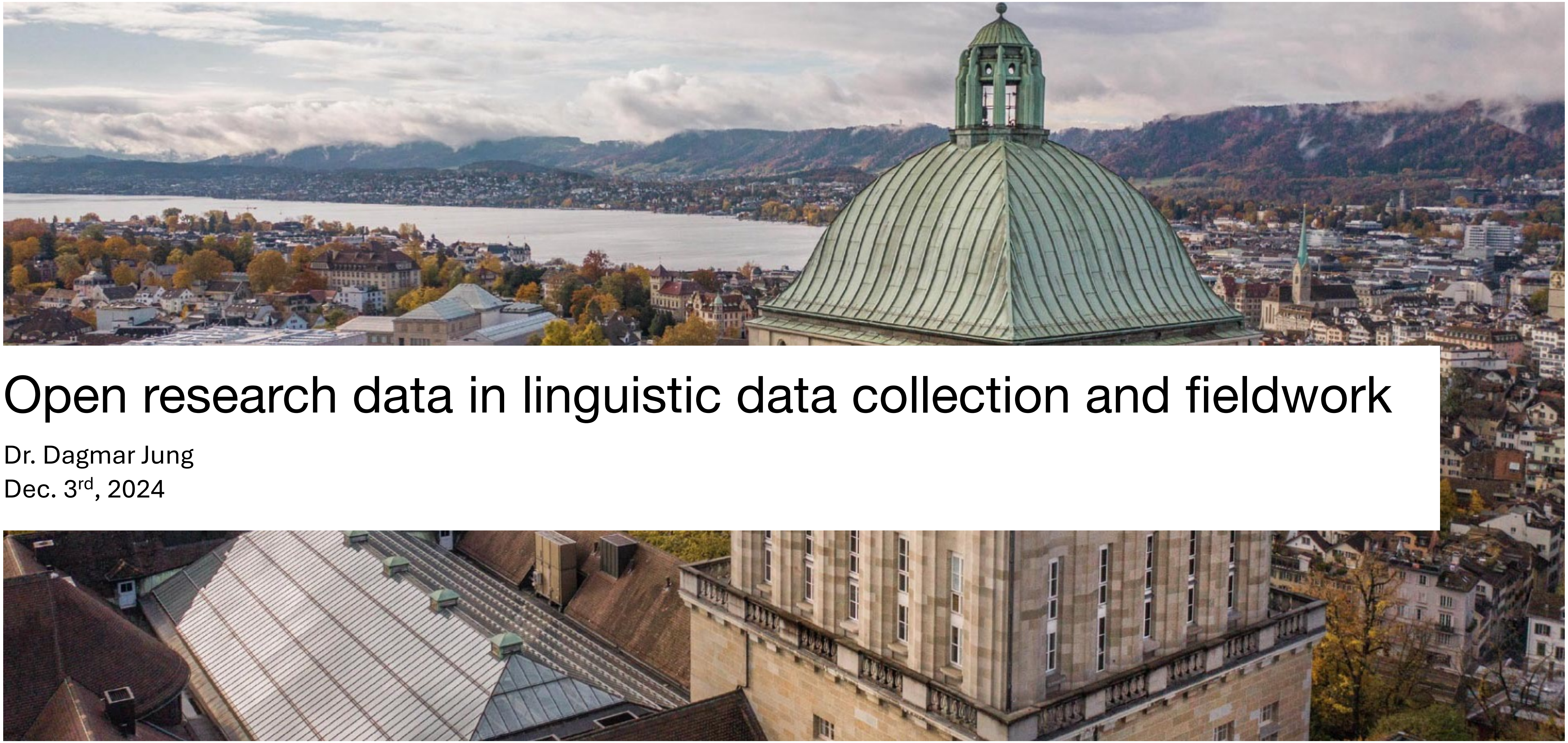
Universität Zürich UZH

Linguistic Research Infrastructure (LiRI) & NCCR@LiRI &
Dept. of Comparative Language Science



# Open research data in linguistic data collection and fieldwork

Dr. Dagmar Jung
Dec. 3rd, 2024

# Topics

Open Science

Linguistic data types

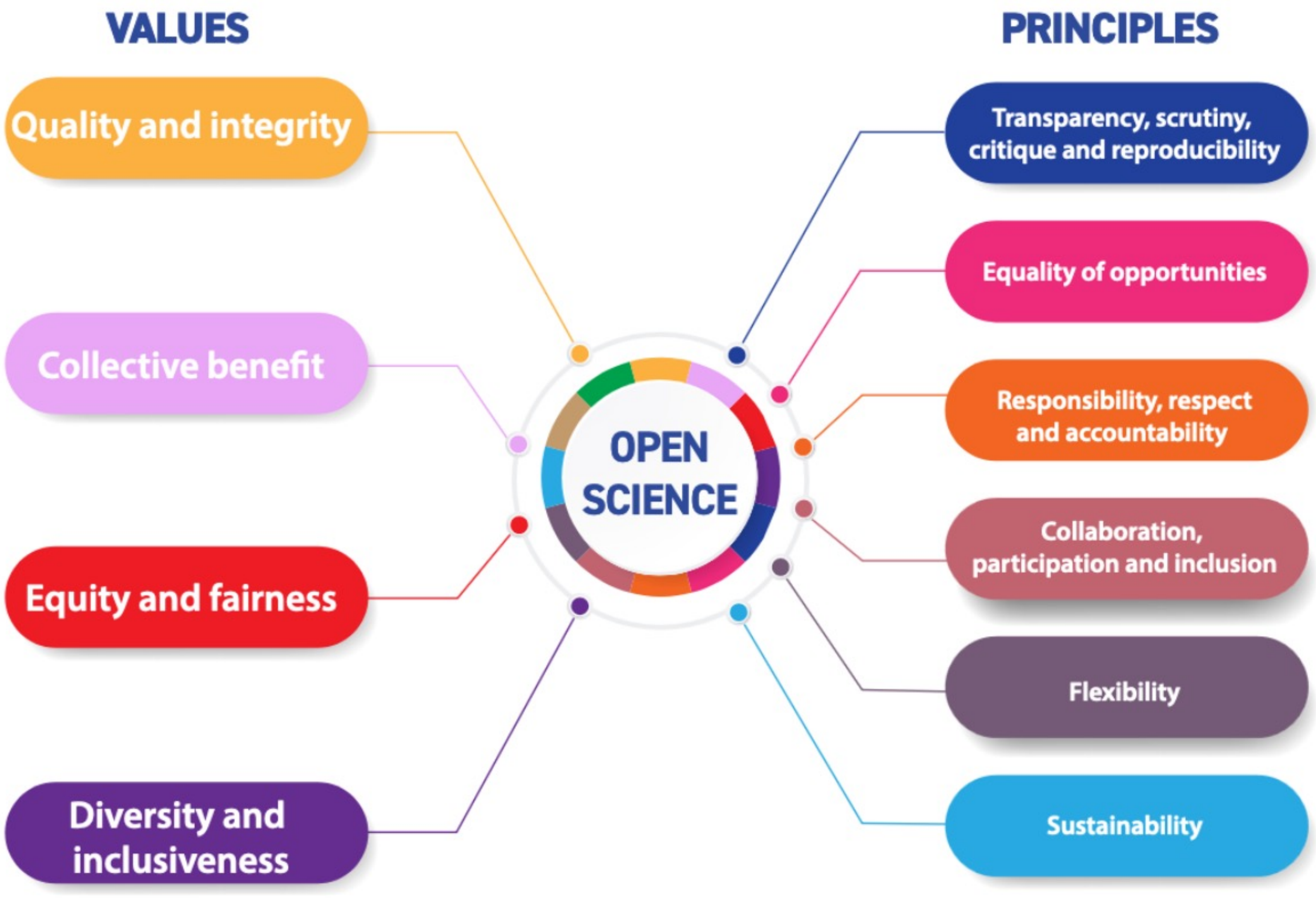
FAIR and CARE principles

Fieldwork data collection

Ethics and Legal considerations

# UNESCO Recommendation on Open Science

"For the purpose of this Recommendation, **open science is defined** as an inclusive construct that combines various movements and practices aiming to **make multilingual scientific knowledge openly available, accessible and reusable for everyone**, to increase scientific collaborations and sharing of information for the **benefits of science and society**, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors **beyond the traditional scientific community**. "


VALUES
Quality and integrity
Collective benefit
Equity and fairness
Diversity and inclusiveness
OPEN SCIENCE
PRINCIPLES
Transparency, scrutiny, critique and reproducibility
Equality of opportunities
Responsibility, respect and accountability
Collaboration, participation and inclusion
Flexibility
Sustainability

- Source: https://unesdoc.unesco.org/ark:/48223/pf0000379949

# Linguistic data collection

**Observational data**

- Audio recordings
- Video recordings
- e.g. for interactional linguistics or language documentation

**Survey methods (metalinguistic knowledge)**

- questionnaires (written)
- interviews (oral)
- e.g. elicitation (systematic survey for language data)

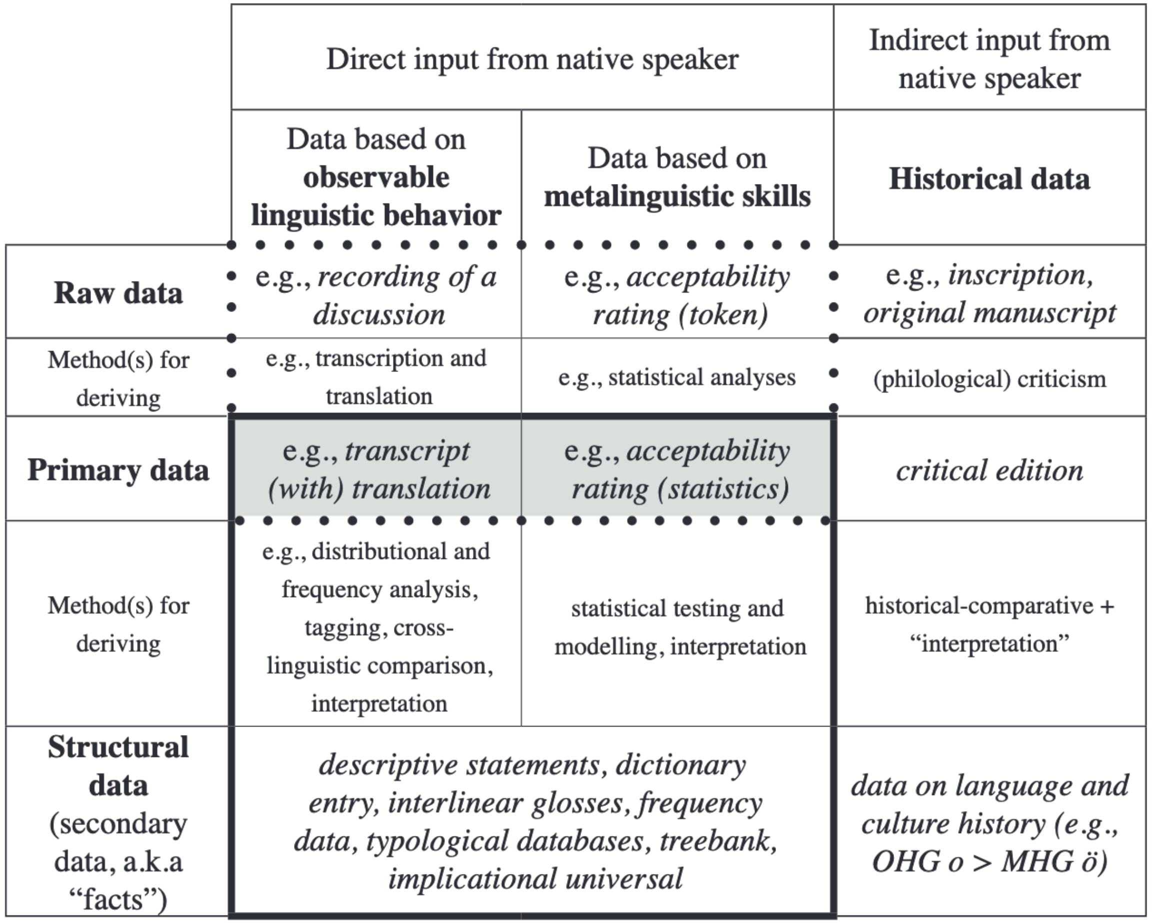
**Experimental data**

- collection of systematic data under controlled conditions
- metrics
- generally in the lab
- sometimes in the field

- Source: Kretzschmar & Voelkel 2021

# Linguistic data types

**Documentary Linguistics and Descriptive Linguistics**



| | Direct input from native speaker | | Indirect input from native speaker |
|---|---|---|---|
| | Data based on **observable linguistic behavior** | Data based on **metalinguistic skills** | **Historical data** |
| **Raw data** | e.g., *recording of a discussion* | e.g., *acceptability rating (token)* | e.g., *inscription, original manuscript* |
| Method(s) for deriving | e.g., transcription and translation | e.g., statistical analyses | (philological) criticism |
| **Primary data** | e.g., *transcript (with) translation* | e.g., *acceptability rating (statistics)* | *critical edition* |
| Method(s) for deriving | e.g., distributional and frequency analysis, tagging, cross-linguistic comparison, interpretation | statistical testing and modelling, interpretation | historical-comparative + "interpretation" |
| **Structural data** (secondary data, a.k.a "facts") | *descriptive statements, dictionary entry, interlinear glosses, frequency data, typological databases, treebank, implicational universal* | | *data on language and culture history (e.g., OHG o > MHG ö)* |

TABLE 5. Basic linguistic data types according to native speaker input (columns) and processing stage (rows)

Source: Himmelmann 2012

# Linguistic data types and Public Access: issues and strategies

- **Descriptive metadata**
- anonymization/de-identification

- **Child Language data**
- minors legally protected
- restricted access
- consent given by legal guardians
- media can be stored offline, metadata anonymized

- **Original texts, Transcripts, Annotations**
- less identifying than multimedia
- intellectual property rights must be respected
- some content may be problematic (-> avoidance of harm)

Source: Seyffedinipur et al. 2019

- **Multimedia**
- personal identifying information
- various potential consequences for

speakers and communities

- restricted, except personal rights cleared

(e.g. via consent) and IPR respected

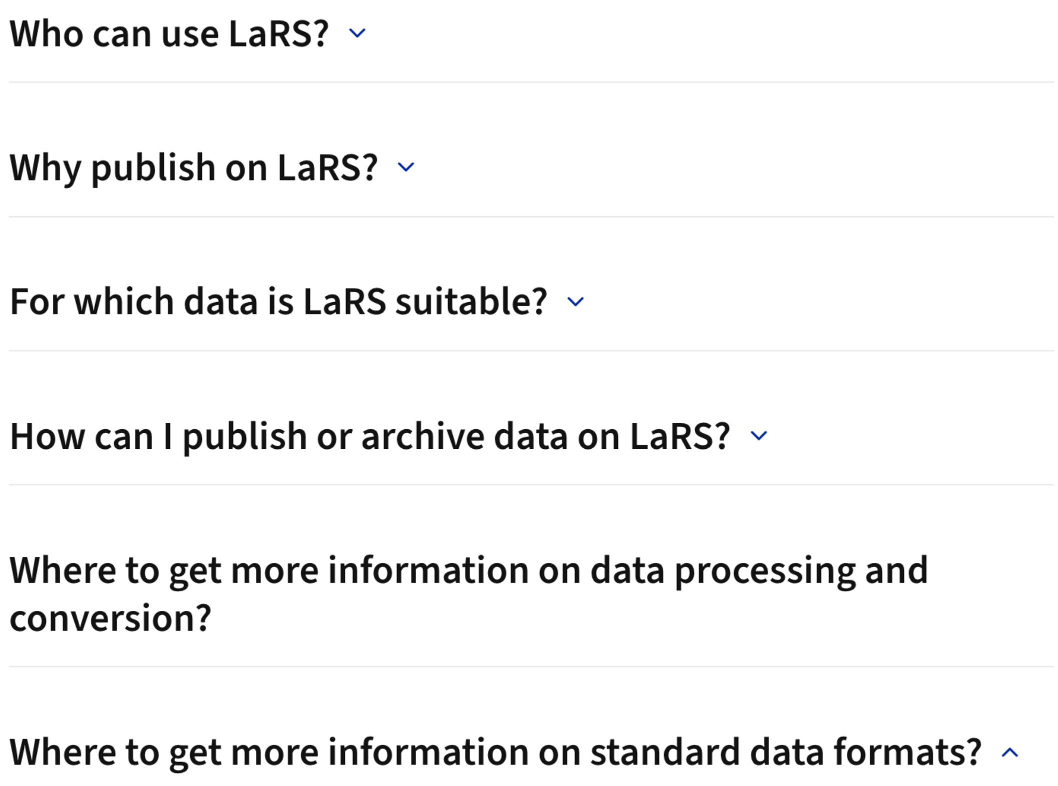
- **Experimental data**
- generally already anonymized
- rights to stimuli must be cleared

- **Sensitive material**
- registered users only

- **Location data**
- geographical coordinates of objects, events
- may be commercially interesting (loggers, poachers, mineral prospectors..)
- restricted access or withholding from archive

- **Legacy materials**
- difficult since no informed consent
- publicly available, but
- acknowledgment of unclear status
- take-down principle if there is grievance

# Language Repository of Switzerland (LaRS) on SWISSUbase



**Who can use LaRS?**

**Why publish on LaRS?**

**For which data is LaRS suitable?**

**How can I publish or archive data on LaRS?**

**Where to get more information on data processing and conversion?**

**Where to get more information on standard data formats?**

- Source: https://www.lars.uzh.ch/en.html

Open "Access": problematic aspects – practical & ethical in fieldwork

**Storage – where are the data?**

**Who has access?**

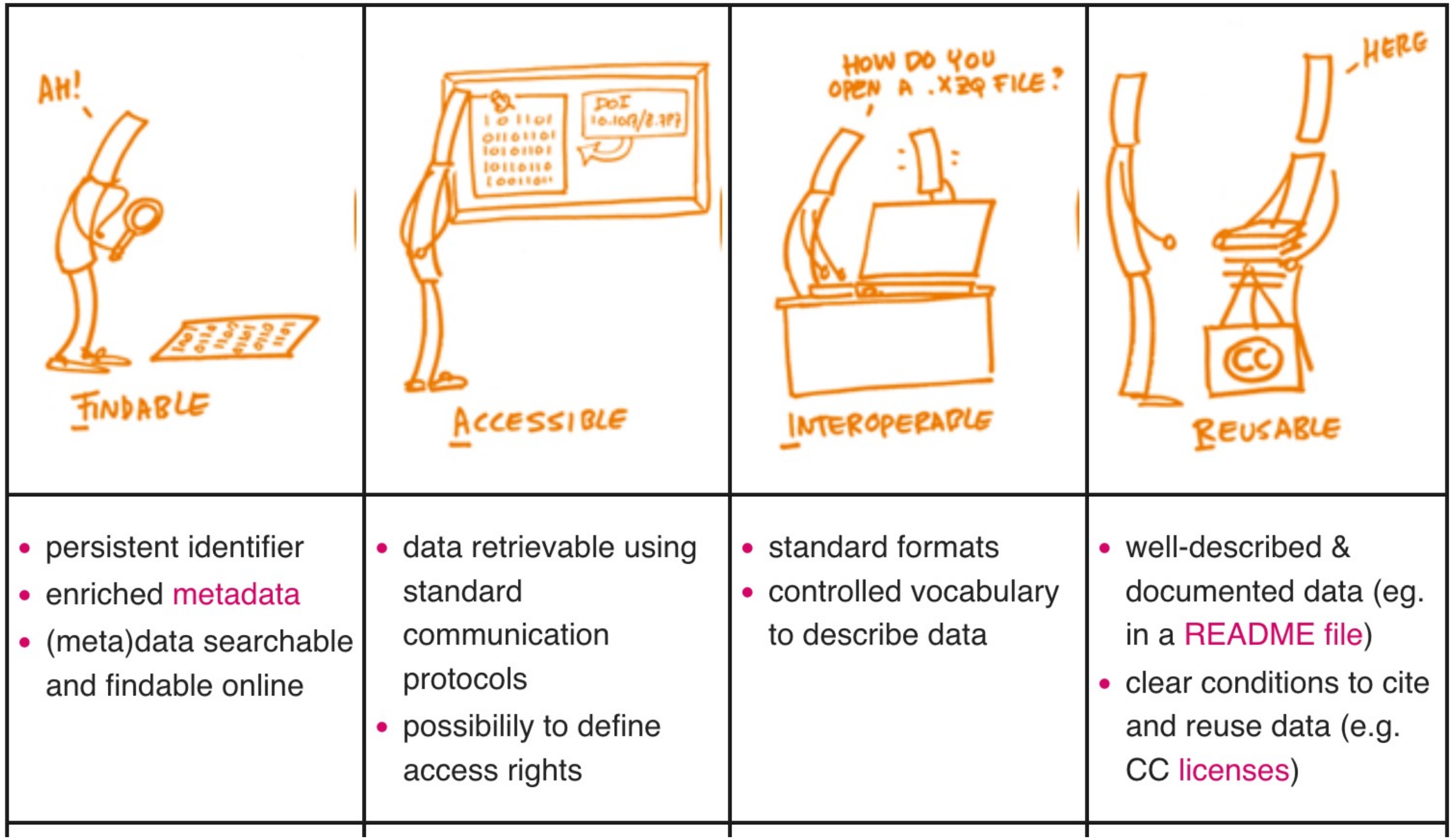
**Who controls access?**

**General issue raised by indigenous scholars: languages are not data sets**

Holton, Leonard & Pulsifer 2022

# FAIR principles in Open Research Data

**Findable, Accessible, Interoperable, Reusable**



| | | | |
|---|---|---|---|
| • persistent identifier<br>• enriched metadata<br>• (meta)data searchable and findable online | • data retrievable using standard communication protocols<br>• possibilily to define access rights | • standard formats<br>• controlled vocabulary to describe data | • well-described & documented data (eg. in a README file)<br>• clear conditions to cite and reuse data (e.g. CC licenses) |

- Source: https://www.unige.ch/researchdata/en/share/fair/

# Research ethics – data collection in the field

Belmont Report (1979)

**Three basic principles for ethical research:**

- respect for persons, beneficence, and justice

**Three elements of ethical research**

- Ownership, control of research
- Extent to which the researched are disinterested subjects of the researcher's activity
- Responsibility of researcher to seek/use knowledge for the benefit of the researched

- Source: Rice 2012

# San Code of Research Ethics

**Respect**

**Honesty**

**Justice and Fairness**

**Care**

**Process**

SAN CODE OF RESEARCH ETHICS

South African San Institute 2017

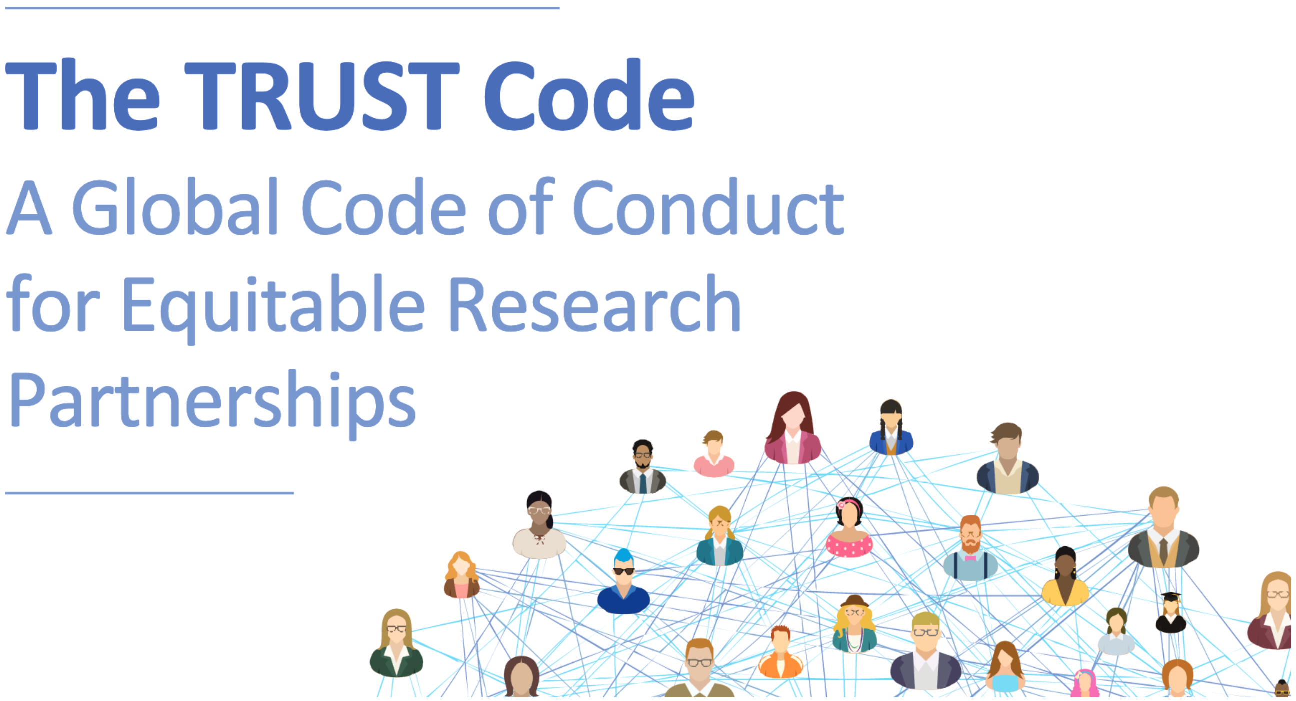
TRUST

Equitable Research Partnerships

Applications for community approval of research involving South African San communities can be applied for at the South African San Council.

- Source: https://www.globalcodeofconduct.org/affiliated-codes/

# globalcodeofconduct.org



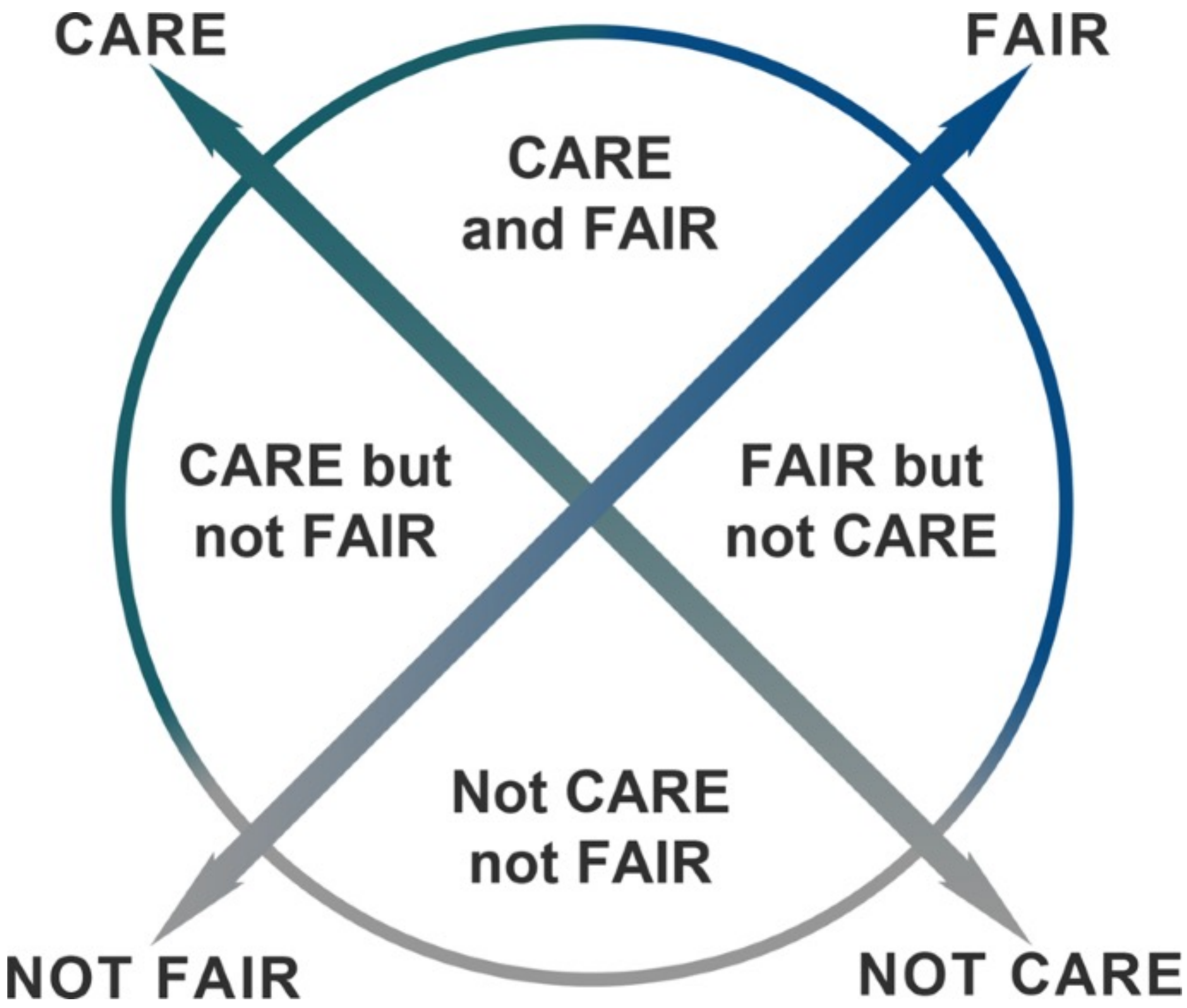
The TRUST Code

A Global Code of Conduct for Equitable Research Partnerships

- Source: TRUST (2018) The TRUST Code – A Global Code of Conduct for Equitable Research Partnerships, DOI: https://doi.org/10.48508/GCC/2018.05

# CARE principles

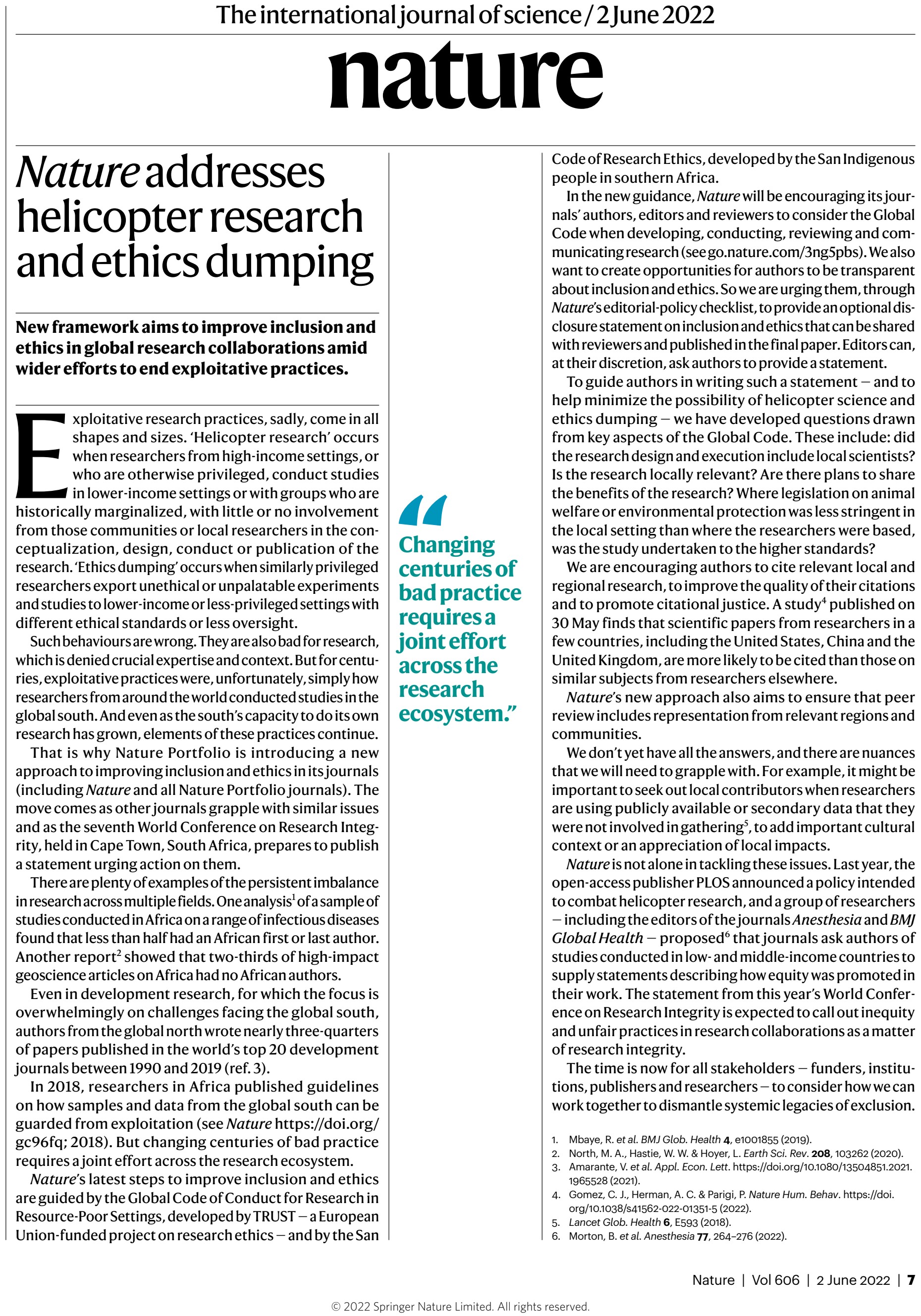
- Global Indigenous Data Alliance 2019
- The ‘CARE Principles for Indigenous Data Governance’ address concerns related to the people and purpose of data; **Collective benefit, Authority to control, Responsibility, and Ethics**
- CARE Principles are designed to be complementary to the FAIR Principles, Findable, Accessible, Interoperable, Reusable, and other mainstream data frameworks, and promote **equitable participation and outcomes from data access, use, reuse, and attribution in contemporary data landscapes**
- Operationalizing Complementary Principles: the Range of Be FAIR and CARE



- Source: Carrol et al. 2021

# Against helicopter research and ethics dumping

"Exploitative research practices, sadly, come in all shapes and sizes. **‘Helicopter research’** occurs when researchers from high-income settings, or who are otherwise privileged, conduct studies in lower-income settings or with groups who are historically marginalized, with little or no involvement from those communities or local researchers in the conceptualization, design, conduct or publication of the research**. ‘Ethics dumping’** occurs when similarly privileged researchers export unethical or unpalatable experiments and studies to lower-income or less-privileged settings with different ethical standards or less oversight."



The international journal of science / 2 June 2022

nature

## *Nature* addresses helicopter research and ethics dumping

**New framework aims to improve inclusion and ethics in global research collaborations amid wider efforts to end exploitative practices.**

Exploitative research practices, sadly, come in all shapes and sizes. ‘Helicopter research’ occurs when researchers from high-income settings, or who are otherwise privileged, conduct studies in lower-income settings or with groups who are historically marginalized, with little or no involvement from those communities or local researchers in the conceptualization, design, conduct or publication of the research. ‘Ethics dumping’ occurs when similarly privileged researchers export unethical or unpalatable experiments and studies to lower-income or less-privileged settings with different ethical standards or less oversight.

Such behaviours are wrong. They are also bad for research, which is denied crucial expertise and context. But for centuries, exploitative practices were, unfortunately, simply how researchers from around the world conducted studies in the global south. And even as the south’s capacity to do its own research has grown, elements of these practices continue.

That is why Nature Portfolio is introducing a new approach to improving inclusion and ethics in its journals (including *Nature* and all Nature Portfolio journals). The move comes as other journals grapple with similar issues and as the seventh World Conference on Research Integrity, held in Cape Town, South Africa, prepares to publish a statement urging action on them.

There are plenty of examples of the persistent imbalance in research across multiple fields. One analysis[1] of a sample of studies conducted in Africa on a range of infectious diseases found that less than half had an African first or last author. Another report[2] showed that two-thirds of high-impact geoscience articles on Africa had no African authors.

Even in development research, for which the focus is overwhelmingly on challenges facing the global south, authors from the global north wrote nearly three-quarters of papers published in the world’s top 20 development journals between 1990 and 2019 (ref. 3).

In 2018, researchers in Africa published guidelines on how samples and data from the global south can be guarded from exploitation (see *Nature* https://doi.org/gc96fq; 2018). But changing centuries of bad practice requires a joint effort across the research ecosystem.

**“Changing centuries of bad practice requires a joint effort across the research ecosystem.”**

*Nature*’s latest steps to improve inclusion and ethics are guided by the Global Code of Conduct for Research in Resource-Poor Settings, developed by TRUST – a European Union-funded project on research ethics – and by the San Code of Research Ethics, developed by the San Indigenous people in southern Africa.

In the new guidance, *Nature* will be encouraging its journals’ authors, editors and reviewers to consider the Global Code when developing, conducting, reviewing and communicating research (see go.nature.com/3ng5pbs). We also want to create opportunities for authors to be transparent about inclusion and ethics. So we are urging them, through *Nature*’s editorial-policy checklist, to provide an optional disclosure statement on inclusion and ethics that can be shared with reviewers and published in the final paper. Editors can, at their discretion, ask authors to provide a statement.

To guide authors in writing such a statement – and to help minimize the possibility of helicopter science and ethics dumping – we have developed questions drawn from key aspects of the Global Code. These include: did the research design and execution include local scientists? Is the research locally relevant? Are there plans to share the benefits of the research? Where legislation on animal welfare or environmental protection was less stringent in the local setting than where the researchers were based, was the study undertaken to the higher standards?

We are encouraging authors to cite relevant local and regional research, to improve the quality of their citations and to promote citational justice. A study[4] published on 30 May finds that scientific papers from researchers in a few countries, including the United States, China and the United Kingdom, are more likely to be cited than those on similar subjects from researchers elsewhere.

*Nature*’s new approach also aims to ensure that peer review includes representation from relevant regions and communities.

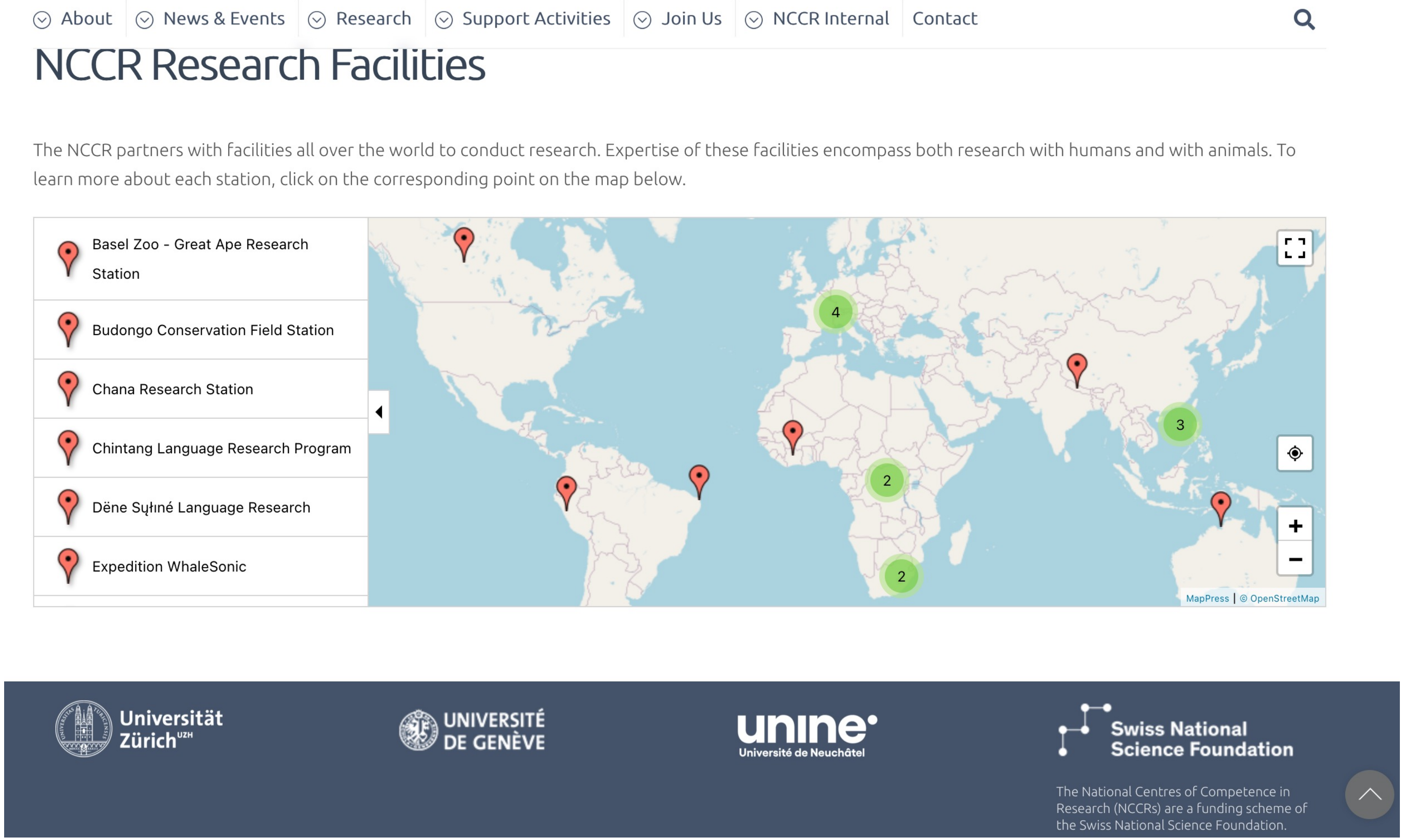
We don’t yet have all the answers, and there are nuances that we will need to grapple with. For example, it might be important to seek out local contributors when researchers are using publicly available or secondary data that they were not involved in gathering[5], to add important cultural context or an appreciation of local impacts.

*Nature* is not alone in tackling these issues. Last year, the open-access publisher PLOS announced a policy intended to combat helicopter research, and a group of researchers – including the editors of the journals *Anesthesia* and *BMJ Global Health* – proposed[6] that journals ask authors of studies conducted in low- and middle-income countries to supply statements describing how equity was promoted in their work. The statement from this year’s World Conference on Research Integrity is expected to call out inequity and unfair practices in research collaborations as a matter of research integrity.

The time is now for all stakeholders – funders, institutions, publishers and researchers – to consider how we can work together to dismantle systemic legacies of exclusion.

1. Mbaye, R. *et al. BMJ Glob. Health* **4**, e1001855 (2019).
2. North, M. A., Hastie, W. W. & Hoyer, L. *Earth Sci. Rev.* **208**, 103262 (2020).
3. Amarante, V. *et al. Appl. Econ. Lett.* https://doi.org/10.1080/13504851.2021.1965528 (2021).
4. Gomez, C. J., Herman, A. C. & Parigi, P. *Nature Hum. Behav.* https://doi.org/10.1038/s41562-022-01351-5 (2022).
5. *Lancet Glob. Health* **6**, E593 (2018).
6. Morton, B. *et al. Anesthesia* **77**, 264–276 (2022).

# International Research: example



## NCCR Research Facilities

The NCCR partners with facilities all over the world to conduct research. Expertise of these facilities encompass both research with humans and with animals. To learn more about each station, click on the corresponding point on the map below.

- Basel Zoo - Great Ape Research Station
- Budongo Conservation Field Station
- Chana Research Station
- Chintang Language Research Program
- Dëne Sųłiné Language Research
- Expedition WhaleSonic

The National Centres of Competence in Research (NCCRs) are a funding scheme of the Swiss National Science Foundation.

Source: https://evolvinglanguage.ch/nccr-research-facilities/

# Fieldwork: data collection across borders and research collaborations

Legal questions

Example: Long-form recordings (audio) in language acquisition studies

- Data transfer across borders
- Data protection – differing standards in different countries enforced

**Map 1:** Countries whose data protection legislation was studied by the LAAC Team (in red) - Argentina, Australia, Bolivia, Brazil, China, Costa Rica, Denmark, Finland, France, Ghana, India, Israel, South Korea, Namibia, Malawi, Mexico, Netherlands, Norway, Papua New Guinea, Solomon Islands, South Africa, Sweden, Tanzania, Timor-Leste, United Kingdom, Uruguay, Vanuatu, and Vietnam.

- Source: Léon et al 2024; Léon & Cristia 2024

# Conclusio

Ethical protocols have to be negotiated and adapted to specific purposes.

Central document in ALL linguistic research projects engaged in data collection: INFORMED CONSENT

Carroll, S.R., Herczog, E., Hudson, M. et al. Operationalizing the CARE and FAIR Principles for Indigenous data futures*. Sci Data 8,* 108 (2021). https://doi.org/10.1038/s41597-021-00892-0

Gary Holton, Wesley Y. Leonard, Peter L. Pulsifer, 2022. "Indigenous Peoples, Ethics, and Linguistic Data", *The Open Handbook of Linguistic Data Management*, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister. Doi: https://doi.org/10.7551/mitpress/12200.003.0008

Léon, M., Meera, S. S., Fiévet, A.-C., & Cristia*,* A. (2024). Long-form recordings in low- and middle-income countries: recommendations to achieve respectful research. Research Ethics, 20(1), 96-111. https://doi.org/10.1177/17470161231199382

Léon, M., & Cristia, A. (2024, September 3). Data Protection Handbook for Long-Form Recording Research: Navigating Data Protection Laws across the Globe. https://doi.org/10.31219/osf.io/dy4wt

Nature Editorial 2022. Nature addresses helicopter research and ethics dumping. *Nature 606, 7 (2022)* https://doi.org/10.1038/d41586-022-01423-6

Rice, Keren. 2012. Ethical Issues in Linguistic Fieldwork. In Thieberger, N (ed.) *The Oxford Handbook of Linguistic Fieldwork*. DOI: 10.1093/oxfordhb/9780199571888.013.0019

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Sebastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, & Tony Woodbury. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation 13*: 545-563

South African San Institute. *San Code of Research Ethics*. In: 2017 [Internet]. Available from: http://trust-project.eu/wp-content/uploads/2017/03/San-Code-of-RESEARCH-Ethics-Booklet-final.pdf

TRUST (2018) *The TRUST Code – A Global Code of Conduct for Equitable Research Partnerships*, DOI: 10.48508/GCC/2018.05