



Old Challenges, New Solutions? Changing DH-Approaches for Historical Newspaper Research

Eva Pfanzelter, Department of Contemporary History, University of
Innsbruck, Austria (Eva.Pfanzelter@uibk.ac.at)

ONB Labs Symposium 2024 «Newspapers as Datasets » (Nov. 25-26, 2024, Vienna)

Introduction

- Humanities with many ethical concerns
- Computer Science adapts models to handle historical complexities
- Libraries as stewards of ethically curated datasets

1. The „Old“ Challenges

- Accessibility of digital archives
- Fragility and preservation
- Search limitations
- Language and typography
- Metadata scarcity
- Readability and structure

2. Changing Approaches: Opportunities and Challenges

- Mass digitization projects
- Enhancement on OCR
- Adapted text mining and data analysis tools
- Crowdsourcing and collaborative platforms
- Europe-wide metadata standards and linked data approaches

- AI, LLMs/VLMs have changed the landscape

3. Implications for DH-Research and Future Directions

- Broadened research opportunities:
 1. Macro and Micro-Level Analysis, Mixed Methods
 2. New Research Questions
 3. Advanced Search and Discovery
 4. Multilingual and Cross-Cultural Studies
 5. Geospatial and Temporal Analysis

4. Ethics and Cultural Sensitivity in Newspaper Research

Ethical Implications in AI-Driven Historical Research

- Historical Biases → biased datasets
- Concern → amplification of existing biases

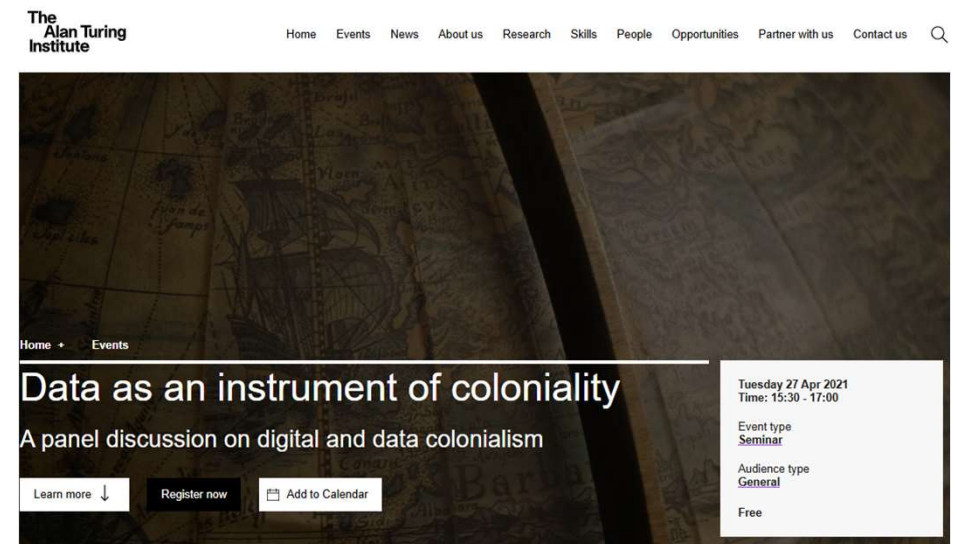
Lit.: Bender, Emily M./Gebru, Timnit u. a., On the Dangers of Stochastic Parrots, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM Digital Library), New York, NY, United States 2021, 10.1145/3442188.3445922.

How big is too big?

Size Doesn't Guarantee Diversity – on the contrary

Static Data / Changing Social Views

LLMs: Encoding Bias



The screenshot shows the event page for 'Data as an instrument of coloniality' on the Alan Turing Institute website. The header includes the institute's name and a navigation menu with links like Home, Events, News, About us, Research, Skills, People, Opportunities, Partner with us, and Contact us. The main content area features the event title, a subtitle 'A panel discussion on digital and data colonialism', and buttons for 'Learn more', 'Register now', and 'Add to Calendar'. A sidebar on the right provides event details: 'Tuesday 27 Apr 2021', 'Time: 15:30 - 17:00', 'Event type: Seminar', 'Audience type: General', and 'Free'.

<https://www.turing.ac.uk/events/data-instrument-coloniality>

About the event

Maps and surveys of "new worlds", passport photos and vaccination cards to control the movement of "impure" bodies, accounting spreadsheets used in plantations of enslaved peoples... all of these technologies suggest that data has always been an instrument of colonialism.

But can the history of European and American colonialism also help us interpret contemporary phenomena like algorithmic racial violence, quarantine apps and vaccination apartheid, the injustices of the gig economy, and disinformation campaigns that threaten our democratic futures? By examining these trends not just in the context of the past few decades, but through the lens of the past 500 years, we can perhaps gain new insights into why theories of capitalist production may not be enough to make sense of the extractivist technologies of today. These technologies need to be understood as manifestations of something deeper, constitutive of colonialism. Only by looking at the histories of colonial extraction and appropriation of land, nature and labor can we understand that our lives are being reconfigured in unprecedented ways, through the medium of data.

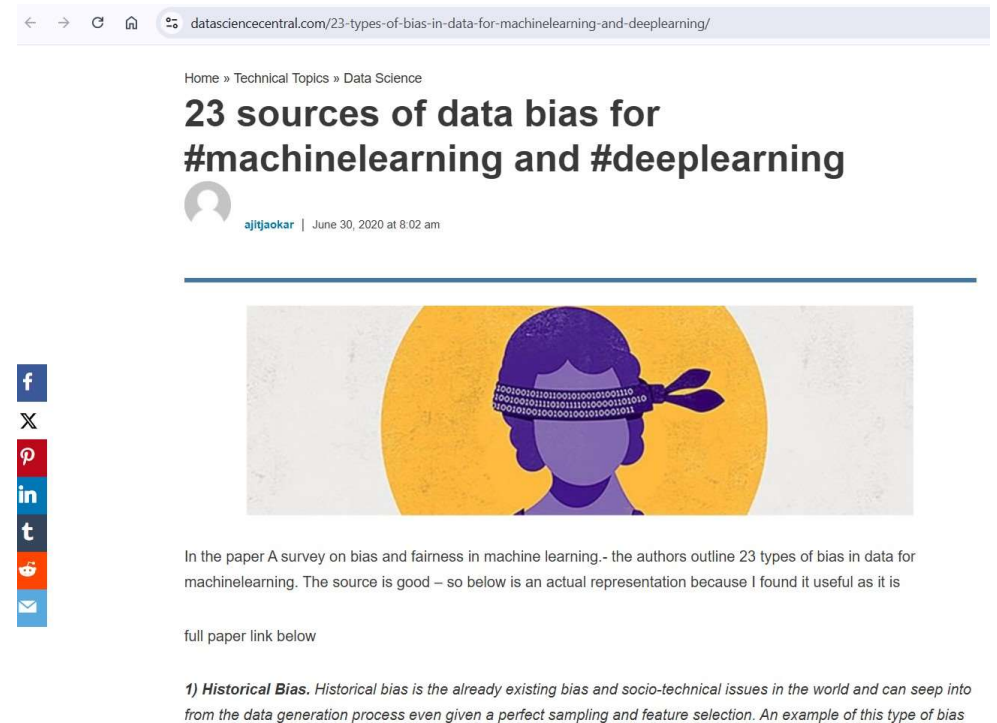
Jump to

About the event
Agenda
Register now
Speakers
Organisers
Links

Further Biases

1. Confirmation bias
2. Selecion bias
3. Historical bias
4. Survivorship bias
5. Availability bias
6. Outlier bias


23 soruces of data bias for #machinelearning and #deeplearning,
<https://www.datasciencecentral.com/23-types-of-bias-in-data-for-machinelearning-and-deeplearning/>



Home » Technical Topics » Data Science

23 sources of data bias for #machinelearning and #deeplearning

ajitjaokar | June 30, 2020 at 8:02 am

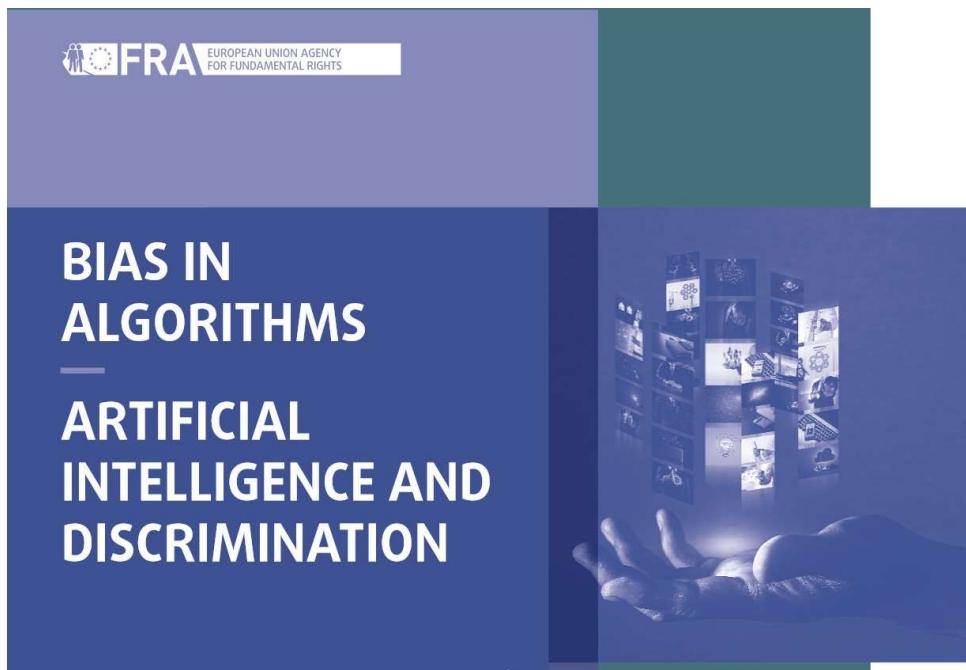


In the paper A survey on bias and fairness in machine learning.- the authors outline 23 types of bias in data for machinelearning. The source is good – so below is an actual representation because I found it useful as it is

full paper link below

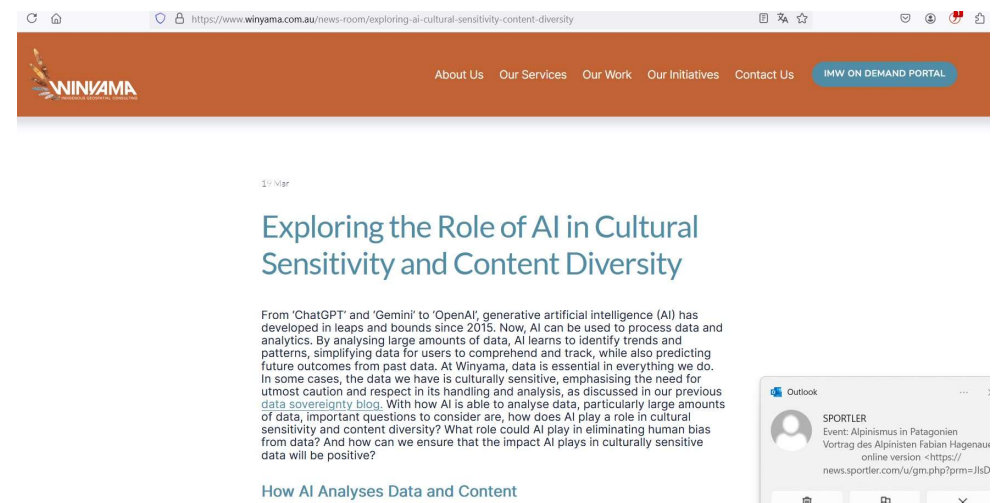
1) Historical Bias. Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection. An example of this type of bias

Further Examples



Bias in Algorithms, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

Exploring the Role of AI in Cultural Sensitivity and Content Diversity,
<https://www.winyama.com.au/news-room/exploring-ai-cultural-sensitivity-content-diversity>



4. Ethics and Cultural Sensitivity in Newspaper Research

Interdisciplinary Collaboration with a Cultural Focus

- Cultural insensitivity in AI outputs
- Misrepresentation
- Lack of contextual understanding
- Sensitive topics, such as colonialism, gender dynamics, or marginalized communities
- Sensitivity beyond the archive:
 - Ethical Digitization Practices
 - Cultural Sensitivity in Data Analysis
 - Interdisciplinary Collaboration
 - Bias Mitigation

Beelen, Kaspar/Lawrence, Jon/Wilson, Daniel C./Beavan, David, Bias and representativeness in digitized newspaper collections: Introducing the environmental scan, in: Digital Scholarship in the Humanities 38 (2023), H. 1, S. 1–22, doi: 10.1093/lc/fqac037.

Contextualizing newspapers

critical engagement with the data, understanding of potential biases, enriching computational analysis with historical context

[British Library Priorities 2023-2030](#): ... our vital interdisciplinary work yeeds to trustworthy innovation → Develop and implement a **new AI Strategy and ethical Guide** to support the next generation of our digital research.

Dagstuhl Report: Computational Approaches to Digitised Historical Newspapers
<https://drops.dagstuhl.de/entities/document/10.4230/DagRep.12.7.112>

Document  <https://doi.org/10.4230/DagRep.12.7.112>     

Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292)

■ Authors ⓘ Maud Ehrmann, Marten Düring, Clemens Neudecker, Antoine Doucet and all authors of the abstracts in this report

► Part of:  Issue: [Dagstuhl Reports, Volume 12, Issue 7](#)
 Volume: [Dagstuhl Reports, Volume 12](#)
 Journal: [Dagstuhl Reports \(DagRep\)](#)

► License:  [Creative Commons Attribution 4.0 International license](#)

► Publication Date: 2023-02-03



■ Cite As

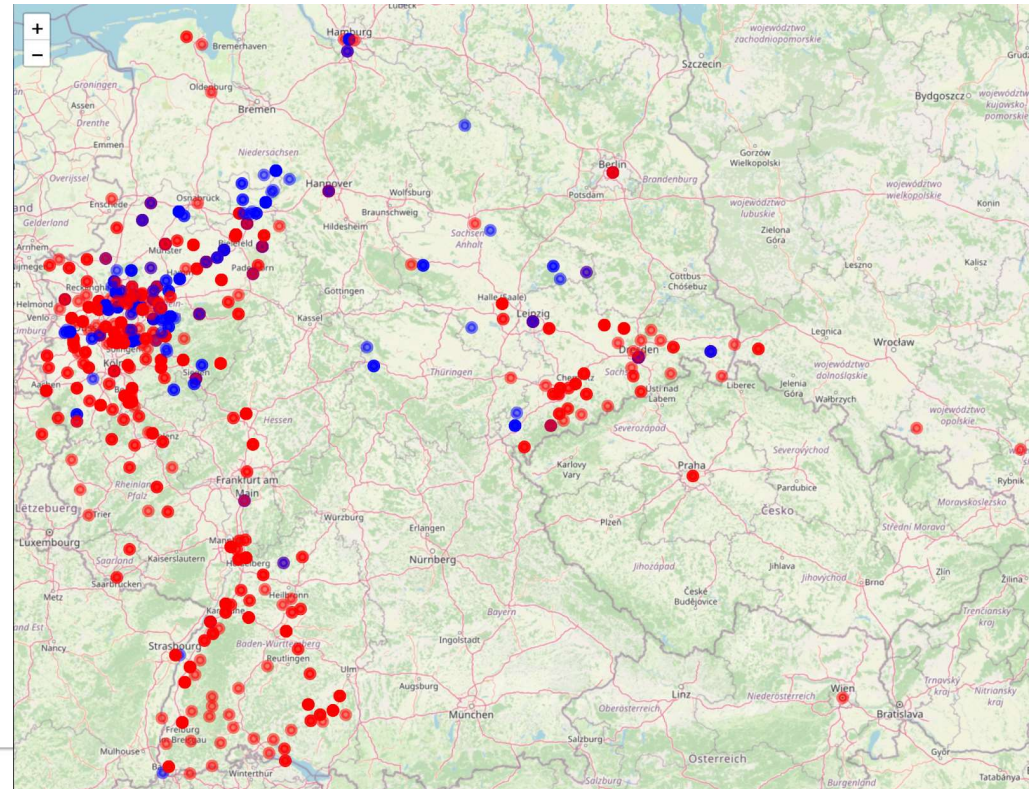
Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet. Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292). In Dagstuhl Reports, Volume 12, Issue 7, pp. 112-179, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2023)

4. Ethics and Cultural Sensitivity in Newspaper Research

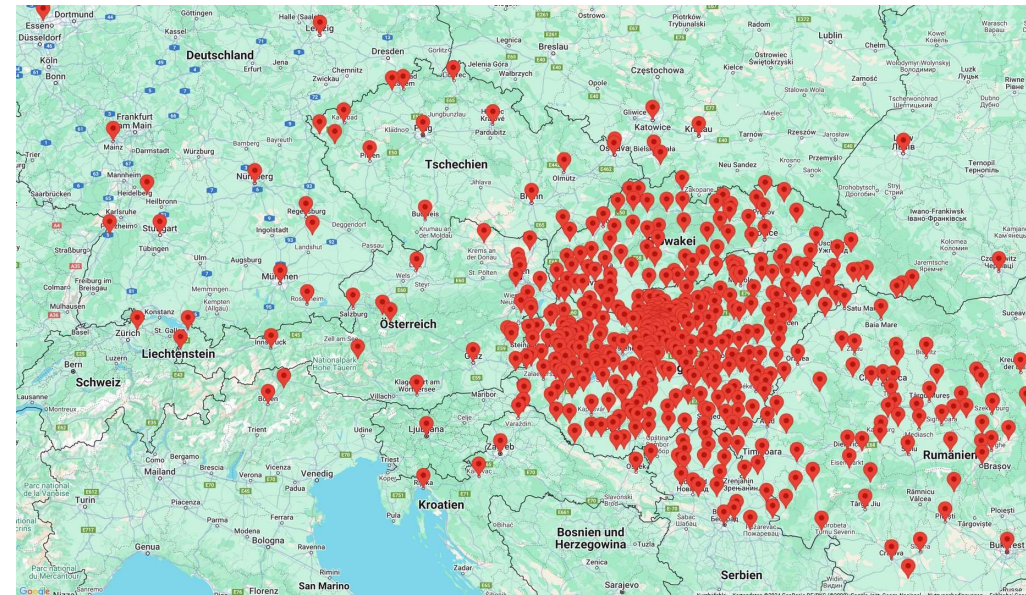
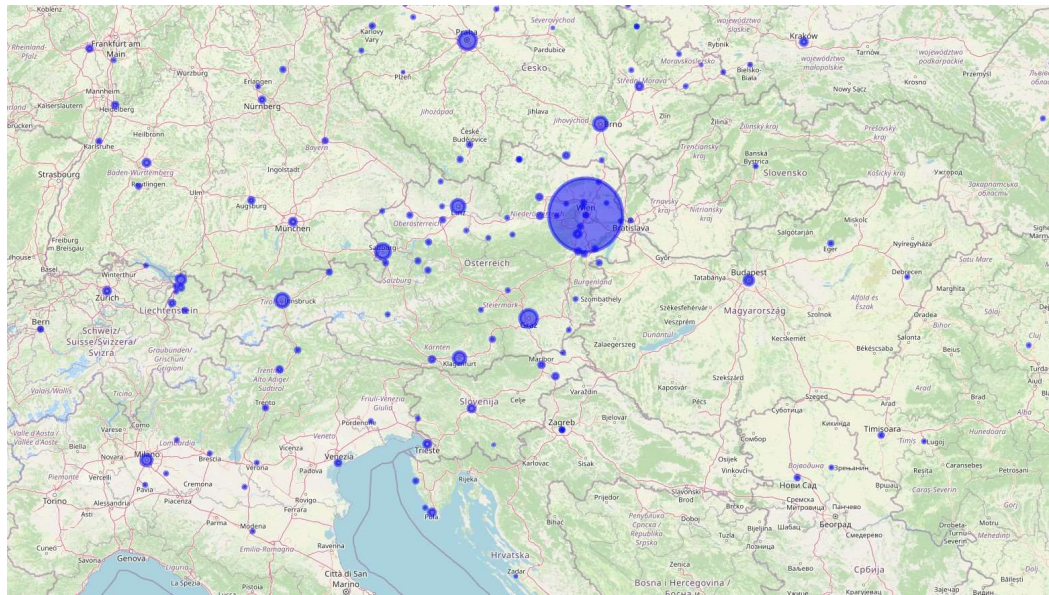
Ethical Digital Archival Practices

- Ethics in Archives = Decisions in Digital Archiving
- Digital archiving = political, cultural, and social biases
- Bias in Archival Descriptions
- Digitization disparities

Visualization of Newspapers collected in Deutsches Zeitungportal, September 2024



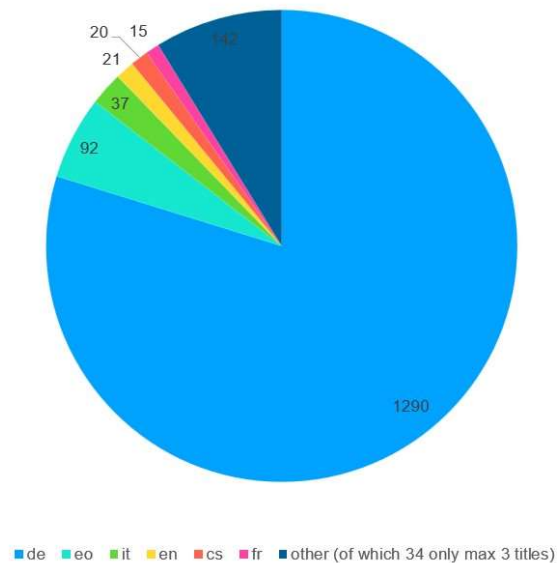
Examples: ANNO & Arcanum.com



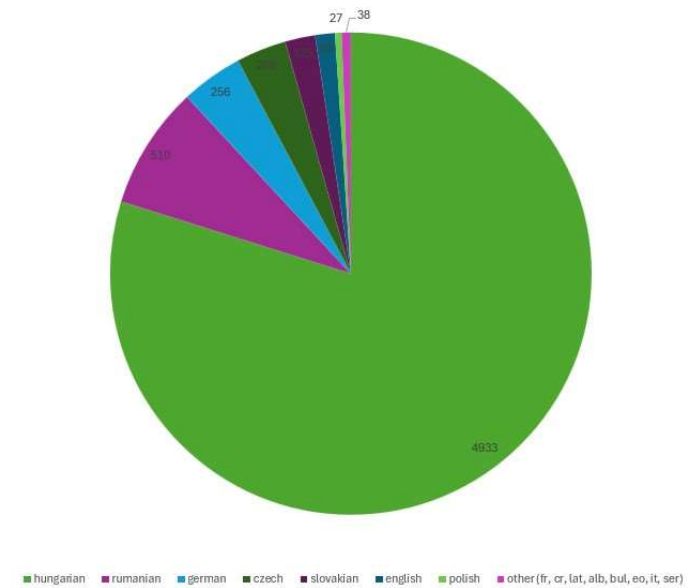
ANNO (left) Arcanum (right), roughly Austria-Hungary 1900

Examples: ANNO & Arcanum.com

number of newspaper titles in ANNO by language

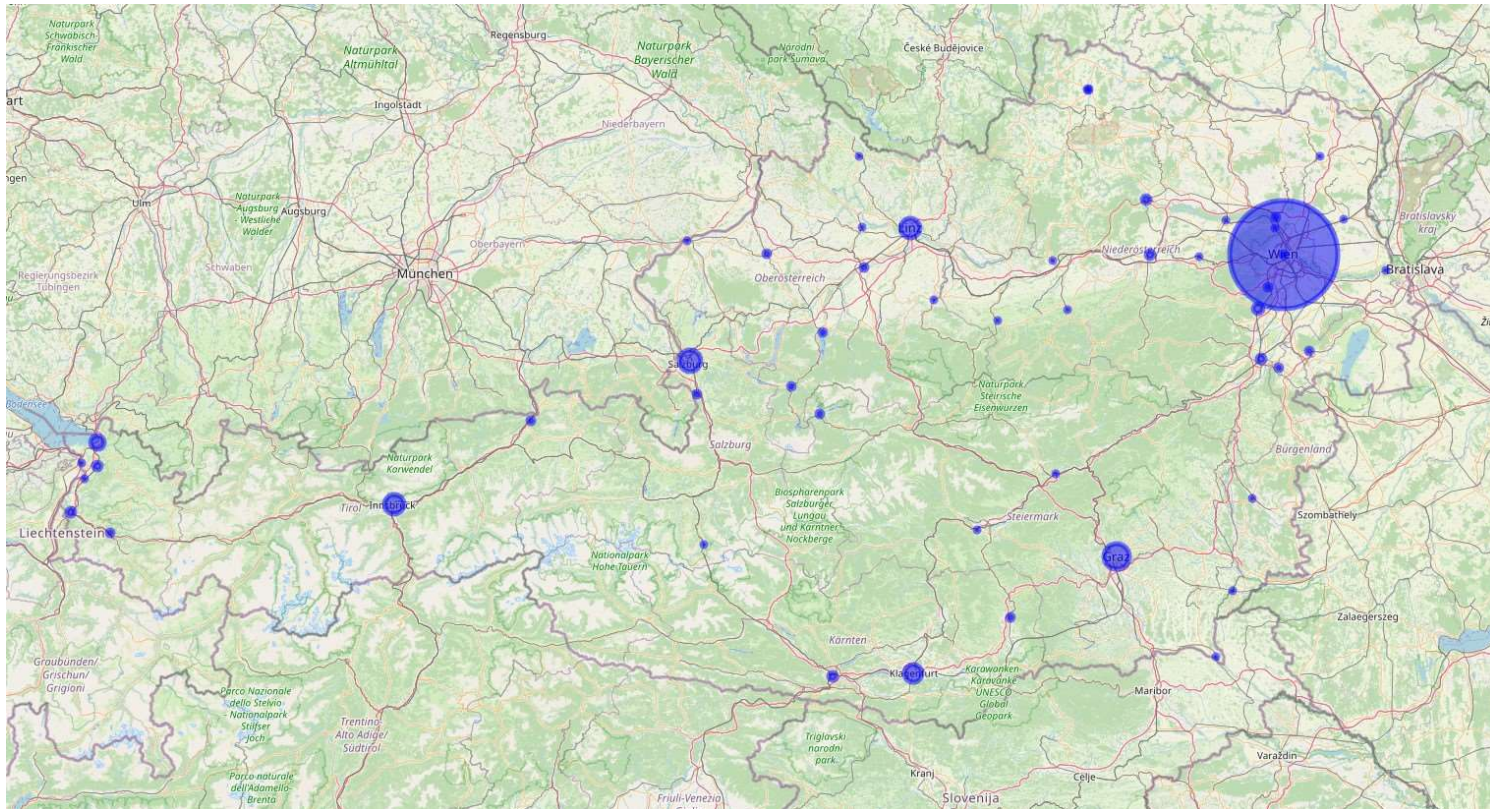


number of newspaper titles in ARCANUM.COM by language



ANNO (left) Arcanum (right) by languages

Examples: ANNO coverage contemporary Austria



ANNO within the borders of contemporary Austria

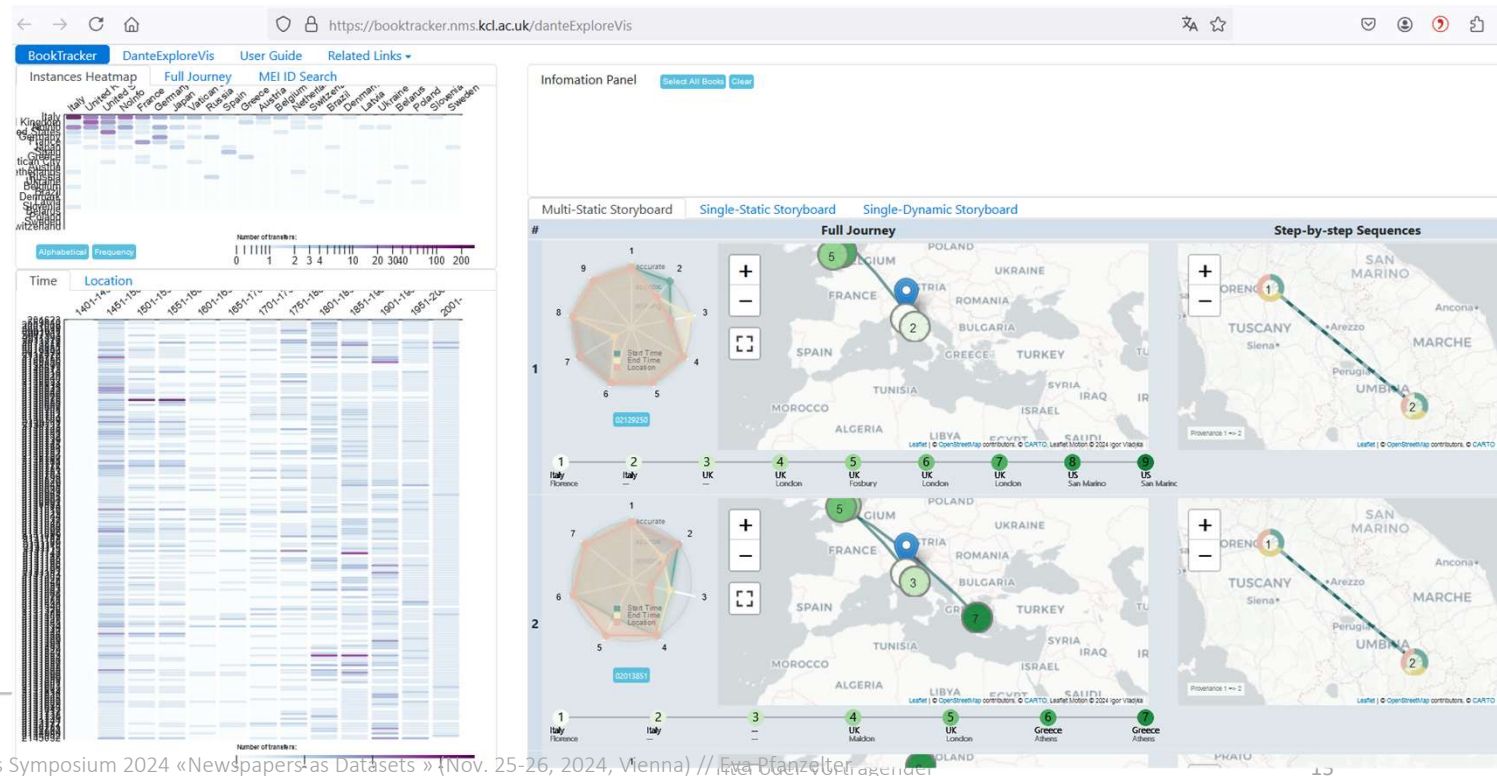
4. Ethics and Cultural Sensitivity in Newspaper Research

Emerging Technologies with Cultural Sensitivity

- LLMs/VLMs open new avenues for research (questions)
- New ways to analyze newspaper data
- Enhance study of mixed media archives (Impresso)
- Risks and biases remain
- Historic language nuances → oversimplification
- Collaboration in curating datasets
- Adapting tools/LLMs to historical contexts

Tool/Algorithm Development with Domain Sensitivity

- Importance of historians'/humanities' expertise in tool design
- Historians' knowledge influences text analysis algorithms and visualization models
- Adjustments to computational tools to account for linguistic evolution and fragmented data

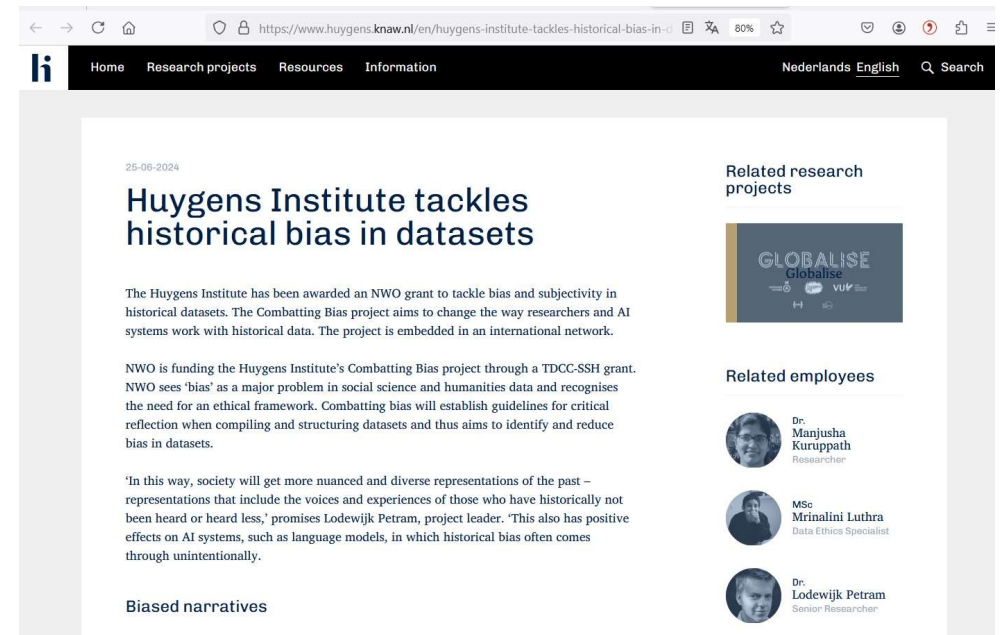
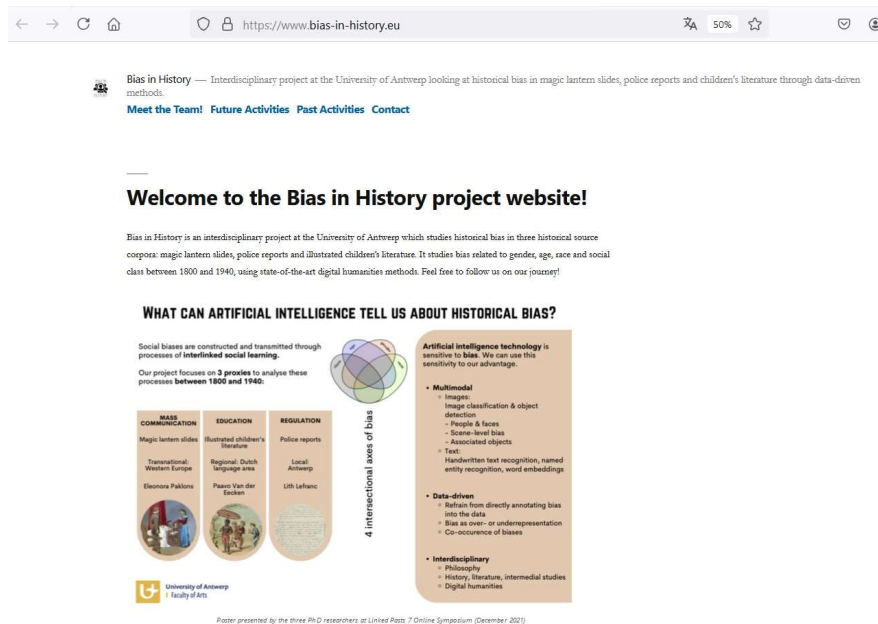


Dante Explore Vis

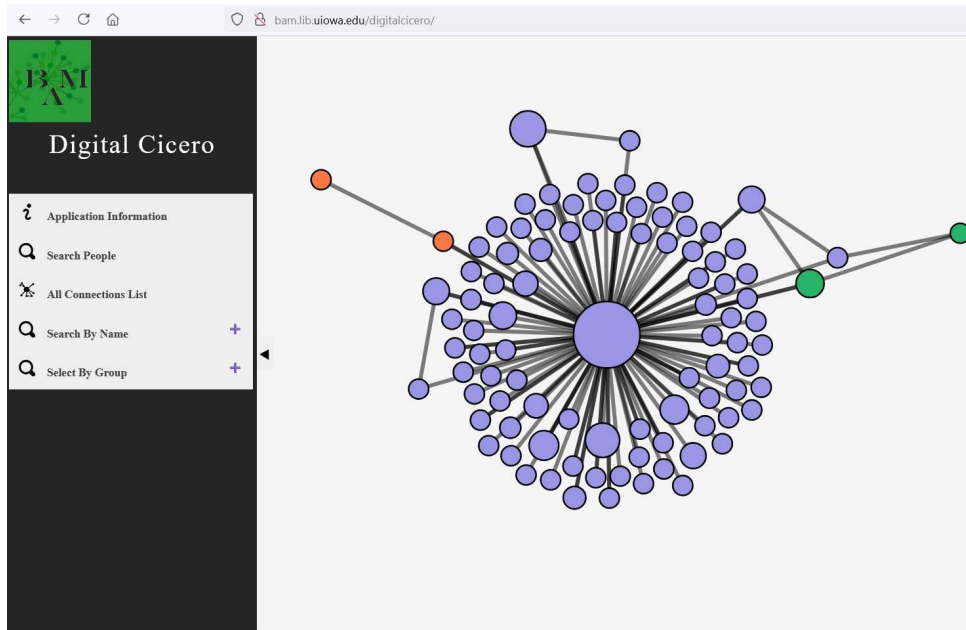
<https://booktracker.nms.kcl.ac.uk/danteExploreVis>

Historical Bias Interpretation

- Bias in Big Data
- Support representativeness and bias by enriching computational analysis
- framework to contextualize the politics of digital heritage preservation



Sentiment analysis for historical texts



Digital Cicero, <http://bam.lib.uiowa.edu/digitalcicero/>

Alcide, <https://alcidedigitale.fbk.eu/>

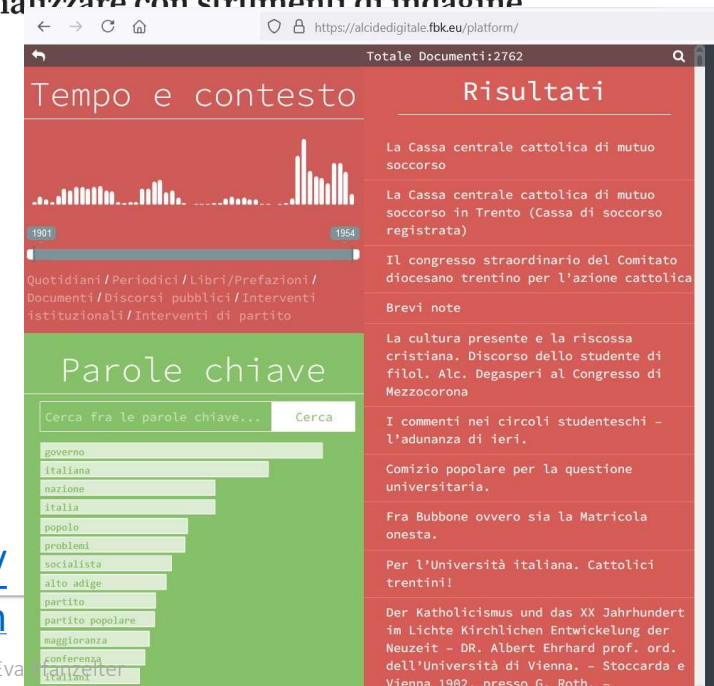
Alcide Plattform, <https://alcidedigitale.fbk.eu/platform>

<https://alcidedigitale.fbk.eu>

Alcide

Un viaggio negli scritti di De Gasperi

Nella vita di Alcide De Gasperi la parola ha avuto un ruolo centrale. Lo testimoniano le migliaia di articoli di giornale, saggi, interventi politici e discorsi istituzionali intorno a cui si è sviluppata la sua attività politica. Scritti e discorsi che oggi possiamo analizzare con strumenti di indagine innovativi.



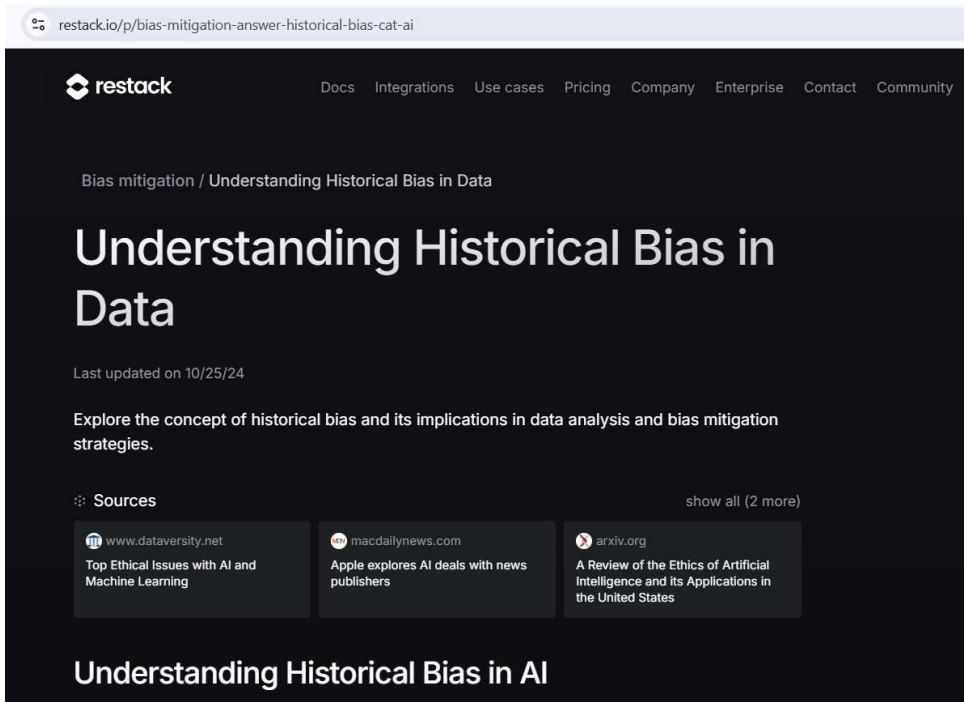
4. Ethics and Cultural Sensitivity in Newspaper Research

Strategies: Interdisciplinary collaboration for bias mitigation

- Metadata Analysis and Annotation for Contextual Clues
- Algorithmic Detection of Language Patterns
- Cross-Referencing Multiple Sources
- Training Data Selection and Model Calibration
- User Warnings and Annotations in Outputs
- Interactive User Tools for Bias Exploration
- Ethics Committees and User Feedback

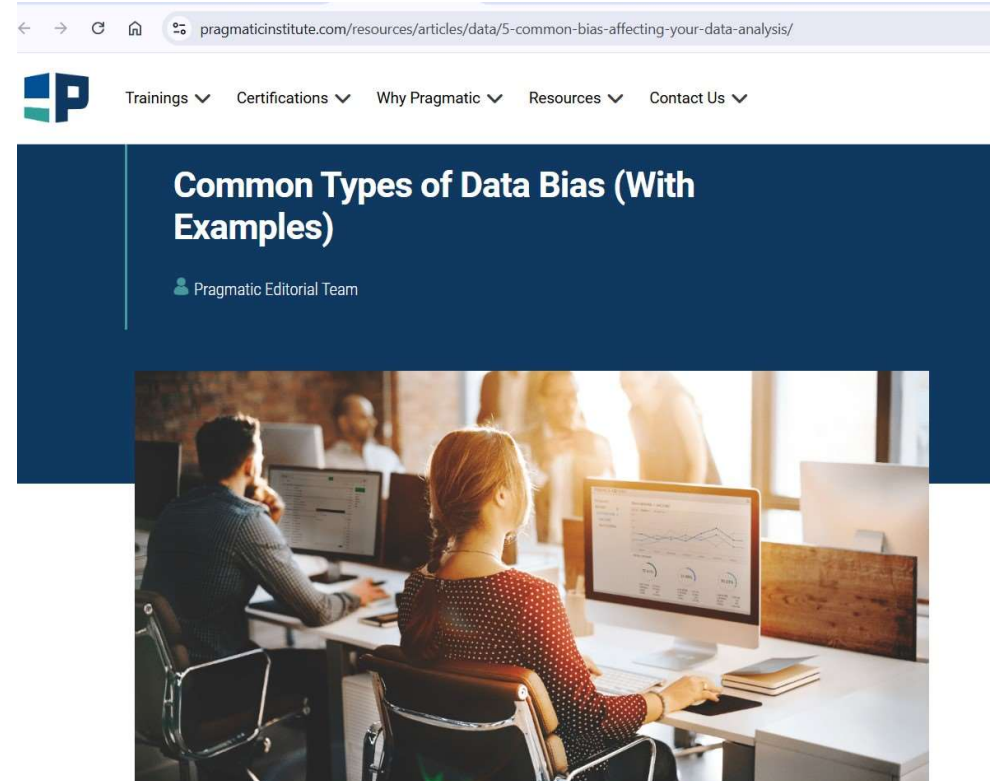
Understanding Historical Bias in Data

Common Types of Data Bias, <https://www.pragmaticinstitute.com/resources/articles/data/5-common-bias-affecting-your-data-analysis/>



The screenshot shows a web browser with the address bar displaying `restack.io/p/bias-mitigation-answer-historical-bias-cat-ai`. The Restack logo is in the top left, and a navigation menu includes Docs, Integrations, Use cases, Pricing, Company, Enterprise, Contact, and Community. The article title "Understanding Historical Bias in Data" is prominently displayed, with a subtitle "Bias mitigation / Understanding Historical Bias in Data". Below the title, it states "Last updated on 10/25/24". A brief description reads: "Explore the concept of historical bias and its implications in data analysis and bias mitigation strategies." A "Sources" section lists three references: `www.dataversity.net` (Top Ethical Issues with AI and Machine Learning), `macdailynews.com` (Apple explores AI deals with news publishers), and `arxiv.org` (A Review of the Ethics of Artificial Intelligence and Its Applications in the United States). A "show all (2 more)" link is also present. At the bottom, the text "Understanding Historical Bias in AI" is visible.

Understanding Historical Bias in Data and AI, <https://www.restack.io/p/bias-mitigation-answer-historical-bias-cat-ai>



The screenshot shows a web browser with the address bar displaying `pragmaticinstitute.com/resources/articles/data/5-common-bias-affecting-your-data-analysis/`. The Pragmatic Institute logo is in the top left, and a navigation menu includes Trainings, Certifications, Why Pragmatic, Resources, and Contact Us. The article title "Common Types of Data Bias (With Examples)" is prominently displayed, with the author "Pragmatic Editorial Team" listed below it. A photograph of people working at computers in an office setting is shown below the text.

4. Ethics and Cultural Sensitivity in Newspaper Research

Increase Accessibility and Cultural Sensitivity

Humanities for All

Mukurtu CMS: An Indigenous Archive and Publishing Tool

Mukurtu is a content management system and digital access tool for cultural heritage, built for and in ongoing dialogue with indigenous communities. Developed and maintained at the Center for Digital Scholarship and Curation at Washington State University, the free and open source platform is designed to meet the particular curatorial and access needs of indigenous peoples. Mukurtu offers the ability to provide differential access to community members and the general public and to create space for traditional narratives and knowledge labels that foreground Indigenous knowledge in the metadata of digitized cultural heritage materials.

PROJECT DIRECTOR(S)

Kimberly Christen

HIGHER ED INSTITUTION(S)

Washington State University
Center for Digital Scholarship
and Curation

LOCATION(S)

Pullman, WA

COMMUNITY PARTNER(S)

Indigenous individuals and
groups

HUMANITIES DISCIPLINE(S)

International and Area
Studies, Library Science /

Humanities for all: Mukurtu Archiv, <https://humanitiesforall.org/projects/mukurtu-an-indigenous-archive-and-publishing-tool>

Sami Archive: <https://digisamiarchives.org/>

← → ↻ 🏠 📄 digisamiarchives.org/#:~:text=The%20Digital%20Access%20to%20the%20Sámi%20Heritage%20... ☆ E

A multi-disciplinary collaboration project Digital Access to the Sámi Heritage Archives working towards the goal of improving accessibility to the Sámi Cultural heritage

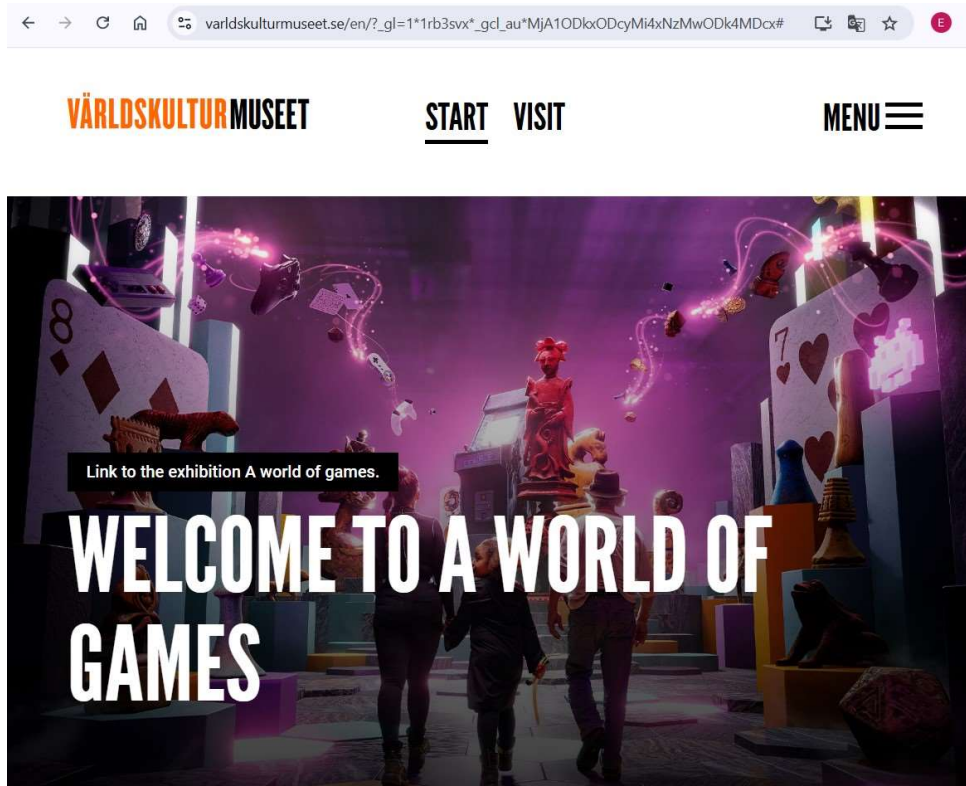
The Digital Access to the Sámi Heritage Archives project

The materials of the Sámi cultural heritage exist in several archives and collections, as due to historical reasons artefacts have also been stored in museums and

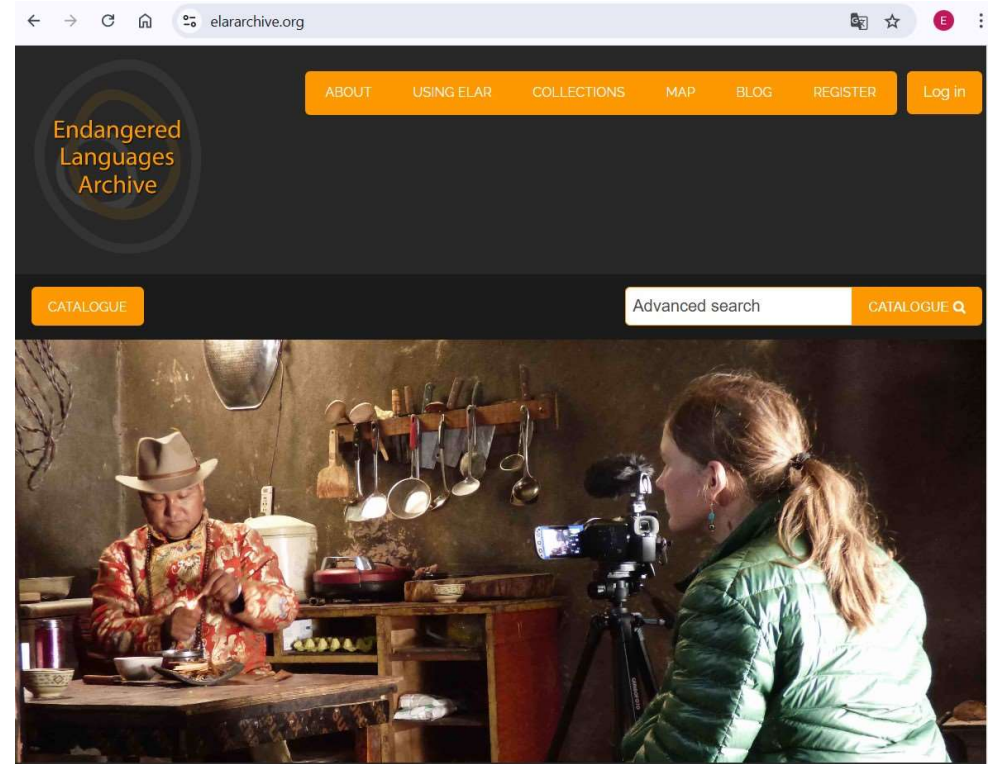


Other examples

Endangered Languages Archive (ELAR):
<https://www.elararchive.org/>



Museum of World Culture,
<https://www.varldskulturmuseerna.se/en/#>



Conclusion

Future challenges

- Preservation and sustainability
- „New“ ethical considerations
- Inclusiveness and accessibility



www.uibk.ac.at