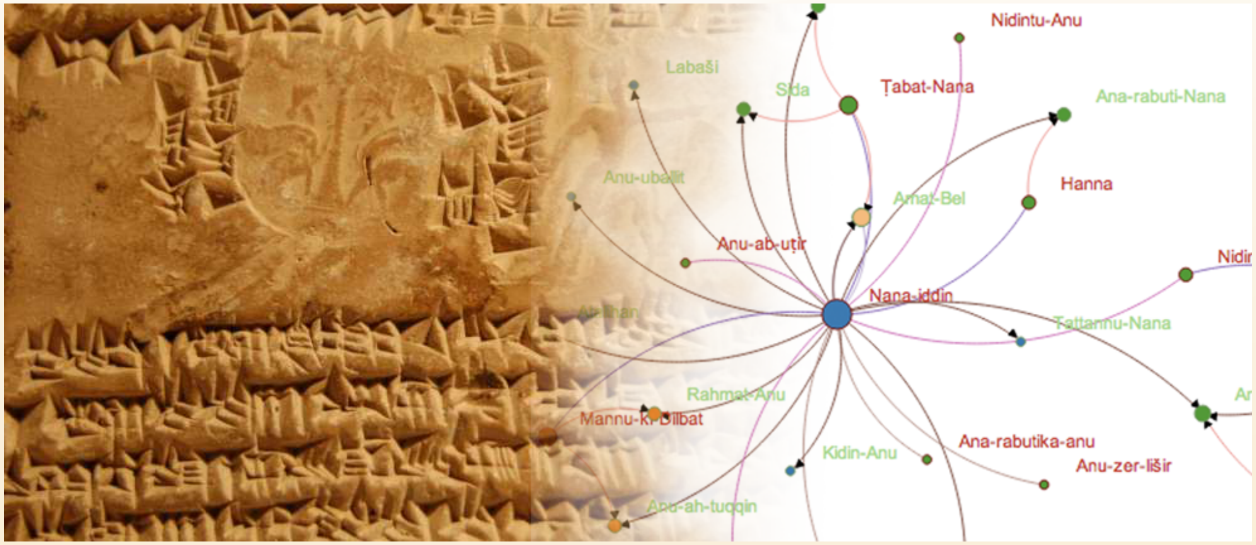
Berkeley Prosopography Services



Ancient Families, Modern Tools

Patrick Schmitz, UCB IST/RIT Dr. Laurie Pearce, UCB NES

DH-Case 2013 (ACM DocEng)
Florence, Italy
10 September 2013

What is Prosopography?

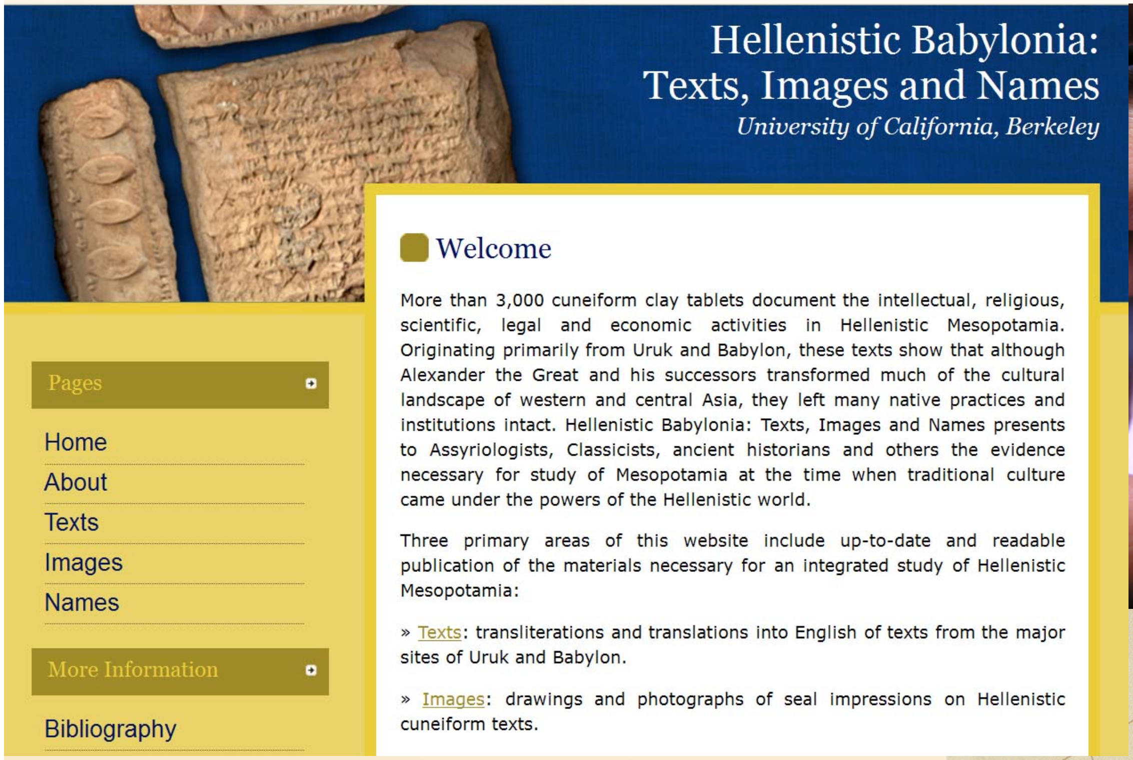
Goals:

- Identify people referenced in corpora: onomasticon
- 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 Content: Hellenistic Uruk



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

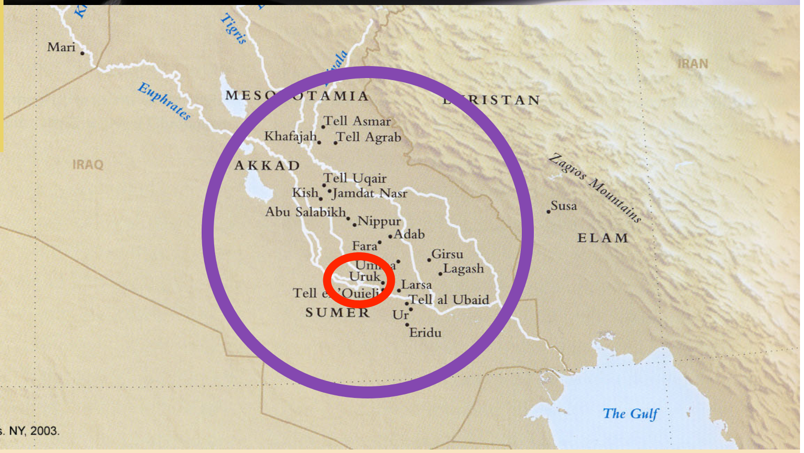
Welcome

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.



J. Aruz, ed. Art of the First Cities. NY, 2003.

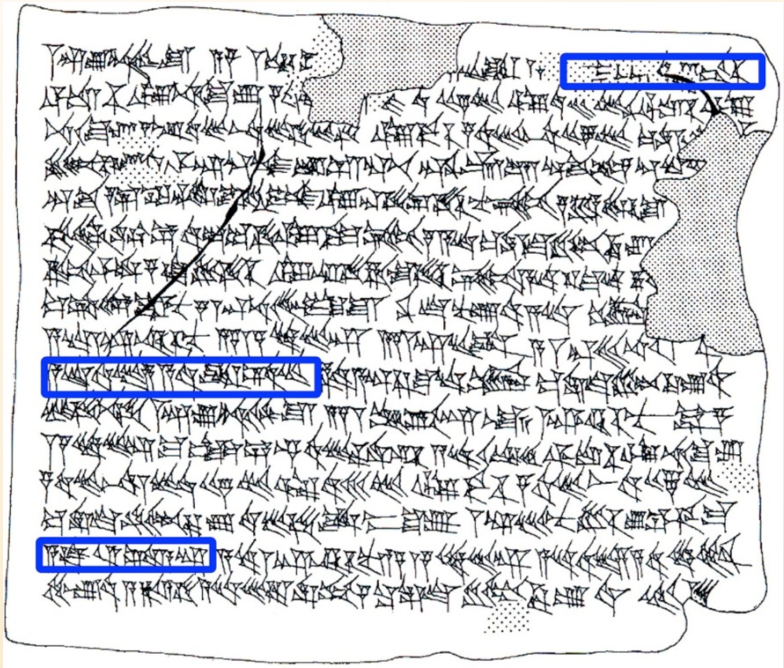
530	**legal texts**
8-20	**name citations/text**
3	**individuals/citation**
10,000	**name instances**

NATIONAL ENDOWMENT FOR THE Humanities

UC BERKELEY IST

Data Mining in Uruk Legal Texts

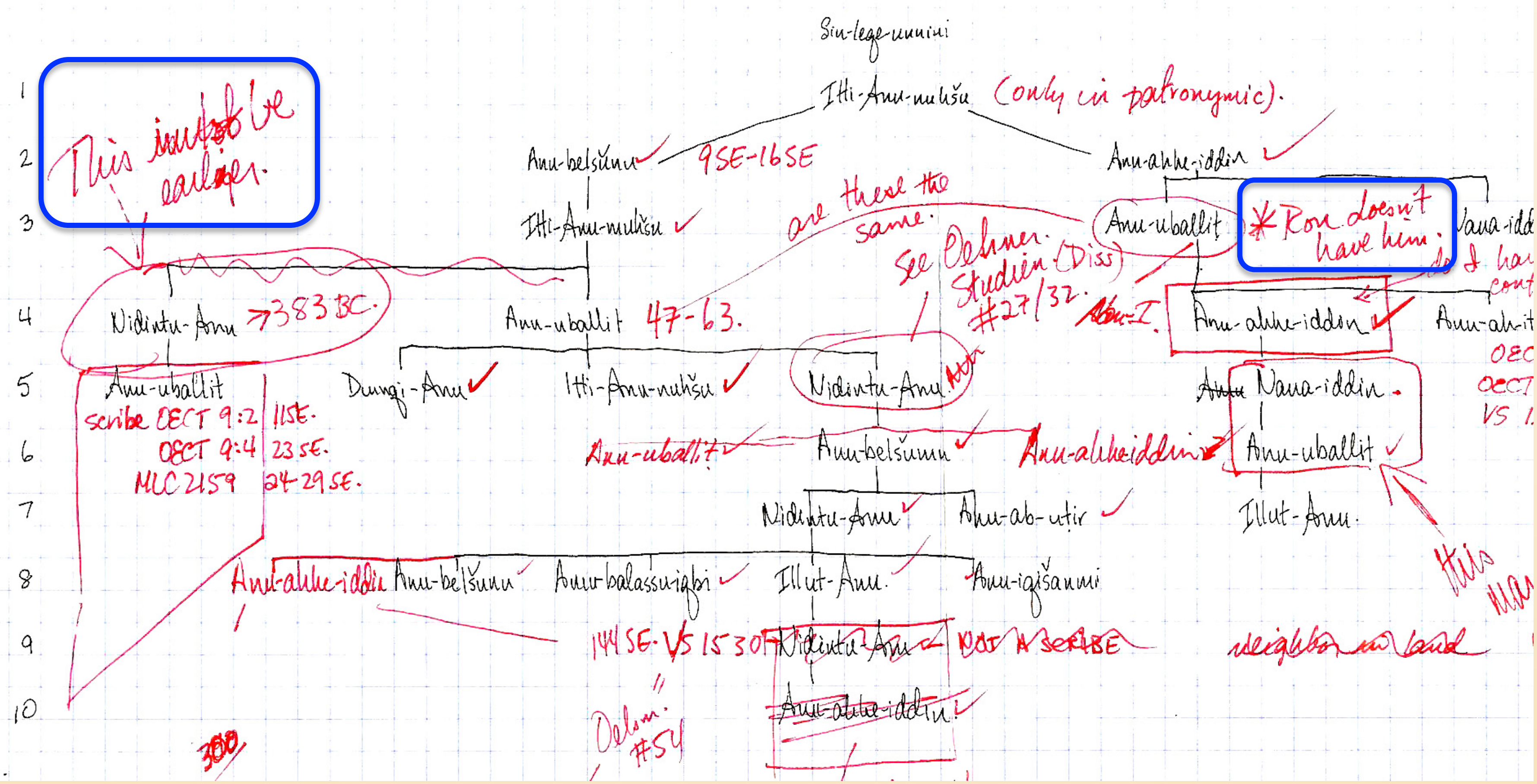
- Boilerplate text
  - repetitive patterns
  - attributes
  - many names!



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

Ancient Texts, Ancient Tools

It takes a lot of time to disambiguate these names and establish the social networks



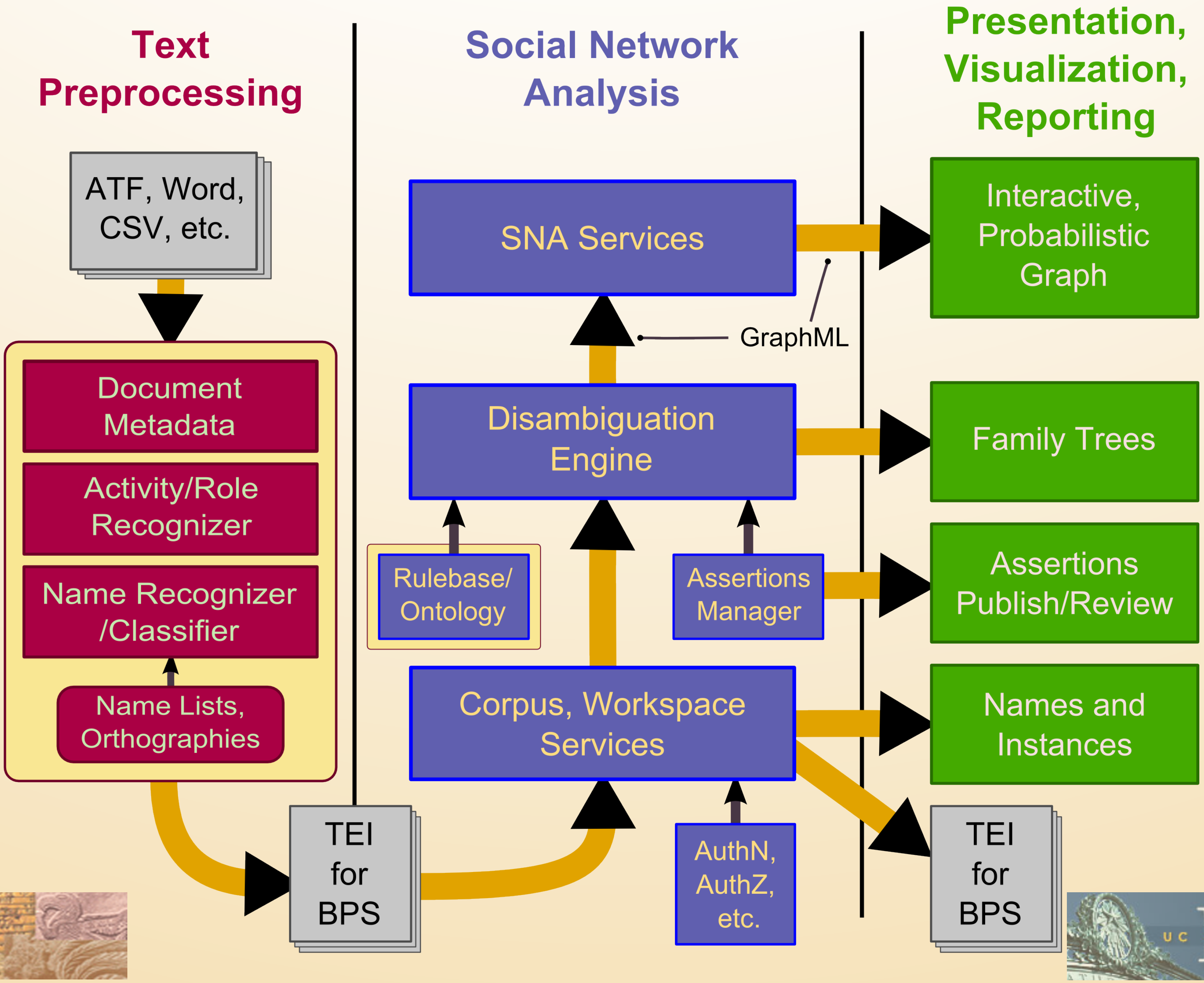
BPS innovations

- Probabilistic model of disambiguation, with extensible, heuristic rules
- Assertions overlay computed model, support uncertainty and disagreement
- Workspaces support hypotheses, build community, track authority
- Digital Humanities application built with software engineering best practices.

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

BPS System Architecture



Text Preprocessing

ATF, Word, CSV, etc.

Document Metadata

Activity/Role Recognizer

Name Recognizer /Classifier

Name Lists, Orthographies

TEI for BPS

Social Network Analysis

SNA Services

GraphML

Disambiguation Engine

Rulebase/ Ontology

Assertions Manager

Corpus, Workspace Services

AuthN, AuthZ, etc.

Presentation, Visualization, Reporting

Interactive, Probabilistic Graph

Family Trees

Assertions Publish/Review

Names and Instances

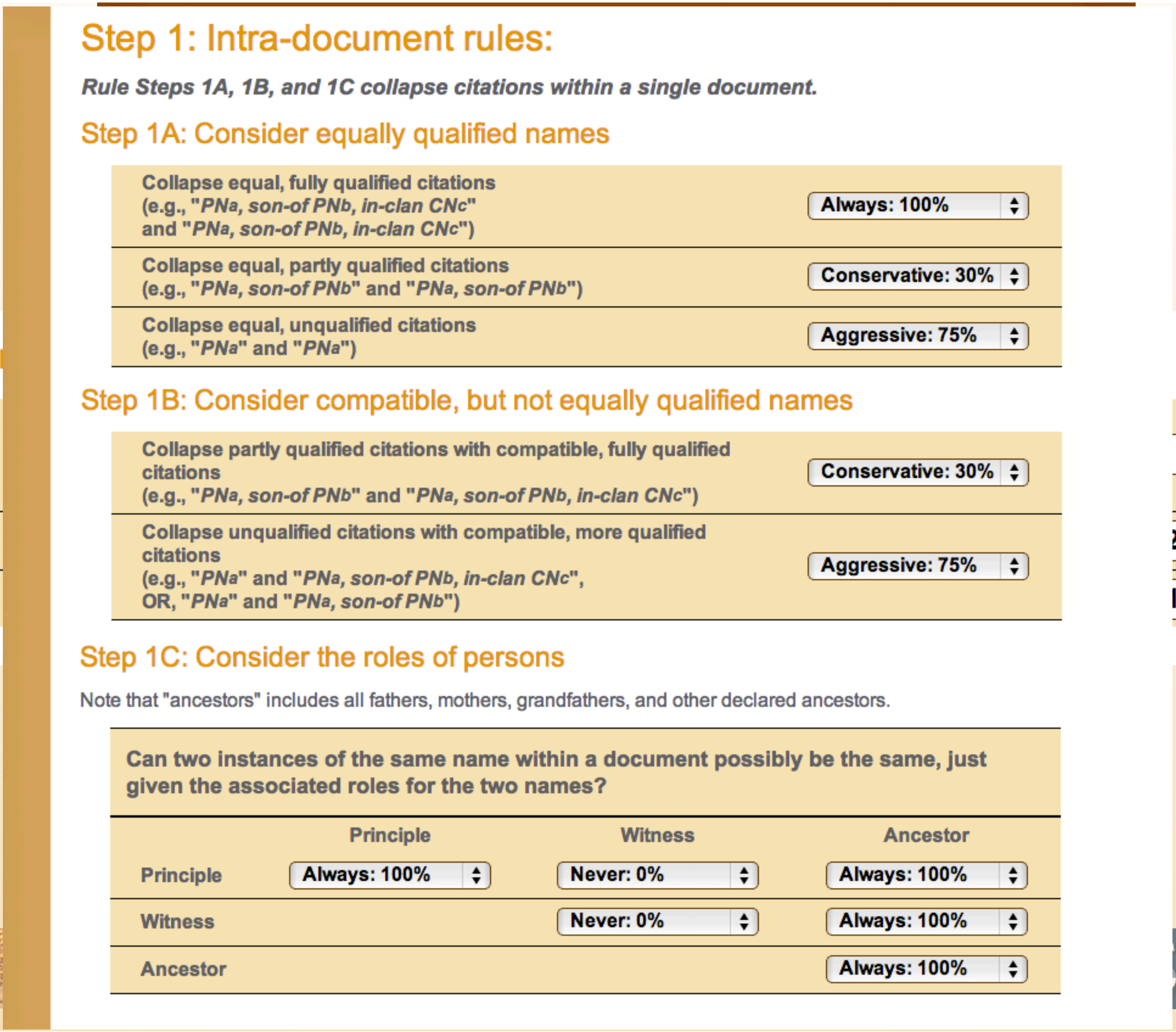
TEI for BPS

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



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., "*PNa, son-of PNb, in-clan CNc*"** **and "*PNa, son-of PNb, in-clan CNc*")**	Always: 100%
Collapse equal, partly qualified citations **(e.g., "*PNa, son-of PNb*" and "*PNa, son-of PNb*")**	Conservative: 30%
Collapse equal, unqualified citations **(e.g., "*PNa*" and "*PNa*")**	Aggressive: 75%

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

Collapse partly qualified citations with compatible, fully qualified citations **(e.g., "*PNa, son-of PNb*" and "*PNa, son-of PNb, in-clan CNc*")**	Conservative: 30%
Collapse unqualified citations with compatible, more qualified citations **(e.g., "*PNa*" and "*PNa, son-of PNb, in-clan CNc*",** **OR, "*PNa*" and "*PNa, son-of PNb*")**	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%

Assertions

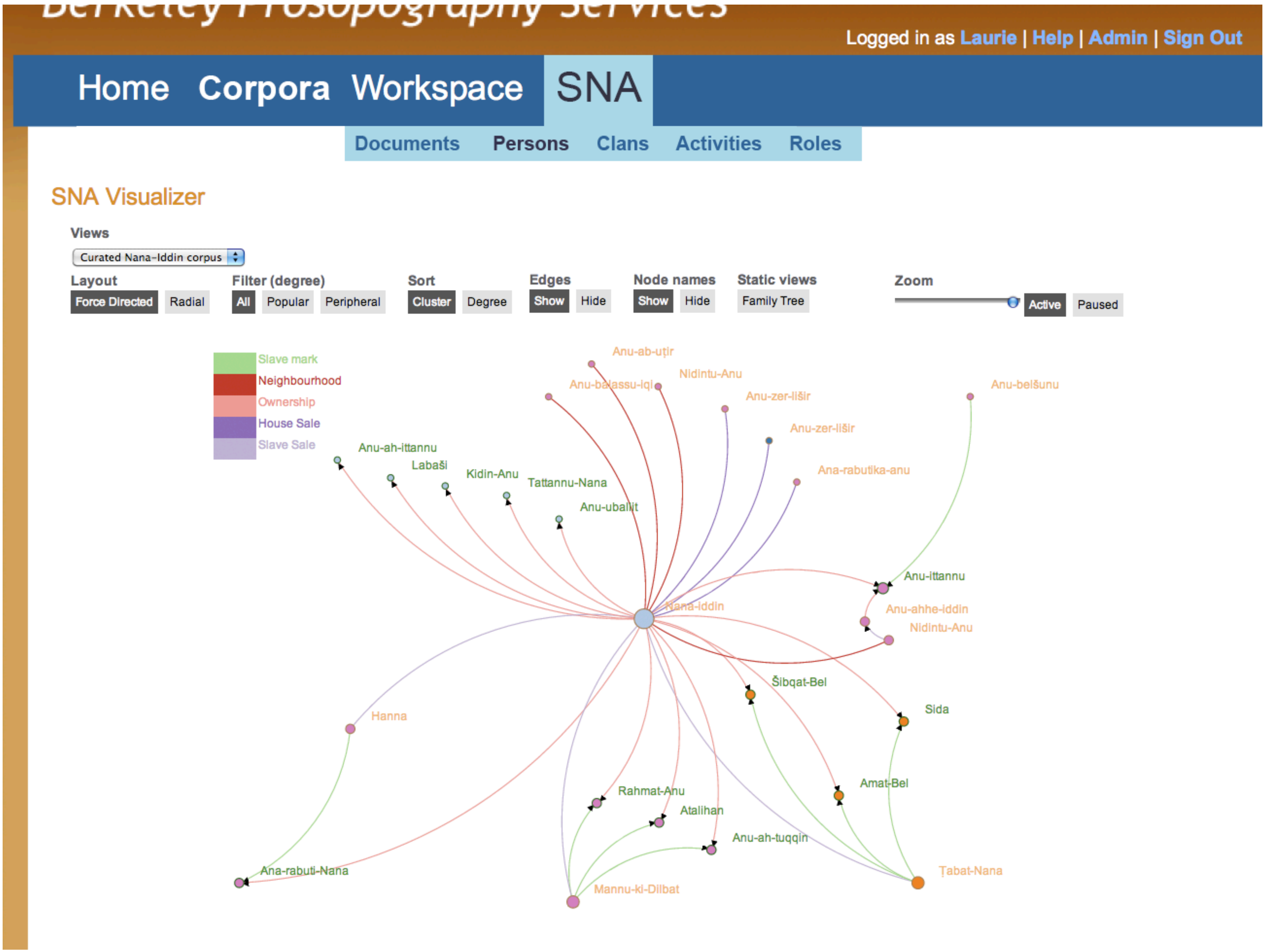
- Assertions integrate directly into model
  - Override disambiguation results
  - Control model (rule) parameters
- Assertions encapsulate judgment by user
- Assertions are sharable
  - Publish-from/Consume-into workspaces
- Assertions expose user-facing aspects (meta-data)
  - Natural language description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

Workspaces

- Gather together corpus and a set of assertions
- Enable experimentation
- Enable collaboration and sharing
    - Some researchers are trusted by others, some are not
    - Students would greatly benefit from seeing the reasoning steps of established researchers
    - Basically, a problem of community curation of a shared resource, but maintaining idiosyncratic views.

Social network analysis

- Services
  - SNA engine (computes metrics, features)
  - Filters and pivots to refine scope
  - Visualization kit
- Features
  - Support probabilistic network
  - Axes and features derived from data
  - Support any incoming data (GraphML), regardless of semantics


Berkeley Prosopography Services
Logged in as Laurie | Help | Admin | Sign Out
Home
Corpora
Workspace
SNA
Documents
Persons
Clans
Activities
Roles

SNA Visualizer

Views: Curated Nana-Iddin corpus

Layout: Force Directed | Radial

Filter (degree): All | Popular | Peripheral

Sort: Cluster | Degree

Edges: Show | Hide

Node names: Show | Hide

Static views: Family Tree

Zoom: Active | Paused

Lessons learned

- Few real standards for corpus exchange
  - TEI useful, but not really standard
  - GraphML *is* widely supported
- Traditional NLP/ML not always appropriate
- Can take a long time to communicate effectively
- Foster serendipity across domains
  - Adding “random” features to play with can benefit
  - Sketches, “confusing” discussions spur revelation

Questions, discussion

http://www.berkeleyprosopography.org

Links below available from About page of site.

- 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)