MARI:

Semi-Automated, Human-in-the-Loop Redaction of Text Corpora

Emma I. C. Peterson, Valerie Zhao, Dan Byrne, Blase Ur
University of Chicago

Motivation

- Sharing qualitative, naturalistic data for the benefit of multiple research groups is difficult
- Protecting participants: redacting identifiable information
  - ○ Most existing *automatic* redaction tools are insufficient
    - ■ Balancing privacy (redacting information) vs research contribution (leaving data untouched)
  - ○ Manual redaction is impractical due to budget and time constraints

Contributions

- **MARI (Mostly Automated Redaction of Identifiers),** a prototype **human-in-the-loop** tool informed by interviews with social science researchers
- Interview findings on redaction and research methods
- Unique combination of PII feature engineering, linguistic analysis, and information-theoretic scoring
  - ○ *Goal:* maximize data utility, cross-discipline generalization, comprehensive redaction coverage

Taxonomy and Examples

Taxonomy Category	Taxonomy Sub-Categories	Examples
Identifiers	Personal Names, Nicknames, Personal Identity, Numbers	*Legally given name, Diminutives, SSN, EIN*
Demographics	Age, Sex, Gender, Pronouns, Sexuality, Race & Ethnicity, Education, Profession, Health Status	*Date of Birth, LQBTQ+, Niche Job, Rare Disease, HIV Positive, Mixed Race*
Locations	Country, State, City, Postal Code, Address, Landmark, Business	*United States, Illinois, Chicago, 5307 S Woodlawn Ave, The Bean, Jimmy's Tap*
Dates & Events	Publicly Recognized Dates & Events, Personal Dates & Events	*Thanksgiving, Christmas, Cancer-Remission Anniversary*
Linguistic Patterns	Regional Dialects, Code-Switching, Unique Vocabularies, Idiosyncratic Expressions	*African American Vernacular English, Scots-English, Spanglish, Parmesan Cheese == "Pasta Sugar"*
Personal Interests & Activities	Traditions, Group Membership, Cultural References, Popular Culture Participation, Hobbies	*University / school traditions, belonging to native tribe or military, member of a small fandom*

Table 1. Our privacy taxonomy consists of hierarchical categories and highlights certain redaction decision thresholds. Classifiers will be constructed per category and tuned to different data presentation types.

System Implementation

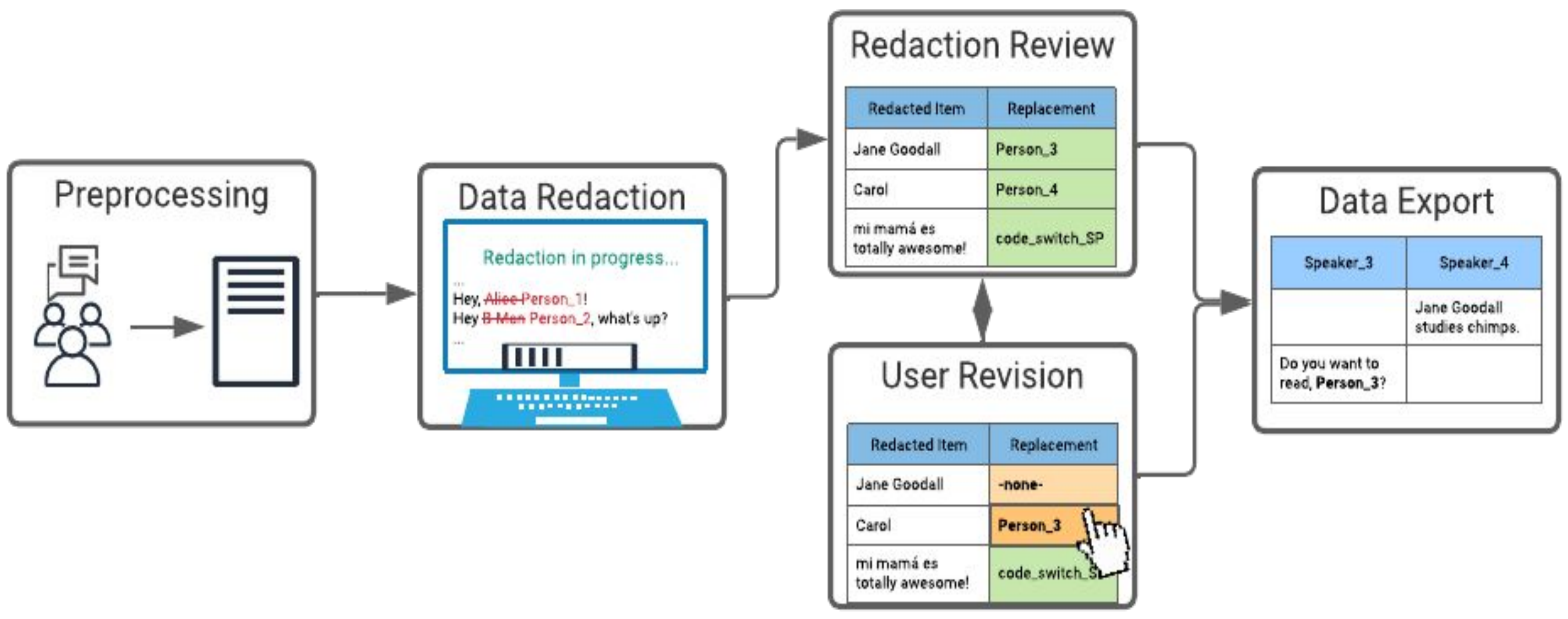

Preprocessing
Data Redaction
Redaction in progress...
...
Hey, ~~Alice~~ Person_1!
Hey ~~B-Man~~ Person_2, what's up?
...
Redaction Review
Redacted Item | Replacement
Jane Goodall | Person_3
Carol | Person_4
mi mamá es totally awesome! | code_switch_SP
User Revision
Redacted Item | Replacement
Jane Goodall | -none-
Carol | Person_3
mi mamá es totally awesome! | code_switch_SP
Data Export
Speaker_3 | Speaker_4
Jane Goodall studies chimps.
Do you want to read, Person_3?

Figure 1. Application flow diagram. The user uploads data, which is then redacted by our system; the user can then revise the automated decisions.

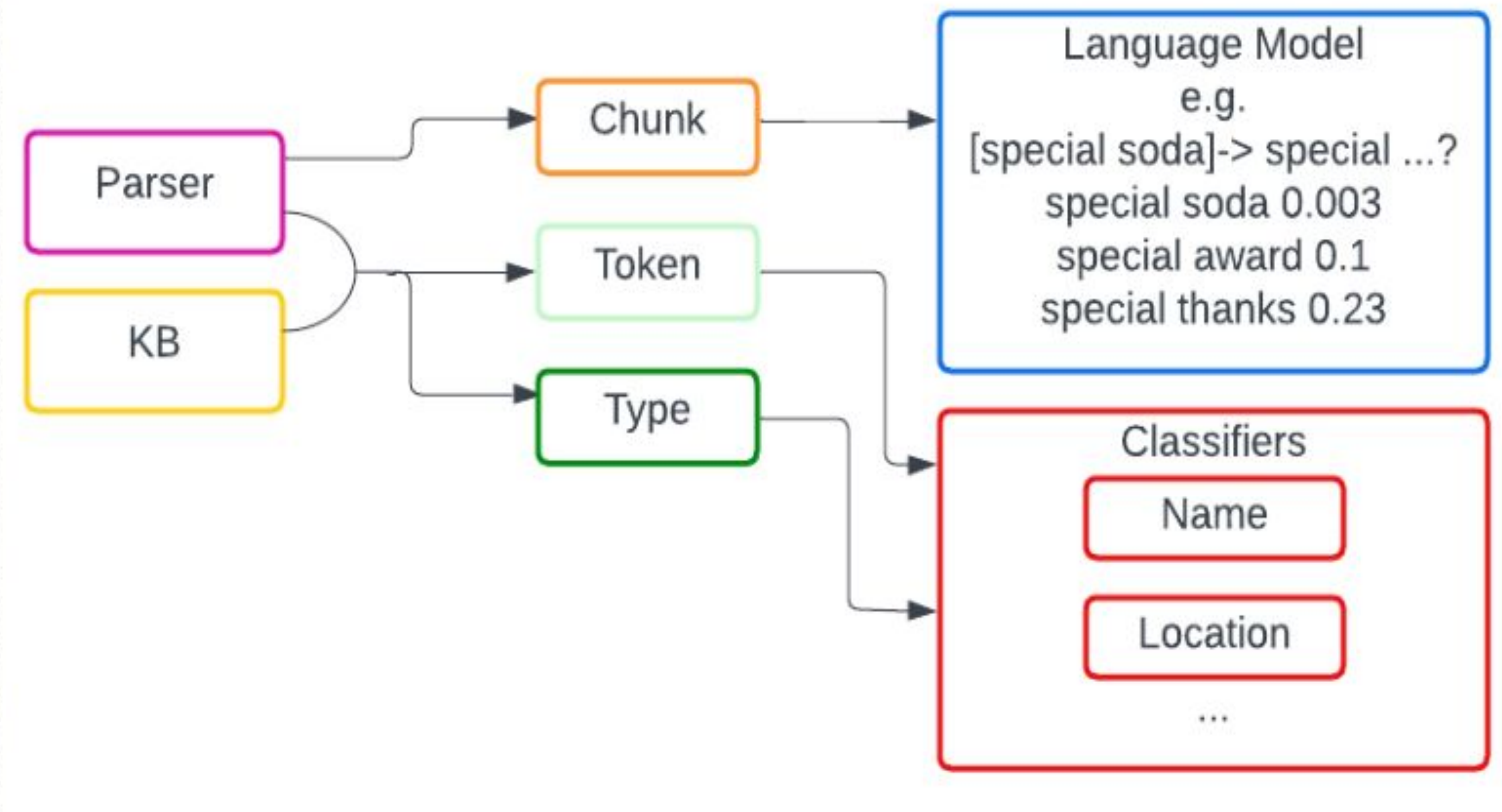


Figure 3. Redaction flow diagram. Data is parsed into different structures. Tokens and types are given to classifiers, whereas chunks are analyzed by language modeling.

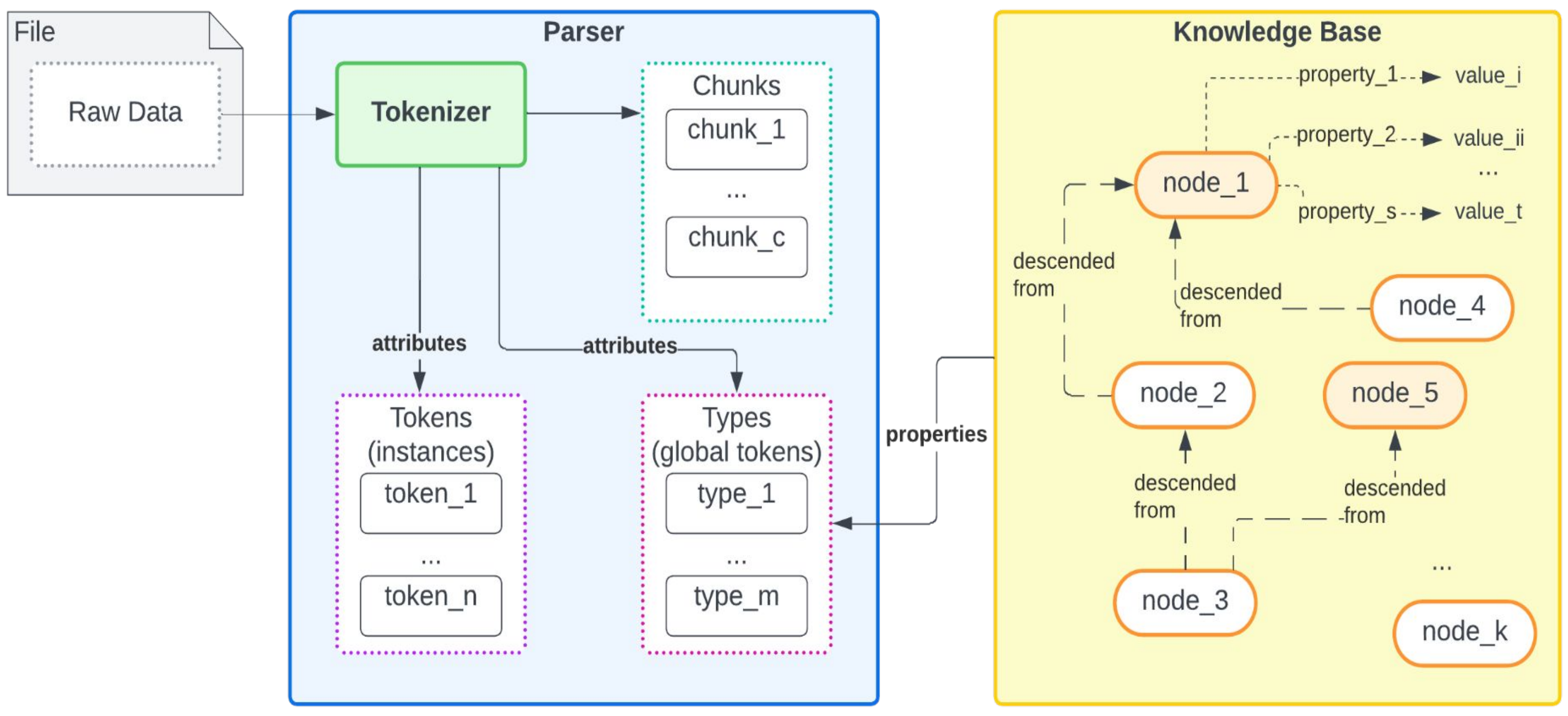
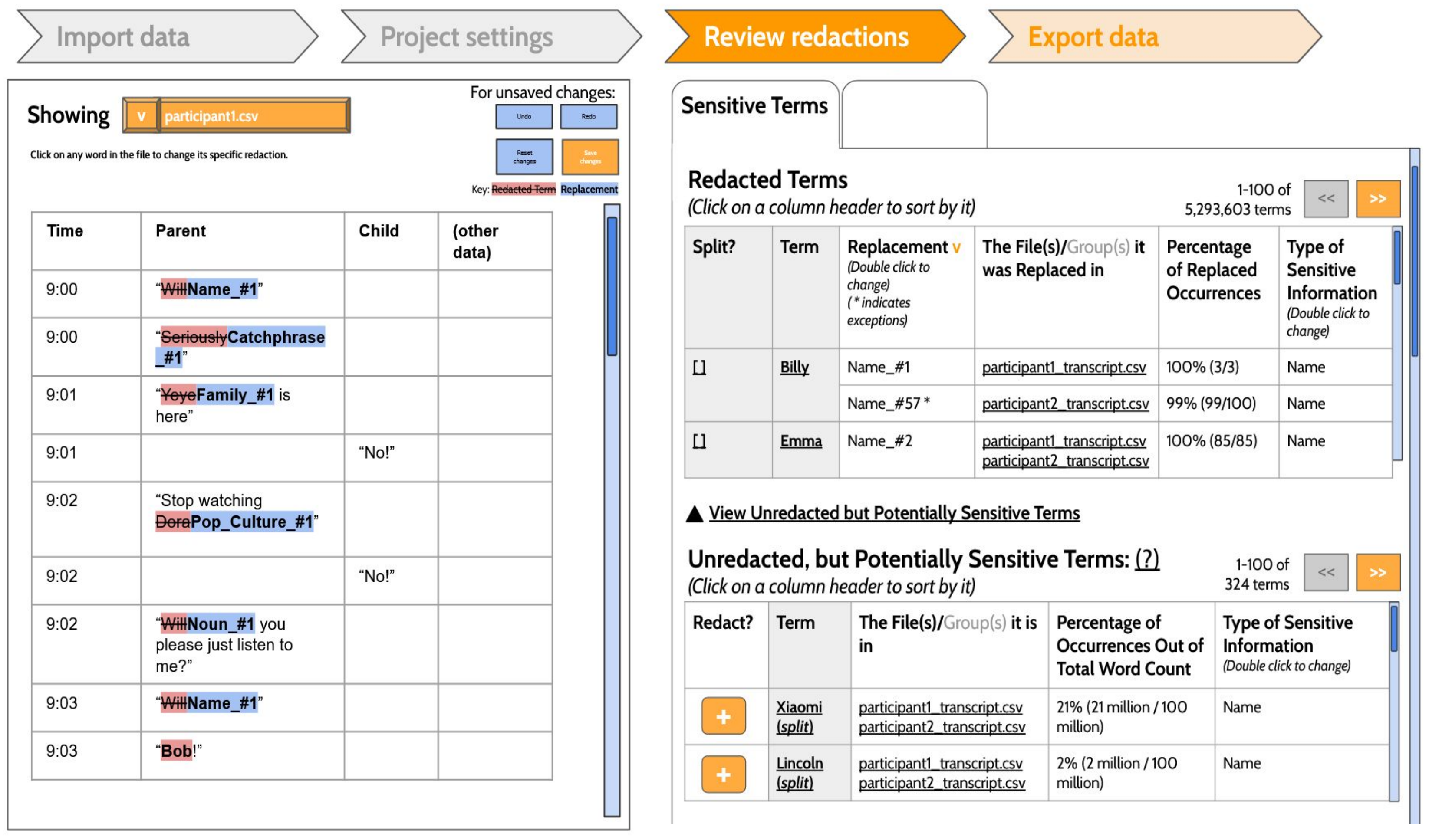


Figure 2. Overview of application pre-redaction processing and connection of types to knowledge base.



Showing participant1.csv

Click on any word in the file to change its specific redaction.

For unsaved changes:

Key: ~~Redacted Term~~ Replacement

Time	Parent	Child	(other data)
9:00	"~~Will~~Name_#1"		
9:00	"~~Seriously~~Catchphrase_#1"		
9:01	"~~Yeye~~Family_#1 is here"		
9:01		"No!"	
9:02	"Stop watching ~~Dora~~Pop_Culture_#1"		
9:02		"No!"	
9:02	"~~Will~~Noun_#1 you please just listen to me?"		
9:03	"~~Will~~Name_#1"		
9:03	"Bob!"		

Sensitive Terms

Redacted Terms
(Click on a column header to sort by it)

1-100 of 5,293,603 terms << >>

Split?	Term	Replacement v (Double click to change) (* indicates exceptions)	The File(s)/Group(s) it was Replaced in	Percentage of Replaced Occurrences	Type of Sensitive Information (Double click to change)
[]	Billy	Name_#1	participant1_transcript.csv	100% (3/3)	Name
		Name_#57 *	participant2_transcript.csv	99% (99/100)	Name
[]	Emma	Name_#2	participant1_transcript.csv participant2_transcript.csv	100% (85/85)	Name

▲ View Unredacted but Potentially Sensitive Terms

Unredacted, but Potentially Sensitive Terms: (?)
(Click on a column header to sort by it)

1-100 of 324 terms << >>

Redact?	Term	The File(s)/Group(s) it is in	Percentage of Occurrences Out of Total Word Count	Type of Sensitive Information (Double click to change)
+	Xiaomi (split)	participant1_transcript.csv participant2_transcript.csv	21% (21 million / 100 million)	Name
+	Lincoln (split)	participant1_transcript.csv participant2_transcript.csv	2% (2 million / 100 million)	Name

Figure 4. UI mockup displaying redacted information to the user. Redacted terms are shown to the user in context both file-wise and term-wise.

This material is based upon work supported by the National Science Foundation under Grant No. CNS-2210193 and by a Meta Privacy-Enhancing Technologies Research Award. We thank Susan Goldin-Meadow, Marisa Casillas, Chenhao Tan, and Ruthe Foushee for helping to guide the project's direction, as well as Oliver Cai and Brianna Zeng for contributing to the implementation of MARI.