Citation Genealogy Analysis

Elevating Research Quality and Wikimedia Impact through Network Analysis and Critical Source Discovery

Charlotte Oertel
MSc Digital Scholarship Graduate
University of Oxford

Keith Chambers
DPhil Mathematics Candidate
University of Oxford

Abstract

The proposed project introduces a novel method to enhance source criticism and quality control in the humanities. Focused on Citation Genealogy Analysis, the project traces the origins, replication, and dissemination of contestable data in historical publications and digital information systems. This method, outlined during the researcher's participation in the Wikimedia Open Science Fellows Program and recently refined in her MSc dissertation at the University of Oxford, combines bibliometric data, humanities process research, and network analysis methods using Linked Data formats.

The project aims to generate central Wikimedia materials, including an Open Educational Resource on Wikiversity, WikiData dataset uploads, open-source code, and revisions for Wikipedia articles. Key objectives include exploring additional applications of the method in Wikisource, as well as data visualization and analysis for existing Wikimedia data. The project also strives to expand collaborative networks with Wikimedia community contributors and academic leaders.

Introduction

Problem Statement:
The proposed project seeks to address a critical challenge in the humanities — the need for enhanced source criticism and quality control to ensure the validity of current information.
The problem is twofold: first, the current literature lacks robust methodology to validate and trace the origins of repeatedly cited "established" information; and second, without interception, there is a risk of continuous propagation of unreliable data.
Grounded in the innovative approach of Citation Genealogy Analysis, the project aims to create accessible documentation and applications for the methodology. These resources are designed to facilitate critical engagement with research data and Wikimedia sources, ultimately fortifying the reliability of historical and digital reference sources.

Wikimedia Projects Importance:
Addressing this problem is of particular significance to Wikimedia's mission of providing accessible and trustworthy knowledge. The Wikimedia community, along with the wider audience relying on its services, stands to benefit substantially from the enhanced credibility of information. With its broad applications, especially in Wikimedia projects like Wikisource, the potential impact on the overall quality of content is substantial,

reinforcing the project's alignment with Wikimedia's goals.

Beneficiaries and Benefits:

- Community Contributors: Improved tools and methodologies for evaluating sources and contributing reliable information.
- Readers: Engaging visualisation and data exploration tools. Informative case studies illustrating historical origins and reflecting the development of information. Increased trust in the accuracy of information, fostering a positive user experience.
- Researchers in the Humanities: Access to analytical methods like Citation Genealogy Analysis, promoting better scholarly practices. Accessible option to share research data through WikiData as an extension of established research processes in the humanities.

Specific Research Questions:

Research Questions or Hypotheses: The proposed research seeks to address the following questions:

- How effective is Citation Genealogy Analysis in tracing the origins and dissemination of contestable data?
- To what extent can this method enhance source criticism and quality control for historic publications (in Wikisource) and digital information systems (in WikiData and Wikipedia)?
- What are the specific challenges and opportunities for implementing this method in Wikimedia projects?
- How can the generated outputs, such as Wikimedia materials, data sets, and open-source code, contribute to the improvement of Wikimedia content?

By addressing these questions, the research aims to provide applicable insights and tools for the Wikimedia community.

Date: June 1, 2024 – May 30, 2025.

Related work

The researcher's dissertation extensively reviewed literature on data quality, citation network research, and existing tools in the humanities. Existing research highlights challenges in bibliometric tools for the humanities, citing limited coverage in citation indexes and lacking chronological layout options. The dissertation proposes alternatives in network visualization tools, adapting coding environment packages like Pyvis and interactive tools such as Jaal. This project extends the dissertation by sharing materials on Wikimedia platforms and aims to apply analysis and visualization in Wikimedia projects.

Methods

Adapt Citation Genealogy Analysis method for Wikimedia implementation and conduct additional case studies based on Wikimedia data.

Expected output

- Review and publish dissertation for open science journal publication to contribute to the scholarly discourse with application examples and discussion for Wikimedia projects

- Write up and publish the workflow as a Wikiversity Open Educational Resource approachable to researchers

- Rework and publish code in open source format (with technical collaborator Keith Chambers, DPhil in Mathematics candidate, University of Oxford)

- Work with Wikimedia community to integrate Citation Genealogy Analysis method with WikiData (potential extension of Cita for Zotero) enhancing research data transfer and visualisation and analysis tools for research community and supporting active contributions to WikiData and Wikimedia projects.

- Incorporate tool options for visualization to strengthen the accessibility and user-friendliness of the method.

Risks

Issues with practical realization, including technical implementation and community support.

Community impact plan

The research aims to engage broader audiences beyond academia by developing educational resources, including a Wikiversity Open Educational Resource, WikiData data set uploads, and open-source code. Emphasizing collaboration with Wikimedia community contributors, academic leaders, and leveraging connections to academic communities in the US, Germany and the UK, the project envisions correcting inaccurate Wikipedia articles, and exploring data visualization applications for existing Wikimedia data. This comprehensive approach is designed to ensure broader impact, enrich the project with diverse perspectives, and facilitate the adoption and utilization of research findings within Wikimedia and academic communities.

Evaluation

Publication of proposed materials, additional application to Wikimedia projects and community.

Budget

Salary or Stipend:

- Lead researcher: £15.00 (950h) = £14250
- Technical collaborator: £15,00 (60h) = £900

Open Access Publishing: £2623 (based on Digital Scholarship in the Humanities Journal)

Conference Participation: £1500

Hospitality expenses: £1000 (meetings with academic community advisors and collaborators)

Overall budget: £20273 = $25724 (as of December 2023)

Prior contributions

· Initial framework developed during Wikimedia Germany Open Science Fellows Program (2020-2021)

· Dissertation for MSc in Digital Scholarship, University of Oxford 2022-2023

· Conference talks at Renaissance Society of America Dublin 2022 and Art Libraries Conference Munich 2022.

· Project talks in digital meetings with US Art Libraries community (National Gallery Library, Washington D.C.) and academic communities in Germany (NFDI4Culture) and the UK (Bodleian Libraries), established network with central information infrastructure

· Established working partnerships within the Wikimedia community, notably collaborating with Diego de la Hera, Cita for Zotero developer

References

Previous contributions:
C. Oertel (2021) WikiData WikiProject Citation Genealogy:
https://www.wikidata.org/wiki/Wikidata:WikiProject_Citation_Genealogy

C. Oertel (2021) Wikiversity Fellow-Programm:
https://de.wikiversity.org/wiki/Wikiversity:Fellow-Programm_Freies_Wissen/Einreichungen/Acceleration_of_quality_in_the_humanities_%E2%80%93_chances_of_open_source_implementation_in_research_and_training

C. Oertel (2023) *Citation Genealogy Analysis. A Network Analysis Approach for Research Quality Assessment in the Humanities*. Dissertation, MSc in Digital Scholarship, University of Oxford, United Kingdom.

Figures

Fig 1. Example of Citation Genealogy Analysis visualization in chronological order, illustrating citation links over time. Based on the Case Study 'The Mass of St. Gregory'.

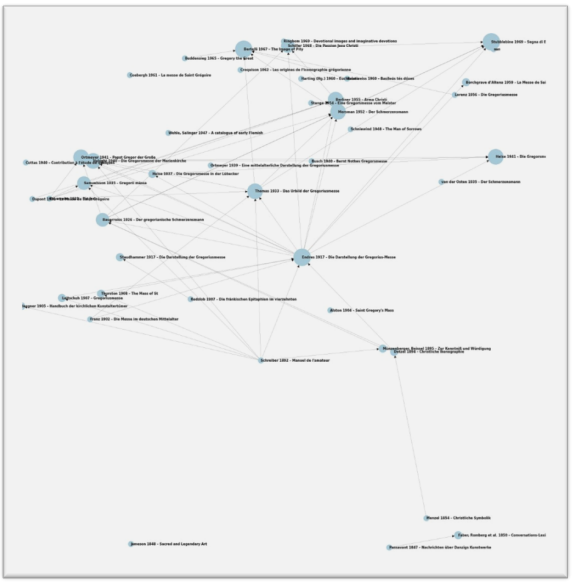

Fig 2. Example of Citation Genealogy Analysis wit Pyvis without chronological order.. Based on the Case Study 'The Mass of St. Gregory'.