Understanding civic and non-profit data through a custom data lifecycle

ANNABEL ROTHSCHILD, Georgia Institute of Technology
CARL DISALVO, Georgia Institute of Technology
AMANDA WOOTEN, Georgia Institute of Technology
BETSY DISALVO, Georgia Institute of Technology

This report details our experience creating a graphic to help track how data flows through our organization, DataWorks. DataWorks specializes in data cleaning and standardization services for civic and non-profits, while simultaneously functioning as a work-training program through which the data wranglers receive both training and a competitive hourly wage. As a result, the way data moves through DataWorks looks different than more traditional data clearinghouses, as those organizations often focus on all steps of the *traditional* data lifecycle. Through recounting our – data wranglers and researchers, with assistance from a design student – efforts to create the data lifecycle graphic, we describe the organization-specific properties of this data flow and theorize how it might apply to other organizations that assisting organizational initial "datafication" and maintenance.

ACM Reference Format:
Annabel Rothschild, Carl DiSalvo, Amanda Wooten, and Betsy DiSalvo. 2022. Understanding civic and non-profit data through a custom data lifecycle. 1, 1 (September 2022), 5 pages.

1 INTRODUCTION

While every organization likely has a domain-specific data flow, we detail the data lifecycle at DataWorks, an early stage data cleaning and standardization (or "preparation" [3]) organization that helps other organizations with their "datafication". Our contribution is sharing both the atypical features of civic (and non-profit data) along with the process we followed to understand how data moves through our organization. In the first section of this case study, we describe the intricacies and hallmark features of civic and non-profit data as they often emerge from low resource technical environments. In the second portion, we present the generation of an organization-specific data flow tracker at DataWorks, highlighting both the challenges of data cleaning and standardization in an inter-domain (or domain-agnostic) space.

2 CIVIC (AND NON-PROFIT) DATA AND ITS DISCONTENTS

There is a longstanding research interest in the use of data, and more broadly, information and communication technologies, in government, civil society, and community contexts [2, 4, 9]. This interest is partly because these contexts are meaningfully different from industry. One commonly noted difference is the relative lack of resources for handling data in such contexts. There are often few people with data skills, less access to technology, and less time and

Authors' addresses: Annabel Rothschild, arothschild@gatech.edu, Georgia Institute of Technology; Carl DiSalvo, cdisalvo@gatech.edu, Georgia Institute of Technology; Amanda Wooten, awooten7@gatech.edu, Georgia Institute of Technology; Betsy DiSalvo, bdisalvo@cc.gatech.edu, Georgia Institute of Technology.

Manuscript submitted to ACM

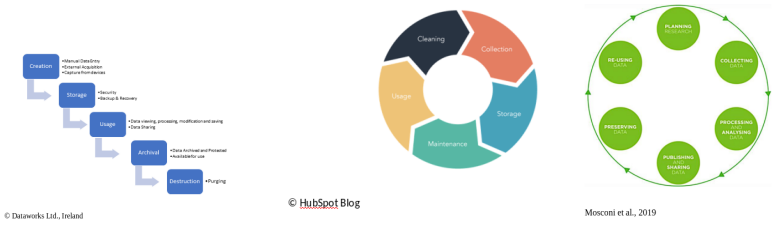

Fig. 1. The three example data lifecycles from other organizations and perspectives. From left: available at https://www.dataworks.ie/5-stages-in-the-data-management-lifecycle-process/, https://blog.hubspot.com/website/data-lifecycle-management, [7].

money to put towards data work. For example, in many city planning departments, expertise in working with data is limited to prior experience with geographic information systems. Even then, that experience is limited to working with data through proprietary software that obfuscates data structures and algorithmic processes. In many grassroots organizations, access to data tools is limited to those that are free or low-cost, which often, in turn, determines a set of limited features.

Furthermore, access to data itself may be limited to what is available in public data sets. In both of these example contexts, data processing and analysis are not core to the organization's work. Instead, data processing and analysis support mission-driven objectives and the values that motivate those missions. In our experience collaborating with partners in government, civil society, and community contexts, time and again, these conditions led to a fractured use of data, with data sets that were often incomplete and riddled with errors. This is **not** a shortcoming of these organizations; this is simply what often happens when organizations that prioritize social relations enter into environments in which decisions are increasingly data determined. It is within these environments that we developed DataWorks. One purpose of DataWorks was to provide data services to government, civil society, and community organizations that might not otherwise have the resources to engage in data work. For example, in 2020, we partnered with the Center for Civic Innovation[1], a non-profit organization in Atlanta that works with residents and other non-profits to advocate for greater transparency in municipal decision-making, as one of their many programs. One project required analyzing whether – and if so, how? – the decisions of the Zoning Review Board took into account community recommendations. To ask and answer this question required reviewing the past decade of agendas and votes from zoning meetings. The agendas and votes were archived as scanned PDF files, making analysis difficult given the difficulty to lift text from them. The Center for Civic Innovation hired DataWorks to transfer the data from every meeting, recommendation, and vote into spreadsheets.

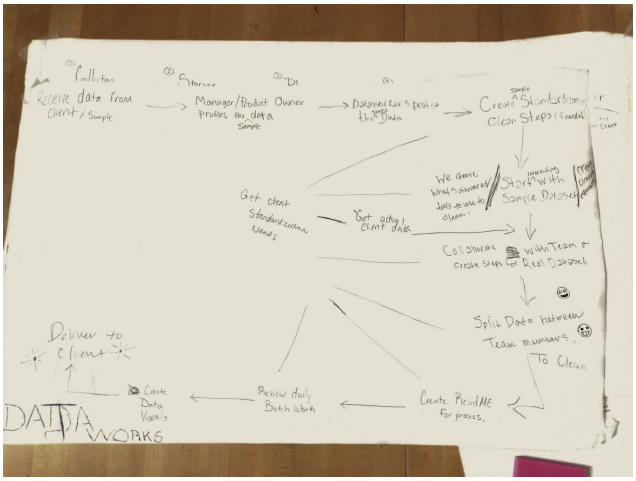

Fig. 2. First iteration of the data lifecycle at DataWorks.

[1]https://www.civicatlanta.org/

This process involved reviewing more than 1,000 pages of PDFs, writing scripts when possible, using manual text entry when needed, and of course sorting, checking, and organizing the data over several months. The process was tedious and exacting, but also necessary the Center for Civic Innovation to have the data is needed; the result was outside of what the Center for Civic Innovation would have been able to do itself.

3 A CUSTOM DATA LIFECYCLE

As a reflection on projects like that for the Center for Civic Innovation, the first author developed a tutorial series that took place in the summer of 2021. In a session facilitated by that author, the data wranglers designed a custom data lifecycle to represent the general flow of data within DataWorks. The custom lifecycle grew out of that tutorial section – on trying to identify organizational overlap with examples of a few common model – an initial engagement of about an hour, with an additional session of roughly an hour for each of the following iterations.

Because DataWorks deals primarily with early stage data processing (cleaning and standardization), these existing models of the data lifecycle (see 1) did not adequately represent our process. While of the projects the data wranglers have worked on include early stage (collection) and later (preliminary analysis) phase work, the most common type of project occurs mostly in an Excel spreadsheet[2] and is comprised of cleaning and standardization. The lifecycle graphic is meant to be used both as a common reference point for tracking project progress, but also to help keep track of important "data moves" [5] and transformations made on current project datasets. With the help of Amanda Wooten, a design student, the data wranglers went through three significant iterations of designs to create a graphic that meaningfully captured the nuance of data flow at DataWorks.

Four data wranglers developed the first version in response to the compare-and-contrast activity, resulting in (2). This design displays the more graph-like (multi-directional flow) shape of the data flow at DataWorks, compared to the cyclical and linear models that are more commonly seen (see 1). Before designing this first version, the wranglers noted both the collaborative nature of data flow at DataWorks – including regular group check-ins – and the prominent role that client relations play. Notably, there are no references in this version to data analysis or visualization which are most often beyond the scope of DataWorks; rather, cleaning, standardizing, and client relations fill that space.

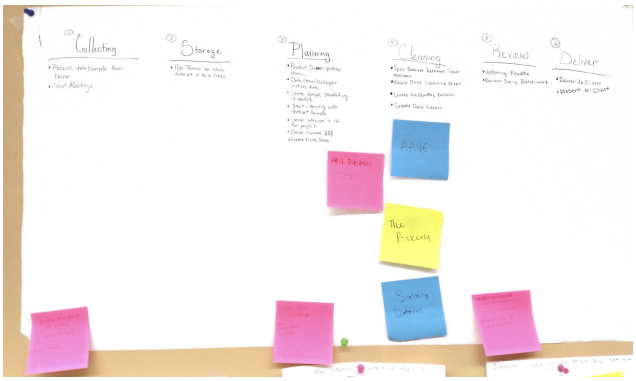

Fig. 3. Second iteration of the data lifecycle at DataWorks.

The second version (3) came shortly thereafter, as a way for the data wranglers to keep track of the stages they identified in the initial version while working on live projects. At this point, the wranglers decided on a partially-linear model (one with six numbered steps) with "planning" and cleaning as the most time-consuming, or task-filled, stages. Planning refers to team strategy meetings and trial runs of proposed data transformations to determine a course of action for the team members to then follow in parallel on different sections of the dataset.

[2]For more details on past projects, see https://dataworkforce.gatech.edu/recent-projects/

With the second version in hand, the wrangler team was joined by Amanda Wooten who assisted them in creating a more formalized version of the lifecycle that could be hung on the wall of the DataWorks office, as a laminated poster than can be annotated with dry-erase markers and sticky notes. The team first met with Amanda for a 1-hour design session and then exchanged feedback over email resulting in the third significant version of the data lifecycle chart (see 4). There are several meaningful changes from the second version, along with the wranglers' decision to remove instructional steps off the lifecycle and on to a secondary checklist that hangs alongside the data lifecycle, with the resulting space free for noting project-specific additional directions or considerations. First, planning and cleaning are given the bulk of space, which demonstrates both DataWorks' unique focus on data cleaning and standardization (colloquially both are grouped into the term "cleaning" in DataWorks office parlance) and an additional step (or deviation) from the more traditional data lifecycle depiction. Second, as a mix of both the linear and cyclical models, while stages generally proceed to left-to-right, the blank space allows for arrows to be drawn back to earlier stages and for projects to clearly hang *between* stages, to denote partial completion or additional steps of ambiguous stage membership.

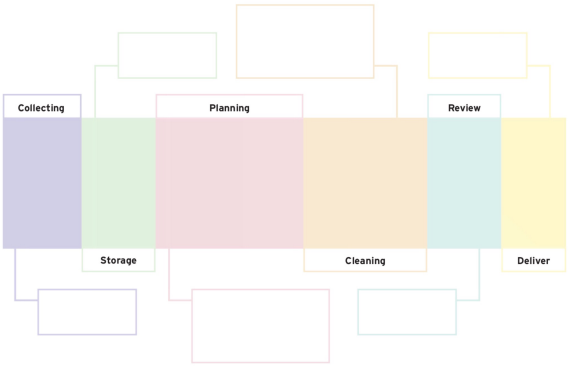

Fig. 4. Third and final version of the data lifecycle chart.

4 DISCUSSION

Why would we bother with a customized data lifecycle? Most data lifecycles, like most business processes in general, reflect a generic approach to work (e.g., [7]). In addition, it's common for such processes to be "handed-down" to workers as prescriptive devices to shape workflow. DataWorks is an organization committed to workplace democracy, so we wanted to design a customized data lifecyle through a collaborative process. We hope that such an approach better supports novice data workers in their growth towards more expertise while also bolstering the agency of the data workers to shape their work environment [1, 6].

Why is the DataWorks perspective unique? The work of DataWorks happens in a mission-driven organization within a university, done by novices of informal data science, and the work environment is being developed collaboratively through participatory methods as the organization grows. In addition to scaffolding broader participation in data work, we also strive to create a culture of data work that centers the workers. One aspect of culture is collaborating with workers to co-create tools and processes that are useful to them, reflective of their work practices and skills; such approaches continue a long-standing participatory design tradition while also refiguring those methods to contemporary data work [6, 8].

What can we add to – and learn from – at the workshop? We hope to participate in this workshop to share our evolving work on creating and using a data lifecycle in DataWork. We believe this is a meaningful contribution because of the distinctive context and processes of DataWorks, as we describe above. At the same time, we hope to learn much from participation in this workshop, such as understanding how data work happens in a range of other environments and how other scholars are engaging ethical and labor issues through their research of data work. Further, at DataWorks we concentrate primarily on early and middle stage data flow processes and are eager to learn from scholars whose work touches on later stages.

REFERENCES

[1] Dinislam Abdulgalimov, Reuben Kirkham, James Nicholson, Vasilis Vlachokyriakos, Pam Briggs, and Patrick Olivier. 2020. *Designing for Employee Voice.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376284

[2] Chris Bopp and Amy Voida. 2020. Voices of the Social Sector: A Systematic Review of Stakeholder Voice in HCI Research with Nonprofit Organizations. *ACM Trans. Comput.-Hum. Interact.* 27, 2, Article 9 (mar 2020), 26 pages. https://doi.org/10.1145/3368368

[3] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2021. Passing the Data Baton: A Retrospective Analysis on Data Science Work and Workers. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb 2021), 1860–1870. https://doi.org/10.1109/TVCG.2020.3030340

[4] Sucheta Ghoshal and Amy Bruckman. 2019. The Role of Social Computing Technologies in Grassroots Movement Building. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 18 (jun 2019), 36 pages. https://doi.org/10.1145/3318140

[5] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11.* ACM Press, 3363. https://doi.org/10.1145/1978942.1979444

[6] Finn Kensing and Joan Greenbaum. 2012. Heritage: Having a say. In *Routledge international handbook of participatory design.* Routledge, 41–56.

[7] Gaia Mosconi, Qinyu Li, Dave Randall, Helena Karasti, Peter Tolmie, Jana Barutzky, Matthias Korn, and Volkmar Pipek. 2019. Three Gaps in Opening Science. *Computer Supported Cooperative Work (CSCW)* 28, 3–4 (Jun 2019), 749–789. https://doi.org/10.1007/s10606-019-09354-z

[8] Randy Trigg and Karen Ishimaru. 2012. Integrating participatory design into everyday work at the Global Fund for Women. *Routledge International Handbook of Participatory Design* (2012), 233–254.

[9] Amy Voida. 2011. Shapeshifters in the voluntary sector: exploring the human-centered-computing challenges of nonprofit organizations. *Interactions* 18, 6 (Nov 2011), 27–31. https://doi.org/10.1145/2029976.2029985