Community-Driven AI Support for Genealogy Research

FEI SHAN, Virginia Polytechnic Institute and State University, USA
KURT LUTHER, Virginia Polytechnic Institute and State University, USA

Genealogists benefit from the availability of digitized historical documents but struggle with the spread of misinformation on genealogy websites. Two causes of misinformation include the large number of users lacking knowledge about effective genealogy research processes, and algorithmic suggestions sharing unreliable research products with these inexperienced users. We conducted an semi-structured interview study with expert and novice genealogists investigating the challenges they face when conducting genealogy research online, including dealing with misinformation. We propose design recommendations for how the genealogist community can leverage human expertise to improve the overall quality of genealogical research supported by artificial intelligence.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

Additional Key Words and Phrases: genealogy, family history, misinformation, sensemaking, community of practice

ACM Reference Format:
Fei Shan and Kurt Luther. 2023. Community-Driven AI Support for Genealogy Research. 1, 1 (May 2023), 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Genealogists who study family history and lineage frequently search through numerous historical documents to find information about ancestors. Traditionally, the search of documents required genealogists to physically visit archives, churches, courthouses, etc. to acquire records. This laborious workflow has changed significantly with the emergence of online genealogy websites like Ancestry.com and FamilySearch.org, which help genealogists to access millions of digitized and transcribed copies of original documents through the Internet. Furthermore, these sites accelerate the genealogical research process with powerful search engines, algorithmic recommendations of relevant information, and collaboration features enabling users to view and connect to others' public family trees. These websites have attracted millions of members [1, 9], and the global genealogy products and services market value is estimated at $5.4 billion in 2023 [4].

While access to genealogical information sources has greatly improved, the propagation of misinformation on these websites poses a major challenge to the genealogist community. One issue is that as these sites become more accessible, they attract novices who lack research training and may acquire and rapidly distribute unvetted information through collaboration tools, such as public, user-generated family trees. As early as the 1990s, the president of American Society of Genealogists asserted that "web-grown genealogist are largely unschooled in research principles," but "empowered" to broadcast their genealogy research product on the Internet, regardless of the research quality [7].

Authors' addresses: Fei Shan, fshan@vt.edu, Virginia Polytechnic Institute and State University, Virginia, USA, 22203; Kurt Luther, kluther@vt.edu, Virginia Polytechnic Institute and State University, Virginia, USA, 22203.

XXXX-XXXX/2023/5-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

A second issue arose when artificial intelligence (AI)–based *research hints* features was introduced by major genealogy websites. Research hints are designed to identify relevant information about an ancestor from the site's databases and recommend it to users for their research. These suggestions may provide irrelevant information (e.g., non-ancestors with similar names) or inaccurate information (e.g., user-generated family trees) that novices tend to accept unquestioningly.

Prior CSCW research by Willever-Farr and Forte [10] reported genealogists' concerns and frustration with misinformation resulting from careless genealogy research and unverified algorithmic suggestions. Their interviewees referred to the genealogy website users who know little about proper research process as "clickologists" and the unreliable research hints as "poison leaves," concerned that such phenomena could damage the overall quality of genealogical research and hinder collaborations within the community [10].

2 INITIAL FINDINGS

We sought to investigate the current state of the misinformation problem in the genealogist community and obtain better understandings of genealogical research online. We conducted a semi-structured interview study with 20 genealogists: 10 experts who are professionals working for individual and institutional clients, developing educational content, and taking leading positions in local and national genealogical organizations; and 10 amateur genealogists who research their own family, did not receive formal training on genealogy research, and have less research experience.

We found that nearly a decade after Willever-Farr and Forte's research [10], misinformation still persists as a major challenge in genealogy. There is a large number of active genealogy hobbyists who are less competent in genealogical research. The AI-generated research hints, despite being helpful to some users and scenarios, may provide misleading information to them. One of our participants described the situation: "Here's a big problem in genealogy community, it's . . . more so in the beginner's stage, where I'll just copy information over and over."

We compared genealogists' research practices across different level of expertise. Expert genealogists appeared to be more knowledgeable about historical documents and context; their evaluation of information is more thorough and critical; and they follow explicit research standards — such as the Genealogical Proof Standard (GPS) [8] — which are unfamiliar to amateurs. Moreover, our participants asserted that the genealogy websites are not well equipped with guidance on how to conduct genealogical research for newcomers. As one of our participants suggested, "I wish Ancestry did a better job of explaining kind of basic methods as you're starting a tree . . . cause that right now, they kinda just have people rely on their hints that could be completely way wrong." As a result, amateur genealogists feel that they need to actively search for educational content outside these websites that is appropriate to their knowledge level.

3 NEXT STEPS

We argue that AI algorithms used by popular genealogy websites, as currently designed, exacerbate genealogy misinformation by encouraging the spread of unverified information rather than leveraging community expertise and standards. This is especially a problem for platforms with large user bases of novices who are not trained to research rigorously and independently. Therefore, inspired by successful efforts in other research-oriented online environments such as Wikipedia [11], we propose a community-driven approach that incorporates human expertise to reduce misinformation and improve the quality of online genealogy research.

First, AI-based algorithms deployed on genealogy websites should better address the diverse needs of the community. In particular, the algorithm could adapt to users of various experience levels (e.g., it may recommend more reliable information, such as primary source materials, to less experienced users). Furthermore, we could introduce human expertise to the algorithmic suggestions (e.g., being vetted by experienced genealogists before

being presented to novices), and the AI algorithm may generalize and learn from expert genealogists' feedback to improve its verification capabilities [6].

Second, user interface of genealogy systems could be designed to scaffold [3] novices in learning genealogy research methods by incorporating genealogy standards such as the aforementioned GPS [8], providing historical context, and proposing evaluation metrics, so novices can approach AI suggestions more critically before accepting them. Third, genealogy websites could improve opportunities for social learning [5]. On one hand, genealogy websites can highlight exemplary research *products* — e.g., family trees constructed and reviewed by expert genealogists — to novice users, similar to featured articles on Wikipedia [2]. On the other hand, the sites could make the research *process* of experienced genealogists more observable to novices, while providing smaller tasks for them to build up their skills.

Genealogy reveals important challenges and opportunities in online communities in terms of sensemaking, collaboration, and misinformation. By researching the practice of genealogists, we hope to generate new insights about how community-driven AI could be developed and deployed to support high-quality research production in this and other domains.

REFERENCES

[1] Ancestry. 2023. Company Facts: Ancestry Corporate. Retrieved September 21, 2023 from https://www.ancestry.com/corporate/about-ancestry/company-facts

[2] Susan L Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work*. 1–10.

[3] National Research Council et al. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. Vol. 1. National Academies Press.

[4] FACT.MR. 2023. Genealogy Products and Services Market to Hit US$ 15.8 Billion by 2033. Retrieved September 21, 2023 from https://finance.yahoo.com/news/genealogy-products-services-market-hit-140000715.html

[5] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

[6] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A review on interactive reinforcement learning from human social feedback. *IEEE Access* 8 (2020), 120757–120765.

[7] Elizabeth S Mills. 1999. Working with historical evidence: Genealogical principles and standards. *National Genealogical Society Quarterly* 87, 3 (1999), 165–84.

[8] Christine Rose. 2014. *Genealogical Proof Standard: Building a Solid Case* (4th ed.). CR Publications.

[9] FamilySearch Wiki. 2022. FamilySearch.org. Retrieved September 21, 2023 from https://www.familysearch.org/en/wiki/FamilySearch.org

[10] Heather L Willever-Farr and Andrea Forte. 2014. Family matters: Control and conflict in online family history production. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 475–486.

[11] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–23.

Received 22 September 2023