Passing On: Reader-Sourcing Gender Diversity in Wikipedia

J. Nathan Matias
MIT Media Lab
MIT Center for Civic Media
75 Amherst St
Cambridge MA, USA
jnmatias@mit.edu

Sophie Diehl
MIT
sophiadiehl@gmail.com

Ethan Zuckerman
MIT Media Lab
MIT Center for Civic Media
75 Amherst St
Cambridge MA, USA
ethanz@gmail.com

Abstract
This paper presents work in progress on "reader-sourcing," an approach to cooperative content creation through the reading behavior of crowds. We present the *Passing On* system, that reader-sources the creation and expansion of Wikipedia articles about women, aiming to support frame changes on women's representation and offer a counter-public for novice Wikipedians. In *Passing On*, browsing and searching creates Wikipedia content when reading about notable women not yet in Wikipedia. This paper presents the design goals, working system, and evaluation plan.

Author Keywords
Wikipedia; Feminist HCI; Crowdsourcing; Civic Media

ACM Classification Keywords
H.5.3. Group and Organizational Interfaces: Computer Supported Cooperative Work

Introduction
Gender disparities in Wikipedia have been widely criticized for their role in reinforcing poorly founded assumptions and expectations for women in many societies. Wikipedia's coverage of women is constrained by broader disparities in society, including the systematic under-representation of women in news media, but Wikipedia may not even match the low

levels of representation found elsewhere [12]. Although Wikipedia includes more biographies of women than any other encyclopedia, gender is a significant predictor of a person's omission from Wikipedia, compared to Britannica [17]. Wikipedians have also organized to grow the small percentage of women editors [10]. Yet women and other newcomers on Wikipedia often struggle to reach full participation [11] [20].

Interventions for change such as day-long edit-a-thons and researcher-in-residence programs address both goals, broadening inclusion and expanding content [19]. Yet these efforts, which engage a few dozen people at a time, struggle to match the scale of available sources. In 2014, the UK Royal Society's Women in Science Edit-a-Thon organized 23 people, helped by 7 experienced Wikipedians, to create new stub articles for 13 women scientists and improve 11 other articles [23]. In the same year, the New York Times published 1084 obituaries [15] and the Guardian published 1613 obituaries [7]. Had the Royal Society edit-a-thon focused only on notable people who died that year, they could only have used roughly 1.5% of available sources to influence 0.3% of Wikipedia's coverage of notable people who died in 2014 [3].

In this paper, we present work in progress on a technique called "reader-sourcing" for creating Wikipedia content and recruiting new editors at scale. To illustrate this technique, we present a working system, *Passing On*, that expands Wikipedia's sources and content from users' reading behavior. Finally, we outline research to evaluate the design.

Reader-Sourcing

The idea of computationally creating content from reading behavior is explored by Bush, where a "trail of interest" left by a single scholar researching a longbow is created and shared [2]. Nelson later argues that computer writing systems should track "early periods of muddled confusion" in a writer's process, including sources read [14]. Systems that support reading as part of writing such as Zotero [4] tend to follow the Bush-Nelson model of systems for experts. This scholar-centered approach is also taken by the RAMP editor, which supports archival staff to merge archival records into Wikipedia as they browse archives [22].

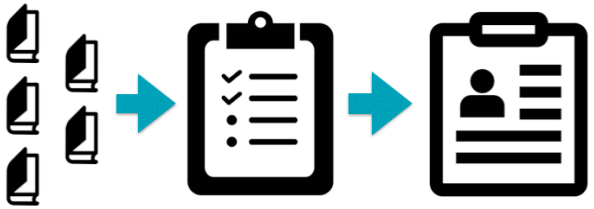

Figure 1: in reader-sourcing, reading behavior and survey answers from many readers are archived and computationally generated into content

Reader-sourcing coordinates the non-expert reading behavior of crowds to create content at scale. Like crowd writing designs in human computation [1][16], reader-sourcing automatically directs and combines contributions of many users to create content. Unlike crowd writing systems, where contributors participate by typing text, reader-sourcing does not require contributors to type: part of the written output is created from the shared record of reading behavior. The generation of content from reader activity makes

reader-sourcing unique from other forms of micro participation like the Wikipedia article feedback tool [9].

Design Goals

Creation of high quality Wikipedia source content through reader-sourcing is the first design goal of the *Passing On* system. As a civic technology focused on social justice [13], *Passing On* has two other equally important design goals.

As a civic technology in the tradition of Feminist HCI, *Passing On* sets out to support a *frame change* among its users [5], inviting them to rethink their understanding of women's representation: recognizing it as (a) a systemic problem in media and (b) an issue open to collective action on Wikipedia.

Finally, *Passing On* also offers an alternative counter-public [21] outside Wikipedia to coordinate these changes. As with the Snuggle system, a tool for socialization on Wikipedia, and the Wikipedia article feedback tool, efficiencies in content productivity are sometimes set aside to foster learning and collective action among contributors [8][6][9]. With *Passing On*, this participation takes place in a space outside Wikipedia where novice contributions are suported and welcome. Learning and collective action goals motivate our use of reader-sourcing rather than automated creation of stub articles from our datasets.

Passing On

Passing On is a reader-sourcing system that organizes casual readers to observe gender disparities in New York Times obituary coverage and to browse notable women's life stories by topic from 20 years of the New York Times. As users read selected sentences from the life stories of notable women and search to learn more about their lives, the sources they find are added to a shared record of sources to support new or improved Wikipedia articles. In the process, users are educated to write new articles from those sources.

System Design: The *Passing On* system is a collaborative web application developed in April 2013 that takes input from news data sources and archives reader behavior to a database for inclusion into Wikipedia articles. Data on notable women is acquired from a dataset of 35,801 obituaries in the New York Times from 1987 and 2007 [18]. 5,674 (15.8%) of these were identified as female using pronoun count methods similar to Reagle and Rhue's quantitative analysis of women's biographies on Wikipedia [17]. *Passing On* also interacts with the archives of the LA Times, Boston Globe, Harper's Magazine, Time Magazine, The Atlantic, Britannica, and Google Books.

User Experience: After an introductory slideshow [*fig 2*], users are shown a chart of obituary gender in the New York Times for 49 topics [*fig 3*]. When users select a topic, *Passing On* shows a report of all obituaries returned by the query, color-coded by that person's presence in Wikipedia (as determined by other users) and the state of information collected by other users [*fig 4*]. When users select a person based on curiosity or impact on Wikipedia, they are shown information about the person, including automatically-selected quotations from the New York Times obituary [*fig 5*]. The quotations are limited to prompt further curiosity.

After selecting a person, all further user actions contribute toward Wikipedia content. When viewing the person's biography on Wikipedia, users are asked to

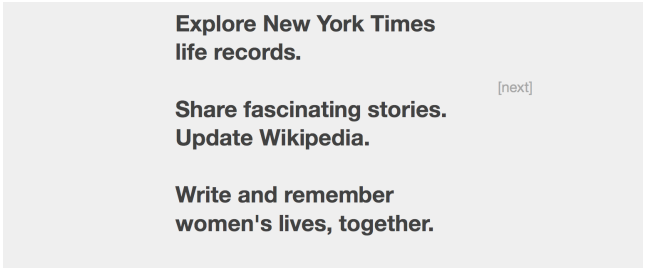

Figure 2: Passing On introductory slideshow

Figure 3: Visualization of topic gender in obituaries

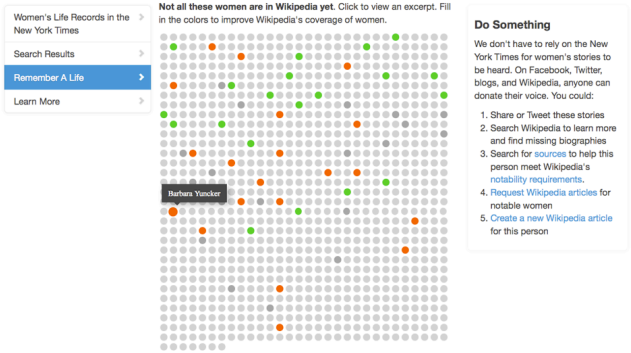

Figure 4: Viewing crowd status and selecting a person

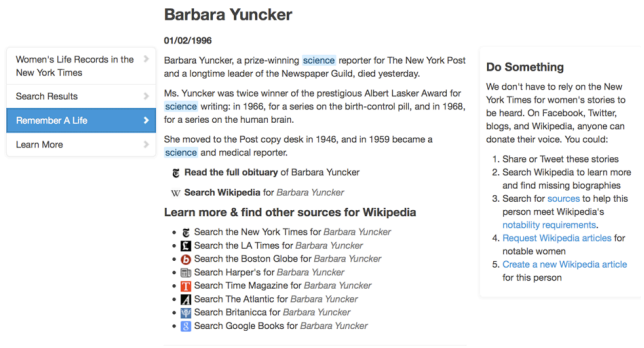

Figure 5: Obituary entry and research task coordinator

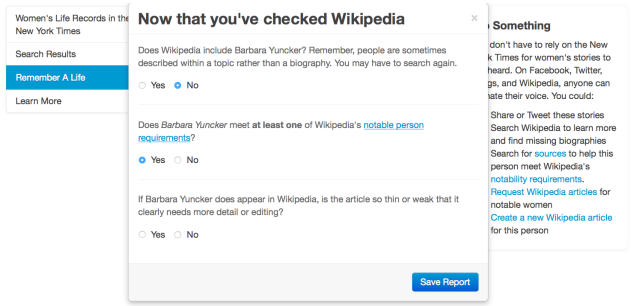

Figure 6: Recording research results

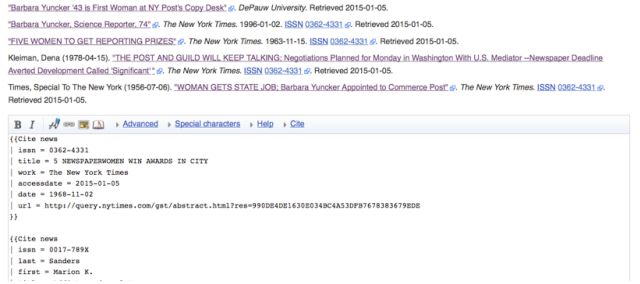

Figure 7: Structured Wiki markup output to Wikipedia

confirm the existence or quality of the Wikipedia entry. When users click to search for the person in the archives of other publications, their success or failure is archived through a click-based survey [*fig 6*]. The status of this combined record is visually displayed back to other users as they make reading choices [*fig 4*]. Finally, *Passing On* converts the resulting lists of sources into wiki markup for a stub article about the person of their interest [*fig 7*].

Although these actions can form a progression of learning for users unfamiliar with Wikipedia, users are not required to carry out more complex actions; the act of reading itself is a substantive contribution.

Future Work and Evaluation

Future field testing of *Passing On* can evaluate the system in respect to its design goals:

Reader-sourcing will be evaluated in comparison with Wikipedia articles sourced by other means. In field tests, the quantity and quality of resulting contributions, as well as the quantity of sources in resulting contributions could be compared to typical biographies.

Frame Transformation among users will be evaluated with qualitative methods similar to Dimond's *Hollaback!* system, which had similar goals [5].

Finally, the effectiveness of *Passing On* as Wikipedia *counter-public* can be evaluated qualitatively through content analysis and interviews, and quantitatively in terms of the survival rate of the new Wikipedians it recruits, relative to other methods.

Future designs could optimize the reader-sourcing process by using passive data collection. Other designs could test reader-sourcing as a method to create content beyond source lists, inviting readers to highlight the most interesting sentences or facts from biographies for quotation and inclusion in Wikipedia.

Acknowledgements

Thanks to Aaron Halfaker of Wikimedia Research and David Riordan of the New York Public Library Labs for their enthusiastic reception of *Passing On* and effort to convince the authors to revisit and publish this work. Thanks also to Tom Steinberg at mySociety and Kate Crawford at Microsoft Research for offering design feedback on the system. CC-by 3.0 licensed clipboard images are by Jardson Almeida and Ilsur Aptukov.

References

[1] Bernstein, M.S., Little, G., Miller, R.C., et al. Soylent: A Word Processor with a Crowd Inside. *UIST '10*, 313–322.

[2] Bush, V. As We May Think. *The Atlantic*, 1945.

[3] "Category:2014_deaths" Wikipedia WikiTools API, accessed Jan 3, 2015. Count: 6,226

[4] Cohen, D.J. Creating Scholarly Tools and Resources for the Digital Ecosystem: Building Connections in the Zotero Project. *First Monday 13*, 8 (2008).

[5] Dimond, J.P., Dye, M., Larose, D., and Bruckman, A.S. Hollaback!: the role of storytelling online in a social movement organization. *CSCW '13*, 477–490.

[6] Geiger, S. Successor systems: the reflexive roles of algorithms in enacting ideological critique. Lecture at Berkman Center for Internet & Society Nov 4, 2014.

[7] Guardian Content API, accessed Jan 3 2015.

[8] Halfaker, A., Geiger, R.S., and Terveen, L.G. Snuggle: Designing for Efficient Socialization and Ideological Critique. *CHI '14*, 311–320.

[9] Halfaker, Aaron, Oliver Keyes, and Dario Taraborelli. 2013. "Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia." CSCW '13.

[10] Hill, B.M. and Shaw, A. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE 8*, 6 (2013)

[11] Lam, S. (Tony) K., Uduwage, A., Dong, Z., et al. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. *WikiSym '11*, 1–10.

[12] Matias, J.N. and Evans, L. Women's representation in media: readers preferences for online news revealed. *the Guardian*, 2012.

[13] Matias, J.N. and Geiger, S. Defining, Designing, and Evaluating Civic Values in Human Computationand Collective Action Systems. *HCOMP '14*, *Citizen-X*

[14] Nelson, T.H. Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference* (ACM '65)

[15] New York Times API, accessed Jan 3, 2015.

[16] Quinn, A.J. and Bederson, B.B. Human Computation: A Survey and Taxonomy of a Growing Field. *CHI '11* 1403–1412.

[17] Reagle, J. and Rhue, L. Gender Bias in Wikipedia and Britannica. *IJOC 5*, 0 (2011), 21.

[18] Sandhaus, Evan. The New York Times Annotated Corpus LDC2008T19. Web Download. Philadelphia: Linguistic Data Consortium, 2008.

[19] Shen, A. How Many Women Does It Take to Change Wikipedia? *Smithsonian*, 2012.

[20] Shaw, A. and Hill, B.M. Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *Journal of Communication 64*, 2 (2014), 215–238.

[21] Squires, C.R. Rethinking the Black Public Sphere: An Alternative Vocabulary for Multiple Public Spheres. *Communication Theory 12*, 4 (2002), 446–468.

[22] Thompson, T.A., Little, J., González, D., Darby, A., and Carruthers, M. From Finding Aids to Wiki Pages: Remixing Archival Metadata with RAMP. *The Code4Lib Journal*, 22 (2013).

[23] Wikipedia:WikiProject Royal Society/Women in Science Wikipedia Edit-a-thon at the Royal Society, March 2014. *Wikipedia, the free encyclopedia*, 2014.