Zurich University
of Applied Sciences

# Navigating Legal Complexities

## Open Research Data & Language Corpora

**Suzanna Marazza**
Università della Svizzera italiana, CCDigitallaw

**Prof. Philipp Dreesen**
ZHAW Digital Discourse Lab

**Dr. Cristina Grisot**
CLARIN-CH

**Marcel Griesinger**
ZHAW Center for Enterprise Law

**Dr. Simon van Rekum**
ZHAW Services Research Data

**Dr. Julia Krasselt**
ZHAW Digital Discourse Lab

Digital Discourse Lab

CLARIN-CH

swissuniversities

School of Applied Linguistics

swissuniversities

Digital Discourse Lab

# Input on Language Corpora as Open Research Data

**Julia Krasselt (ZHAW Digital Discourse Lab)**

Zurich University
of Applied Sciences

# Open Science as paradigm shift

## The Open Science Movement

- the Open Science Movement (Open Access + Open Research Data) leads to a paradigm shift in scientific research

- Facilitating access and reuse of research data (ORD) leads to an increase in the value of data

**(1) transparency** and **replicability** of research results, fostering scientific integrity

**(2) interdisciplinary exchange**, fostering innovative and impactful research

- Swiss National Open Research Data Strategy + associated action plan (swissuniversities, ETH Domain, SNFS, SAGW/ASSH)

Zurich University
of Applied Sciences

# ORD Challenges for Language Corpora

## Open Research Data and Language Corpora

- Open Research Data means
  - adherence to FAIR principles: findable, accessible, interoperable, reusable (Wilkinson et al., 2016)
  - „[…] as open as possible, as protected as necessary“ (Swiss National ORD Strategy, 2021, p. 6)

- With regard to language data (i.e., spoken or written utterances, multimodal data), ORD principles pose the following challenges

1) Heterogeneity of researchers working with language data needs to be considered
2) Legal restrictions (copyright and data protection) impact FAIR principles

- Publishing corpora on conventional data repositories is not appropriate and, in most cases, legally questionable

Zurich University
of Applied Sciences

# Swiss-AL as an ORD platform

## The Perspective of the ZHAW Digital Discourse Lab

- **„Swiss-AL: Linguistic ORD Practices for Applied Sciences“** (Swiss Open Research Data Grants, Track B, matching funds by swissuniversities and ZHAW)

- Swiss-AL: large multilingual corpus of Swiss public communication hosted and developed at ZHAW (journalistic articles, media releases, blogs, news,…)

### 1. Heterogeneous user groups, research practices and interests

Which practices (e.g., data preparation, processing, visualization, dissemination) do researchers from different disciplines use when working with language data?

### 2. Open Research Data Cycle

How can Swiss-AL be developed as a dynamic family of corpora with a flexible processing pipeline and an interactive workbench?

### 3. Legal Aspects

How can language data be made publicly accessible while complying with copyright and data protection laws?

# Swiss-AL as an ORD platform

## Legal aspects: Challenges related to language data

- the majority of primary data collected in linguistic corpora
  - is subject to **copyright protection** (e.g., newspaper articles, press releases) and cannot be easily shared (e.g., uploaded on a repository)
  - contains sensitive data (i.e., data relating to an identified or identifiable natural person[1], e.g., social media posts), i.e., are subject to **data protection regulations**
- differing regulations on a national and international level

**→ How can the Open Science Movement be aligned with copyright and data protection law (e.g., Federal Act on Data Protection, Federal Act on Copyright and Related Rights)?**

[1] Federal Act on Data Protection, Art. 3, https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en

Zurich University
of Applied Sciences

# Swiss-AL as an ORD platform

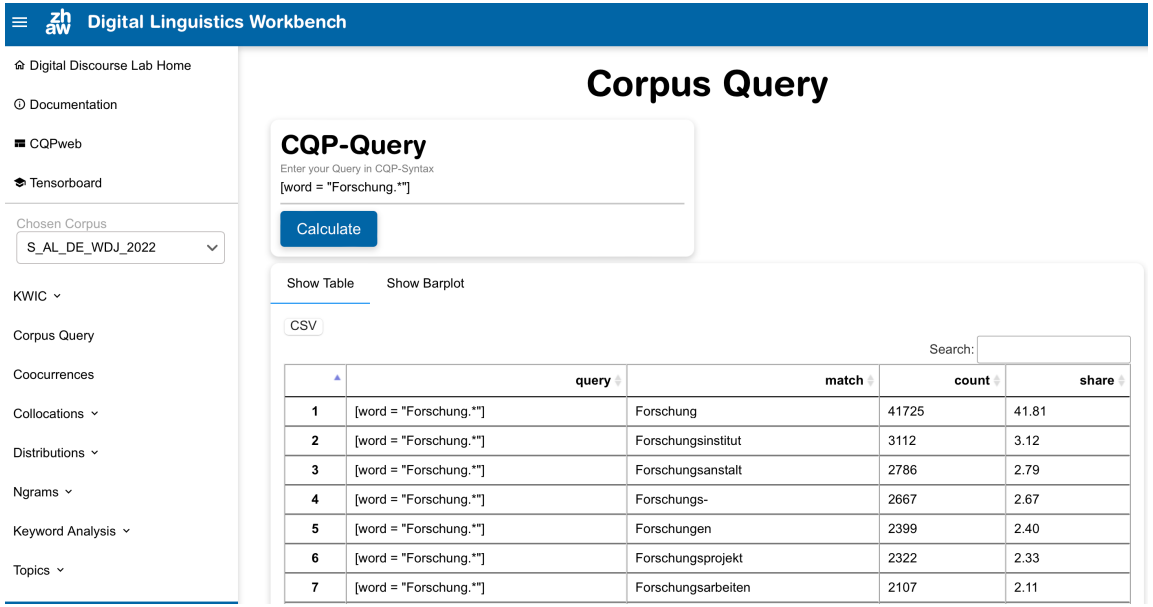
## Possible Solutions for Swiss-AL

1. Aggregated, frequency based data and language models
2. Metadata only
3. Modes to display excerpts from the corpus, e.g., shuffling and randomization
4. Manage access rights

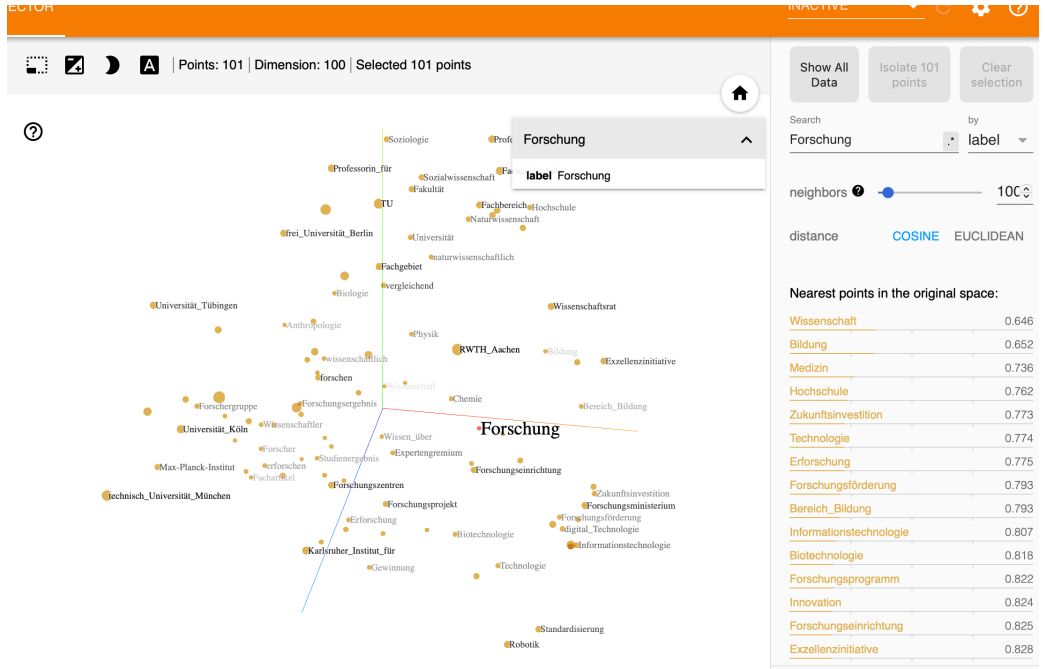
Zurich University
of Applied Sciences

# Swiss-AL as an ORD platform

## Aggregated, frequency based data and language models




| | query | match | count | share |
|---|---|---|---|---|
| 1 | [word = "Forschung.*"] | Forschung | 41725 | 41.81 |
| 2 | [word = "Forschung.*"] | Forschungsinstitut | 3112 | 3.12 |
| 3 | [word = "Forschung.*"] | Forschungsanstalt | 2786 | 2.79 |
| 4 | [word = "Forschung.*"] | Forschungs- | 2667 | 2.67 |
| 5 | [word = "Forschung.*"] | Forschungen | 2399 | 2.40 |
| 6 | [word = "Forschung.*"] | Forschungsprojekt | 2322 | 2.33 |
| 7 | [word = "Forschung.*"] | Forschungsarbeiten | 2107 | 2.11 |



# Swiss-AL as an ORD platform

Zurich University
of Applied Sciences

## Possible Solutions

1. Aggregated, frequency based data and language models
2. Metadata only (e.g., source, publication date, url)
3. Modes to display excerpts from the corpus, e.g., shuffling and randomization
4. Manage access rights

Zurich University
of Applied Sciences

# Swiss-AL as an ORD platform

## Possible Solutions

1. Aggregated, frequency based data and language models
2. Metadata only
3. Modes to display excerpts from the corpus, e.g., shuffling and randomization
4. Manage access rights

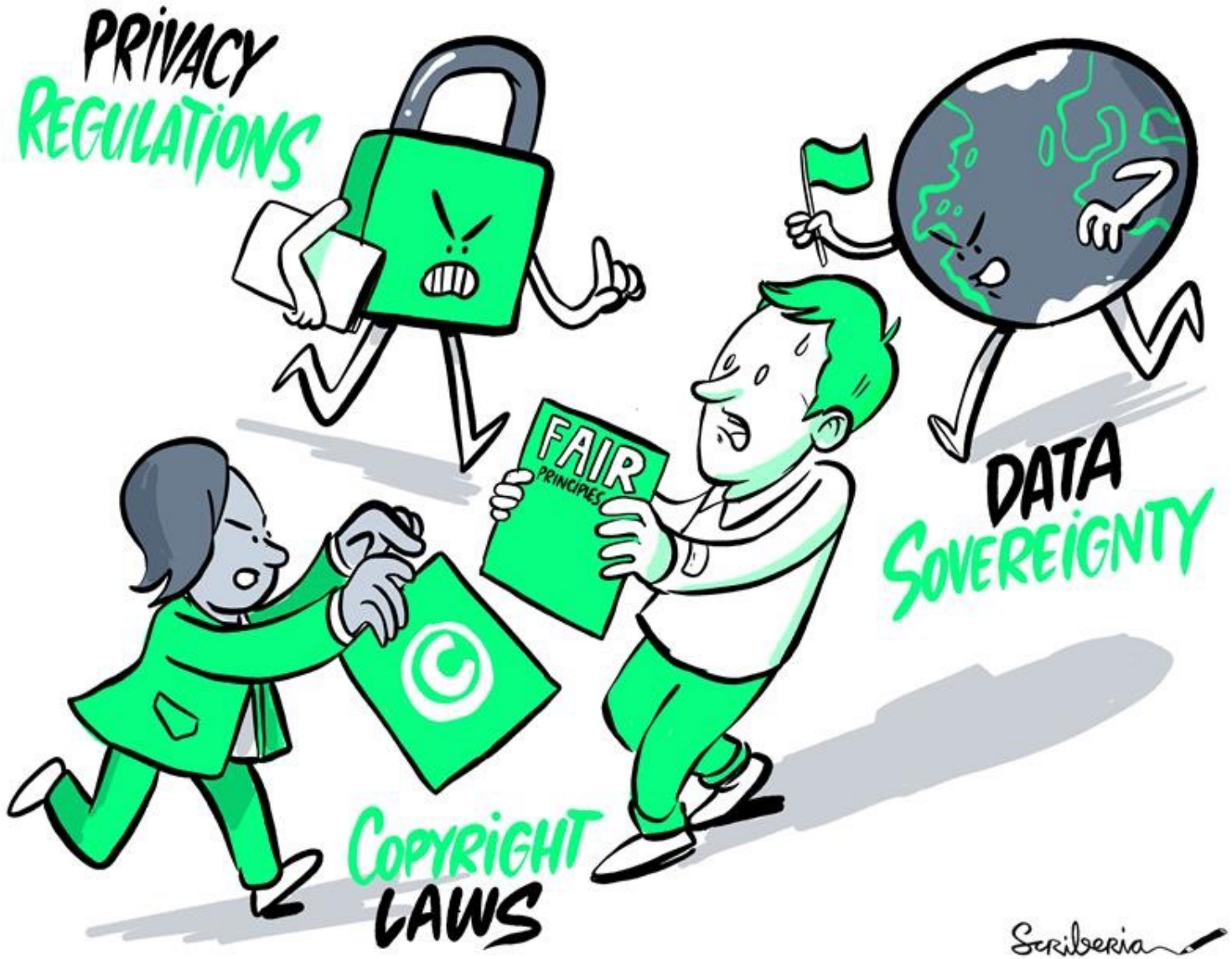
Zurich University
of Applied Sciences

# Swiss-AL as an ORD platform

**Possible Solutions**

1. Aggregated, frequency-based data and language models
2. Metadata only
3. Modes to display excerpts from the corpus, e.g., shuffling and randomization
4. Manage access rights

**But: solutions do not fully comply with needs formulated by researchers (e.g., the need to access whole documents for qualitative coding) and with FAIR principles**



Zurich University
of Applied Sciences

# Contact

digitaldiscourselab.linguistik@zhaw.ch

https://www.zhaw.ch/en/linguistics/business-services/digital-discourse-lab/

https://www.zhaw.ch/en/linguistics/research/swiss-al-linguistic-open-research-data-practices-for-applied-sciences/

Zurich University
of Applied Sciences

# Navigating Legal Complexities

## Open Research Data & Language Corpora

**Suzanna Marazza**
Università della Svizzera italiana, CCDigitallaw

**Prof. Philipp Dreesen**
ZHAW Digital Discourse Lab

**Dr. Cristina Grisot**
CLARIN-CH

**Marcel Griesinger**
ZHAW Center for Enterprise Law

**Dr. Simon van Rekum**
ZHAW Services Research Data

**Dr. Julia Krasselt**
ZHAW Digital Discourse Lab

Digital Discourse Lab

CLARIN-CH

swissuniversities