TRACKING AND REPRESENTING THE PROVENANCE OF GENDER DATA IN THE DIGITAL HUMANITIES

WORKING TITLE

LISA POGGEL

I will conduct a quantitative meta-analysis of how gender is represented in digital humanties (DH) projects.
The aim is to identify successful approaches and best practices for tracking and representing the provenance of historical gender data in a linked data setting

ABOUT ME

I'm a PhD student and research assistant in digital humanities at Freie Universität Berlin with a background in history and political science.

Affiliation

Freie Universität Berlin, Germany
Chair of Digital Humanities

Supervisor

Frank Fischer

Introduction

Digital humanities projects working with prosopographical data face a dilemma: historical gender data is inaccurate, messy, and mostly binary, but leaving it out means rendering gender as a social category of difference invisible. Approaches to tracking and representing the provenance of historical gender data vary and standardization is needed to improve interoperability and interpretability.

The problem

When it comes to managing gender data, common challenges and beliefs in the DH domain seem to be:

- In the particular historical context we are concerned with, gender is a binary social category and can be modelled as such.
- We cannot fact check gender data for historical (dead) individuals.
- Gender data is a byproduct of our reconciliation workflow, we do not control how gender is recorded in the GND (Integrated Authority File) or other authority files.
- If we do not resort to names and pronouns to infer gender, we cannot represent gender at all. Better include it falsely than leave it out.
- We want to model gender accurately and we know specialized vocabularies, but we can't use them because modern categories such as gender identity are not applicable to the historical context.
- Determining a historical person's gender is often an interpretative task. We cannot track this kind of contextual provenance automatically and we do not have capacities to model this sort of provenance for every data item manually.

Background: Treatment of gender in standards often used in the DH domain

CIDOC CRM
- Entity *E76 Gender* removed in 2001
- Gender usually modelled via property *P2 has type*
- Alternatives discussed but complex, f.e. via gender assignment event [1]

TEI
- Encoding texts in XML-TEI is a common step in DH projects
- Standard approach is @sex attribute with values M, F, O (other), N (none)
- But any external standard can be used
- Alternative approaches exist but complex: f.e. via <trait type="gender"> [2]

GND
- Allowed values for gender in GND entities "male" and "female", without reference
- Other values allowed in free form in different field only if reference is provided

Of course, not all is bad...

HOMOSAURUS
- Expansive linked data vocabulary for contemporary LGBTQ terms
- Not the only one, f.e. GSSO ontology

WiGeDi WIKIDATA
- Wikimedia-funded research project, currently reevaluates Wikidata model:
- Require references for gender statements
- Define standards for references
- Remove *P21 sex or gender*
- Separate gender identity and modality

1

Quantitative content analysis of DH projects

How is gender represented in DH datasets?

Is provenance information provided?

Which gender categories are employed?

Which standards, ontologies, vocabularies are used?

Phase 1

Goals and methodology

The goal is to take stock: How many projects actually work with gender data, how many employ linked data technologies, and which standards are used? The projects are identified by scraping and parsing books of abstracts of DH conferences for links to DH projects, which are pre-filtered. Out of 13,000 URLs, a random sample of (so far) n=500 is drawn, and each URL is evaluated by a coder. Inter-coder reliability scores are calculated to assess whether coding decisions align.

Some VERY preliminary findings

The following tables and the stacked bar plot present a few findings (obviously still too few to be representative) from the first 500 annotated links, 35 of which contained gender data.

Gender expressions	Count	Percentage
Gendered pronouns and/or nouns	15	44%
Binary categories	13	38%
Three categories (incl. "unknown", "other", ...)	4	12%
More than three categories	2	6%

Provenance	Count	Percentage
External databases or encyclopaedias	8	35%
Archival Sources	7	30%
Inferred from gendered pronouns, names or nouns	4	17%
Research literature	2	9%
Provided by relatives or friends	1	4%
Not specified	1	4%

Fig. 1: Absolute counts and percentages of gender expressions by availability of provenance information

So far, only 5 out of 18 digital humanities datasets with structured data followed a standard to represent gender data. Only three projects employed linked data technologies.

2

Interview stakeholders in selected DH projects

What are typical data integration and reconciliation workflows?

Why are certain standards and modelling approaches adopted/rejected?

Which types of provenance are tracked and how?

What are common requirements, constraints?

Phase 2

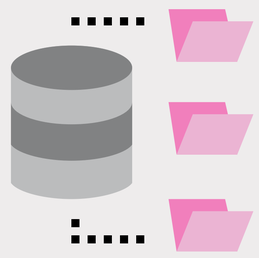
Goals and methodology

The goal is to identify successful (and unsuccessful) approaches to tracking and representing the provenance of historical gender data. This section of my thesis will focus on the subset of DH projects that work with linked open data technologies, particularly in a Wikibase environment. The interview method will draw upon requirements elicitation techniques as well as expert interview methods from the social sciences.

I will then evaluate the interview material and formulate recommendations for DH projects looking to standardize the representation and management of provenance information for historical gender data.

Create repository documenting best practices and sample workflows for tracking and representing the provenance of gender data in the DH domain.

The precise scope and purpose of the repository are not entirely clear yet. The repository should ideally raise awareness to the issue and provide resources for DH projects working with historical gender data in a linked data setting. It could also document successful visualization and user interface design strategies for making gender data provenance transparent in a linked open data context, especially to non-experts.



Why I am here

During the summer school, I hope to broaden my perspective beyond the digital humanities and explore shared challenges in managing provenance data in a linked data setting with participants from other disciplines.

References:
[1] Andrews, Tara et al. (2024), Gender Assignment as an Event - a Contemporary Approach for the Adequate Depiction of Historical Gender Categories, https://academic.oup.com/dsh/article/39/1/5/7577820
[2] Flanders, Julia (2021), Gender in the Machine. Representing Gender in Digital Publication Frameworks, https://www.db- thueringen.de/receive/dbt_mods_00048960.
[3] Samuel, John et al. (2023), Modelling Gender on Wikidata, https://www.wikidata.org/wiki/Wikidata:Events/Data_Modelling_Days_2023.

FREIE UNIVERSITÄT BERLIN