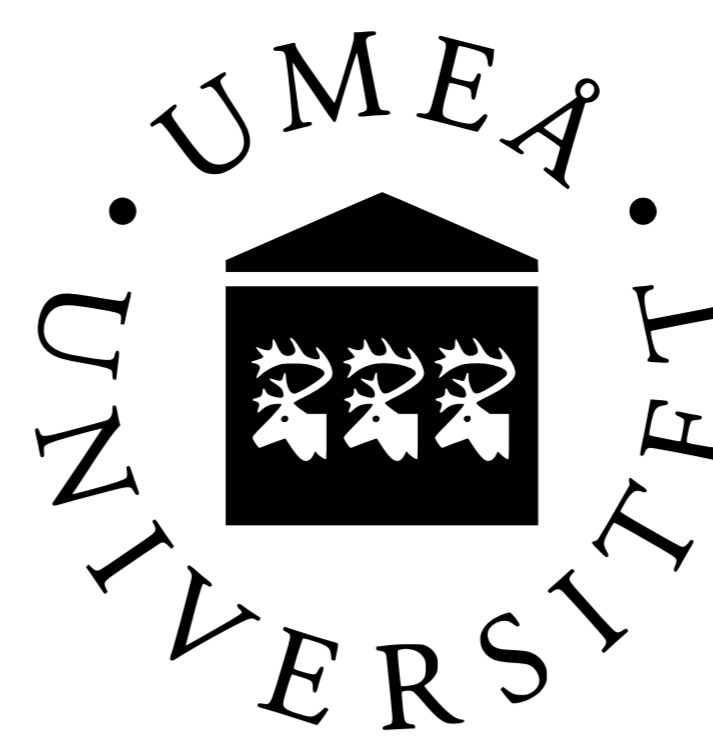


Pseudonymization Categories across Domain Boundaries



Maria Irena Szawerna¹, Simon Dobnik², Ricardo Muñoz Sánchez¹, Therese Lindström Tiedemann³, Xuan-Son Vu⁴, and Elena Volodina¹

¹Språkbanken Text, University of Gothenburg, Sweden, ²CLASP, University of Gothenburg, Sweden,

³University of Helsinki, Finland, ⁴Umeå University, Sweden and DeepTensor AB

Background

- **Personal Identifiable Information (PII)** complicates sharing linguistic data.
- We explore existing methods mitigating this issue with a focus on **pseudonymization**.
- We also identify types of PII in corpora representing various domains.

Questions

1. What are the differences in PII between tagsets and domains?
2. Is a universal tagset possible?

Materials and Methods

We have chosen to investigate eight different tagsets to compare the kinds of tags appearing in them:

ID	Paper	Domain
Anonymization		
1	Adams et al. (2019)	Chat
2	Pilán et al. (2022)	Legal
3	Accorsi et al. (2012)	SMS
4	Bråthen et al. (2021)	Medical
Pseudonymization		
5	Megyesi et al. (2018, 2021)	L2 essays
6	Eder et al. (2019, 2020, 2022)	E-mail
7	Alfalahi et al. (2012)	Medical
8	Dalianis (2019)	Medical

Table 1: The analyzed tagsets and their domains.

When attempting to determine the applicability of one of those tagsets we chose to try to annotate data from the following domains and sources:

ID	Source	Domain	Language
A	Private	Medical	Swedish
B	Enron Corp and Cohen	E-mails	English
C	Pilán et al. (2022)	Legal	English
D	Szawerna (2023)	Memoir	Polish
E	Ahrenberg et al. (2020)	Blogs	Swedish
F	Twitter Mix, n/a, (2020)	Tweets	Swedish
G	Ahrenberg et al. (2020)	Web news	Swedish
H	Volodina et al. (2022)	L2 Essays	Swedish
I	McAuley and Leskovec (2013)	Reviews	English

Table 2: The analyzed datasets, their domains and languages.

Results

- There is variety in what is covered by various tagsets and at what level of detail.
- Some PII appear almost universally (e.g. city names, personal names), while other types can be domain-specific (e.g. urls) or simply rare (e.g. middle names).
- Many tagsets have a heterogenous *miscellaneous* category.
 - Nevertheless, there are types of PII that are potentially frequent enough to warrant their own categories, e.g. hashtags or events.
- The distribution of PII categories varies across domains.

Tag	Domains
firstname_male	1, 2, 3, 4, 5, 6, 7, 8
firstname_female	1, 2, 3, 4, 5, 6, 7, 8
firstname_unknown	1, 2, 3, 4, 5, 6, 7, 8
initials	1, 2, 3, 4, 5, 6, 7, 8
middlename	1, 2, 3, 4, 5, 6, 7, 8
surname	1, 2, 3, 4, 5, 6, 7, 8
foreign	1, 2, 3, 4, 5, 6, 7, 8
area	1, 2, 3, 4, 5, 6, 7, 8
city	1, 2, 3, 4, 5, 6, 7, 8
geo	1, 2, 3, 4, 5, 6, 7, 8
country	1, 2, 3, 4, 5, 6, 7, 8
place	1, 2, 3, 4, 5, 6, 7, 8
region	1, 2, 3, 4, 5, 6, 7, 8
street_nr	1, 2, 3, 4, 5, 6, 7, 8
zip_code	1, 2, 3, 4, 5, 6, 7, 8
school	1, 2, 3, 4, 5, 6, 7, 8
work	1, 2, 3, 4, 5, 6, 7, 8
other_institution	1, 2, 3, 4, 5, 6, 7, 8
transport_name	1, 2, 3, 4, 5, 6, 7, 8
transport_nr	1, 2, 3, 4, 5, 6, 7, 8
age_digits	1, 2, 3, 4, 5, 6, 7, 8
age_string	1, 2, 3, 4, 5, 6, 7, 8
date_digits	1, 2, 3, 4, 5, 6, 7, 8
day	1, 2, 3, 4, 5, 6, 7, 8
month_digit	1, 2, 3, 4, 5, 6, 7, 8
month_word	1, 2, 3, 4, 5, 6, 7, 8
year	1, 2, 3, 4, 5, 6, 7, 8
phone_nr	1, 2, 3, 4, 5, 6, 7, 8
email	1, 2, 3, 4, 5, 6, 7, 8
url	1, 2, 3, 4, 5, 6, 7, 8
personid_nr	1, 2, 3, 4, 5, 6, 7, 8
account_nr	1, 2, 3, 4, 5, 6, 7, 8
license_nr	1, 2, 3, 4, 5, 6, 7, 8
other_nr_seq	1, 2, 3, 4, 5, 6, 7, 8
extra	1, 2, 3, 4, 5, 6, 7, 8
prof	1, 2, 3, 4, 5, 6, 7, 8
edu	1, 2, 3, 4, 5, 6, 7, 8
fam	1, 2, 3, 4, 5, 6, 7, 8
sensitive	1, 2, 3, 4, 5, 6, 7, 8

Table 3: SweLL tags and the papers (Table 1) they correspond to (in bold).

Tag	Domains
username	1, 2, 3, 4, 5, 6, 7, 8
password	1, 2, 3, 4, 5, 6, 7, 8
IP address	1, 2, 3, 4, 5, 6, 7, 8
product	1, 2, 3, 4, 5, 6, 7, 8
facility	1, 2, 3, 4, 5, 6, 7, 8
nationality	1, 2, 3, 4, 5, 6, 7, 8
work of art	1, 2, 3, 4, 5, 6, 7, 8
language	1, 2, 3, 4, 5, 6, 7, 8
unit	1, 2, 3, 4, 5, 6, 7, 8
med/chem entity	1, 2, 3, 4, 5, 6, 7, 8
sports team	1, 2, 3, 4, 5, 6, 7, 8
known group	1, 2, 3, 4, 5, 6, 7, 8
known figure	1, 2, 3, 4, 5, 6, 7, 8
fictional figure	1, 2, 3, 4, 5, 6, 7, 8
healthcare unit	1, 2, 3, 4, 5, 6, 7, 8
demographic attribute	1, 2, 3, 4, 5, 6, 7, 8
duration	1, 2, 3, 4, 5, 6, 7, 8
quantity, value	1, 2, 3, 4, 5, 6, 7, 8
nickname	1, 2, 3, 4, 5, 6, 7, 8
belief	1, 2, 3, 4, 5, 6, 7, 8
political views	1, 2, 3, 4, 5, 6, 7, 8
sexuality, gender identity	1, 2, 3, 4, 5, 6, 7, 8
ethnicity	1, 2, 3, 4, 5, 6, 7, 8
health	1, 2, 3, 4, 5, 6, 7, 8
patronymic/other name	1, 2, 3, 4, 5, 6, 7, 8

Table 4: Generic non-SweLL tags and the papers (Table 1) they correspond to (in bold).

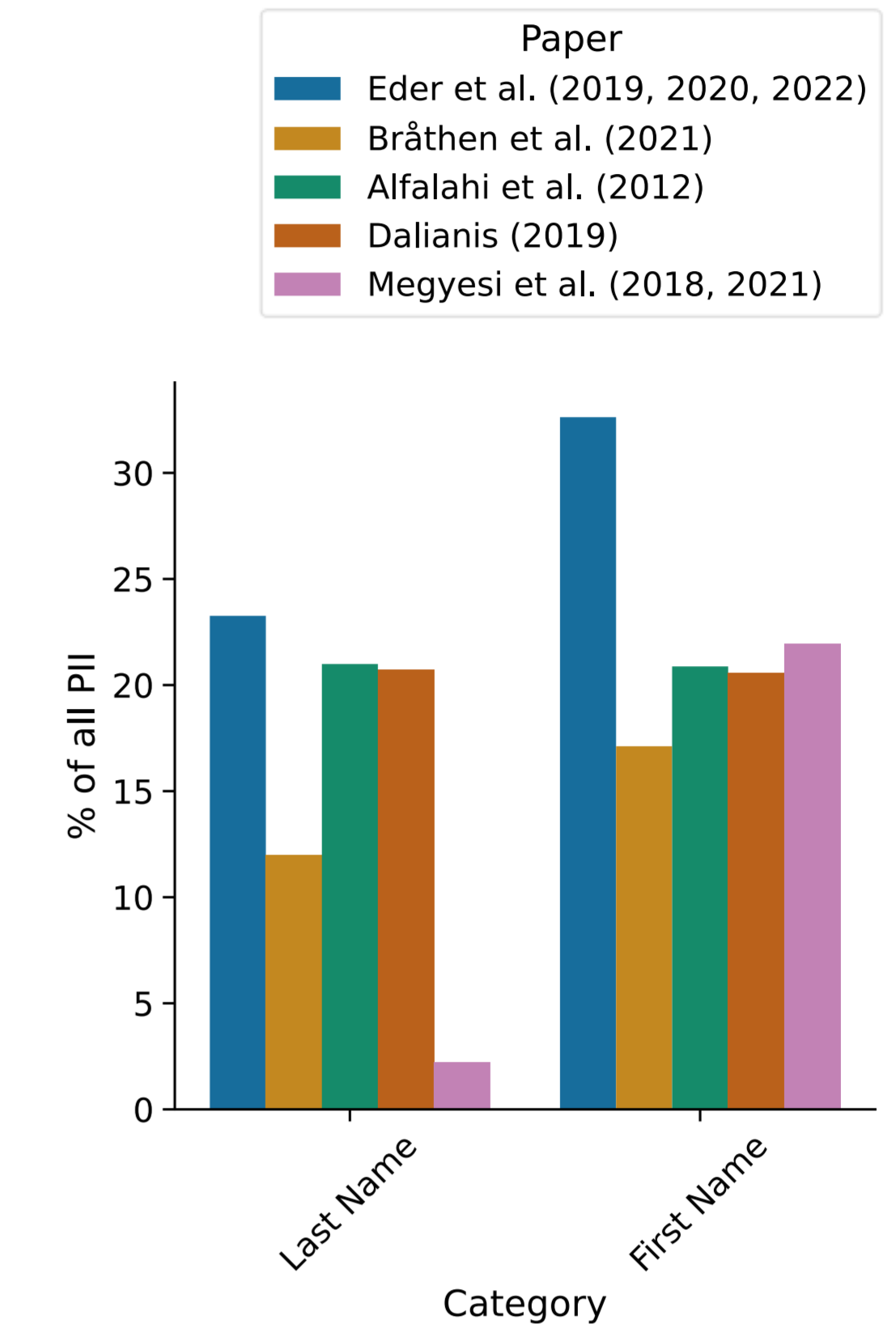


Figure 1: An example of the uneven tag distribution across domains.

Tag	Domains
firstname_male	A, B, C, D, E, F, G, H, I
firstname_female	A, B, C, D, E, F, G, H, I
firstname_unknown	A, B, C, D, E, F, G, H, I
initials	A, B, C, D, E, F, G, H, I
middlename	A, B, C, D, E, F, G, H, I
surname	A, B, C, D, E, F, G, H, I
foreign	A, B, C, D, E, F, G, H, I
area	A, B, C, D, E, F, G, H, I
city	A, B, C, D, E, F, G, H, I
geo	A, B, C, D, E, F, G, H, I
country	A, B, C, D, E, F, G, H, I
place	A, B, C, D, E, F, G, H, I
region	A, B, C, D, E, F, G, H, I
street_nr	A, B, C, D, E, F, G, H, I
zip_code	A, B, C, D, E, F, G, H, I
school	A, B, C, D, E, F, G, H, I
work	A, B, C, D, E, F, G, H, I
other_institution	A, B, C, D, E, F, G, H, I
transport_name	A, B, C, D, E, F, G, H, I
transport_nr	A, B, C, D, E, F, G, H, I
age_digits	A, B, C, D, E, F, G, H, I
age_string	A, B, C, D, E, F, G, H, I
date_digits	A, B, C, D, E, F, G, H, I
day	A, B, C, D, E, F, G, H, I
month_digit	A, B, C, D, E, F, G, H, I
month_word	A, B, C, D, E, F, G, H, I
year	A, B, C, D, E, F, G, H, I
phone_nr	A, B, C, D, E, F, G, H, I
email	A, B, C, D, E, F, G, H, I
url	A, B, C, D, E, F, G, H, I
personid_nr	A, B, C, D, E, F, G, H, I
account_nr	A, B, C, D, E, F, G, H, I
license_nr	A, B, C, D, E, F, G, H, I
other_nr_seq	A, B, C, D, E, F, G, H, I
extra	A, B, C, D, E, F, G, H, I
prof	A, B, C, D, E, F, G, H, I
edu	A, B, C, D, E, F, G, H, I
fam	A, B, C, D, E, F, G, H, I
sensitive	A, B, C, D, E, F, G, H, I

Table 5: SweLL tags and the domains (Table 2) they correspond to (in bold).

Conclusions

- We strive towards a **universal tagset**, acknowledging that while none of the existing tagsets comprehensively covers all types of PII found in texts, it is not unreasonable to pursue such a standard.
- Ideally, the universal tagset would have a hierarchical structure
- Such a tagset would have to be regularly revised, in sync with new kinds of potentially personal information emerging.



For the full paper and the references, please scan the QR code

For additional information, please contact:
 Maria Irena Szawerna
 maria.szawerna@gu.se or mormor.karl@svenska.gu.se
<https://mormor-karl.github.io/>