



GdR TAL CNRS « Traitement Automatique des Langues et les Humanités Numériques » (Nov. 7, 2024, La Rochelle)

Innsbruck, Austria (Eva.Pfanzelter@uibk.ac.at)

Unlocking the past interdisciplinarily

- 1. Enrichment in Collaboration
- 2. Unique Challenges Computer Science Faces in History
- 3. Understanding the Intellectual Rigor of Historical Research



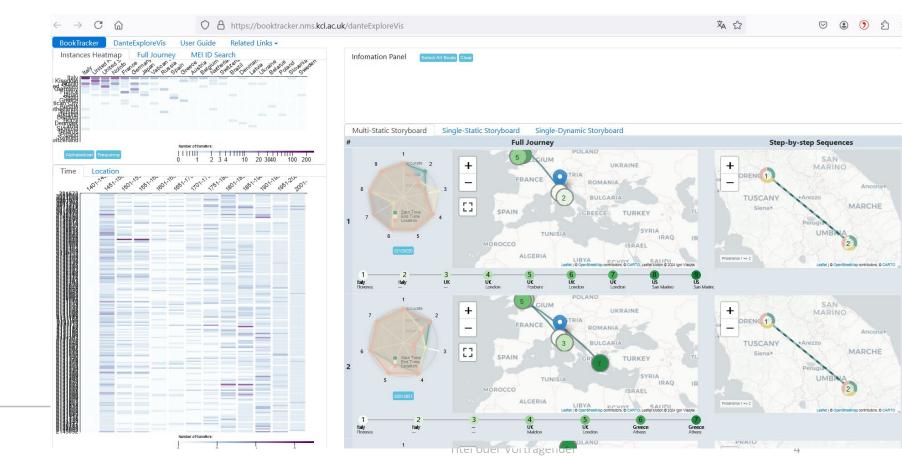
1. Enrichment in Collaboration

How Historians and Computer Scientists Benefit Each Other



a. Tool Development with Domain Sensitivity

- Importance of historians' 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.

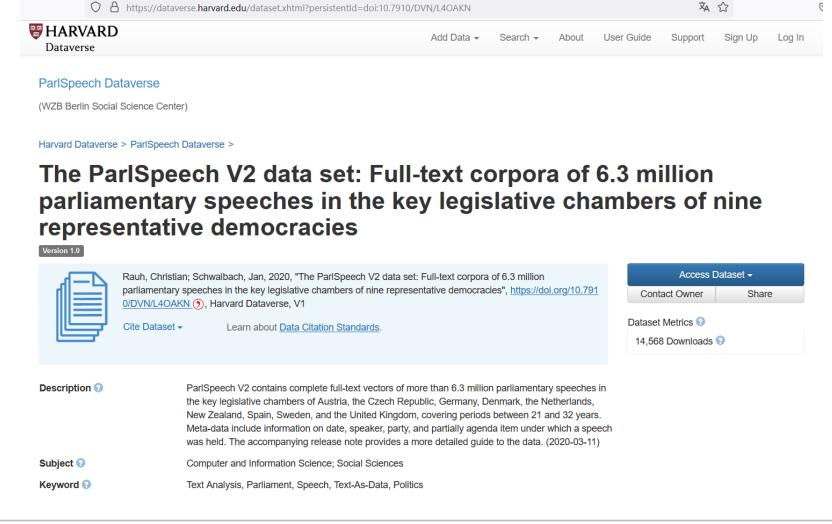


Dante Explore Vis

https://booktracker.nms.kcl.ac.u k/danteExploreVis



Example





ParlSpeech V2:

Other Examples

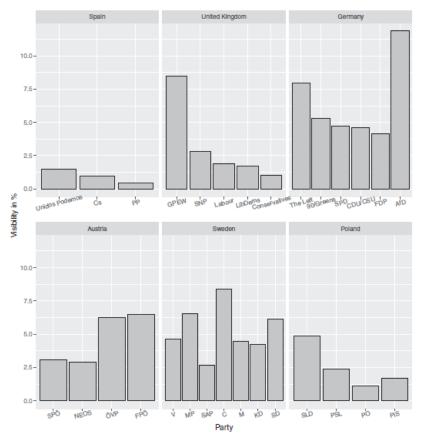


Figure 1. Migration-related status posts as a proportion of the total.

Note: Data are percentages and shown grouped by country. Parties are shown left to right in order of the Chapel Hill leftright ideological score. For an overview see Table A1 in the Appendix.

Political migration discourses on social media:

https://www.tandfonline.com/doi/full/10.1080/1369183X.201

9.1665990



Analysing complexity of parliamentary speeches:

https://www.awendsjo.com/posts/complex language/



https://www.awendsjo.com/posts/complex_language/

Albert Wendsjö research posts

Analysing complexity of parliamentary speech

September 23, 2023

Introduction

In this post I'll walk through the steps of an analysis of parliamentary speech. I'll be using data from ParlSpeechV2 to assess the complexity of how politicians talk in parliaments, as well as analyse predictors of the complexity of parliamentary speech. In this post I go through how to use ParlSpeech data, and how to merge it with other sources such as ParlGov and EveryPolitician. I will then go through how to evaluate the complexity a speech using text-as-data methods, some exploratory analysis and visualization, and some statistical inference

Content

- Data Preparation
- Evaluating the Complexity of Language
- Exploratory Analysis
- Statistical Analysis

Data preparation

ParlSpeech data

To start of I'll be using data from ParlSpeechV2, which is a dataset collected by Rauh and Schwalbach (2020). It contains parliamentary data from eight western countries, with a coverage of about 21-32 years depending on the country. To simplify matters we'll only use the Swedish data this time, but this could be

b. Ethics and Cultural Sensitivity

- Sensitive topics, such as colonialism, gender dynamics, or marginalized communities
- Ethics in Archives = Decisions in Digital Archiving
- Digital archiving = political, cultural, and social biases
- Interdisciplinary approaches
- Sensitivity beyond the archive:
 - Ethical Digitization Practices
 - Cultural Sensitivity in Data Analysis
 - Interdisciplinary Collaboration
 - Bias Mitigation

Dagstuhl Report: Computational Approaches to Digitised Historical Newspapers

https://drops.dagstuhl.de/entities/document/10.4230/DagRep.12.7.112





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





Examples

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.

Kimberly Christen

HIGHER ED INSTITUTION(S)

Washington State University
Center for Digital Scholarship
and Curation

LOCATION(S)

Pullman, WA

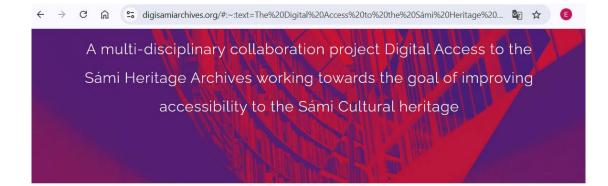
COMMUNITY PARTNER(S)

Indigenous individuals and groups

HUMANITIES DISCIPLINE(S)
International and Area
Studies, Library Science /

PROJECT DIRECTOR(S)

Sami Archive: https://digisamiarchives.org/



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



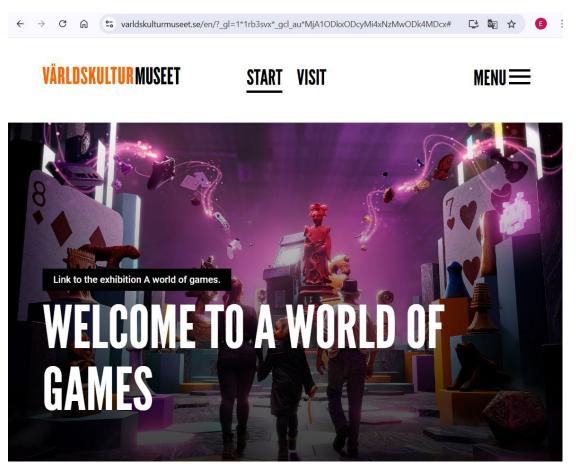
Humanities for all: Mukurtu Archiv,

https://humanitiesforall.org/projects/mukurtu-an-indigenous-

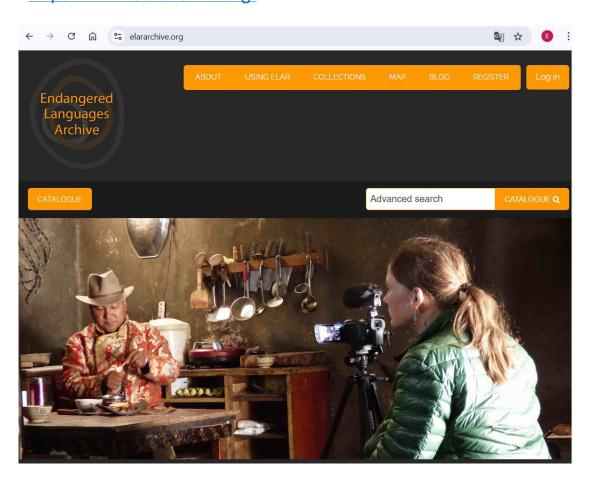
archive-and-publishing-tool



Other examples



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



Museum of World Culture,

https://www.varldskulturmuseerna.se/en/#



c. Modelling Uncertainty and Complexity

Digitizing History

- Sources are often incomplete, fragmented, or contradictory
- Unlike the structured data sets computer scientists are used to, historical data is handling uncertainty
- provisional evidence and ambiguous interpretations, which may shift as new discoveries are made



Examples

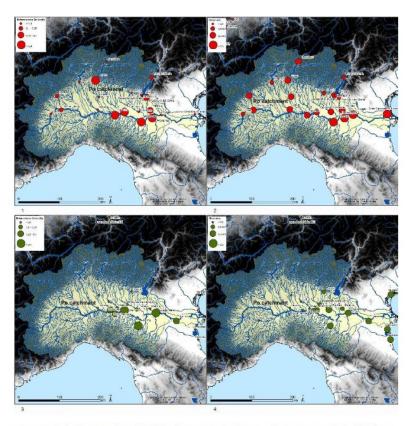
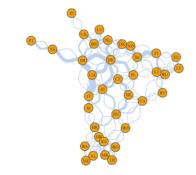
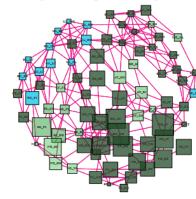


Fig. 4 Nodes in the network model of riverine transport in the Po catchment: 1 In period I (1st-5th cent. CE) sized according to their betweenness centrality. – 2 In period I (1st-5th cent. CE) sized according to their closeness centrality. – 3 In period II (6th– early 11th cent. CE) sized according to their betweenness centrality. – 4. In period II (6th– early 11th cent. CE) sized according to their closeness centrality. – (Cartography and Data Collection L. Werther, Universität Jena; Network Analysis J. Preiser-Kapeller, ÖAW).

Statistical Monitoring of European Cross-Border Physical Electricity Flows Using Novel Temporal Edge Network Processes: https://arxiv.org/abs/2312.16357



(a) The edge thickness implicates the strength of the flow.



(b) The size of the nodes implicates the strength of the electricity exchange and the colour its proportion of electricity generated with renewable energy sources (light green corresponds to the higher proportion, blue corresponds to the missing information).

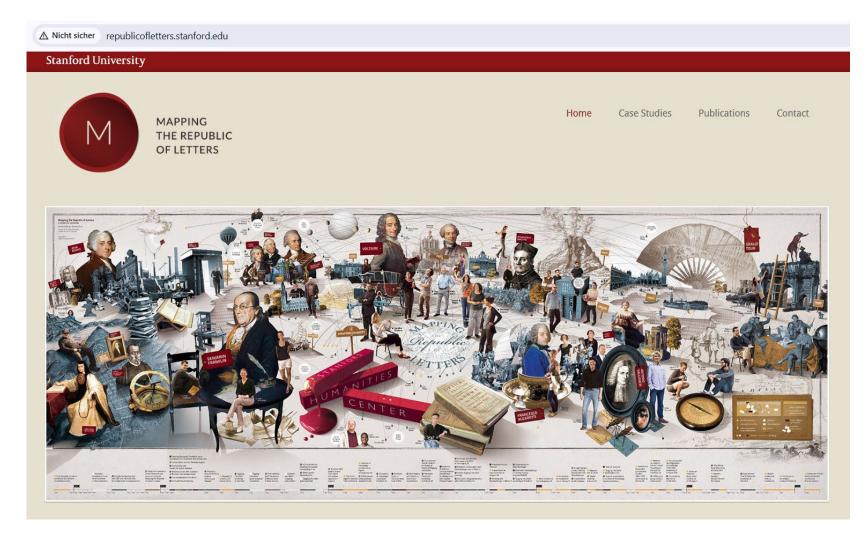
Fig. 8: Conventional representation of the cumulated cross-border physical electricity flow in the year 2019 (a) and new representation (b) following the description in Section 3.1.2.

Connecting Harbours: A Comparison of Traffic Networks Across Ancient and

Medieval Europe: https://arxiv.org/pdf/1611.09516



Example







2. Unique Challenges Computer Science Faces in Digital History

How to deal with uncertainty

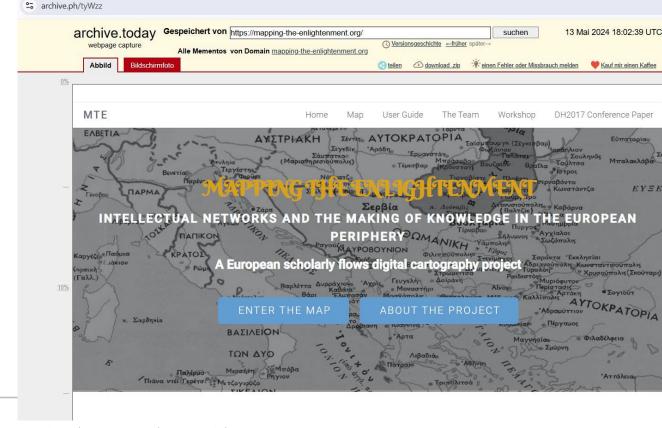


a. 1. Inconsistent, Fragmented, and Ambiguous Data

- Technical expertise and understanding the context
- Preserving gaps
- Train machine learning models
- Uncertainty detection, e.g., in Historical Databases

Example: Mapping the enlightenment:

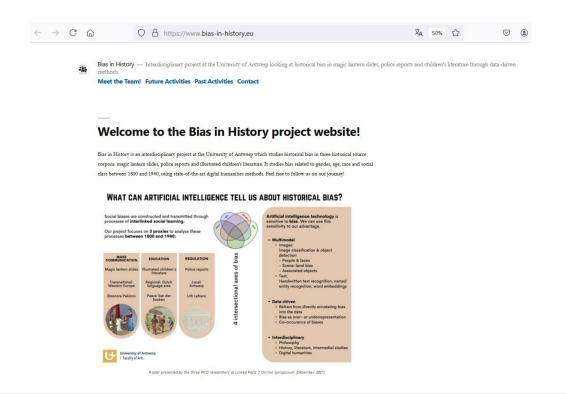
https://archive.ph/tyWzz

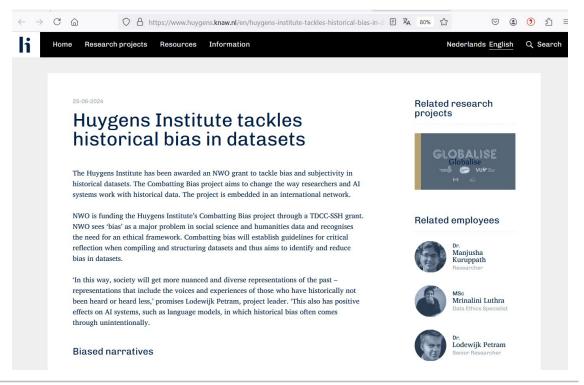


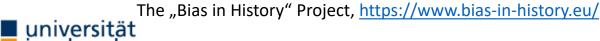


b. Historical Bias Interpretation

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

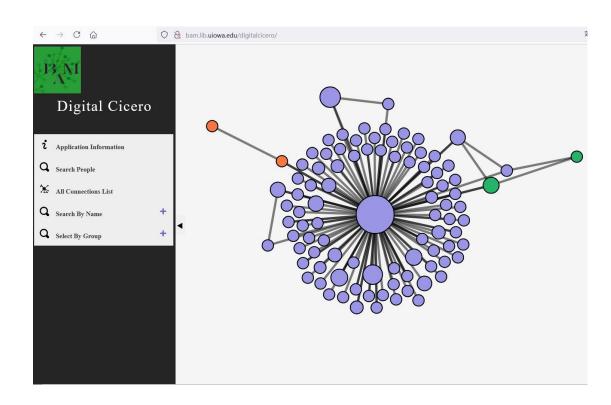






Alcide

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.



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

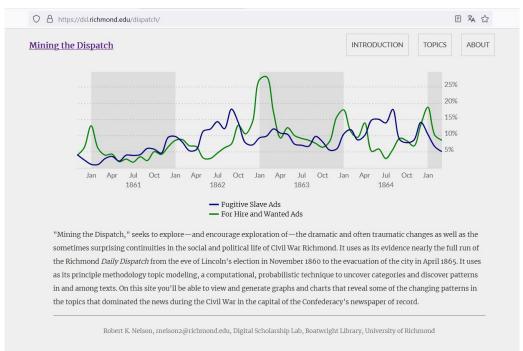
Alcide, https://alcidedigitale.fbk.eu/

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

universität innsbruck

c. Contextualizing Data Beyond Patterns

- Historical reasoning
- Contextualization
- Historical Causal inference



Mining the Dispatch, https://dsl.richmond.edu/dispatch/

The Digital Panopticon, https://www.digitalpanopticon.org/



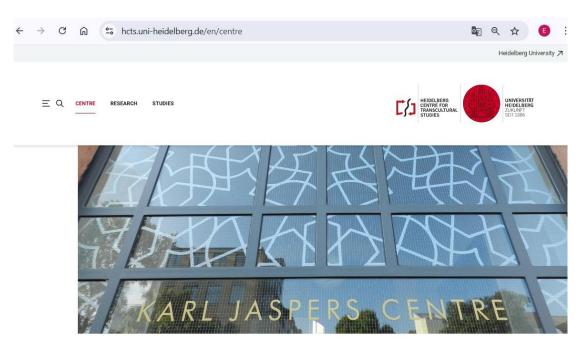


Examples



The French Revolution Digital Archive, https://frda.stanford.edu/

Heidelberg Center for Transcultural Studies, https://www.hcts.uni-heidelberg.de/en/research/projects



CENTRE

The Heidelberg Centre for Transcultural Studies was founded in 2013 as host to the Cluster of Excellence "Asia and Europe in a Global Context: The Dynamics of Transculturality". Since the end of the Cluster Asia and Europe in 2019, the HCTS has continued to foster research and teaching with a transcultural agenda.





3. Understanding the Rigor of Historical Research

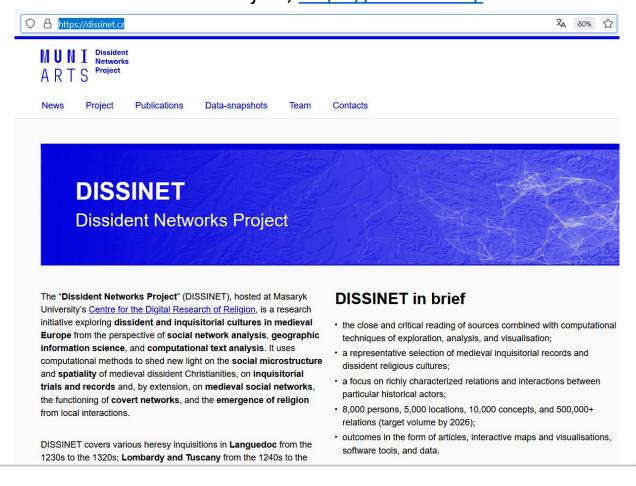
How to do digital source criticism



a. Source Criticism

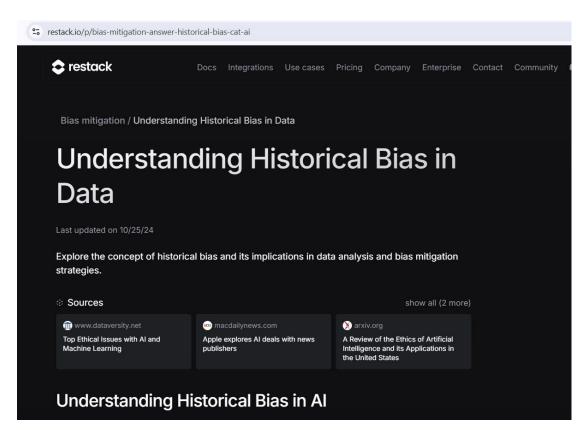
- Authenticity of Sources
- Read between the lines
- External and internal source criticism
- Distortions over time
- Changes in language
- Source Critical data models and visualizations and multimodal context

Dissident Networks Project, https://dissinet.cz/





Examples: Understanding Historical Bias in Data



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

Commont Types of Data Bias,

https://www.pragmaticinstitute.com/resources/articles/data/5-common-bias-affecting-your-data-analysis/





Further Examples

- 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





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

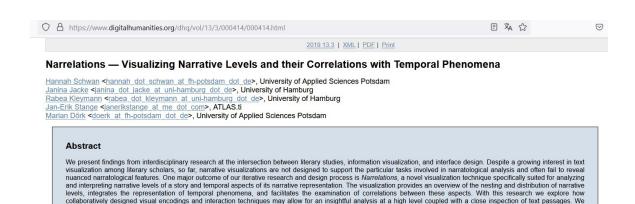
full paper link below

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



b. Narrative Construction

- Historians as Digital Storytellers
- Multimodal narration

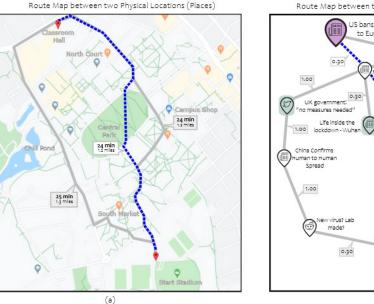


discuss prior work relevant to our research objectives and explain the specific characteristics of narrative levels and temporal aspects of narrative representation. After describing the research process and design principles, we apply the visualization on a test corpus of eight annotated German short stories and demonstrate its heuristic value for literary



analyses and interpretations. In particular, we explore the intricate connections between the literary content of the novellas and their narrative form

Narrative Maps, https://arxiv.org/abs/2009.04508



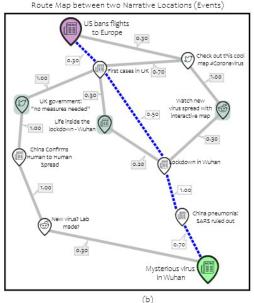


Fig. 1. Illustration of the Route Map Metaphor. (a) A route map showing how to get from Start Stadium to Classroom Hall. (b) A narrative route map showing how to get from the Mysterious Virus in Wuhan event (starting event \bigcirc) to the US banning flights to Europe (ending event \bigcirc). We highlight some representative landmarks \bigcirc for each route in the narrative map.

Narrelations,

https://www.digitalhumanities.org/dhq/vol/13/3/000414/000414.html



Examples in Tableau



While data visualization often conjures thoughts of business intelligence with button-down analysts, it's usually a lot more creative and colorful than you might think. There are many wide-ranging applications from business dashboards to public health visualizations to pop culture trend breakdowns. Great and beautiful data visualization requires graphic design and storytelling skills in addition to great analysis skills.

In this article, we're going to highlight some of the most influential, most interesting, and most revealing visualizations out there. We'll look at some notable historical examples first and then fast forward and discuss some more contemporary visualizations. Also, be sure to check out our <u>detailed guide to data visualization</u> or check out some of our <u>favorite examples</u>.

In this article, we'll cover:

- What is data visualization?
- · The best data visualization examples
- 1. Napoleon March Map
- 2. 1854 Broad Street Cholera Outbreak Map
- 3. Causes of Mortality in the Crimean War
- 4. New Chart of History

Data is Beautiful,
https://www.tableau.com/visualizat
ion/data-visualization-examples



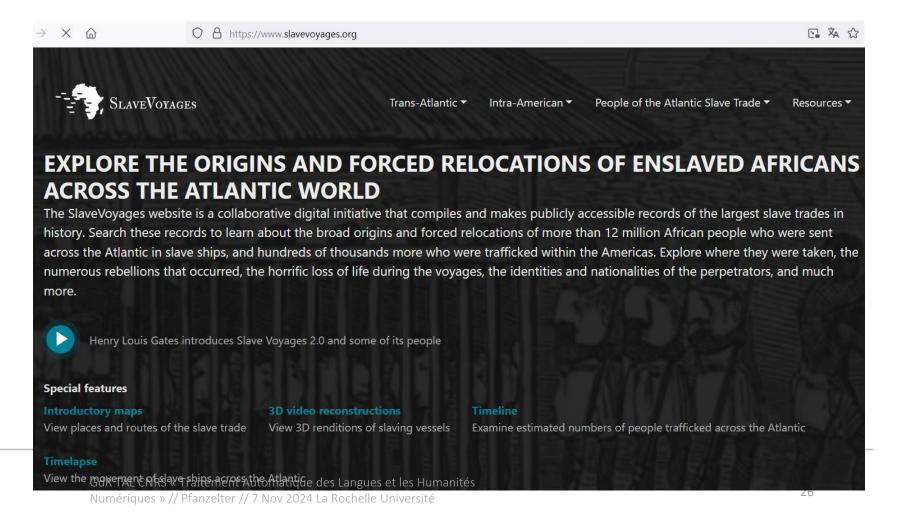
3. Temporal and Cultural Sensitivity

- What one time was true, may no longer be so
- Also: diverse perceptions of time

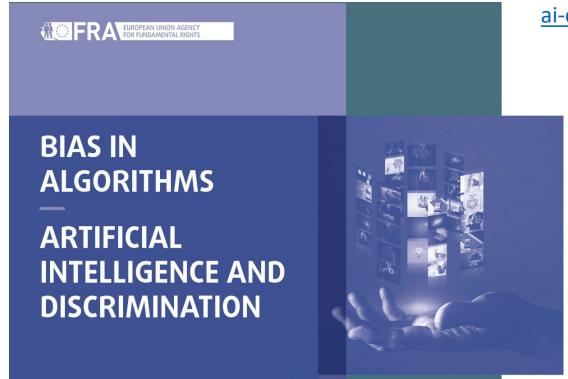
The Transatlantic Slave Trade Database,

https://www.slavevoyages.org





Examples – again bias



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



Bias in Algorithms,

https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-

bias-in-algorithms en.pdf



Conclusion: Do historians still need CS?

- Collaboration is important
- Interdisciplinarity can unlock the past
- Mitigating historical bias = not repeat the mistakes of the past?



