SweLL
& legal aspects

Elena Volodina

WG5 meeting, Bolzano, September, 7, 2017

RIKSBANKENS JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

SweLL

Research infrastructure for Swedish as a Second Language

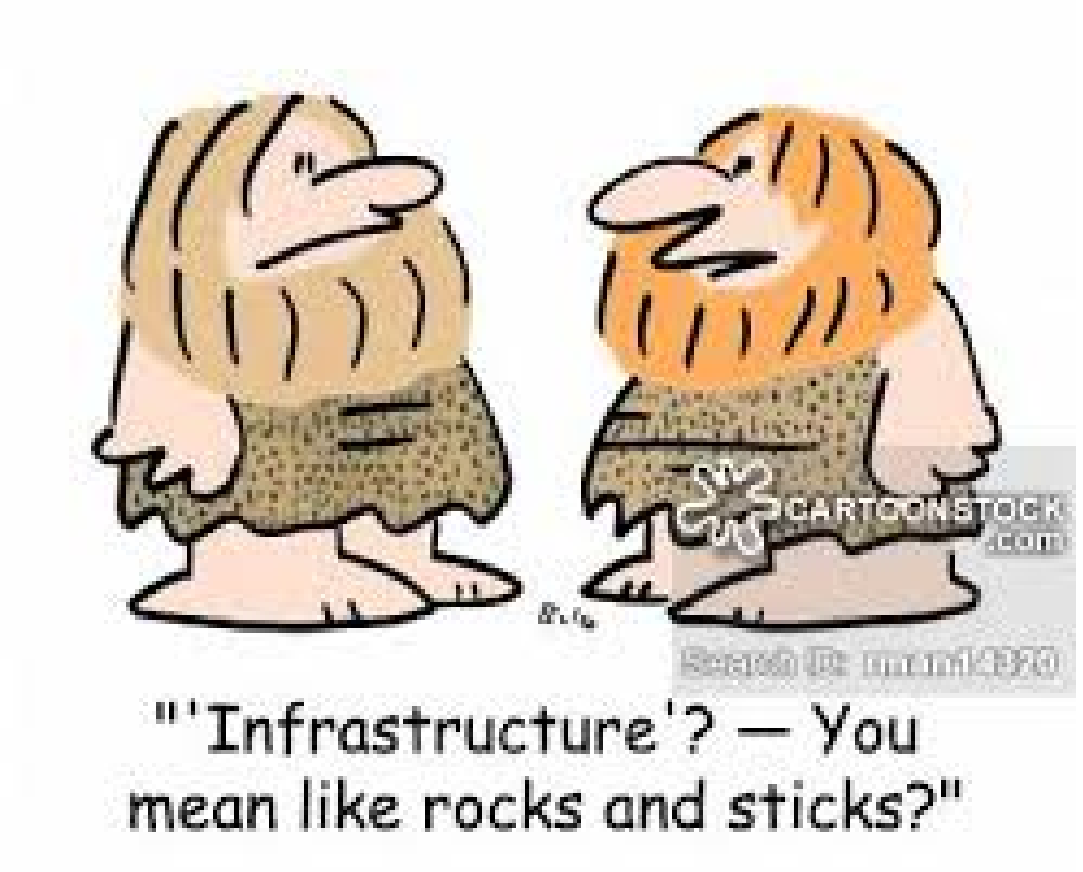
Elena Volodina

Lena Granstedt, Julia Prentice, Monica Reichenberg,
Beata Megyesi, Mats Wirén, Gunlög Sundberg

Key terminology

SweLL	**Swe**dish **L**earner **L**anguage
L2	Second (and foreign) language

What is infrastructure?



An electronic research infrastructure

- (free accessible) data in electronic format
- technical platform for exploring data, including tools and algorithms for data analysis, and visualization
- a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
- a network of experts in the relevant disciplines, incl. legal and ethical questions

Data collection

What data?

- L2 essays (16 years+)
- Exercise logs (16 years+)
    - incl. demographic metadata: age, gender, L1, residence time in Sweden, education level, etc.
    - incl. task metadata: date/term, topic, level of performance/grade, genre, etc.

Relevant laws & regulations

the core ones

- Personal Privacy Act, PPA (*Personuppgiftslagen*)
- Ethical Review Act (*Etikprövningslagen*)
- Open access law (*Tryckfrihetsförordningen*)
- Copyright (*Upphovsrättslagen*)

Different priority, and pretty conflicting

What is “personal information”?

Personal Privacy Act (PPA)

- **Personal information**: Any kind of information that directly or indirectly can be linked to a physical person (alive one).
  - Agreements (signed)
  - Hand-written versions *versus* ip-addresses
  - Demographic metadata: age, gender, L1, residence time in Sweden (month/year of arrival), education level, other languages
  - Essay metadata: date, school, class, teacher, topic, genre, grade, level, additional information/handouts

- If possible to identify a person
  - → applications to PPA
  - → sketch scenarios where the data can be used

Ethical Review Act

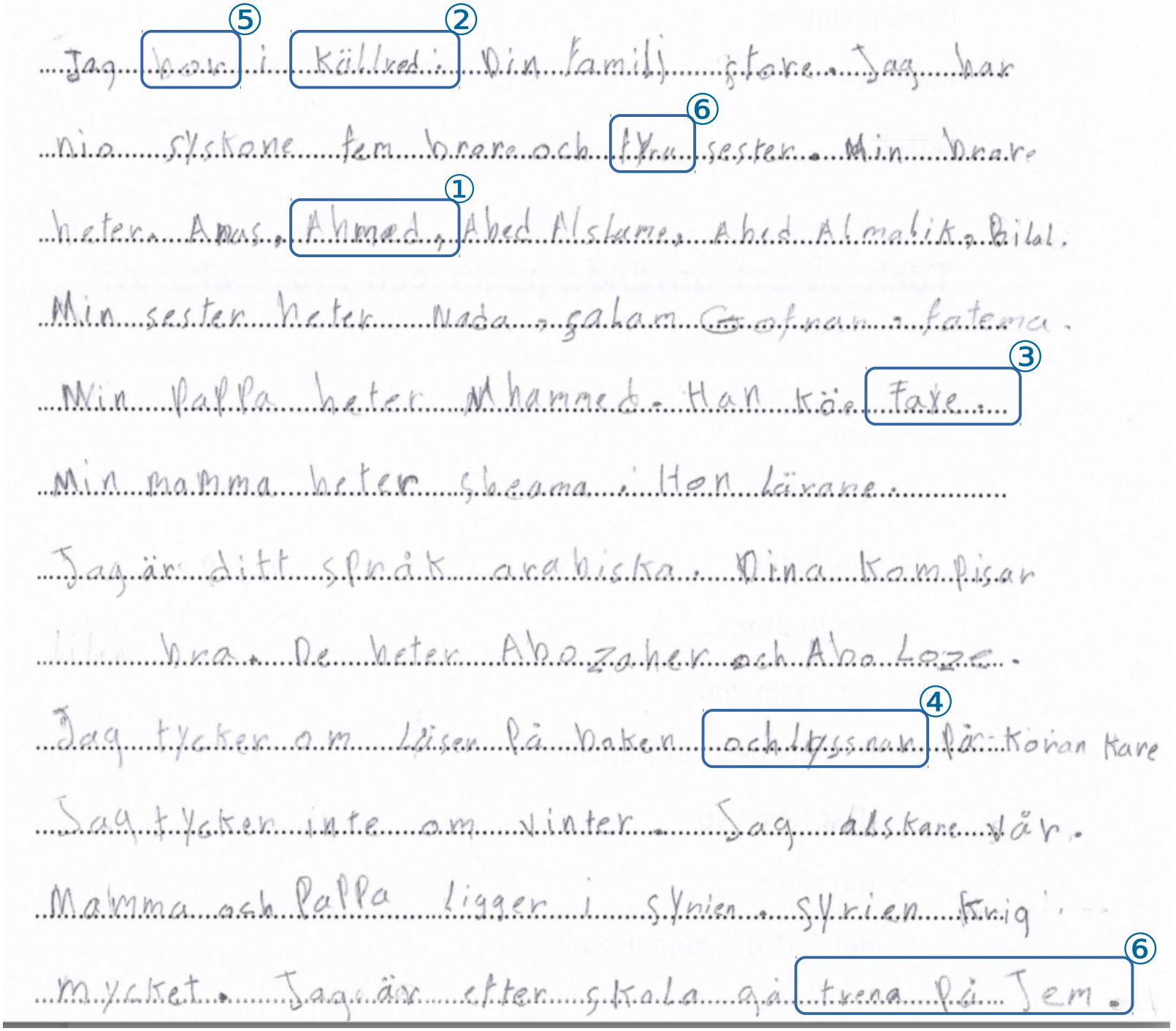
- **Ethnic, religious and political “risks”**: if the data can reveal a person's (alive one) ethnic background, religious or political beliefs.
    - Agreements (signed)
    - Hand-written versions *versus* ip-addresses
    - Demographic metadata: age, gender, L1, residence time in Sweden (month/year of arrival), education level, other languages
    - Essay metadata: date, school, class, teacher, topic, genre, grade, level, additional information/handouts

- If possible to identify a person
    - → applications to Ethical Review Board
    - → sketch all scenarios where the data can be used

FACTS

Level: A1

Topic:
Presentation/
Om mig

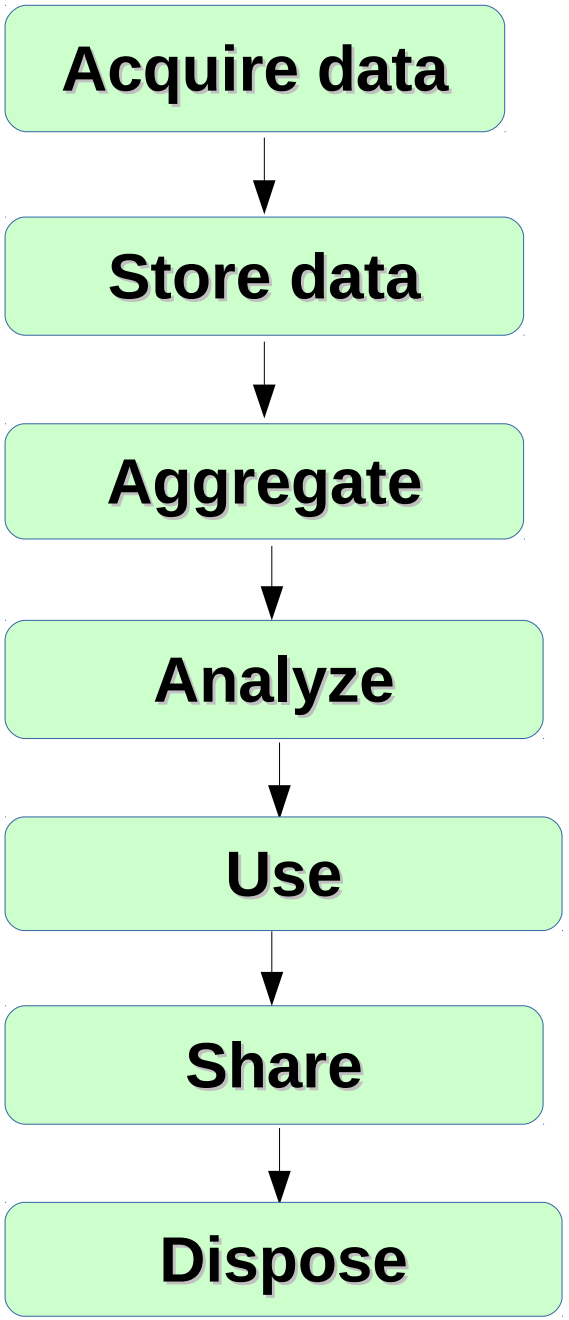


Implications?

sketch ALL scenarios
where the data can be used

Counter-
productive?

non-PPA approach

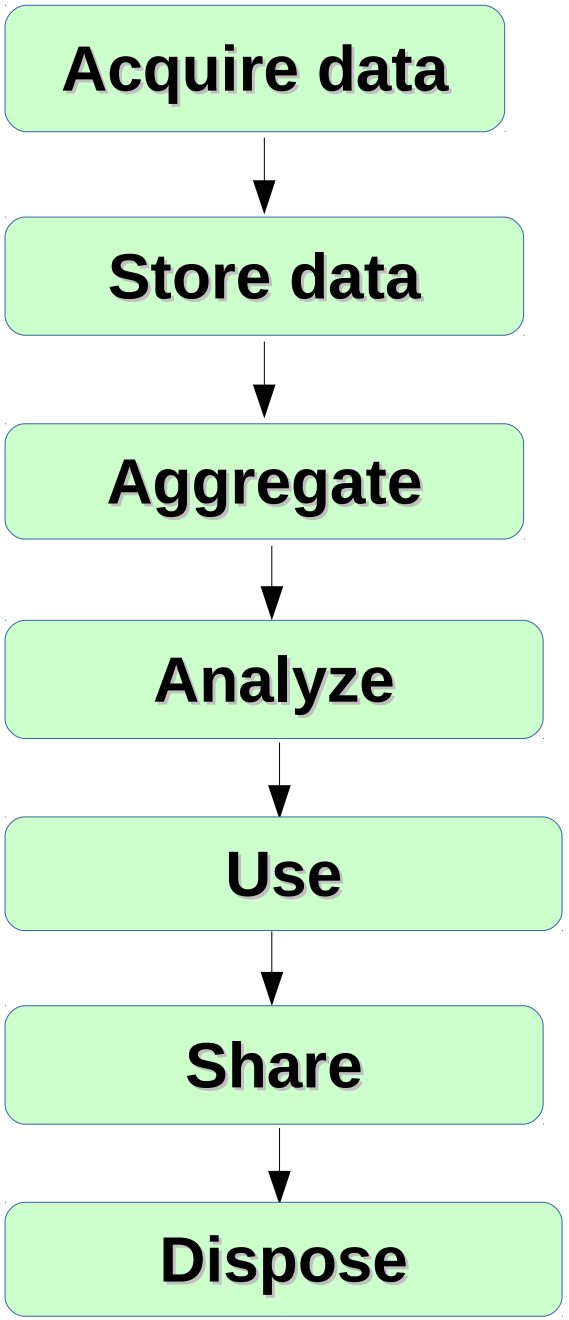


When prepared

→ for lawyers' scrutiny

Model from *Building digital trust: The role of data ethics in the digital age*

non-PPA approach



Implications

- for kind of applications (PPA, Ethical Review)
- for portal development/metadata
- for the overall data handling flow
- who can have accees to data
- for data usage scenarios

Model from *Building digital trust: The role of data ethics in the digital age*

SweLL corpus (2013-2016)

core data

SweLL workflow

Essay collection *ongoing* → **CEFR-grading** *teachers* → **Digitization** → **Manual registration** *editor in Lärka* → **Automatic annotation** *Korp pipeline*

Learner variables
Collected through permits*

Student variables:
- Age/birthyear
- Gender
- Mother tongue(s)
- Residence time in Sweden
- Education level

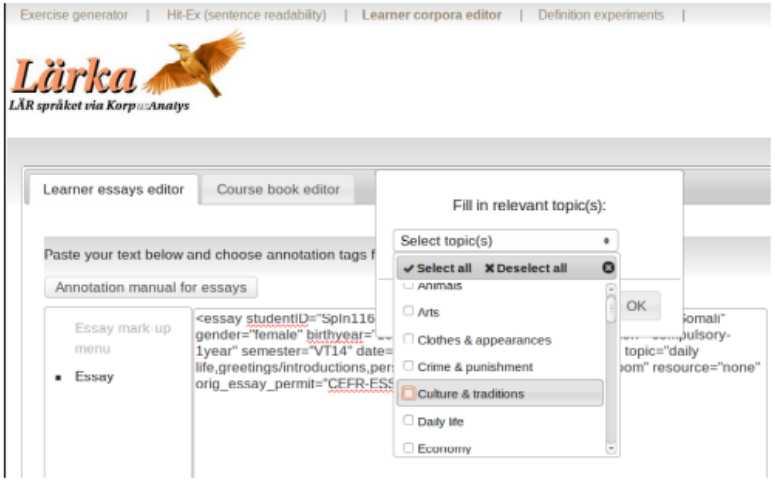
Essay variables
- Assigned CEFR level
- Essay setting (exam/home)
- Use of extra materials
- Academic term and date
- (Title, topic, genre, grade)

Assessors
Minimum of two trained assessors

Inter-annotator agreement
- A degree to which several annotators agree about assigning attributes
- Reported for SW1203 subcorpus
- Krippendorff's alpha for pairwise agreement = 0.80
- **0.80 = good annotation quality** (Artstein & Poesio 2008)

SweLL digitization principles

1. **Do not revel author identity**
   * revealing names → replace with *NN*
   * addresses → replace with *NN-street*
2. **Do not correct errors**
   * if several interpretations possible → make *positive assumption*, i.e. that the learner made no mistake
3. **Preserve illegible handwriting**
   * each illegible letter → replace with @
   * stricken text → leave out



* *http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/tillstand eng-24042013 v03.pdf*

SweLL corpus

core data

Sub-corpus	A1	A2	B1	B2	C1	Un-known	Total
Tisus	-	-	-	27	78	-	105
Sw1203	-	-	33	45	11	1	90
SpIn	16	83	42	2	-	1	144
Total	16	83	75	74	89	2	**339**

The Uppsala Corpus of Student Writings

reference corpus

Level	Age	School level and curriculum	Number of essays	Number of tokens	Tokens per essay
C-3	9	Compulsory, Lpf94 + Lgy11	91	8,644	95
C-5	11	Compulsory, Lpf94	66	13,121	199
C-6	12	Compulsory, Lgr11	47	17,741	377
C-9	15	Compulsory, Lgr94 + Lgr11	249	137,689	553
US-1	16	Upper Secondary, Lgy11	131	76,521	584
US-3	18	Upper Secondary, Lgy11	410	347,836	848
GY-3	18	Upper Secondary, Lpf94	1,506	1,055,468	701
Total			2,500	1,657,020	663

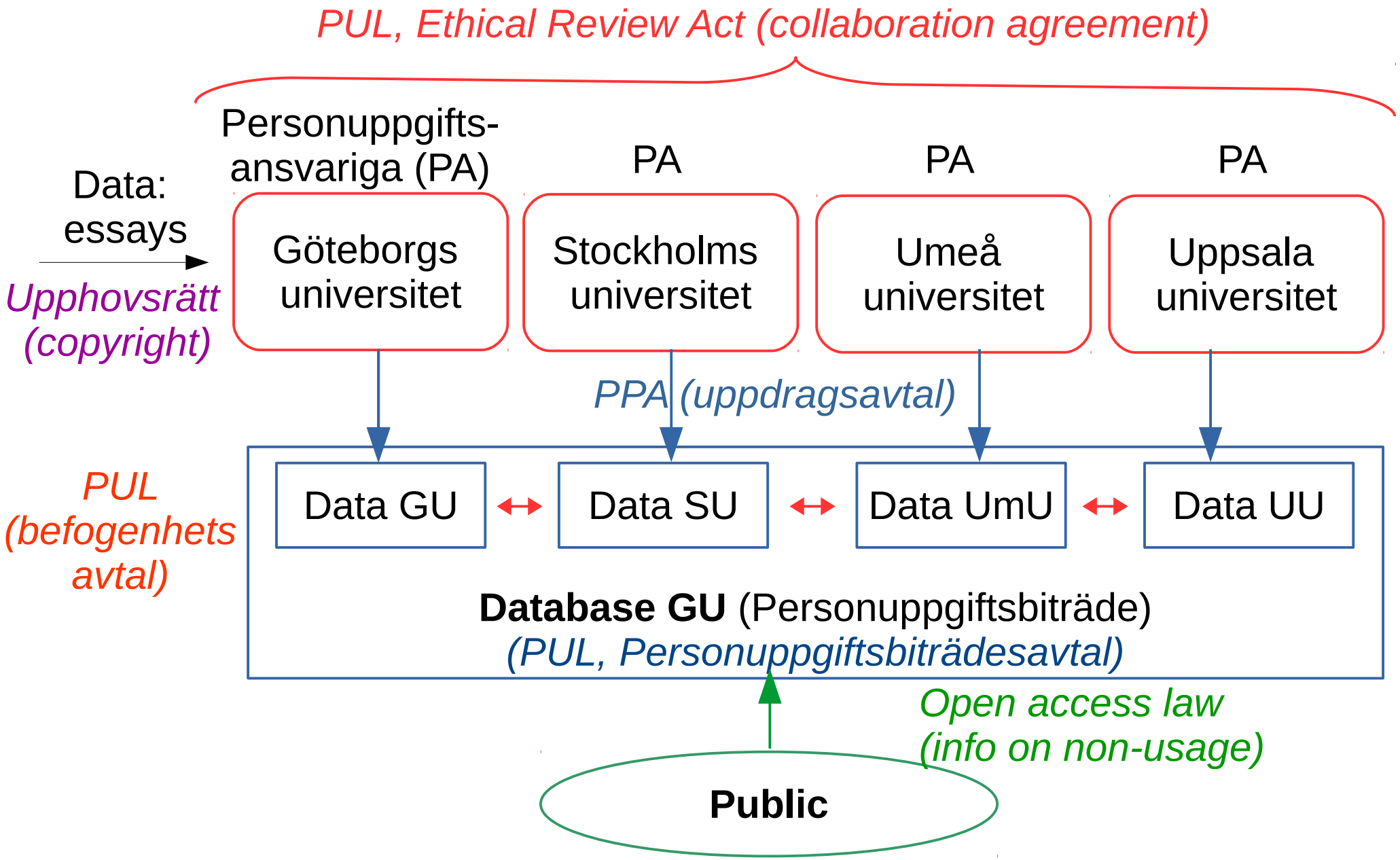
Table 1: Distribution of the subset of texts by school year, given as number of texts, sentences and tokens, and average number of tokens per essay used in the pilot study.

Handwritten essays	Printed essays
Transcription	Scanning-conversion-editing
Coding	Coding
Proofreading and final editing	Proofreading and final editing

Table 2: Preparation of the essays.

More “personal information” terms

- **Personuppgiftsansvarig**: Den som ensam eller tillsammans med andra bestämmer ändamålen med och medlen för behandlingen
- **Personuppgiftsbiträde**: Den som behandlar personuppgifter för den personuppgiftsansvariges räkning.
- **Tredje man**: Någon annan än den registrerade, den personuppgiftsansvarige, personuppgiftsombudet, personuppgiftsbiträdet och sådana personer som under den personuppgiftsansvariges eller personuppgiftsbiträdets direkta ansvar har befogenhet att behandla personuppgifter.


PUL, Ethical Review Act (collaboration agreement)
Personuppgifts-
ansvariga (PA)
PA
PA
PA
Data:
essays
Upphovsrätt
(copyright)
Göteborgs
universitet
Stockholms
universitet
Umeå
universitet
Uppsala
universitet
PPA (uppdragsavtal)
PUL
(befogenhets
avtal)
Data GU
Data SU
Data UmU
Data UU
Database GU (Personuppgiftsbiträde)
(PUL, Personuppgiftsbiträdesavtal)
Open access law
(info on non-usage)
Public

Documents we need

the ones I know

- *Samtycke* (permit)
- *PUL application* (to GU lawyer)
- *Application to Rthical Review Board*
- *Samverkansavtal* (between project partners)
- *Personuppgiftsbirädesavtal* (all partners with GU)
- *Uppdragsavtal* (all partners with GU)
- *Befogenhetsavtal* (between all partners)
- *Information to the public* (in case some researchers would want to get access to the data)

Implications for research infrastructures

- Do we need new data legislation principles?
  - On a national level?
  - On an international level?
  - Give a mandate to the lawyers to help avoid situations where all research/usage scenarios will need to be defined