Berkeley Summer Institute](#) (DHBSI), Escobar and Pearce received fellowships to investigate these challenges as a team. At “Data Workflows and Network Analysis”, taught by Chris Church (Assistant Professor of History at the University of Nevada, Reno and one of the founders of UC Berkeley’s [DH Working Group](#)), Escobar and Pearce explored OpenRefine, a tool for pre-processing and cleaning data, and Gephi, an open source tool for network analysis. After working with an example network graph of attendees’ research interests, Escobar began investigating semantic networks in cuneiform recipe texts.

Escobar explained that many Akkadian words appearing in technical texts remain undefined by scholars. Escobar's corpus consists of cuneiform recipes composed between the second and first millennium BCE. While some recipes describe technical processes for activities like making colored glasses that imitate precious stones, manufacturing perfumed oils, and dyeing wool in exotic colors, they may be indistinguishable from recipes containing magical or imaginary ingredients. To better understand these undefined words and how they function, Escobar plans to visualize a semantic network in Gephi. Escobar will investigate these unknown words by graphing their common substitutions, as well as verbs, nouns, and modifiers applied to them. Escobar plans to present his work at the [Digital Humanities Faire](#) poster session on Wednesday, April 13th.

MORE RESOURCES:

- [Network Analysis Resources](#)
- [Berkeley Prosopography Services](#)
- Training
 - [DHSI at the University of Victoria](#) (June)
 - [The Digital Humanities at Berkeley Summer Institute](#) (August)

TAGS BERKELEY PROSOPOGRAPHY SERVICES NEAR EASTERN STUDIES

NETWORK ANALYSIS DHSI (DIGITAL HUMANITIES SUMMER INSTITUTE)

DIGITAL HUMANITIES AT BERKELEY SUMMER INSTITUTE (DHBSI)

All contents © Digital Humanities at Berkeley unless otherwise specified.

Site designed by Agile Humanities in association with Intelligent Machines.



DIGITAL HUMANITIES AT BERKELEY

Berkeley Prosopography Services and the Tebtunis Papyri



[Berkeley Prosopography Services](#), a tool for historical social network analysis originally developed for use with an Uruk text corpus from the Hellenistic period (331-46 BCE), is currently in the process of being generalized for use with a variety of corpora. Project co-directors and DH Fellows Laurie Pearce, Lecturer in Near Eastern Studies, and Niek Veldhuis, Professor in Near Eastern Studies, and technical lead, Patrick Schmitz, Associate

Director of Research IT are taking a variety of approaches to developing a robust, reusable tool. With the support of an NEH Digital Humanities Implementation Grant and a [Social Science Matrix research seminar grant](#), the team has gathered scholars and software developers from a variety of relevant projects to discuss issues like disambiguation rules, chronology, and software integration. Participants have been drawn from groups such as the [Center for the Tebtunis Papyri](#), [the SNAC Project](#), [the Perseids Project](#), [Trismegistos](#), [the Coptic Scriptorium](#), and [the Old Assyrian Social Network 1950-1750 BCE](#).

In 2015, Micaela Langellotti, a Postdoctoral Researcher at the [Center for the Tebtunis Papyri](#), worked with Pearce to prepare the first test corpus for the new Berkeley Prosopography Services, with support from a collaborative research grant from Digital Humanities at Berkeley. Langellotti's corpus contains 16 months of records from the notary office of the Egyptian village of Tebtunis in years 445 and 446 AD. The ledgers include daily records of contracting parties and the nature of their transactions, and the corpus provides a rich snapshot of socioeconomic life in Roman Egypt. During her work at King's College London, Langellotti compiled the records and built a database for organizing and accessing them.

As Langellotti worked with Pearce, she realized that using the data with BPS would require a reformulation of her database. Databases are not neutral containers of information, Langellotti emphasized. "As I worked with Laurie, I worked to understand the aim of my own database, define its attributes, and determine which attributes were useful for disambiguation. You need to ask, 'What question does your database intend to answer?'" Though Langellotti had gathered large amounts of data on her topic, her database tables were not well suited for network analysis. When Langellotti initially gathered the data, she was interested in an overview of socioeconomic activities in Tebtunis. As Langellotti's research interests shifted towards the thousands of people listed in her corpus and their various interconnections, she ran into several problems. Though Langellotti's original database helped her ask questions about contracts, it made it difficult to trace persons, both as they appeared as contracting parties and other mentions of them as, for instance, neighbors or guarantors. While she was able to facet her database by contract type or date, she could not create a facet for all buyers of

houses or all transactions by members of a particular family.

Through a process of consultation with Pearce and attending several Matrix prosopography seminars, Langellotti reformatted her database around a new vocabulary of roles. Instead of focusing on individual contracts, this database represents persons by their roles as buyers, sellers, lessor, lessees, neighbors, etc. Langellotti pointed out that though these roles are straightforward to define, there are tough decision to make as they are applied to the reality of a complex society. Langellotti pointed to slavery as a prime example of this problem. How can slaves be represented in a way that captures both their status as objects and their role as people who are connected with various households? In her original contract-centered database, slaves only appeared in the text of contracts. Because they were objects and not contracting parties, slaves could not be studied as persons who moved between households. In this new person-centered formulation of the database, slaves are represented as persons with attributes such as, “role: slave” and a transaction type (“slave of”) that connects them to another person in the database.

Langellotti advises researchers to take a proactive role in interrogating their database and considering its architecture. Though databases can contain vast amounts of information, the ability to use and access that data can be greatly constricted by a mismatch between architecture and the requirements of a scholar’s research questions and analysis tools.

Interested in building a database for your research? [Get in touch with a digital humanities consultant to discuss database design](#) or see [other consulting services offered by Digital Humanities at Berkeley](#).

Image: Langellotti participates in a seminar hosted by Berkeley Prosopography Services and Social Science Matrix

TAGS BERKELEY PROSOPOGRAPHY SERVICES CENTER FOR THE TEBTUNIS PAPYRI
NETWORK ANALYSIS DATABASES NEAR EASTERN STUDIES
SOCIAL SCIENCE MATRIX


 Search

Research IT (RIT) provides research computing technologies, consulting and community for the Berkeley campus. Our goal is to advance research through IT innovation.

PROGRAMS	SERVICES	PARTNERSHIPS	PROJECTS	NEWS	ABOUT
----------	----------	--------------	----------	------	-------

Social Science Matrix sponsors Research Seminar on Berkeley Prosopography Services

Submitted by [Patrick Schmitz](#) on February 16, 2016

The [Social Science Matrix](#) is sponsoring a year-long research seminar focused on helping researchers with the reconstruction of social contexts, modeling the relationships between people based upon available data and probabilistic models. The seminar, “Developing Tools and Collaborations in Prosopographical and Historical Social Network Research Environments” is co-organized by Laurie Pearce (Near Eastern Studies) and Patrick Schmitz (Research IT), who lead [Berkeley Prosopography Services](#) (BPS).

Prosopography is a research method focused on the analysis of data, primarily personal names preserved in historical documentation, to identify relationships and patterns of social interaction. The word prosopography derives from the Greek prosopoeia, or "face created," reflecting that this methodology helps “put a face on” individuals about whom little, beyond a few basic facts, is known. It supports reconstruction of the networks of interaction and activity in which people lived and worked. Prosopography is an important tool for the study of all kinds of past societies. The BPS toolkit facilitates management and analysis of prosopographic data.

This research seminar — a continuation of a [2014 Matrix prospecting seminar](#) — is leveraging the input of on- and off-campus partners to expand the toolkit for application to a broader set of domains, and to integrate BPS with other tools used in the analysis of text corpora.



Participants in the seminar include researchers from the [Department of Near Eastern Studies](#), the [Social Networks and Archival Context Project](#) (with partners at the [School of Information](#)), and [The Center for the Tebtunis Papyri](#), as well as external partners, such as the [Perseids Project at Tufts University](#), the [Coalition for Networked Information](#), and the [University of Pacific's Coptic Scriptorium](#).

Goals of the research seminar include:

- To explore common approaches for prosopography-based research that can be employed across domains.
- To identify reusable components of the Berkeley Prosopography Services' toolkit, and elicit user-driven needs for future expansion of BPS.

For more information, visit the [BPS Website](#), or contact project leads [Laurie Pearce](#) or [Patrick Schmitz](#).

Tags: [Berkeley Prosopography Services](#) [Digital Humanities](#)

Program: [Digital Humanities](#)

Project: [Berkeley Prosopography Services](#)

Partnership: [Arts & Humanities Division](#)

Technology @ Berkeley

Email us: research-it@berkeley.edu



[@research_it_ucb](#)



Research IT (RIT) provides research computing technologies, consulting and community for the Berkeley campus. Our goal is to advance research through IT innovation.

PROGRAMS	SERVICES	PARTNERSHIPS	PROJECTS	NEWS	ABOUT
----------	----------	--------------	----------	------	-------



Semantic Network Analysis and Cuneiform Intellectual History

Submitted by [Quinn Dombrowski](#) on April 22, 2016

For Eduardo Escobar, a PhD student in the Near Eastern Studies department, technology provides both a key theoretical concept and a set of practical tools for his dissertation on cuneiform “recipes” from between the second and first millennium BCE. These “recipes” are documents that share a predictable structure for transmitting procedural knowledge from an expert to a novice. To support and present his research claims and related findings about his corpus, Escobar leverages technology as it is understood in a modern context, through his adoption of tools and methods associated with digital humanities.

As one component of his dissertation, Escobar will develop a digital critical edition for a subset of his corpus, and publish it through the [ORACC \(Open Richly Annotated Cuneiform Corpus\) consortium](#), which has developed widely-adopted standards for the online publication of cuneiform texts. The critical edition will include network analysis tools to decipher probable meanings of technical terms that appear in recipe texts. These terms have remained opaque even in comprehensive reference works such as The Chicago Assyrian Dictionary. The interface Escobar is developing will allow scholars to visualize recipes within a semantic network of related texts, examine the various contexts where a particular term appears, and use the intersection of those contexts to better ascertain the meaning of poorly understood terms.

The Digital Humanities at Berkeley program, a partnership between the Dean of Arts and Humanities and Research IT, has supported Escobar in developing the methodological and technical skills necessary to implement the network analysis aspect of his dissertation. In Fall 2014, Escobar met with digital humanities consultants to discuss which course at the Digital Humanities Summer Institute (DHSI) at the University of Victoria would be the best fit for his research interests. Escobar was awarded a tuition scholarship for DHSI, and in June 2015 he joined colleagues from around the world in the course “Data, Math, Visualization, and Interpretation of Networks: An



Introduction”. In August 2015, Escobar deepened his knowledge of network analysis by attending the course “Data Workflows and Network Analysis” at the inaugural Digital Humanities at Berkeley Summer Institute (DHBSI).

While the application of network analysis in Escobar’s dissertation is focused on semantic networks of technical terms, his broader research interests have allowed him to explore the method’s use in other contexts as well. In Spring 2015, Escobar received a small research grant to work with Laurie Pearce, a lecturer in Near Eastern Studies, on a project that explores the intersection of their individual research agendas. This project examines how and where the astronomer authors of Escobar’s recipes appear within the historical social networks Pearce has developed through her [Berkeley Prosopography* Services](#) project. Though prosopographical work is performed in a variety of fields, the problem of constructing biographies is particularly prominent when studying ancient societies, due to the relative scarcity of available texts. While these biographies may be incomplete, they are enriched by understanding the contexts in which these figures move about: what kinds of texts did they write? In what records do they appear? Who writes about them? Who were they holding transactions or contracts with? Escobar’s research on Babylonian astronomers is enhanced by understanding how they also participated in society as legal scholars, authors of literary works, and owners of land.

The transformative impact that digital tools and methodologies have had on Escobar’s research has inspired him to develop new tools that will support both research and teaching in his discipline. In December 2015, he received a grant from the Student Technology Fund to build a script that will scan cuneiform texts for the most frequently used signs, outputting a statistical dataset of those signs most used by scribes. This tool will have significant applications for cuneiform pedagogy and research. “Students learning cuneiform for the first time are often intimidated by the vast range of choices they face when transliterating a text from the original script. Advanced scholars of cuneiform, in contrast, have often relied on analog tools for analyzing text data, thus, lacking the time to analyze substantial text corpora,” explained Escobar. “In both cases, a ‘plug-and-play’ tool that decodes cuneiform signs to their base values will prove itself invaluable to both novices and experts on campus and off.”

** A prosopography is the result of (a) the collecting of name instances in texts, along with related data, such as family relationships, and attributes, such as titles, roles, age, and gender; (b) the disambiguating those name instances into individuals; and (c) the presentation of the results in a structured, searchable format. The creation of a prosopography is not an end in itself, but rather a methodology for collecting and organizing data that supports complex and interesting research questions.*

Tags: [Digital Humanities](#) [Berkeley Prosopography Services](#) [prosopography](#) [DHBSI](#) [Researcher Profile](#)

Program: [Digital Humanities](#)

Partnership: [Arts & Humanities Division](#)

Technology @ Berkeley

Email us: research-it@berkeley.edu



[@research_it_ucb](#)

Berkeley Prosopography Services ([BPS](#)) successfully demonstrated the recently implemented end-to-end integration of a toolkit for prosopographical research last week. In a two-day workshop organized as part of a Social Science Matrix-sponsored Research Seminar for AY 2015-2016, the BPS team and research partners from across the United States had the opportunity to view and test drive digital tools that facilitate the disambiguation of namesakes in text corpora and visualization of associated social networks, regardless of language, script, chronological framework, or corpus contents.

BPS is an open-source prosopographical toolkit that leverages heuristics and workflows familiar to humanities researchers, computes social network metrics, and generates interactive visualizations of the biological and social connections that link documented individuals in text corpora. The tools provide a dynamic means of researching historical communities documented in legal, administrative, and literary texts and archives.

Cuneiformists, papyrologists, and classicists shared data sets and explored the outcomes as the results of the disambiguation process were presented in a dynamic SNA graph visualization. BPS particularly acknowledges the support of the [Center for the Tebtunis Papyri](#), its director Todd Hickey, graduate student Caroline Cheung, and recent post-doc Micaela Langellotti. The accompanying image of an SNA visualization depicts the networks that BPS detected in a papyrus register that is a component of Langellotti's study of the Grapheion archive. Partners from Tufts University and the [Perseids Project](#) explored potential integration with the SNA engine as a complement to the pedagogical workflow and goals of the Journey of the Hero project.

The participants' exploration of the toolkit, which is still under development, led to the identification of features necessary and desirable for future development. Attendees agreed that BPS has the potential to support innovative research agendas.

Patrick Schmitz, Associate Director of Research IT and BPS technical lead said, "It was very satisfying to see this community of scholars using the tools we have been working on. The discussions and ideas for additional functionality were really exciting, and a real validation of our ongoing collaboration to build better tools for humanists."

Adam Anderson, BPS research partner and incoming DH Postdoctoral Fellow said,

“Participating with the BPS project has been a transformative experience, and proof that building networks is the key to success.”

BPS will have a tutorial and demonstration corpus on its website in the near future. The team and research partners are committed to ongoing conversation, sharing of data, both as a scholarly best practice and as a means of providing a range of evidence against which to test implemented and planned features.

A collaboration between the Berkeley's [Research IT](#) group and the [Department of Near Eastern Studies](#), the BPS team consists of:

Prof. Niek Veldhuis, Near Eastern Studies, UC Berkeley, PI; Dr. Laurie Pearce, Near Eastern Studies, UC Berkeley, co-PI and project manager; Patrick Schmitz, Associate Director, Research IT, technical lead; Davide Semenzin, core-developer; r. Terri-Lynn Tanaka, Near Eastern Studies, UC Berkeley, communication and project analyst; Caroline Cheung, Ancient History and Mediterranean Archaeology, UC Berkeley, GSR.

In addition to the support of the Social Science Matrix Research Seminar, BPS has received funding from an NEH Digital Humanities Implementation grant and a [DH@Berkeley](#) Mellon grant for Capacity Building.

TAGS BPS

All contents © Digital Humanities at Berkeley unless otherwise specified.

Site designed by [Agile Humanities](#) in association with [Intelligent Machines](#).


 Search

Research IT (RIT) provides research computing technologies, consulting and community for the Berkeley campus. Our goal is to advance research through IT innovation.

PROGRAMS	SERVICES	PARTNERSHIPS	PROJECTS	NEWS	ABOUT
----------	----------	--------------	----------	------	-------



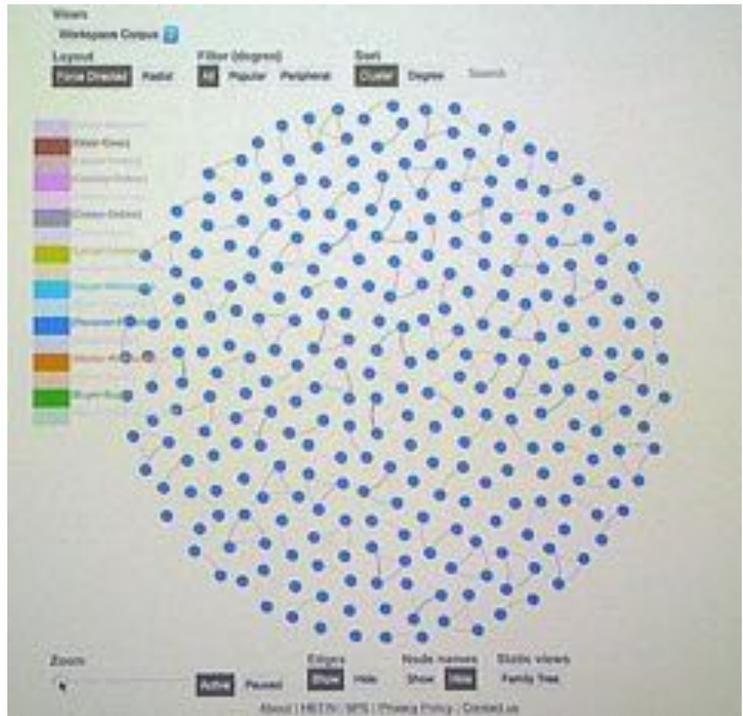
Berkeley Prosopography Services Successfully Demonstrates New Toolkit

Submitted by [Quinn Dombrowski](#) on May 11, 2016

Berkeley Prosopography Services (BPS) successfully demonstrated the recently implemented end-to-end integration of a toolkit for prosopographical research last week. In a two-day workshop organized as part of a Social Science Matrix-sponsored Research Seminar for AY 2015-2016, the BPS team and research partners from across the United States had the opportunity to view and test drive digital tools that facilitate the disambiguation of namesakes in text corpora and visualization of associated social networks, regardless of language, script, chronological framework, or corpus contents.

BPS is an open-source prosopographical toolkit that leverages heuristics and workflows familiar to humanities researchers, computes social network metrics, and generates interactive visualizations of the biological and social connections that link documented individuals in text corpora. The tools provide a dynamic means of researching historical communities documented in legal, administrative, and literary texts and archives.

Cuneiformists, papyrologists, and classicists shared data sets and explored the outcomes as the results of the disambiguation process were presented in a dynamic SNA graph visualization. BPS particularly acknowledges the support of the [Center for the Tebtunis Papyri](#), its director Todd Hickey, graduate student Caroline Cheung, and recent post-doc Micaela Langellotti. The accompanying image of an SNA visualization depicts the networks that BPS detected in a papyrus register that is a component of Langellotti's study of the Grapheion archive. Partners from Tufts University and the [Perseids Project](#) explored potential integration with the SNA engine as a complement to the pedagogical workflow and goals of the Journey of the Hero project.



[Read more on the Digital Humanities at Berkeley blog.](#)

Tags: [Social Sciences Matrix](#) [Berkeley Prosopography Services](#)

Program: [Digital Humanities](#)

Project: [Berkeley Prosopography Services](#)

Partnership: [Arts & Humanities Division](#)

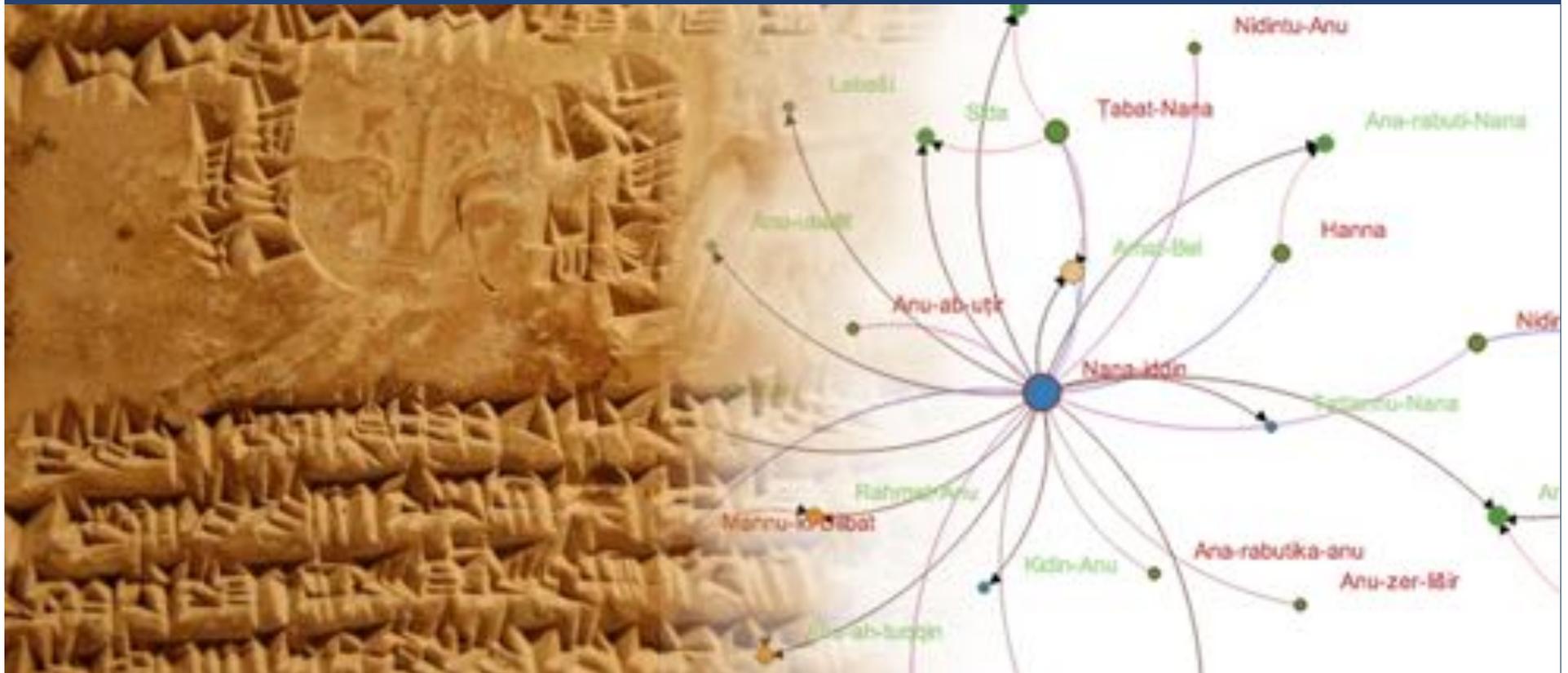
Technology @ Berkeley

Email us: research-it@berkeley.edu

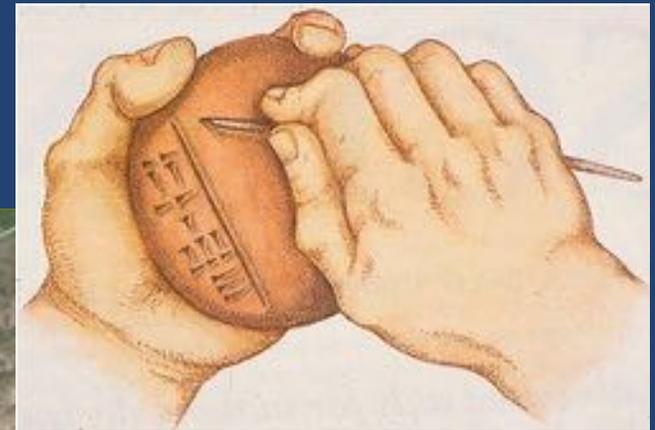
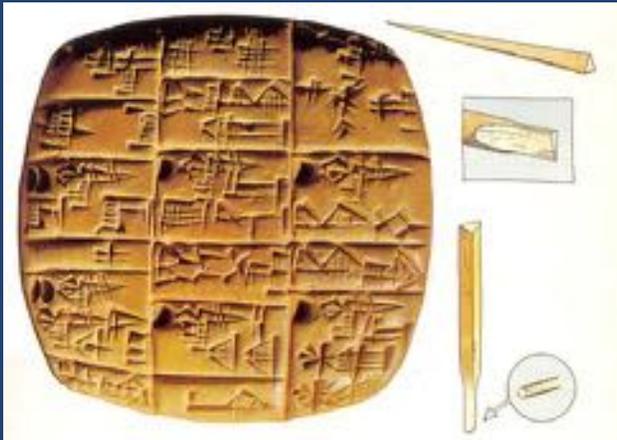


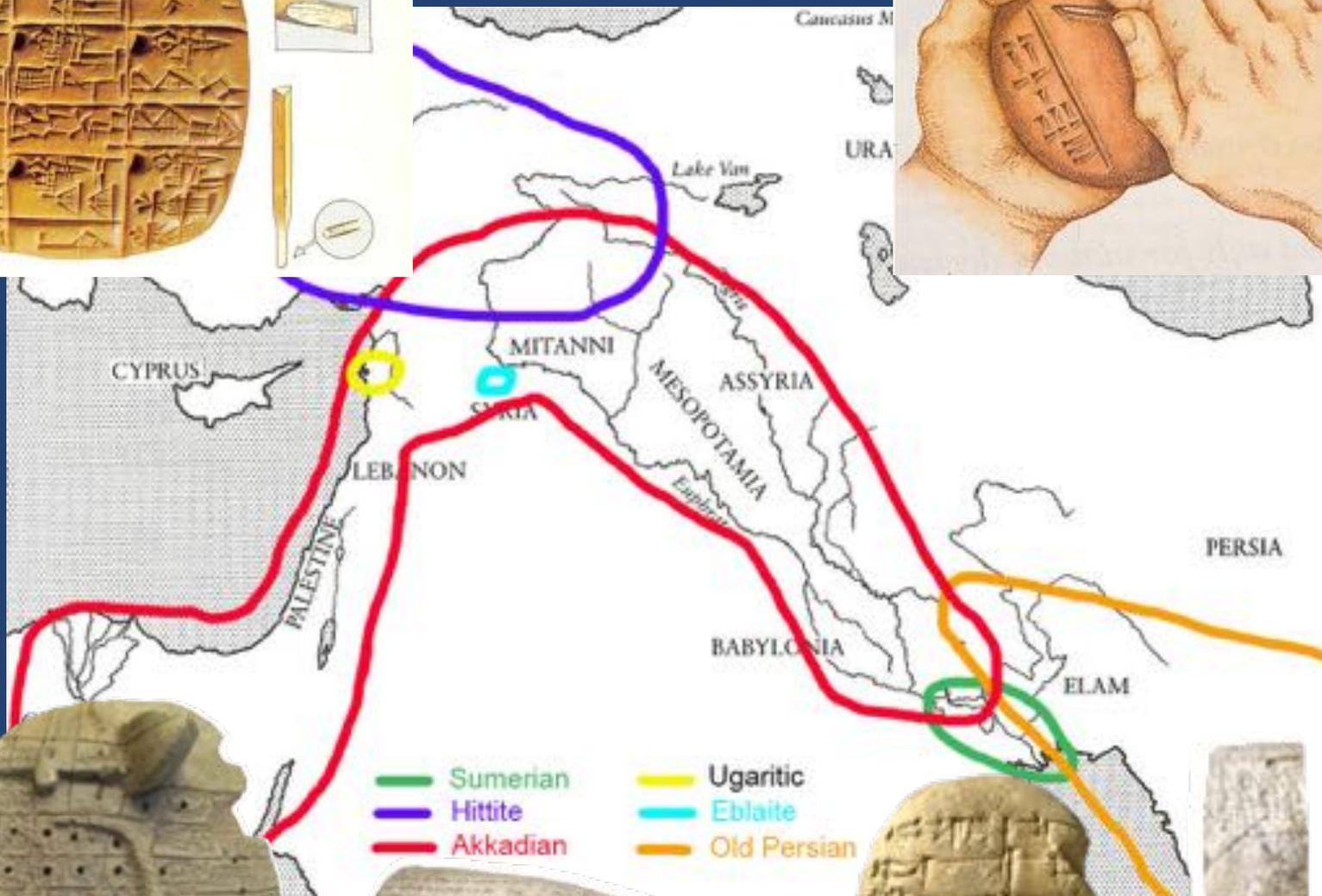
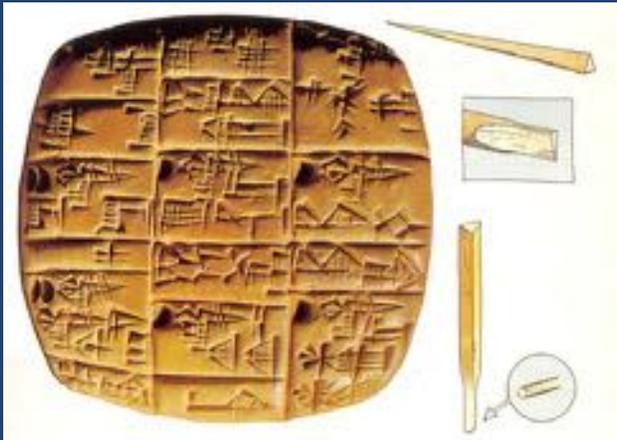
[@research_it_ucb](#)

Social Networks and Ancient Texts



A Presentation to Dennis Feehan's
Data Science Connector Class
21 November 2016
Laurie Pearce, Eduardo Escobar
Department of Near Eastern Studies
University of California, Berkeley





meaning of archaic sign	Uruk IV ca. 3400	Uruk III ca. 3200	ED III ca. 2400	Ur III ca. 2000	Old Assyrian ca. 1900	Old Babylonian ca. 1700
SAG "head"						
NINDA "ration"						
GU ₁ "disbursement"						
AB ₂ "cow"						
APIN "plow"						
KI "locality"						

At around 1500, scribes began to write and read cuneiform at an orientation rotated 90° counterclockwise from the norm

Middle Assyrian ca. 1200	Neo-Babylonian ca. 600

Paleography of cuneiform from ca. 3500 – 500 BC

Assyriology in the Digital Age

- digital catalog
cdli.ucla.edu

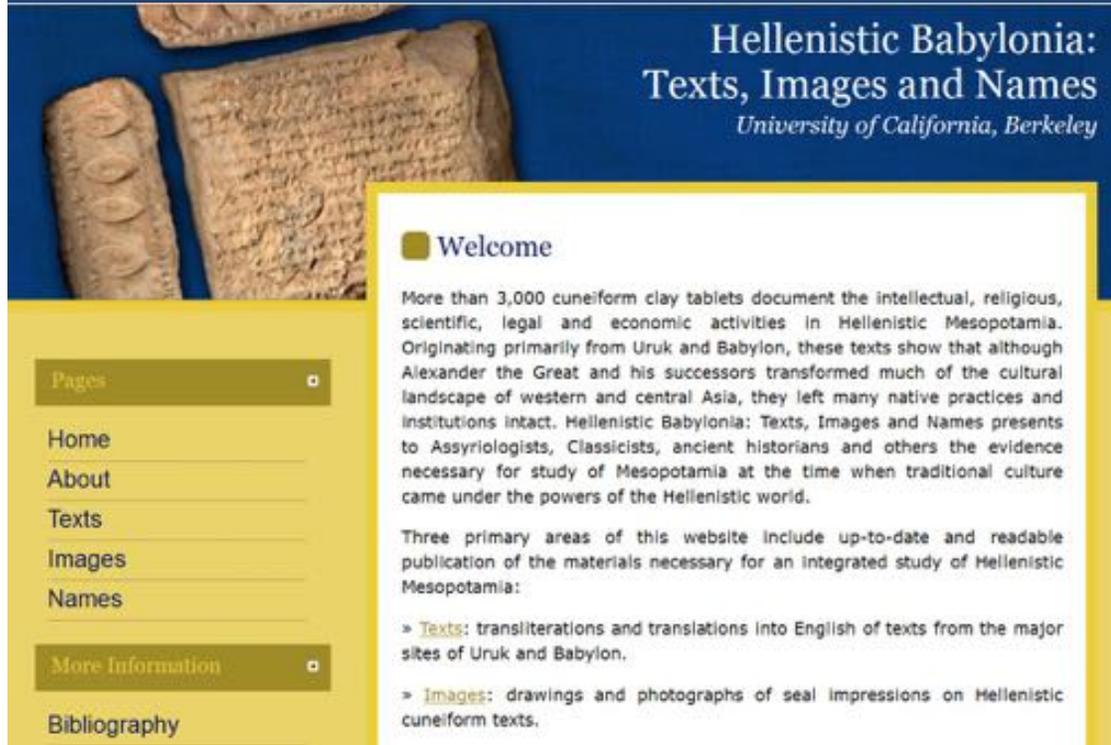


- digital publication
 - exportable format
 - shared standards
 - open access

oracc.org



Project Content: Hellenistic Uruk



**Hellenistic Babylonia:
Texts, Images and Names**
University of California, Berkeley

Welcome

More than 3,000 cuneiform clay tablets document the intellectual, religious, scientific, legal and economic activities in Hellenistic Mesopotamia. Originating primarily from Uruk and Babylon, these texts show that although Alexander the Great and his successors transformed much of the cultural landscape of western and central Asia, they left many native practices and institutions intact. Hellenistic Babylonia: Texts, Images and Names presents to Assyriologists, Classicists, ancient historians and others the evidence necessary for study of Mesopotamia at the time when traditional culture came under the powers of the Hellenistic world.

Three primary areas of this website include up-to-date and readable publication of the materials necessary for an integrated study of Hellenistic Mesopotamia:

- » **Texts:** transliterations and translations into English of texts from the major sites of Uruk and Babylon.
- » **Images:** drawings and photographs of seal impressions on Hellenistic cuneiform texts.

Pages

- Home
- About
- Texts
- Images
- Names

More Information

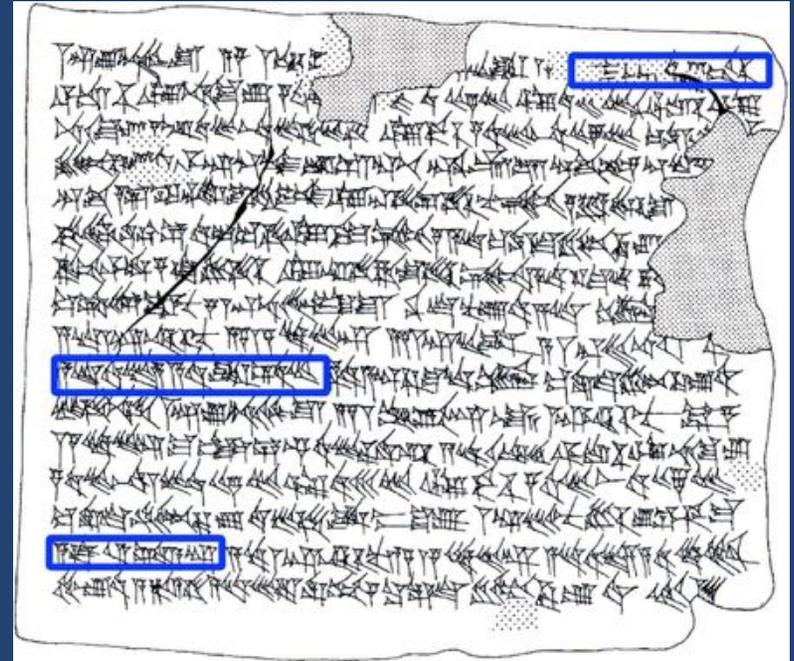
- Bibliography



~500 legal texts
8-20 name citations/text
3 individuals/citation
10,000 name instances

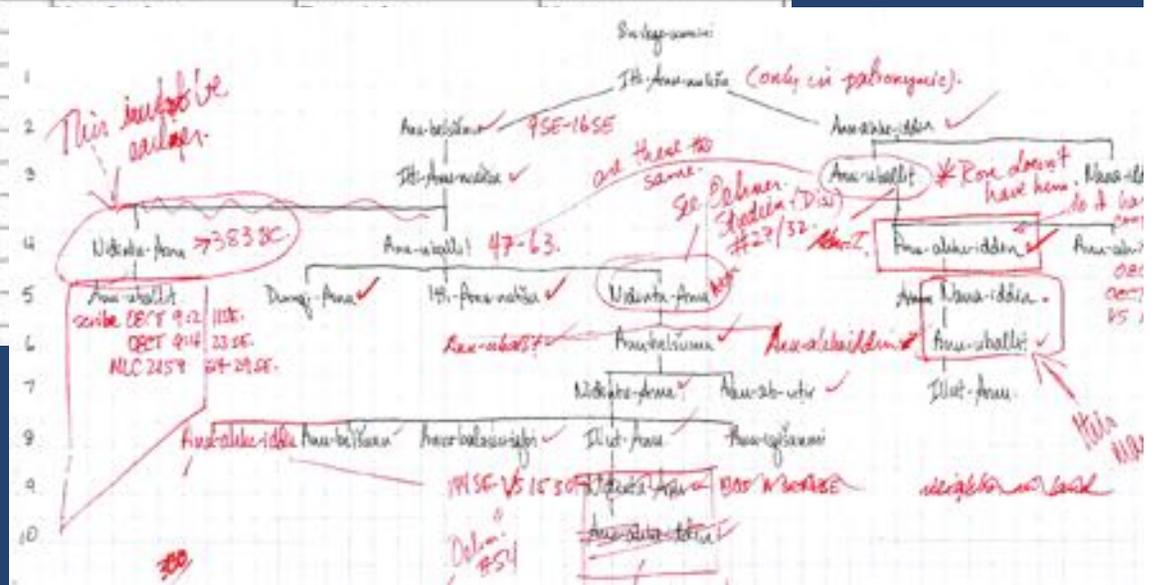
Data Mining in Uruk Legal Texts

- Boilerplate text
 - repetitive patterns
 - attributes
 - many names!
- Onomastic data
 - standard naming pattern:
A / son of B / son of C // descendant of D
 - papponymy: name child for (male) ancestor



What a mess!

Text	Lines	PN	Role	FN	GFather	Ancestor
MLC 2178	28	Balaṣu	witness	Bassiya	Ištar-šum-ereš	Ekur-zakir
MLC 2178	28f	Anu-ah-ittannu	witness	Tattannu		Ekur-zakir
MLC 2178	29	Nidintu-Anu	witness	Anu-uballiṣ		Ah'utu
MLC 2178	30	Labaši	witness	Anu-balassu-iqbi		Ekur-zakir
MLC 2178	30f	Rabi-Anu	witness	Anu-zer-iddin		Ekur-zakir
MLC 2178	31f	Anu-mar-ittannu	witness	Anu-ab-ušur		Hunzu
MLC 2178	32	Arad-adešu	witness	Nidintu-Anu	Anu-ahhe-iddin	
YOS 20 74 (=MLC 2179)	1, 13	Idat-Anu	seller	Dumqi-Anu	Anu-ah-ušabši	Ekur-zakir
YOS 20 74 (=MLC 2179)	20	Anu-ahhe-iddin	witness	Nidintu-Anu	Anu-belšunu	Ah'utu
YOS 20 74 (=MLC 2179)	21	Tanittu-Anu	witness	Anu-ittannu	Anu-ab-ušur	Gimil-Anu
YOS 20 74 (=MLC 2179)	24	Ušuršu-Anu	witness	Nana-iddin	Anu-zer-iddin	Ekur-zakir
YOS 20 74 (=MLC 2179)	25	Nidintu-Anu	witness	Anu-ah-ittannu	Illut-Anu	Hunzu
YOS 20 74 (=MLC 2179)	26	Ṭab-Anu	witness	Illut-Anu	Anu-zer-iddin	Kuri
YOS 20 74 (=MLC 2179)	27	Dannat-Belti	witness	Labaši	Rihat-Anu	
MLC 2180+AoF 05 08 24	28f	Tattannu	witness	Dumqi-Anu	Tattannu	Hunzu
MLC 2180+AoF 05 08 24	29f	Nana-iddin	witness	Arad-reš		Hunzu
MLC 2180+AoF 05 08 24	33	Nidintu-Anu	witness	Anu-zer-iddin		Luštammar-Adad
MLC 2180+AoF 05 08 24	32	Anu-uballiṣ	witness			
MLC 2180+AoF 05 08 24	33	Šamaš-ittannu	witness			
YOS 20 27 (=MLC 2181)	r3,r4f	Nidintu-Anu	seller			
YOS 20 27 (=MLC 2181)	r8	Anu-ab-ušer	witness			
YOS 20 27 (=MLC 2181)	r8f	Nidintu-Anu	witness			
YOS 20 27 (=MLC 2181)	r9	Kidin-Anu	witness			
YOS 20 27 (=MLC 2181)	r10	Anu-uballiṣ	witness			
YOS 20 27 (=MLC 2181)	r10f	Anu-ah-iddin	witness			
YOS 20 27 (=MLC 2181)	r11	Nana-iddin	witness			
YOS 20 27 (=MLC 2181)	r12	Rihat-Anu	witness			
YOS 20 27 (=MLC 2181)	r12f	Nidintu-Anu	witness			



Common workflow across corpora/disciplines

NRAD

- **Name**
- **Activity**
- **Role**
- **Document**

Rules & probabilities

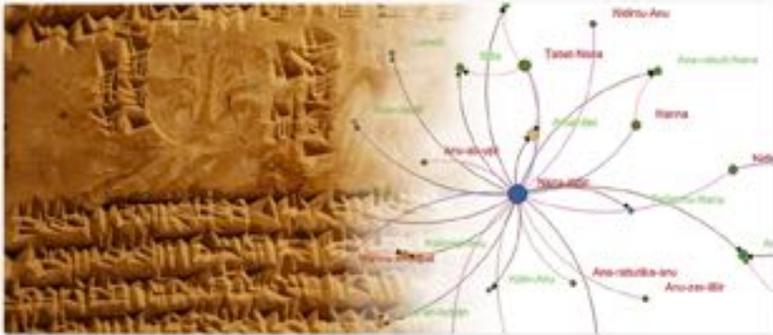
- corpus expertise
- patterns & text structure
- features w/in text
- features across texts

Berkeley Prosopography Services

bps: berkeley prosopography services

HOME BPS TEAM NEWS DOCUMENTATION USE BPS

About BPS: Welcome!



Berkeley Prosopography Services (BPS) is an open-source digital toolkit that supports prosopographical research and generates interactive visualizations of the biological and social connections that link documented individuals. It provides a dynamic and heuristic tool for researching historical communities documented in legal and administrative archives.

BPS is an innovate digital toolkit.

The *architecture of BPS* is *corpus agnostic*. That means that the toolkit components are built to work with general categories of data, not specific pieces of data. This opens the possibility of researchers using the toolkit to explore any textual corpus, regardless of the language of the

Search

QUESTIONS? SUGGESTIONS?

Contact us [here](#).

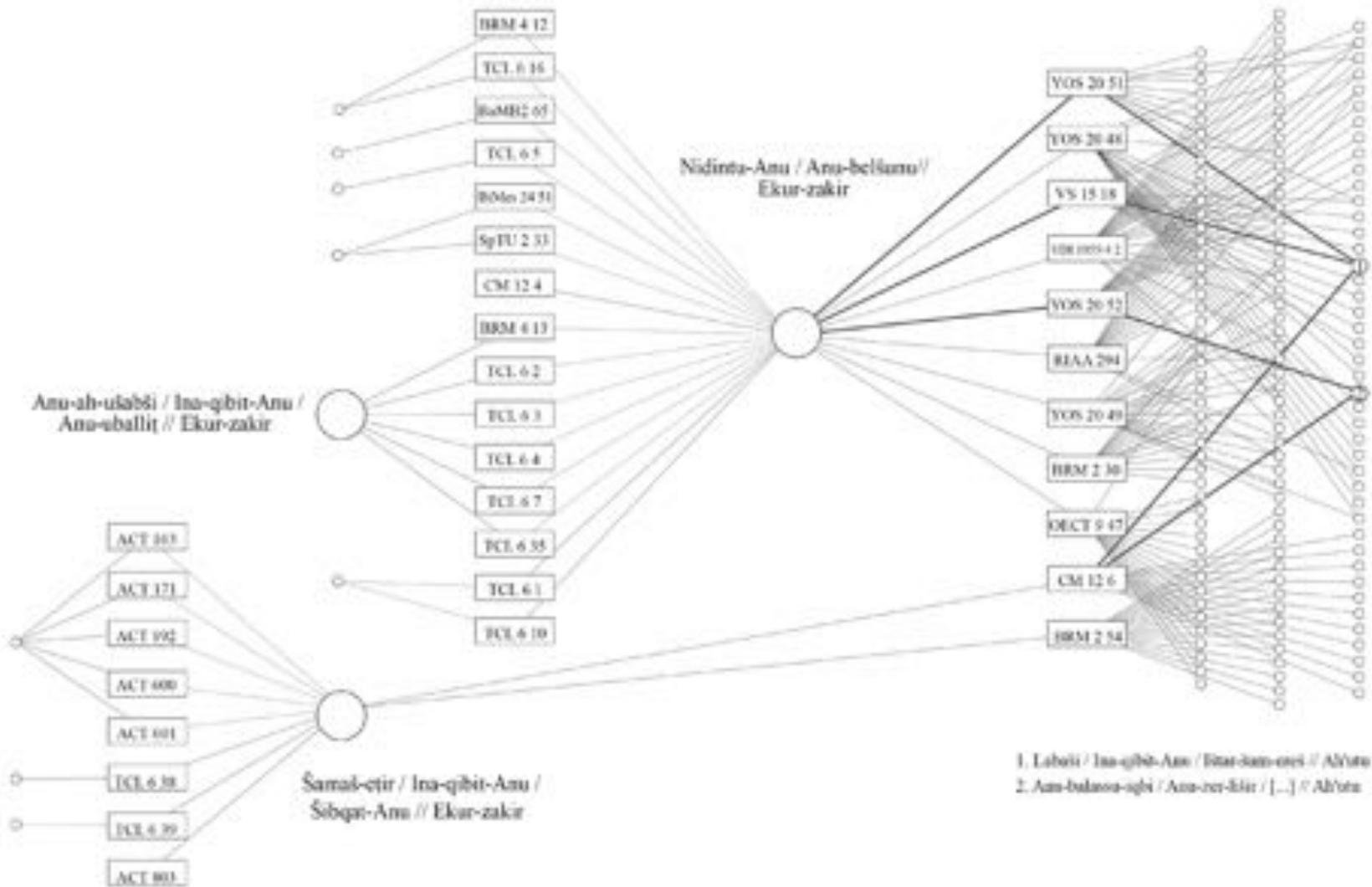
BPS HAS RECEIVED GENEROUS SUPPORT FROM:



Research IT
Advancing Research@Berkeley

<http://berkeleyprosop.digitalhumanities.berkeley.edu/>





Welcome to BPS!

There's some background on the project on our [About](#) page, and links to the related project homes are in the footer, below.

This site and the tools are still under development, but you are welcome to look around and see what's here. You can register for a basic account, with read-only access. Or, if you just want to browse a little, you can login in as "Reader" using the password "reader" to get read-only access (surprise!).

If you want to work with BPS with your corpus, you'll need to ask for more access. If you know either of [Laurie or Patrick](#), just send us an email with a description of your interest and any corpora you are working on, and we'll create a login for you. You can also contact us using the [Contact Us](#) page.

If you click on the Corpora link in the main navigation bar, you will see a list of the corpora that have been added. Once you are a registered user, you'll have a workspace as well, where you can import a corpus, configure the parameters and rules for our disambiguation engine, and see BPS in action!

Thanks for visiting.

Berkeley Prosopography Services

[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

Login

Username *

Password *

Remember me on this computer

[Forgot your password?](#)

[About](#) | [HBITN](#) | [BPS](#) | [Privacy Policy](#) | [Contact us](#)

Welcome to BPS!

There's some background on the project on our [About](#) page, and links to the related project homes are in the footer, below.

This site and the tools are still under development, but you are welcome to look around and see what's here. You can register for a basic account, with read-only access. Or, if you just want to browse a little, you can login in as "Reader" using the password "reader" to get read-only access (surprise!).

If you want to work with BPS with your corpus, you'll need to ask for more access. If you know either of [Laurie](#) or [Patrick](#), just send us an email with a description of your interest and any corpora you are working on, and we'll create a login for you. You can also contact us using the [Contact Us](#) page.

If you click on the Corpora link in the main navigation bar, you will see a list of the corpora that have been added. Once you are a registered user, you'll have a workspace as well, where you can import a corpus, configure the parameters and rules for our disambiguation engine, and see BPS in action!

Thanks for visiting.

Corpora Management

Click on a corpus name for details, and to import.

Corpus Name	# Docs	Description	
Test 5	2		Delete Corpus
Test corpus demo	2		Delete Corpus
Nielsen DB	7		Delete Corpus
HBTIN Matrix Demo	6		Delete Corpus
Langellotti corpus	125		Delete Corpus
Joth	13		Delete Corpus
Nielsen2	0	foo	Delete Corpus
HBTIN-Corpus2	3		Delete Corpus
Nielsen Reloaded	7		Delete Corpus
Bridget-fail-corpus	13	this seems to actually load after all...	Delete Corpus
Bridget-fail-corpus-2	13		Delete Corpus
Demo 2	2		Delete Corpus
Demo 1	7		Delete Corpus
nociantest2	2		Delete Corpus
Nana-iddin Demo	8		Delete Corpus
Anu-ah-ušabši Demo	3		Delete Corpus
test-idav	1		Delete Corpus

Add a New Corpus

Corpus Name	Description
<input type="text"/>	<input type="text"/>

Berkeley Prosopography Services

Logged in as Laurie | [Help](#) | [Admin](#) | [Sign Out](#)

Home Corpora Workspace

[Documents](#) [Person-Names](#) [Clan-Names](#) [Admin](#)

Showing 8 Documents in Corpus: Nana-iddin Demo

Document	Publication	Notes	Date
BRM 2 01			(?)
BRM 2 25			(?)
CM 12 01			(?)
VDI 1965/4 1			(?)
VS 15 03			(?)
VS 15 11			(?)
VS 15 23			(?)
VS 15 30			(?)

[About](#) | [HBTIN](#) | [BPS](#) | [Privacy Policy](#) | [Contact us](#)

Home Corpora Workspace

Corpus Document Details

[Return to Corpus details](#)

Document	Publication	Notes	Date
BRM 2 01			(?)
See also: CDLI Oracc TEI Image Line art			

17 Name-Role-Activity instances in Document:

Name	Normalized Form	Role	Activity	XML ID
Anu-ab-upur	-	Witness	Lease	
Anu-ahhe-bulluq	-	Father	Lease	
Anu-ah-tuqqin	-	Witness	Lease	
Nidintu-Anu	-	Father	Lease	
Nana-iddin	-	Witness	Lease	
Kidin-litar	-	Father	Lease	
Qisti-Anu	-	Witness	Lease	
Ina-qibit-Anu	-	Father	Lease	
Ša-Anu-iššu	-	Witness	Lease	
Ina-qibit-Anu	-	Father	Lease	
Anu-uballit	-	lessee	Lease	
Kidin-Anu	-	Father	Lease	
Mušezištu	-	neighbor	Lease	
Nana-iddin	-	lessor	Lease	
Tarittu-Anu	-	Father	Lease	
Ubar	-	scribe	Lease	
Šerki-Anu	-	Father	Lease	

[Return to Corpus details](#)

Home Corpora Workspace

Documents Person-Names Clan-Names Admin

68 Person-Names in Corpus:

Filter by Role: Filter by Gender:

Name	Gender	# Documents	Total Instances
(Missing Name)	unknown	2	4
Amal-Marduk	unknown	1	1
Ana-rabui-Nana	unknown	1	1
Ana-rabulka-Anz	unknown	2	2
Ana-rabulifu	unknown	1	1
Anu-ab-usur	male	3	4
Anu-ab-uter	male	1	2
Anu-ah-iddin	male	2	3
Anu-ah-itannu	male	4	5
Anu-ah-luggin	unknown	1	1
Anu-ah-ulabli	male	4	7
Anu-ahhe-bulut	male	1	2
Anu-ahhe-iddin	unknown	4	4
Anu-belassu-igbi	unknown	1	2
Anu-bellunu	male	6	13
Anu-bullissu	unknown	1	1
Anu-ikur	unknown	2	2
Anu-itannu	unknown	1	1
Anu-mar-itannu	unknown	1	1
Anu-mukin-apil	male	1	2
Anu-qilan	unknown	2	2
Anu-sballit	male	7	17
Anu-ulalim	unknown	1	1
Anu-zer-bri	male	1	2
Anu-zer-iddin	male	4	7
Anu-zer-Mir	male	2	6
Apla	male	1	2
Balatu	male	2	3
Bel-ab-usur	unknown	1	1
Dumci-Anu	male	1	3
Eriba	male	1	2
Harin-Esi	unknown	1	1
Hanna	unknown	1	1
Ilut-Anu	male	1	2

Home Corpora Workspace

Documents People Clans Settings Admin SNA

Set Model Parameters

Background on the model

The BPS analyzer will try to disambiguate among citations using the same name(s). To do this, it will basically model a new citation-person for each name it finds in a document (including fathers, grandfathers, etc. that are mentioned as qualifiers to the named actors). Then, it will attempt to collapse some of those citation-persons to get to the set of actual (real world) persons mentioned in all the corpus documents. Each citation-person is compared to other citation-persons, and a set of rules is applied to determine how likely it is that the two citations are the same person. The analyzer proceeds in two steps: first it considers all the citation-persons within each single document (intra-document), and then it considers the citation-persons across the entire corpus (inter-document).

When comparing two citation-persons, the analyzer will first require that there is no conflicting information about the two citation-persons - e.g., if they have different declared fathers, they will be considered as distinct, and will not be collapsed. The rules below allow you to configure whether specific roles must be considered to be distinct, and to control how strong the likelihood that two persons with partial matching name information are the same real world person.

General settings:

Number of qualifications (father/grandfather/ancestor/clan) in addition to forename required to consider a name citation "fully qualified"	<input type="text" value="2"/>
Assumed typical length of active business life (years)	<input type="text" value="25"/>
Assumed typical separation of generations (years)	<input type="text" value="15"/>

Step 1: Intra-document rules:

Rule Steps 1A, 1B, and 1C collapse citations within a single document.

Step 1A: Consider equally qualified names

Collapse equal, fully qualified citations (e.g., "PN _a son-of PN _b , in-clan CN _c " and "PN _a son-of PN _b , in-clan CN _c ")	<input type="text" value="Always: 100%"/>
Collapse equal, partly qualified citations (e.g., "PN _a son-of PN _b " and "PN _a son-of PN _b ")	<input type="text" value="Conservative: 30%"/>
Collapse equal, unqualified citations (e.g., "PN _a " and "PN _a ")	<input type="text" value="Aggressive: 75%"/>

Step 1B: Consider compatible, but not equally qualified names

Collapse partly qualified citations with compatible, fully qualified citations (e.g., "PN _a son-of PN _b " and "PN _a son-of PN _b , in-clan CN _c ")	<input type="text" value="Conservative: 30%"/>
Collapse unqualified citations with compatible, more qualified citations (e.g., "PN _a " and "PN _a son-of PN _b , in-clan CN _c ", OR, "PN _a " and "PN _a son-of PN _b ")	<input type="text" value="Aggressive: 75%"/>

Step 1C: Consider the roles of persons

Note that "ancestors" includes all fathers, mothers, grandfathers, and other declared ancestors.

Step 1: Intra-document rules:

Rule Steps 1A, 1B, and 1C collapse citations within a single document.

Step 1A: Consider equally qualified names

Collapse equal, fully qualified citations (e.g., "PN _a son-of PN _b in-clan CN _c " and "PN _a son-of PN _b in-clan CN _c ")	Always: 100% -
Collapse equal, partly qualified citations (e.g., "PN _a son-of PN _b " and "PN _a son-of PN _b ")	Conservative: 30% -
Collapse equal, unqualified citations (e.g., "PN _a " and "PN _a ")	Aggressive: 75% -

Step 1B: Consider compatible, but not equally qualified names

Collapse partly qualified citations with compatible, fully qualified citations (e.g., "PN _a son-of PN _b " and "PN _a son-of PN _b in-clan CN _c ")	Conservative: 30% -
Collapse unqualified citations with compatible, more qualified citations (e.g., "PN _a " and "PN _a son-of PN _b in-clan CN _c ", OR, "PN _a " and "PN _a son-of PN _b ")	Aggressive: 75% -

Step 1C: Consider the roles of persons

Note that "ancestors" includes all fathers, mothers, grandfathers, and other declared ancestors.

Can two instances of the same name within a document possibly be the same, just given the associated roles for the two names?			
	Principle	Witness	Ancestor
Principle	Always: 100% -	Never: 0% -	Always: 100% -
Witness		Never: 0% -	Always: 100% -
Ancestor			Always: 100% -

SNA Visualizer

Views

Workspace Corpus

Layout

Force Directed Radial

Filter (degree)

All Popular Peripheral

Sort

Cluster Degree

Search



Zoom

Active Paused

Edges

Show Hide

Node names

Show Hide

Static views

Family Tree

SNA Visualizer

Views

Workspace Corpus

Layout

Force Directed

Circle

Filter (degree)

All

Popular

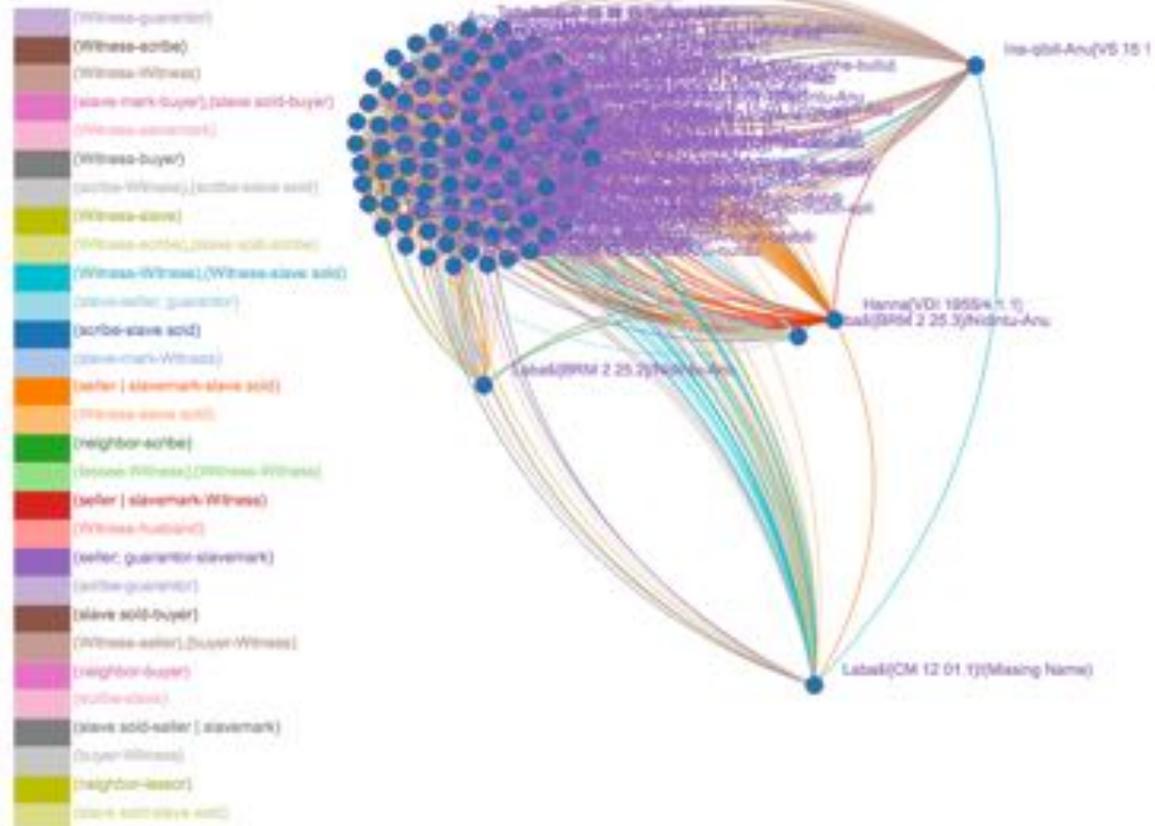
Peripheral

Sort

Cluster

Degree

Search



Zoom



Active

Paused

Edges

Show

Hide

Node names

Show

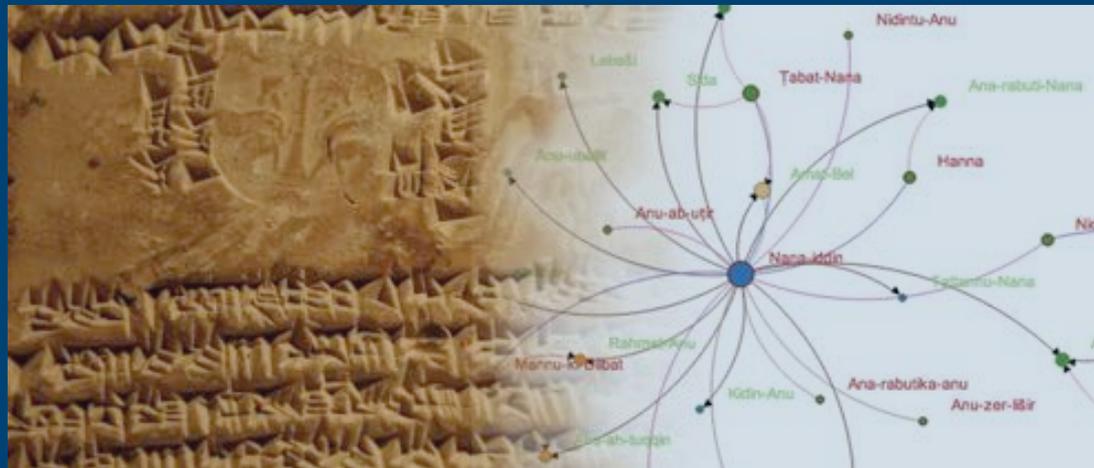
Hide

Static views

Family Tree

Berkeley Prosopography Services (BPS)

a toolkit supporting humanities research

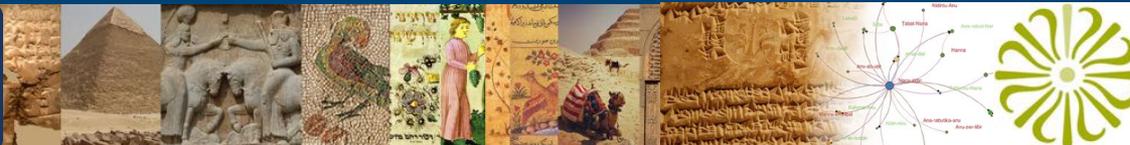


*UCB I-School Friday Afternoon Seminar
April 14, 2017
Laurie Pearce, Near Eastern Studies
Patrick Schmitz, Research IT*



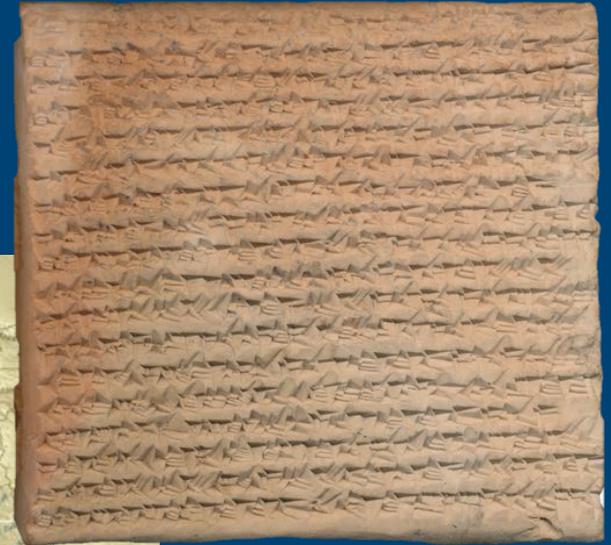
What is prosopography?

- Goals
 - identify people referenced in text corpora
 - build genealogies: family lineages
 - recover relationships: social networks
- Dependencies
 - scope and condition of media and data
 - disambiguation of namesakes
 - finding family relations
 - recognizing activities and roles
 - controlling chronological framework



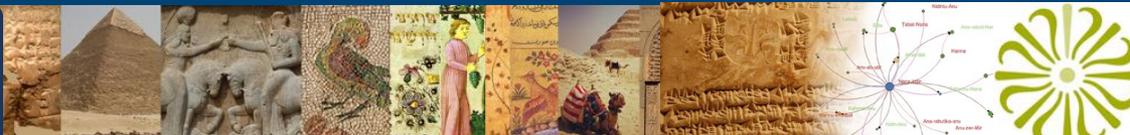
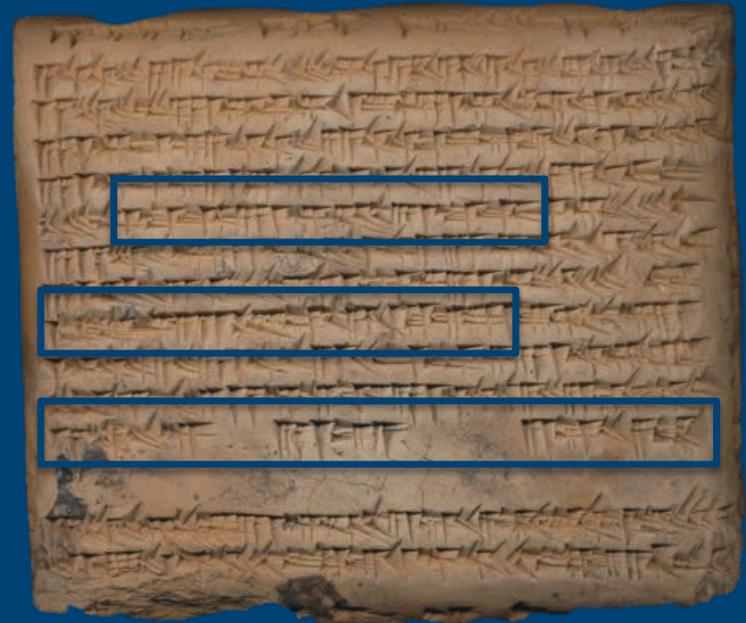
Project context: Hellenistic Uruk legal texts

- ~700 texts, c. 330-100 BCE
- contracts: slave, real estate, prebends
- 8-20 name citations/text
- 3 individuals/citation
- > 10k name instances



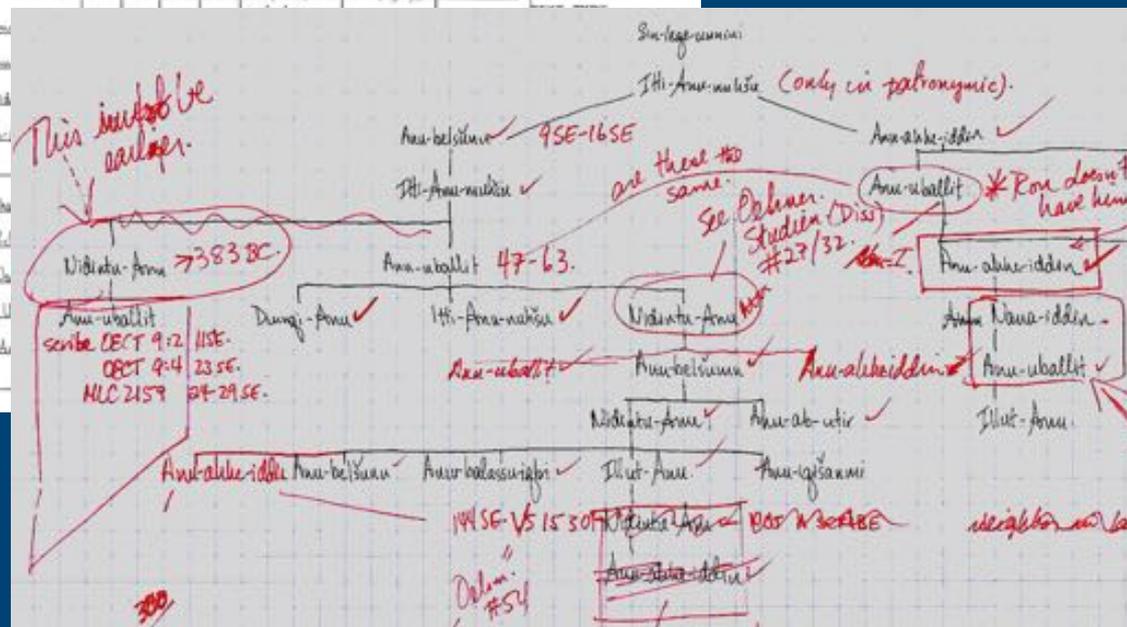
Data mining in Uruk legal texts

- Boilerplate text
 - repetitive patterns
 - attributes
 - many names
- Onomastic data
 - standard naming pattern
 - **A** son of **B** son of **C** descendant of **D**
 - papponymy: naming for (male) ancestor
 - “they all have the same name!”



BPS: identify & replicate familiar workflows

PAGE 1 OF FORM										PAGE 2 OF FORM	
No.	NAME	SX	FATHER	GRANDFATHER	Age	Role	Tel	Rel	LINES	SEAL	TEXT REF.
5005	Tolat-Anu		Nidanta-Anu		500	sc			30		DECT 9-12 (Bairasid)
5006	Nana-iddin		Anu-nubinsigbi	Anu-ahhe-iddin	L	s		same of 500	11, 15, 18, 22		DATE AS IS 2. 9. 18. 122
5007	Elah-Anu		Ubar			L			3		ERA sc
5008	Nana-iddin		Anu-ahhe-iddin			L			4		JULIAN DATE 54 23/07/122
5009	Tolat-Anu		Silat-Anu	Damat-Belli	L	b		same of 506 see # 530, 531	10, 15, 16, 18, 20		MULER
5010	Tarbit-Anu		Tolat-Anu	Damat-Belli	L	b		see of 506 see of 506 see of 506, 507	10, 13, 15, 18, 20		Anu-ahhe-iddin
5011	Damat-Belli		Tolat-Anu								
5012	Balatu		Damat-Belli	Anu							
5013	Anu-nubinsigbi		Ubar	Nidanta-Anu							
5014	Sakharita		Anu-ahhe-iddin	Sa							
5015	Damat-Belli		Reakaba								
5016	Tolat-Idin		Anu-bellu	Bib							
5017	Reakaba		Nidanta-Anu	B							
5018	Sunor-Idin		Damat-Anu	N							
5019	Tolat-Anu		Anu-ahhe-iddin	L							
5020	Anu-Si		Ubar-sunor-iddin	A							



DB and viz, c. 1985 AD



Prosopography ≈ NRAD

- **names**
 - naming formula
 - relationships
- features on persons:
 - **roles**, titles, gender
- features of & on **texts**:
 - genre, origin, date

Name

in a

Role

in an

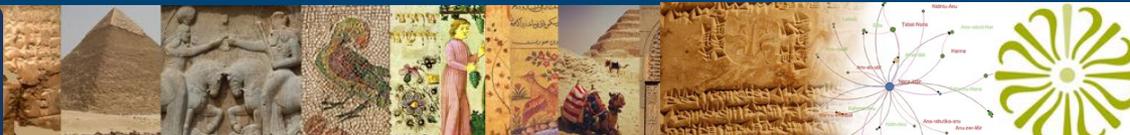
Activity

in a **Document**



Progress and Challenges

- + Community building
 - BPS & Social Science Matrix
 - research partners across disciplines / institutions
 - refine / expand understanding workflows
- + Documentation: improvements based on feedback
- Staffing: access to dev and UCD/UX staff



Progress and Challenges

Disambiguation

- + role matrix functionality
- + role matrix complexity
- long tail filtering



Progress and Challenges

Visualization

- + integration into toolkit
- resource allocation: impact on filtering
- real estate
- ? relevance to workflow
- ? semantic filtering: issues with external viz/graph semantics



Progress and Challenges

Workspace

- not fully implemented
- + documentation of assertions:
 - scholarly discourse
 - methodological transparency
 - preservation
- + pedagogic value



Progress and Challenges

Documentation

- + refined documents based on user experience with toolkit





Welcome to BPS!

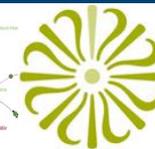
There's some background on the project on our [About](#) page, and links to the related project homes are in the footer, below.

This site and the tools are still under development, but you are welcome to look around and see what's here. You can register for a basic account, with read-only access. Or, if you just want to browse a little, you can login in as "Reader" using the password "reader" to get read-only access (surprise!).

If you want to work with BPS with your corpus, you'll need to ask for more access. If you know either of [Laurie](#) or [Patrick](#), just send us an email with a description of your interest and any corpora you are working on, and we'll create a login for you. You can also contact us using the [Contact Us](#) page.

If you click on the Corpora link in the main navigation bar, you will see a list of the corpora that have been added. Once you are a registered user, you'll have a workspace as well, where you can import a corpus, configure the parameters and rules for our disambiguation engine, and see BPS in action!

Thanks for visiting.



Workspace: Settings

Berkeley Prosopography Services

[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

dev.berkeleyprosopography.org

[Documents](#) [People](#) [Clans](#) **Settings** [Admin](#) [Visualizer](#) [Visualizer 2.0](#)

Set Model Parameters for Workspace: My Workspace

Background on the model

The BPS analyzer will try to disambiguate among citations using the same name(s). To do this, it will basically model a new citation-person for each name it finds in a document (including fathers, grandfathers, etc. that are mentioned as qualifiers to the named actors). Then, it will attempt to collapse some of those citation-persons to get to the set of actual (real world) persons mentioned in all the corpus documents. Each citation-person is compared to other citation-persons, and a set of rules is applied to determine how likely it is that the two citations are the same person. The analyzer proceeds in two steps: first it considers all the citation-persons within each single document (*intra-document*), and then it considers the citation-persons across the entire corpus (*inter-document*).

When comparing two citation-persons, the analyzer will first require that there is no conflicting information about the two citation-persons - e.g., if they have different declared fathers, they will be considered as distinct, and will not be collapsed. The rules below allow you to configure whether specific roles must be considered to be distinct, and to control how strong the likelihood that two persons with partial matching name information are the same real world person.

General settings:

Number of qualifications (father/grandfather/ancestor/clan) in addition to forename required to consider a name citation "fully qualified"

2

Assumed typical length of active business life (years)

15

Research IT

Advancing Research@Berkeley



NATIONAL ENDOWMENT FOR THE

Humanities

Berkeley Prosopography Services | 135

Workspace: Settings

Berkeley Prosopography Services



[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

dev.berkeleyprosopography.org

[Documents](#) [People](#) [Clans](#) [Settings](#) [Admin](#) [Visualizer](#) [Visualizer 2.0](#)

General settings:

Number of qualifications (father/grandfather/ancestor/clan) in addition to forename required to consider a name citation "fully qualified"

Assumed typical length of active business life (years)

Assumed typical separation of generations (years)



Workspace: Pattern matching in names

Berkeley Prosopography Services

[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

dev.berkeleyprosopography.org

[Documents](#) [People](#) [Clans](#) [Settings](#) [Admin](#) [Visualizer](#) [Visualizer 2.0](#)

Step 1: Intra-document rules:

These rules collapse citations within a single document.

Step 1A: Consider equally qualified names

Collapse equal, fully qualified citations
(e.g., "PN_a, son-of PN_b, in-clan CN_c"
and "PN_a, son-of PN_b, in-clan CN_c")

Always ▾

Collapse equal, partly qualified citations
(e.g., "PN_a, son-of PN_b" and "PN_a,son-of PN_b")

Mostly ▾

Collapse equal, unqualified citations
(e.g., "PN_a" and "PN_a")

Mostly ▾

Step 1B: Consider compatible, but not equally qualified names

Collapse partly qualified citations with compatible, fully qualified citations



Name in **Role** in Activity in Document

Berkeley Prosopography Services



[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

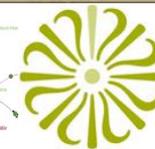
dev.berkeleyprosopography.org

[Documents](#) [People](#) [Clans](#) [Settings](#) [Admin](#) [Visualizer](#) [Vizualizer 2.0](#)

Step 1C: Consider the roles of persons

Can two instances of the same name within a document possibly be the same, just given the associated roles for the two na

	Witness	unknown	slave-mark	slave sold	slave	seller	scrib
buyer	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>
co-owner	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>
divider	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>
guarantor	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="Maybe"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>
lessee	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>
lessor	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>
neighbor	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="No"/>
Principal	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>	<input type="text" value="No"/>	<input type="text" value="Yes"/>	<input type="text" value="No"/>



Workspace: View documents

Berkeley Prosopography Services



[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

dev.berkeleyprosopography.org

[Documents](#) [People](#) [Clans](#) [Settings](#) [Admin](#) [Visualizer](#) [Vizualizer 2.0](#)

Showing 15 Documents in Workspace:

Document	Publication	Notes	Date
BRM 2 25			(?)
TCL 13 229			(?)
VDI 1955/4 1			(?)
VS 15 11			(?)
VS 15 23			(?)
VS 15 30			(?)
VS 15 31			(?)
VS 15 36			(?)
VS 15 48			(?)
VS 15 50			(?)
VS 15 51			(?)
YOS 20 2			(?)
YOS 20 4			(?)
YOS 20 5			(?)
YOS 20 95			(?)



Workspace: View name instances

Berkeley Prosopography Services



[Register](#) | [Help](#) | [Sign In](#)

[Home](#) [Corpora](#) [Workspace](#)

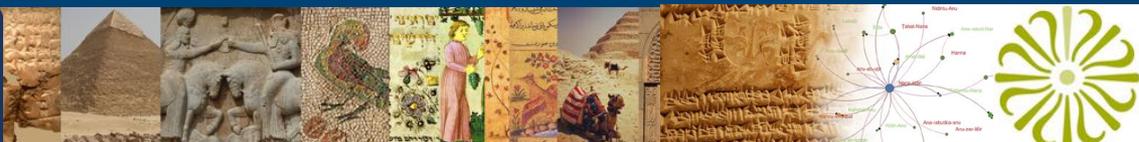
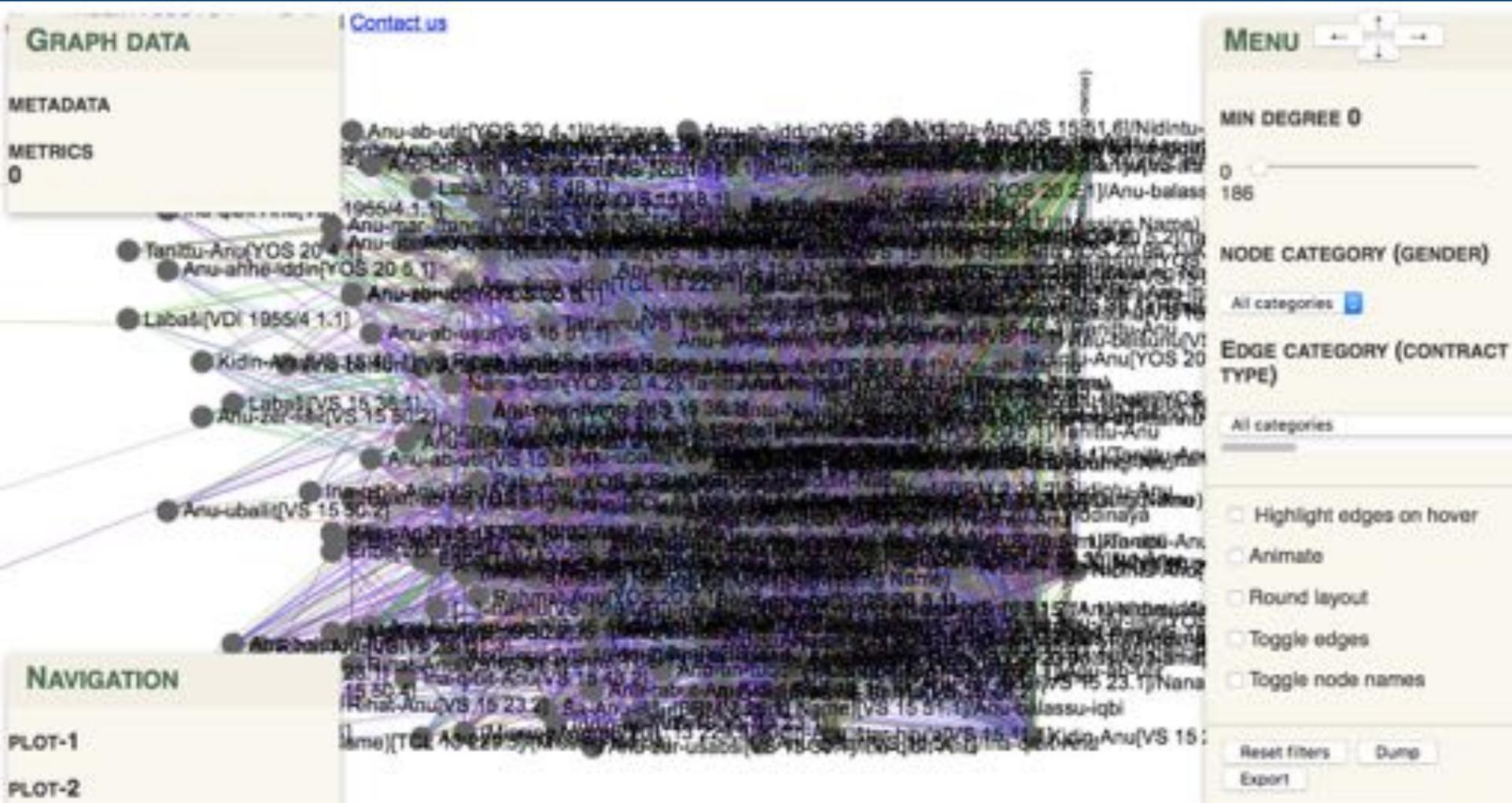
dev.berkeleyprosopography.org

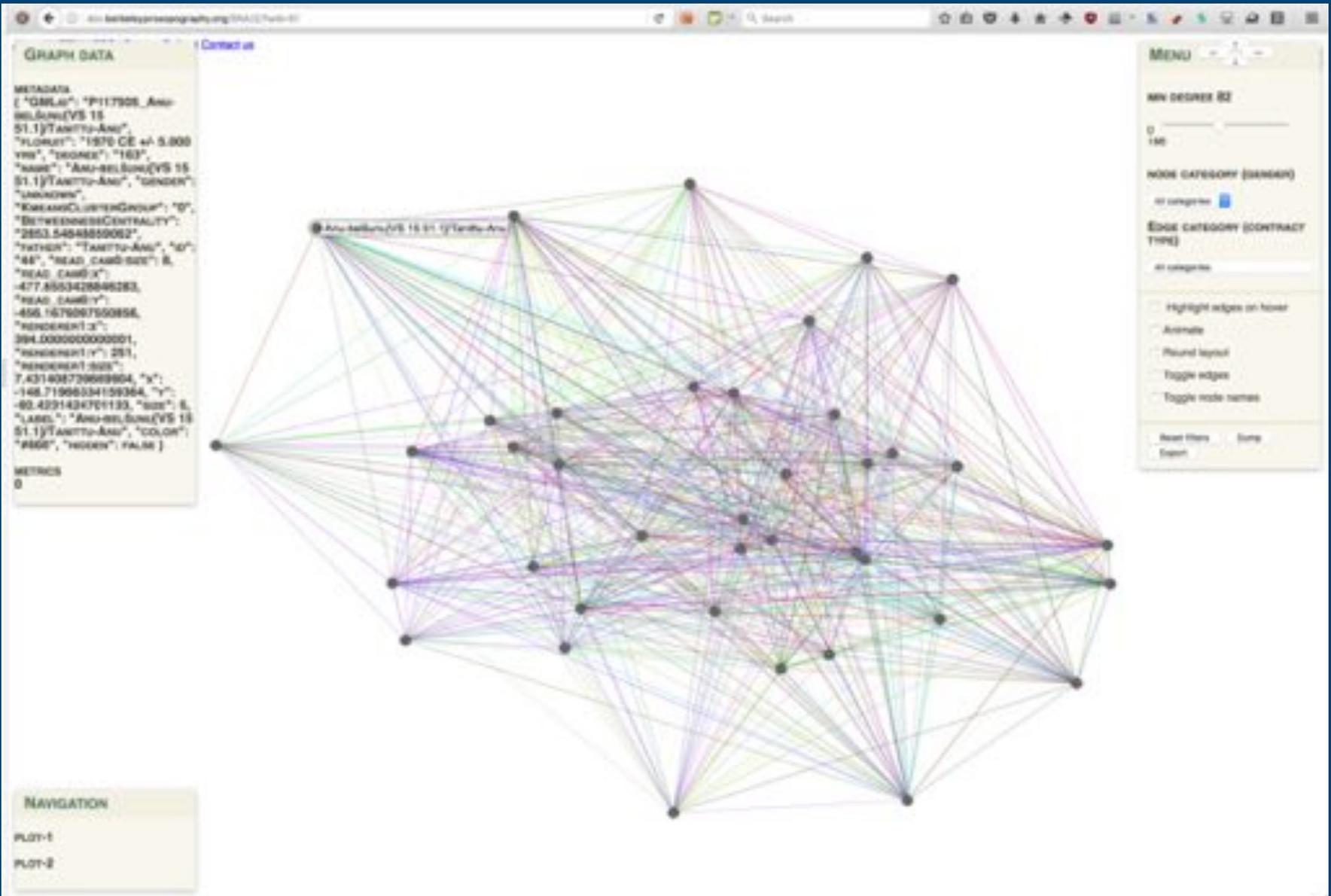
[Documents](#) [People](#) [Clans](#) [Settings](#) [Admin](#) [Visualizer](#) [Vizualizer 2.0](#)

Showing 427 Persons for Workspace:

Anu-belšunu[VS 15 31.1]/Nidintu-Anu	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 36.1]/Tanittu-Anu	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 48.1]	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 48.2]/Ina-qibit-Anu	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 48.3]/Tanittu-Anu	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 50.1]	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 51.1]/Tanittu-Anu	1970 CE +/- 5.000 yrs
Anu-belšunu[VS 15 51.2]	1970 CE +/- 5.000 yrs
Anu-bullissu[VDI 1955/4 1.1]	1970 CE +/- 5.000 yrs
Anu-ikšur[VS 15 48.1]/Anu-ahhe-iddin	1970 CE +/- 5.000 yrs
Anu-ikšur[VS 15 51.1]/Kidin-Anu	1970 CE +/- 5.000 yrs
Anu-ikšur[VS 15 51.2]/Anu-ah-Itannu	1970 CE +/- 5.000 yrs
Anu-ikšur[YOS 20 4.1]/Anu-ah-Itannu	1970 CE +/- 5.000 yrs
Anu-ikšur[YOS 20 5.1]/Anu-ah-Itannu	1970 CE +/- 5.000 yrs







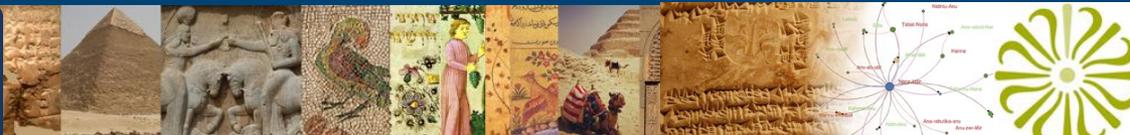
Thank you!

Berkeley Prosopography Services (BPS)
dev.berkeleyprosopography.org

for more info contact:

Laurie Pearce lpearce@berkeley.edu

Patrick Schmitz pschmitz@berkeley.edu



Computational Assyriology & Reproducible Research

Research IT Reading Group

October 4, 2018

Laurie Pearce, PhD

Near Eastern Studies, UC Berkeley



- <http://oracc.museum.upenn.edu/cams/gkab/literacies/index.html>

Computational Assyriology & Reproducible Research

collaborations between
Assyriologists, Astronomers,
and IT specialists

@

University of California, Berkeley

David Bamman, Gil Breger, Jason Moser, Laurie
Pearce, Francesca Rochberg,
Patrick Schmitz, Joanne Tan, Niek Veldhuis

&

Ludwig-Maximilians-Universität München

Nathan Morello, Jamie Novotny, Karen Radner,
Frauke Weiershaeuser

UCB-LMU collaboration

collaborative, transdisciplinary,
computationally engaged research

- **Berkeley**
 - Berkeley Prosopography Services
 - Computational Assyriology
- **LMU**
 - Specialized research team
 - NA domain expertise
 - SAAo (oracc.org/saao/corpus)

Computational Assyriology

- counting words



Computational Assyriology

- counting words
- best practices
 - software
 - methods and tools
 - documentation
- asking good questions

Computational Assyriology

- counting words
- best practices
 - software
 - methods and tools
 - documentation
- asking good questions

Reproducible Research

- replicable
- reproducible

Initial project corpus: NA letters

- ~3,850 letters / astrological reports (= SAA), 744-612 BC
- 8 Assyrian kings, provincial governors, courtiers
- 250 individual correspondents w. Assyrian kings
- replication and reproducibility of Parpola's LAS dossiers
 - experienced epigrapher, recognized authority
 - workflow & decision-making processes not documented
 - *explicitly* intended as preliminary assessment
 - assessments *canonized* but unchecked

Heuristics for assembling letter dossiers

– astronomical observations

- solar /lunar eclipse, movement of planets: compute absolute dates

– prosopography

- link individuals to dated outside sources: Eponym List, Eponym Chronicle, royal inscriptions, admin texts

– “scribal hands”

- observing and grouping diagnostic characteristics of the shapes of individual cuneiform signs, orthographies

Astronomy as chronological tool

- explicit evidence
 - eclipses
 - planetary position/movement
 - constellations
 - synodic phenomena
- implicit evidence
 - “farmer”
- challenges
 - intercalation
 - 29 / 30 day months
 - constellation boundaries

Explicit data in LAS 70 = SAA 10 73

(P334877)

Addāru 27, eponymy of Šulmu-Bēl-lašme

(= 26-03-669 BC, Parpola)

MUL.UDU.IDIM.[GUD].UD

DUMU—LUGAL ṣu^{*1}-ú

ina ŠÀ MUL.[LÚ.ḪUN].GÁ

in-[na]-ṣar¹

MUL.dil-bat [ina KÁ].DINGIR.RA.KI

ina É—ṣAD^{*1}-[šú in]-na-mar

^d30 ina ITI.BARAG

UD-mu ú-ṣal^{*1}-lam^{*}

♀ visible in ṽ

♀ visible in Babylon

☾ completes day in Nisannu

Assumptions in LAS 70 = SAA 10 73

Addāru 27, eponymy of Šulmu-Bēl-lašme
(= 26-03-669 BC, Parpola)

- prince resident in Babylon, visit king, his father
- Mercury visible in Aries; at same time
- *Venus said to be visible in Babylon but not in Nineveh*
- written at turn of year
- Addāru will be 30 days
- day on which written considered auspicious

Astronomical challenges

- ina Nisannu **rēš šatti** ^dSîn ina ūmi ušallam
- 30-day month before or in named month?
- limitations of tools available to Parpola
- updated algorithms in Alcyone/Stellarium
 - NA intercalation not standardized

Alcyone: JPL algorithms

Alcyone Astronomical Tables 3.1 - Lunar Eclipses at Babylon -700 - 600

File Edit View Settings Chart Help

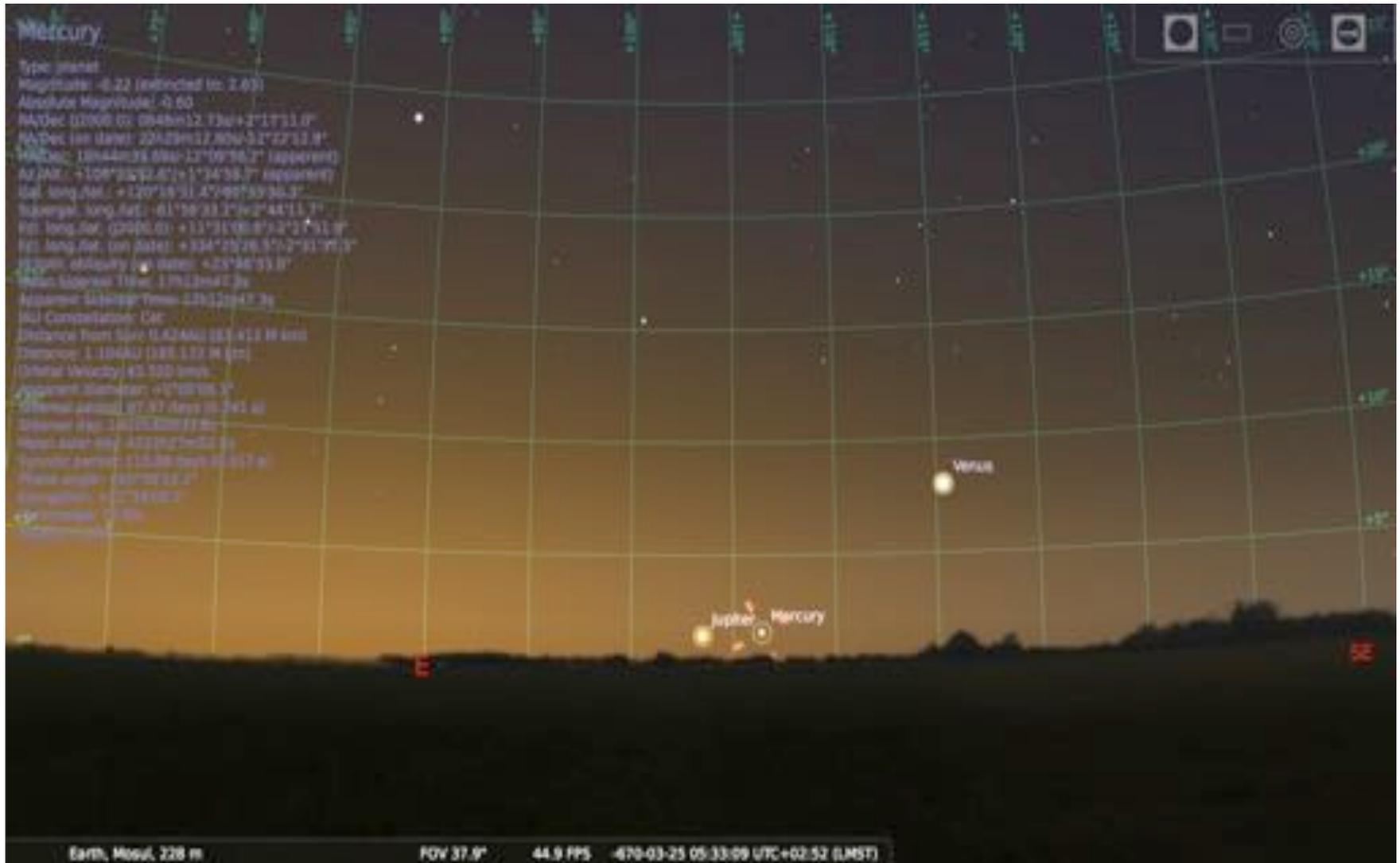
Local circumstances of lunar eclipses - Babylon (Jag) [-700.00 - 600.00]

Date	Duration	Semi	delta T	Magnitude (pen.)	Magnitude (Limb)	Eclipse type	Penumbral phase begins		Partial phase begins		Total phase begins		Maximum eclipse		Total phase ends		Partial phase ends		Penumbral phase ends		Duration		
							Time	Moon alt.	Time	Moon alt.	Time	Moon alt.	Time	Moon alt.	Time	Moon alt.	Time	Moon alt.	Time	Moon alt.	Time	Moon alt.	Time
670-01-07	-0204	22	05.36.20	0.512	-0.538	penumbral	14:40	52° 44'	—	—	—	—	10:22	54° 14'	—	—	—	—	18:04	60° 57'	03:24	—	—
670-02-01	-0207	65	05.36.11	0.186	-0.832	penumbral	07:52	42° 32'	—	—	—	—	18:46	44° 00'	—	—	—	—	18:46	42° 50'	01:53	—	—
669-01-27	-0202	32	05.36.04	1.894	-0.857	partial	23:05	28° 37'	00:09	13° 38'	—	—	11:42	17° 33'	—	—	03:34	14° 32'	04:18	24° 17'	05:34	03:06	—
669-07-23	-0276	37	05.35.58	1.836	-0.872	partial	08:34	30° 10'	11:08	24° 52'	—	—	12:36	1° 24'	—	—	14:28	8° 36'	15:43	20° 33'	05:49	03:19	—
668-01-06	-0270	42	05.35.48	2.117	-1.050	total	03:47	26° 22'	14:42	27° 38'	15:40	40° 28'	16:36	38° 44'	17:11	64° 28'	18:23	74° 37'	19:05	75° 17'	05:17	03:28	00:10
668-07-12	-0264	47	05.35.40	2.424	-1.402	total	00:21	30° 28'	11:24	32° 18'	12:23	47° 38'	13:18	17° 11'	14:04	27° 22'	23:23	14° 42'	16:23	24° 17'	06:10	03:46	00:20
667-07-01	-0252	57	05.35.25	1.090	-0.058	partial	03:57	5° 17'	16:46	21° 20'	—	—	18:36	24° 49'	—	—	16:45	26° 26'	18:36	31° 42'	04:39	00:17	—
666-01-22	-0246	29	05.35.10	1.210	-0.257	partial	08:27	32° 22'	12:46	36° 57'	—	—	18:39	34° 30'	—	—	18:33	32° 34'	20:51	27° 56'	04:24	01:46	—
665-09-11	-0229	39	05.36.55	2.574	-1.633	total	05:06	40° 20'	05:36	34° 58'	10:17	02° 24'	11:44	04° 07'	13:33	47° 18'	13:28	57° 54'	14:22	59° 10'	05:36	03:30	00:35
664-02-24	-0211	34	05.36.02	1.920	-0.402	partial	07:27	42° 22'	08:36	52° 21'	—	—	10:03	11° 30'	—	—	11:17	47° 32'	12:36	7° 26'	03:08	02:12	—
663-03-22	-0206	21	05.36.26	0.428	-0.621	penumbral	23:42	12° 30'	—	—	—	—	01:38	8° 22'	—	—	—	—	12:34	17° 18'	03:12	—	—
663-04-20	-0205	39	05.36.24	0.222	-0.796	penumbral	03:23	7° 18'	—	—	—	—	14:33	20° 18'	—	—	—	—	15:40	24° 07'	03:18	—	—
663-09-14	-0190	26	05.36.17	0.496	-0.467	penumbral	00:17	21° 38'	—	—	—	—	11:37	12° 18'	—	—	—	—	13:07	1° 42'	03:00	—	—
663-10-14	-0209	64	05.36.18	0.326	-0.648	penumbral	20:25	46° 20'	—	—	—	—	11:42	34° 12'	—	—	—	—	22:59	22° 23'	03:34	—	—
662-03-16	-0204	31	05.36.09	1.796	-0.664	partial	23:35	21° 38'	00:32	4° 47'	—	—	02:53	11° 46'	—	—	03:34	02° 57'	04:36	41° 37'	03:45	03:02	—
661-03-27	-0202	41	05.36.14	2.676	-1.620	total	06:27	7° 34'	11:34	17° 14'	00:20	47° 28'	10:28	27° 58'	14:31	02° 22'	16:12	40° 21'	16:24	40° 31'	06:52	03:46	00:45
661-08-21	-0276	46	05.35.46	2.646	-1.628	total	04:19	14° 17'	15:20	24° 28'	16:20	32° 52'	17:10	34° 00'	17:39	42° 03'	18:00	44° 17'	20:00	43° 12'	03:41	03:40	00:39
660-02-16	-0270	51	05.35.38	1.323	-0.330	partial	09:01	30° 21'	10:23	43° 08'	—	—	11:11	7° 10'	—	—	12:28	24° 07'	13:46	13° 41'	04:46	02:03	—
660-08-11	-0264	56	05.35.30	1.353	-0.292	partial	21:10	24° 36'	22:15	8° 12'	—	—	12:46	0° 43'	—	—	10:38	4° 30'	12:17	18° 21'	03:02	01:45	—
660-01-07	-0259	23	05.35.24	0.733	-0.234	penumbral	14:02	50° 38'	—	—	—	—	15:46	51° 53'	—	—	—	—	17:36	60° 45'	03:34	—	—
660-02-06	-0258	61	05.35.23	0.124	-0.841	penumbral	00:46	7° 30'	—	—	—	—	11:32	1° 50'	—	—	—	—	12:19	7° 11'	01:34	—	—
660-06-21	-0240	38	05.35.00	2.076	-1.063	total	11:31	24° 22'	11:15	11° 41'	13:38	1° 28'	13:17	4° 46'	14:17	9° 47'	15:39	16° 26'	16:45	24° 26'	05:36	03:24	00:38
660-12-16	-0215	41	05.34.52	2.124	-1.472	total	14:19	32° 48'	15:24	46° 12'	16:28	62° 18'	17:14	71° 16'	18:00	76° 42'	19:01	75° 14'	20:10	68° 30'	05:51	03:41	00:32
660-12-06	-0212	51	05.34.57	1.315	-0.136	partial	07:24	74° 49'	16:24	71° 30'	—	—	20:08	62° 20'	—	—	20:52	57° 54'	22:42	34° 28'	03:08	01:29	—
660-08-01	-0217	58	05.34.29	0.800	-0.076	penumbral	18:50	52° 22'	—	—	—	—	18:44	26° 42'	—	—	—	—	20:38	24° 26'	03:48	—	—
660-06-26	-0219	34	05.34.19	0.741	-0.879	penumbral	15:16	12° 48'	—	—	—	—	16:56	14° 16'	—	—	—	—	17:59	11° 47'	03:26	—	—

⌂

Moon -700.00 - 600.00 UT+3h⁺ Morrison and Stephenson (2004) Babylon (Jag) Refraction

Stellarium



Astronomy meets paleography

LAS 14

- Observation: occultation of Jupiter by moon
 - “Back of the moon should be shown to a eunuch who has a sharp eye: there is less than a span left to close. He should sta[nd] in shadow to observe”
- Assignment of scribe
 - paleography: Issar-šum-ereš
 - orthography and phraseology: Akkullanu

Possibilities of paleography

- Changes in paleography / orthography
- Disambiguation of namesakes
- New connections
- Regional paleography / orthography

Identify most significant signs

Paleography

\d (displaced)
\m (missing) →
\t (tilted)
\p (additional)
\y (other)
@v (other or tilted, LU₂* and TA*)

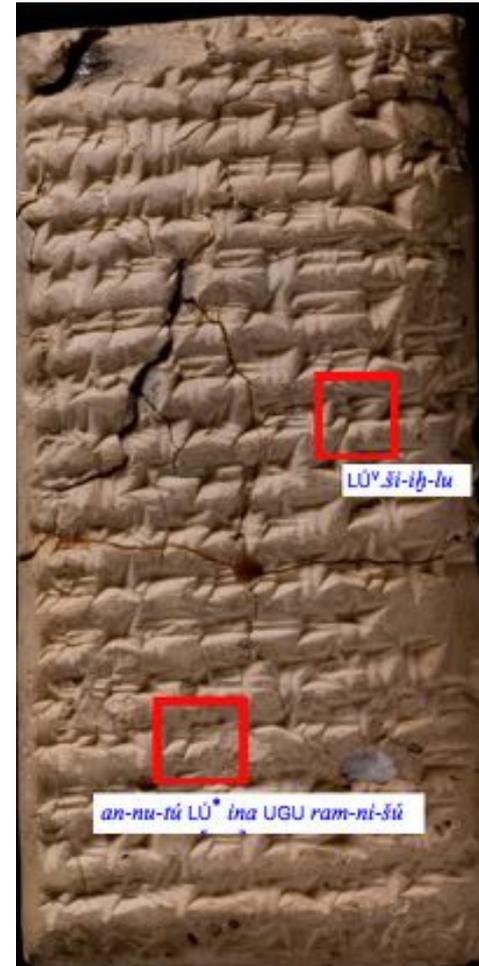
intentional vs. casual?

relation with other signs influence modification?

E.g., LU₂ as determinative or as logogram

Orthography

- li / li₂
- plene orthographies le-(e-)mur
- GID₂ v. SUD



SAA 1 205 r. 8, 14

Variations: individual tendencies or corpus-wide conventions?

- *li*₂ restricted to forms like *be-li*₂ vs. *li* everywhere else?
 - contextual variation, not scribal tendency to use variations within same forms
 - complementary distribution
- *ia* / *ia*₂ used freely in forms like *EN-ia*/*EN-ia*₂
 - indication of scribe's preference
 - free or mixed distribution.



Greetings formulas

- **abat šarri**
- **ana šarri béliya urdaka PN**
- **ana šarri béliya urdaka PN lú šulmu ana šarri béliya**
- **ana šarri béliya urdaka PN + PN lú šulmu ana šarri béliya**
- ana šarri béliya urdaka PN lú šulmu ana šarri béliya DN (+DN) likrubu**
- ana šarri béliya urdaka PN lú šulmu ana šarri béliya šulmu ana ON (+ON)**
- ana šarri béliya urdaka PN lú šulmu ana šarri béliya šulmu ana ON (+ON)**
(libbu ša) šarri béliya (adanniš) lú ſab (šu)
- ana PN/LÜ-^{*} béliya (urdaka) PN**
- IM PN ana PN/LÜ-^{*}**



SAA 05 074. Mule Express not Available (NL 062; ND 2367; SAA 19 161)

<ul style="list-style-type: none"> ○ 1 a-na LUGAL be-li-ia ○ ARAD-ka maḥ-de-e ○ lu-u DI-mu a-na MAN EN-ia 	<p>(1) To the king, my lord: your servant Mahdē. Good health to the king, my lord!</p>
<ul style="list-style-type: none"> ○ 4 ki-mi li-ú ša ni-^rda¹-nu-mi ○ 5 LÜ-^rgur-bu¹-ti ša il-^rak-[an]-^rni¹ 	<p>(4) (As to) the mule express that we provide and the royal bodyguard who com[es] citing a royal order that he should go as far as Šabirešu — he has used up the [...] in my possession!</p>
<ul style="list-style-type: none"> ○ 6 ma-a [a]-bat MAN ši-i-ti ○ 7 ma-a a-di URU.šá-bi-ri-šá ○ 8 lil-lik ANŠE¹ ša ina <IGD>-ia ○ 9 ^rug-da¹-me-^rra¹ LUGAL be-li ○ 10 ú-da a-di URU.šá-bi-ri-šá ○ 11 ú-re-u la ú-^rkal¹-[la] ○ 12 ša il-^rak]-ú-[ni] 	<p>(6) The king, my lord, knows that I do not main[tain] a team (to go) as far as Šabirešu; the ones that go do not return. My teams are used up; the king, my lord, should know (this).</p>

Tools and Methods

1. JSON files of ORACC texts: Python / Jupyter notebooks
2. identify most significant signs and syllables
3. encode that information per SAA letter into text vector; clustering algorithm to group text vectors into clusters
4. compare clustering to the classifications drawn on basis of features like sender location or dossier

Text Vectors

- set up text vector by determining the distribution of one selection of sign form or syllable versus another
 - e.g., letter 1: 5 instances of šu, 15 instances of šu₂

	šu	šu ₂
Letter 1	0.25	0.75

Clustering

- K-Means algorithm to cluster text vectors
- algorithm groups SAA letters into clusters where total “distance” between letters and centers of clusters is minimal

Problems with paleography

- sign characteristics not readily apparent
- insufficient data set size
- base line not established for clusters
- challenge of verifying / augmenting the data set

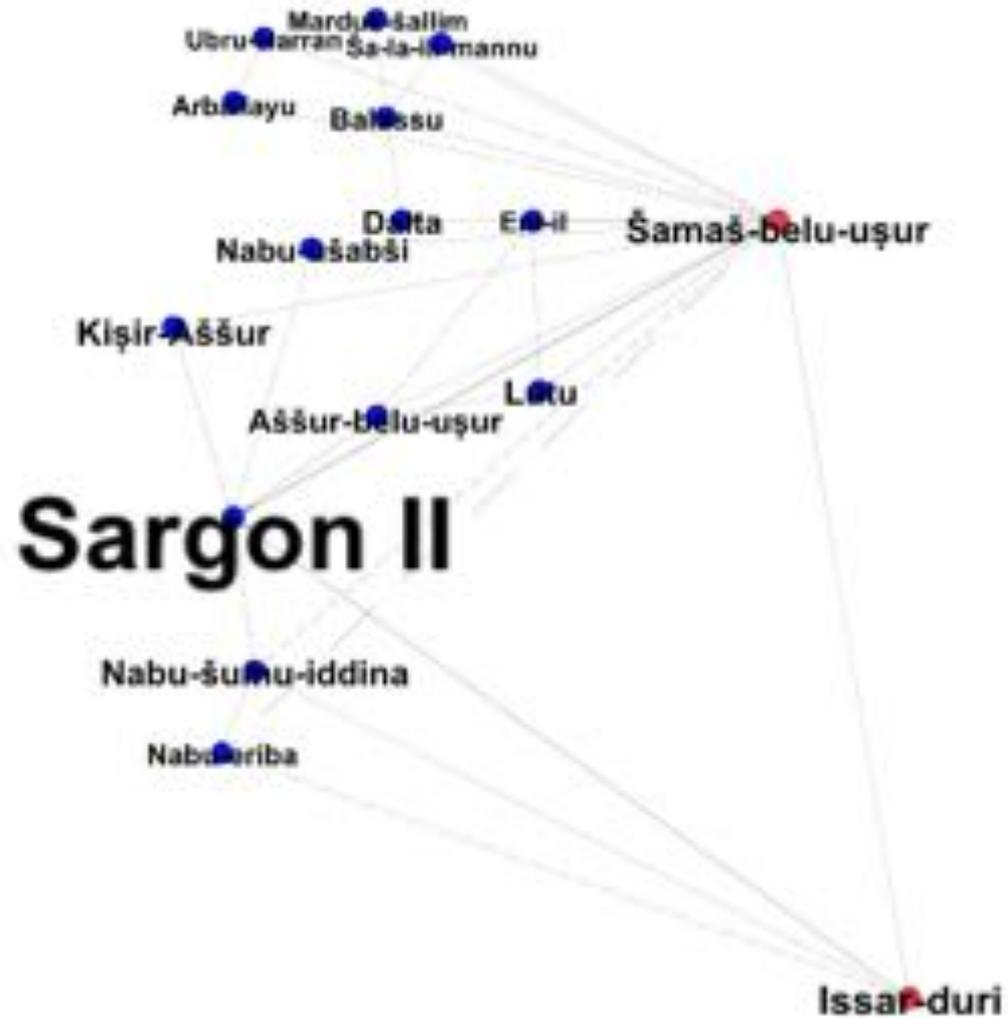
Social Network Analysis

- accept Parpola's identification of Sargon-related letters
- 15 top-tier officer contemporaries of Sargon, not named
- connections between senders / recipients / mentioned individuals
- difficulty of defining roles within in the letters
- all connected through Sargon
- 2nd degree connections to those people \approx 93-94% of network

Social Network Analysis

- consider last 6-7% to see how/to whom connected.
- chief treasurer \approx 33% of network (= one of the 15-20 guys, but at 2 levels down). Would confirm stability of Parpola's base corpus,
- All datable individuals: enemies, eponyms, top-level officials (25-30 persons). At def 2, yesterday was 50-60 ties to Sargon.

Social Network Analysis



Realigning the collaboration

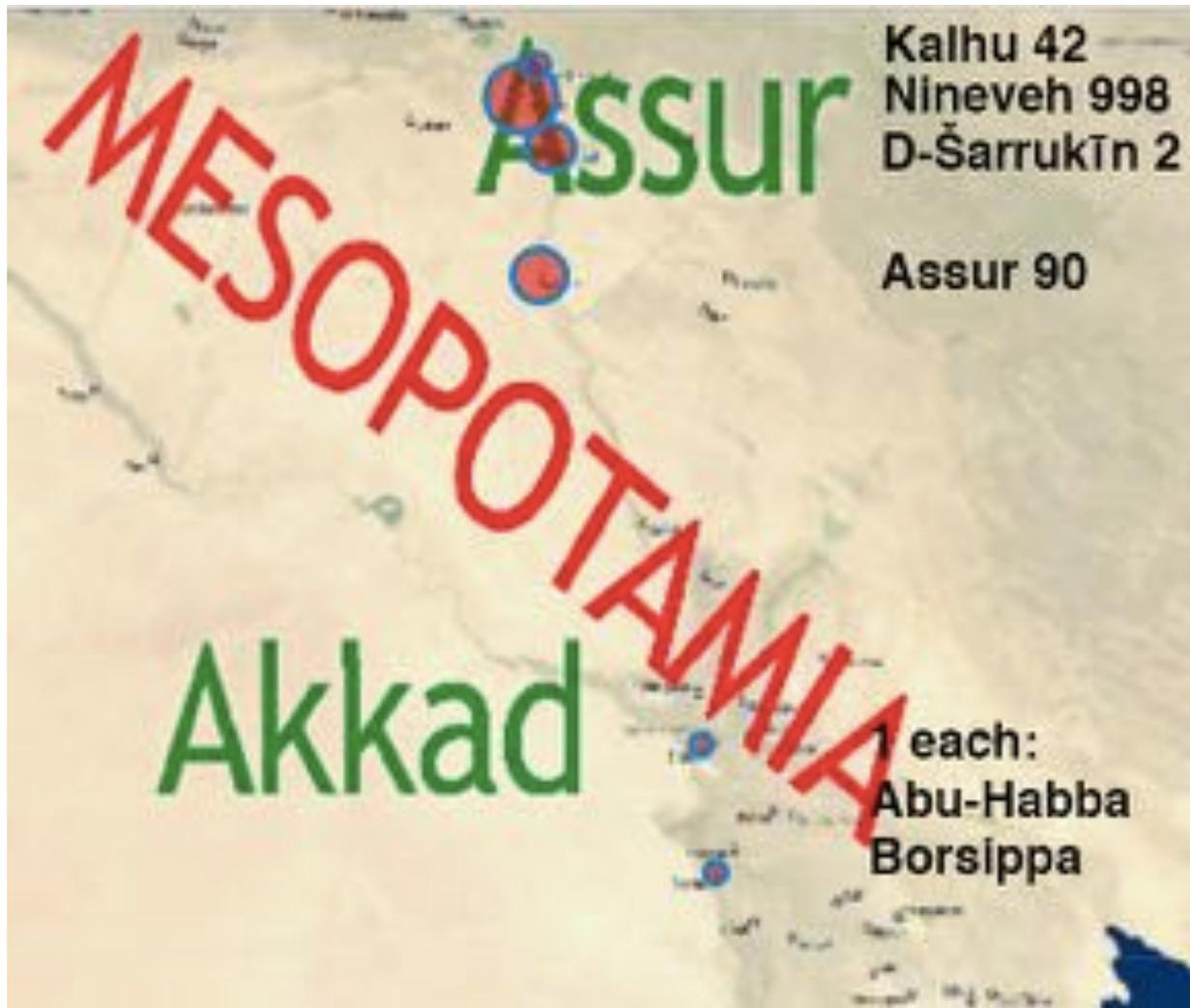
- SNA letters, inscriptions
- vocabulary usage pools in which corpus
- n-grams across the corpus, vocab clustering, corpus typical phraseology
- lexical corpus incomplete in DCCLT

Corpus building

- lexical corpus incomplete in DCCLT (oracc.museum.upenn.edu/dcclt)

The screenshot shows the website's navigation and content. At the top is a blue header with the title "Digital Corpus of Cuneiform Lexical Texts". Below it is a "Main menu" section with a yellow underline. The menu items are "Home", "Lexical Lists: Periods" (highlighted in yellow), and "Lexical Lists: Typology". To the right of the menu is a breadcrumb trail: "Home » Lexical Lists: Periods » Assyrian". The main content area is titled "Assyrian Lexical Texts" and contains the text "FORTHCOMING".

Clusters of lexical texts: NA

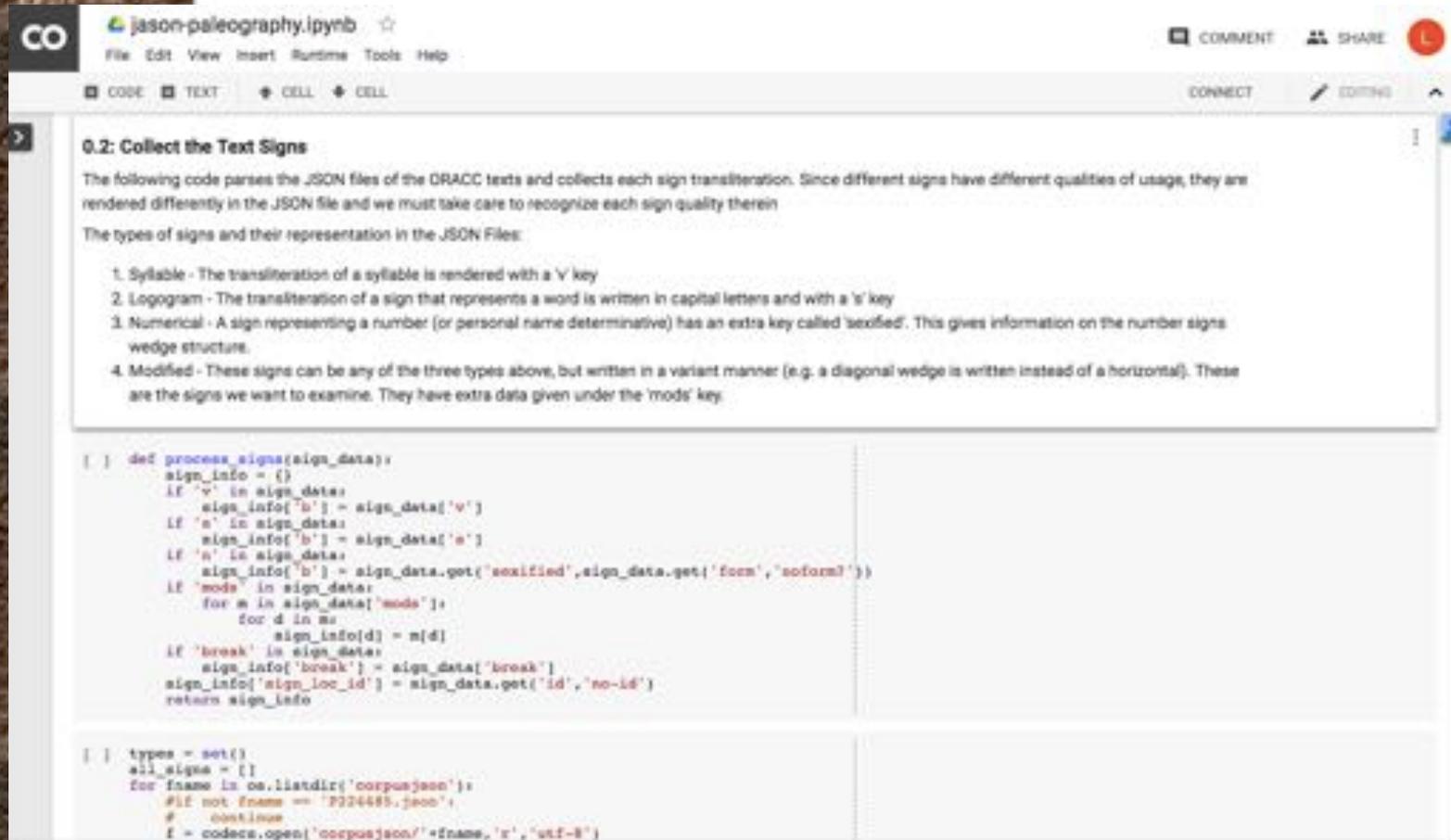


Reproducibility and replicability

- data
- code
- documentation

Reproducibility and replicability

- code
- documentation



The screenshot shows a Jupyter Notebook interface with the title 'jason-paleography.ipynb'. The notebook content includes a section titled '0.2: Collect the Text Signs' with explanatory text and a list of sign types. Below this is a Python code cell defining a function 'process_signs' and a loop that iterates over files in a directory named 'corpusjson'.

```
[ ] def process_signs(sign_data):
    sign_info = {}
    if 'v' in sign_data:
        sign_info['b'] = sign_data['v']
    if 's' in sign_data:
        sign_info['b'] = sign_data['s']
    if 'n' in sign_data:
        sign_info['b'] = sign_data.get('sexified', sign_data.get('form', 'soform'))
    if 'mods' in sign_data:
        for m in sign_data['mods']:
            for d in m:
                sign_info[d] = m[d]
    if 'break' in sign_data:
        sign_info['break'] = sign_data['break']
    sign_info['sign_loc_id'] = sign_data.get('id', 'no-id')
    return sign_info

[ ] types = set()
all_signs = []
for fname in os.listdir('corpusjson'):
    #if not fname == '7224485.json':
    #    continue
    f = codecs.open('corpusjson/'+fname, 'r', 'utf-8')
```

Future directions: astronomy

- plotting sky images, comparisons to LAS
- identify / make transparent LAS assumptions
- collect scientific critique(s) of tools

Future directions: paleography

- Parpola's methods not reproducible
- selective paleographic annotations
- reread every single NA letter, apply annotations
- review A dossiers: base parameters
- without ground truth or training set
can't do computational analysis:
lacking sufficient data set size

Future directions

Social network analysis:

**Corpus building /
annotation @berkeley and
@lmu.edu**



Thank you!

@berkeley.edu

gilbreger, jason_moser, lpearce,
pschmitz, joanne.tan, veldhuis

@lmu.de

nathan.morello, Jamie.Novotny,
K.Radner, F.Weiershaeuser