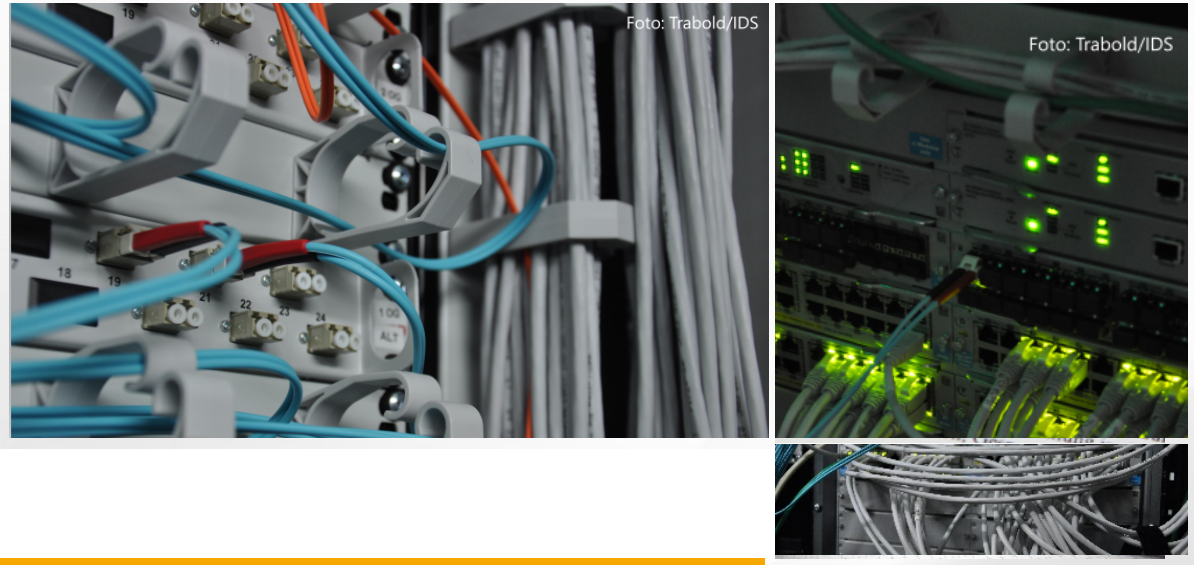
IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE



Paweł Kamocki, Andreas Witt

ETHICAL ISSUES IN LANGUAGE RESOURCES AND LANGUAGE TECHNOLOGY — TENTATIVE TAXONOMY

Leibniz Gemeinschaft

ETHICAL ISSUES IN LR AND LT

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- often invoked, but rarely discussed
- **lack of clear definitions** of what is/should be ethical in LR/LT
- **other frames of reference are insufficient** (different scopes)
  - ‘scientific ethos’ (Mertonian norms)
  - technoethics
  - AI ethics
- **law ≠ ethics**
  - e.g. would anyone publish Goethe’s poems as their own?
- need for a **LR & LT Code of Ethics**

TENTATIVE TAXONOMY: THE FIVE PRINCIPLES

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Privacy

Property

Equality

Transparency

Freedom

PRIVACY

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **stakeholders (data providers, users) should be protected against disproportionate intrusion and allowed to keep certain information secret;**

- privacy ≠ data protection
- conception phase (‘privacy by design and by default’)
  - LT tools: no excessive data collection, no unsolicited interaction, user control over privacy-sensitive functions and features
  - LR: granularity of data collection forms, no necessarily intrusive questions
- creation phase and use phase
  - use pseudonymisation/anonymisation techniques, even data deletion
  - ‘ethical’ use of data relating to deceased persons (not covered by the GDPR)

PROPERTY

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **intellectual and cultural property should be handled with respect, in compliance with applicable law, ensuring that any potential harm (evaluated from the owner's perspective) is outweighed by collective benefit;**

- CARE principles
- creation phase
  - 'ethical' use of IP-protected data (e.g. 'diligent search' for rightsholders of orphan works)
  - 'ethical' use of cultural property (e.g. indigenous languages) — community involvement

EQUALITY

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **no group of stakeholders or contributors should be directly or indirectly discriminated against;**

- conception phase
  - data selection for LR: representativeness and balance to avoid discriminatory effect
  - selection of people for certain tasks (e.g. fieldwork)

TRANSPARENCY

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **stakeholders should be informed about the main principles of, and given a possibility to learn the details about the functioning of LT;**
- **LT outputs should be clearly marked as such.**
- conception phase
  - proper documentation of the conception process
- creation phase
  - information of data contributors, with the possibility to learn about the details
- use phase
  - e.g. MT outputs, chatbots, etc. should be transparently marked
- evaluation phase
  - evaluation should be based on transparent criteria

FREEDOM

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **data providers should be free to contribute their data to LR & LT, and, to a reasonably practicable extent, to change their mind at any later stage**
- **human intervention should be necessary and decisive in any process involving the use of LT the outcome of which may seriously impact the user.**

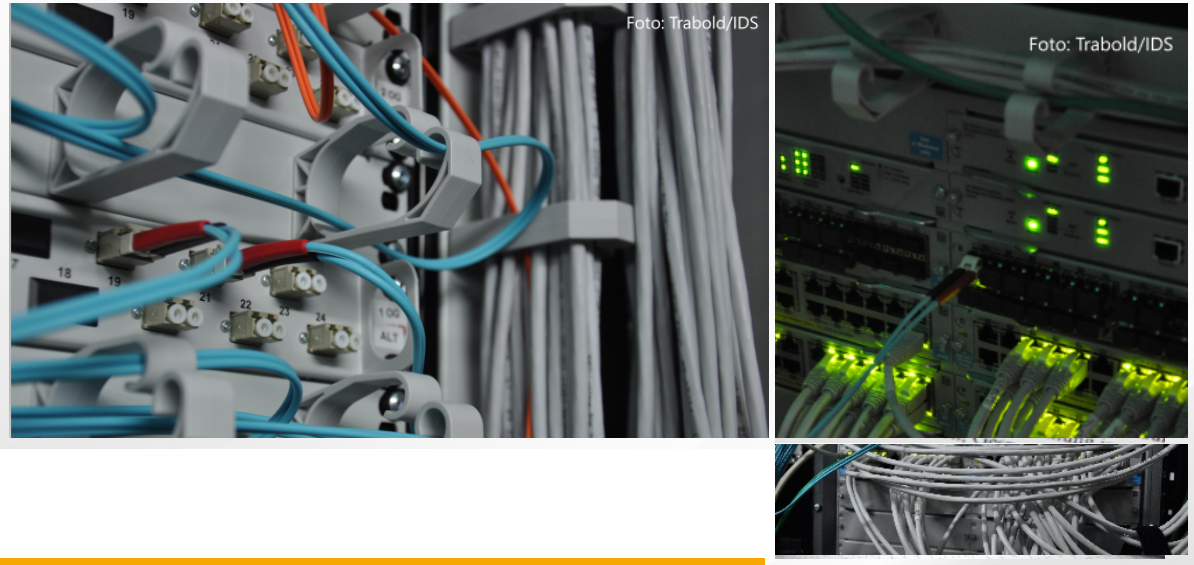
- creation phase
  - data contributors should be given the possibility to withdraw their data (also non-personal)
- use phase
  - LT should not be used to make important decisions without human intervention

CONCLUSION

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- **common frame of reference for evaluation of LR&LT projects**
- **spark a debate on LR & LT Code of Ethics**
  - feel free to contact us at: kamocki | witt@ids-mannheim.de
- **use as metadata**

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE



Paweł Kamocki, Andreas Witt

ETHICAL ISSUES IN LANGUAGE RESOURCES AND LANGUAGE TECHNOLOGY — TENTATIVE TAXONOMY

Leibniz Gemeinschaft