TU Delft

Delft University of Technology

Personal Data Comics

A Data Storytelling Approach Supporting Personal Data Literacy

Gomez Ortega, Alejandra; Bourgeois, Jacky; Kortuem, Gerd

DOI
10.1145/3630970.3630982

Publication date
2024

Document Version
Final published version

Published in
CLIHC 2023 - 11th Latin American Conference on Human Computer Interaction

Citation (APA)
Gomez Ortega, A., Bourgeois, J., & Kortuem, G. (2024). Personal Data Comics: A Data Storytelling Approach Supporting Personal Data Literacy. In *CLIHC 2023 - 11th Latin American Conference on Human Computer Interaction* Article 2 (ACM International Conference Proceeding Series). ACM. https://doi.org/10.1145/3630970.3630982

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Personal Data Comics: A Data Storytelling Approach Supporting Personal Data Literacy

Alejandra Gómez Ortega
Delft University of Technology
Delft, The Netherlands
a.gomezortega@tudelft.nl

Jacky Bourgeois
Delft University of Technology
Delft, The Netherlands
j.bourgeois@tudelft.nl

Gerd Kortuem
Delft University of Technology
Delft, The Netherlands
g.w.kortuem@tudelft.nl

ABSTRACT

Most people interact with digital technologies that collect personal data about their behavior and experiences, leaving behind a data trail. The data within this trail is abstract and difficult to interpret; still, people often need to decide about its collection and distribution. Hence, it is paramount to support personal data literacy, for which data visualization approaches have been successful. These approaches focus mostly on data from single sources (e.g., IoT devices at home) or types (e.g., menstrual logs) and fail to capture people's situated knowledge. We hypothesize that creating data comics can address these limitations and support people in developing personal data literacy. In this paper, we explore how non-data experts create personal data comics, starting from simple data visualizations, and investigate their effectiveness and engagement in the context of pregnancy. Doing so, we identify comic elements that facilitate the autonomous exploration of personal data and provide design recommendations to support independent data comic creation.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Personal Data; Data Literacy; Data Visualization; Data Comics;

ACM Reference Format:
Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2023. Personal Data Comics: A Data Storytelling Approach Supporting Personal Data Literacy. In *XI Latin American Conference on Human Computer Interaction (CLIHC 2023), October 30–November 01, 2023, Puebla, Mexico.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3630970.3630982

1 INTRODUCTION

Personal data is part of daily life. Every time a person uses or interacts with a digital product or service (e.g., an internet browser) she leaves behind a digital trace (e.g., search history) which is scattered around, yet, deeply entangled with her (i.e., relating to and reflecting specific events and aspects of her life) [10]. It contains meaningful but decontextualized information about her (e.g., her search queries) and additional information could be inferred from it (e.g., her location, interests, and worries). When individual data traces are combined, they constitute a personal data trail [12]. Most people are unaware of the nature of the data within these trails and are often uninformed or ill-informed about the data collection practices around them [4, 12].

Traces of personal data are increasingly used for scientific research across various contexts and domains; as researchers invite people to share or donate their data through digital apps and platforms[1]. In practice, people request and obtain a copy of their data from a data controller (i.e., public or private service provider) and then share or donate it. Although this is a voluntary process and begins with informed consent, prior research has demonstrated how people who share or donate personal data are often not aware of the sensitive or intimate information that is part of their data and is transferred [10]. Hence, when a person transfers her data she could inadvertently and unintentionally reveal sensitive and intimate details about herself and her social relationships.

To support better-informed decisions around sharing and donating personal data, it is critical to promote personal data literacy. This term has different meanings in different contexts, but broadly it is about the knowledge and skills a person must have to successfully manage and navigate the (personal) data ecosystems of which she is a part [11, 22]. Informed by previous literature, in this paper we refer to personal data literacy as understanding (1) the practices and infrastructures behind personal data collection (e.g., how is my data collected and stored? how can I access it?) [8, 11], (2) the content and characteristics of personal data (e.g., what is in my data?) [8, 14], and (3) the entanglements between data and people (e.g., how does personal data reflect my behavior and experiences? how is personal data related to me?) [8, 12].

Data visualization has been used successfully to foster personal data literacy across several contexts and types (e.g., [8, 13, 19]). Nonetheless, data visualization approaches are often limited to a specific type of data (e.g., sensor data), and a specific context (e.g., physical activity). Hence, they often disregard the 'trail' that results from combining data from multiple sources and types. Moreover, they elicit yet fail to capture rich and contextual insights from people's situated knowledge. This enriches the data, helps people better understand their entanglements with data, and reduces the possibility of incorrect assumptions or inferences being made [9]. One way this limitation has been addressed is through annotated data visualizations (e.g., [3, 13]), a combination of non-sequential

This work is licensed under a Creative Commons Attribution International 4.0 License.

CLIHC 2023, October 30–November 01, 2023, Puebla, Mexico
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1657-7/23/10.
https://doi.org/10.1145/3630970.3630982

[1]This process has been enabled by the European General Data Protection Regulation (GDPR); particularly the 'right to access' [7] which allows people to request a copy of their data and (re)use it for their purposes.

data visualizations with unstructured textual elements added organically by people as they relate the data to their experiences. The process of annotating a data visualization often requires support from experts (e.g., a researcher or interviewer) who guide people through the interpretation of the data and invite reflections with probing questions. Yet, expert support is not necessarily available when one is at home pondering whether to share or donate personal data.

We propose personal data comics, a visual storytelling method of sequential images and structured textual elements [24], as an alternative to annotated data visualizations which may overcome some of the limitations described above. First, integrating data from multiple sources and types, as previous research illustrates how data comics invite multiple ways to represent, communicate, and combine the data [1, 21]; second, eliciting and capturing situated knowledge as Bach et al. [1] and Zhang et al. [23] demonstrate data comics invite to create connections between data visualizations and the context of the data through narrative. We hypothesize data comics support individuals in independently developing personal data literacy. However, previous research (e.g., [1, 20, 21, 23]) has mainly focused on data comics created by *data experts* (e.g., professional designers, data scientists, and illustrators) and it remains unknown to what extent they can be created independently by *non-data experts* (i.e., people with no prior experience analyzing or visualizing data).

In this paper, we investigate the following research question: **How effective and engaging are data comics in fostering personal data literacy?** Specifically, we aim to understand: **(RQ1)** How does independently creating data comics support *non-data experts* in developing personal data literacy? and **(RQ2)** How effective and engaging are data comics when compared to annotated data visualizations? To investigate our research questions, we conducted two complementary studies focused on the specific context of pregnancy, where the data trail is made up of various types of data from different sources, increasingly used for research. First, we qualitatively explore the process of *creating* data comics with three pregnant participants who brought their pregnancy-related data (Section 2). Second, we compare the personal data comics with the annotated data visualizations, in terms of effectiveness and engagement through a quantitative between-subject study with 34 participants (Section 3). Our contribution is threefold, we (1) illustrate how non-data experts create personal data comics, (2) describe the modalities of interaction with personal data that originate from the independent creation of personal data comics, and (3) evaluate how personal data comics are perceived and understood by the general public.

2 STUDY 1: CREATION OF PERSONAL DATA COMICS

With this study, we aimed to investigate how independently creating data comics supports *non-data experts* in developing personal data literacy (RQ1). We invited three *non-data expert* participants to create personal data comics about their data and their pregnancy (i.e., a specific life event entangled in, and reflected by, the multiple sources of data). We acknowledge the limited sample size of this study, but highlight the participants' engagement in collecting and sharing intimate data about their pregnancy and the rich and meaningful insights it enabled.

2.1 Participants

We used snowball sampling to reach out to pregnant people across different online communities and invited them to participate in our research. Three women in their early thirties volunteered to participate. None of them had previous experience working with data or data visualization. Participants used three digital products, as they normally would, for a period of one month: (1) web browser, (2) fitness tracker, and (3) pregnancy tracker. Resulting in three different types and formats of data: (1) browsing history, (2) physical activity logs, (3) pregnancy self-reports.

2.2 Procedure

The individual comic-creation sessions lasted between 65 and 95 minutes and took place via Zoom using the online whiteboard tool Miro. We choose this setup because of the flexibility of the tool to zoom in and out, drag and drop, and link elements. Prior to each session, we asked participants to obtain a copy of their data and share them with us. We provided video tutorials explaining how to request and obtain a copy of their data from each data controller (e.g., Google, Ovia Pregnancy App). With the data, we created several data visualizations over three periods of time: (1) a day, (2) a week, and (3) a month. In doing so, we aimed to highlight the temporality of data. In addition, we created 'the personal data trail' comic[2] describing the practices and infrastructures behind data collection to facilitate understanding and familiarize participants with the data comic format. Each session comprised two main activities. First, we invited participants to read through and discuss "the personal data trail" comic, with this activity, we aimed to familiarize participants with the format of data comics. Second, we invited participants to create their personal data comics, by dragging and dropping different elements into an empty Miro board and thinking aloud. For this activity, we provided participants with comic elements as material, including characters portraying different emotions[3], narrative texts, vignettes, and speech bubbles. In addition, we provided them with data visualizations, which are commonly used as a starting point for creating data comics (e.g., [1, 24]). We invited participants to create three comics, each representing a time: a day, a week, and a month. We were not involved in the comic creation process other than answering specific questions about the online whiteboard tool Miro.

2.3 Analysis and Results

We recorded the audio of each comic-creation session and transcribed it using MS Office 365. We analyzed the transcripts as well as the personal data comics created by each participant, using reflexive thematic analysis [5, 6], within a constructionist framework. We coded the entire dataset using ATLAS.ti, reviewed the codes, and grouped them into themes: (1) *data is hidden in plain sight*, (2) *comics drive personal data stories*, and (3) *data comics transform data*.

[2]The 'Personal Data Trail' comic in supplementary material.
[3]Characters generated with the Comicgen API https://gramener.com/comicgen/

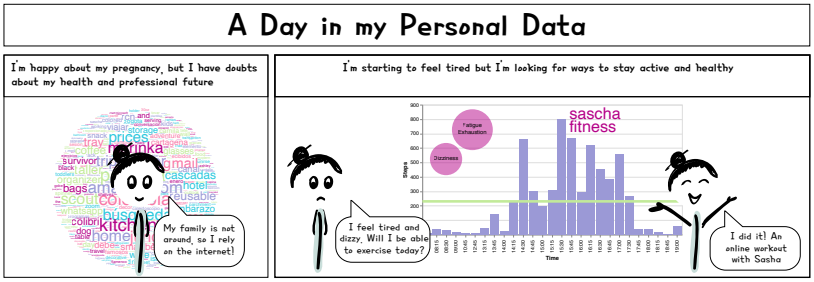


Figure 1: A segment of a personal data comic created by P3. Shown with her permission

2.3.1 Data is Hidden in Plain Sight. Prior to our study, participants did not know about the possibility of obtaining a copy of their data or the practicalities of the process. In this way, data is hidden behind regulations, such as the GDPR, which are not well known, and privacy policies or similar documents that are not easily digestible. Yet, it is available upon request and it is relatively easy to access. All participants found the process of obtaining a copy of their data surprisingly easy. *"When you first told me that I needed to download my data, I thought that I needed to hire a computer scientist and that it was going to be very difficult. Then I did it step by step and it was fast, it was easy." (P3)* Nonetheless, none of the participants managed to open the different files they obtained by requesting a copy of their data. They either assumed these would be too technical and complicated (P1, P2) or tried to open them with incompatible software (P3). These results echo previous research (e.g., [4]) describing how the files and formats in which data is returned to data subjects are difficult to understand and manipulate. In this way, data is hidden even more, behind files and formats that are not easy to navigate; although it is technically possible to navigate them; *"I did not open it* [the file]. *I just assumed that I wouldn't understand what it looks like." (P2)*

Overall, participants initially perceived the notion of data, and its manifestation in digital form, as intimidating and foreign. Data, even if personal and deeply entangled with the participants and their pregnancies, was initially perceived as unfamiliar. Hence, the personal nature of the data is also hidden, although it becomes visible through interpretation and exploration. That participants initially perceived their data as unfamiliar and intimidating illustrates their personal data literacy at the beginning of the interview. In addition, it highlights the need to design and develop interactions and interventions that convey what is possible (e.g., obtaining a copy of your personal data) and facilitate people's engagement with their data.

2.3.2 Comics Drive Personal Data Stories. The process of creating personal data comics is one of approaching personal data as a probe that elicits memories and reflections on lived experiences. This process is initiated from and guided by the data's temporal dimension. Time mediates a two-way relationship between the data and the participant's lived experiences and invites questions (e.g., where was I? What was I doing?). *"The searches* [browsing history logs] *for that week are from when I had to inform my boss that I am pregnant, I had to enter various platforms and pages." (P3)* The sequential and narrative format of the comic invites participants to create stories with and from the data, which is contextualized and enriched through characters and comic elements such as captions and speech bubbles. These elements support participants in connecting and relating data of different types and from different sources with their lived experiences, which are intertwined. In addition, they invite the participants to (re)interpret and give meaning to the data. Is logging the symptom 'joint pain' an indicator of 'joint pain' or '*unusual* joint pain'? How much pain is joint pain? *"I was like oh I need to take it easy today* [Monday]. *Especially because on Sunday I did a lot of steps. And Monday I had joint pain, yeah, and joint pain is interesting because I only record it if it really hurts because obviously, I've got a lot of joint pain. But only if it's unusual, you know? Well, uh, we could do like a little arrow between the joint pain and the steps, yeah." (P2)*

The resulting personal data comics are then a visual story. Perceived as beautiful, entertaining, and fun. Unlike data, which hides behind its complexity, data comics are clear and easy to digest, *"if I was just reading this and these were written as paragraphs it wouldn't be as engaging or as clear"* (P1). Participants appreciated the experience of creating the comics as an easy and engaging way to explore their data. In addition, they appreciated the data comics themselves, reminders of their experiences during a month of their pregnancy. *"The data presented in the way it did and using it to make a comic invited me to reflect on what has been going on in my life as well as the pregnancy. It makes a lovely diary/blog as a memory of this really special time as well." (P2)*

2.3.3 Data Comics Transform Data. The comic format invites participants to select, filter, combine, and compare different types of data and augment data with contextual insights represented through captions and other narrative elements. Hence, the process of creating data comics transforms the data, by integrating it with other data and elements and augmenting it with relevant contextual information (e.g., what "joint pain" really means). We briefly illustrate (Fig. 2) and describe the strategies that were used by the participants throughout this process:

- **Cropping the Data**: Through cropping fragments (e.g., a period of time) or specific elements (e.g., a category of data) of a data visualization, participants identified and segmented the salient or relevant aspects of the data for a given topic or activity. This is illustrated in Fig. 2.a, where the daily steps visualization is cropped to show only the time of day the participant walked to dinner with her partner and walked back.

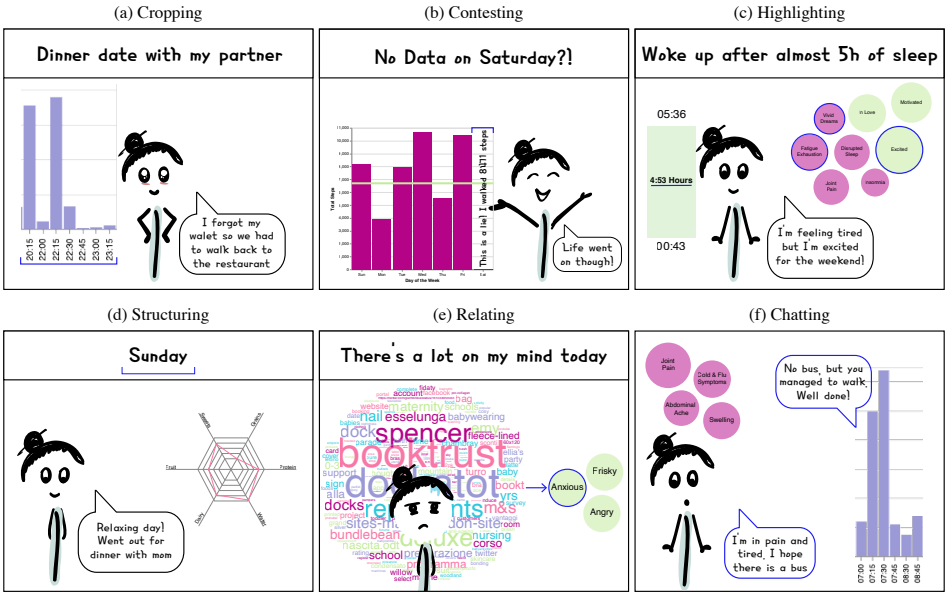


Figure 2: Examples of comic vignettes created by the participants and the different modalities of interacting with data (blue). Shown with permission.

- **Contesting the Data**: While creating the data comics, participants realized that the data, represented through data visualizations, was sometimes incomplete (e.g., missing data) or incorrect (e.g., wrong values). They used comic elements, such as captions, speech bubbles, and annotations, to contest and correct the data. This is illustrated in Fig. 2.b, where the text *"this is a lie, I walked 8471 steps"* is added to the last day of the weekly steps visualization; showing a bar with fewer than 200 steps.
- **Highlighting the Data:** Participants used lines, shapes (e.g., circles and stars), and other elements (e.g., onomatopoeia) to highlight the most important values of a data visualization or a vignette combining multiple data visualizations. This is illustrated in Fig. 2.c, where the total sleep time is highlighted together with (lack of) sleep-related symptoms such as vivid dreams, fatigue or exhaustion, and mood. All of these are important to emphasize the key message of the vignette, the "tiredness" that the participant experienced that day.
- **Structuring the Data:** Participants used several design patterns [1] (e.g., narrative, temporal, faceting) in the data comics to structure the data and the comics. In doing so, they used a dimension (e.g., time) or category (e.g., emotions) of the data to shape the story. This is illustrated in Fig. 2.d, where time (i.e., day of the week) in the caption is used to drive the temporal sequence of the comic.
- **Relating the Data:** While creating the data comics, participants identified relationships between different types of data (e.g., sleep and browsing history logs), as well as between the data and their lived experiences (e.g., *"this is my first pregnancy and it makes me feel doubts. But my family is not around, so I search online, it is easier."* (P3)). Participants used comic elements such as arrows[4], captions, and speech bubbles to underline these relationships. This is illustrated in Fig. 2.e, where the participant draws an arrow between her browsing history logs, representing all the things in her mind, and her feeling of anxiety.
- **Chatting with the Data:** While creating the comic participants approached the data as an entity with something to say about them and their pregnancy experiences (e.g., evidence or confirmation that something happened). This is illustrated in Fig. 2.f, where the participant represents her morning steps data visualization both as accounting for something that happened (i.e., she walked to work) and "telling" her something about it through a speech bubble.

2.4 Summary

Due to its nature, understanding, interpreting, and manipulating personal data is not a trivial process for non-data experts. The process of creating personal data comics enables participants to demystify the perceived complexity around personal data and understand the way data is intertwined with their lived experiences. Through this process, participants were able to (1) identify relevant data sources, (2) relate them to each other and their lived experiences, and (3) correct or oppose the data when necessary. In doing so, they created personal narratives that augment the data with relevant contextual information. This process and its output (i.e., the data comics) were perceived as clear, entertaining, and fun.

[4]These were not initially provided by us as part of the material.

3 STUDY 2: EVALUATION OF PERSONAL DATA COMICS

With this study, we aimed to investigate how effective and engaging are data comics (i.e., sequential images and data representations with structured textual elements) and annotated data visualizations (i.e., non-sequential data visualizations with unstructured textual elements). To measure effectiveness, we use a combination of three factors: efficiency (i.e., time spent working with the visual) [2, 25], accuracy (i.e., correct understanding of the message) [21, 25], and immediate recall (i.e., ability to maintain and retrieve information) [18, 21]. To measure engagement we use the user engagement scale [16], which understands engagement in terms of focused attention (i.e., time spent looking at the visual) [16–18], perceived usability (i.e., degree of control and effort) [16, 17], aesthetic appeal (i.e., attractiveness), and reward (i.e., appreciation for the visual and perceived success) [17]. We developed the following hypotheses:

- $H_{effectiveness-efficiency}$: we expect participants to derive specific information from the visual faster after reading the data comic than the annotated data visualization. Informed by previous research (e.g., [21, 24]), we anticipate that the linear and sequential structure of the data comic, where information is divided into chunks, improves efficiency.
- $H_{effectiveness-accuracy}$: we expect participants to have a more correct understanding of the data after reading the data comic than the annotated data visualization. Informed by previous research (e.g., [21, 24]), we anticipate that the narrative style of the comic will help improve understanding.
- $H_{effectiveness-recall}$: we expect participants to have a more correct immediate recall of the data after reading the data comic than the annotated data visualization. Informed by previous research, (e.g., [2, 24]), we expect the comic format, which groups information into panels with unique information and visuals, to facilitate immediate recall.
- $H_{engagement}$: we expect participants to rate the data comics as more engaging than the annotated data visualization. Similar to Zhao et al. [24], who compared data comics to infographics and concluded that data comics were more engaging, we expect the comic to be perceived as more engaging and more fun.

3.1 Participants

After performing a power analysis ($\alpha = 0.05, \beta = 95\%$), we recruited a non-probabilistic self-selected sample of thirty-four participants (17 identified as women and 17 as men) via the online research platform Prolific. Data comic evaluation studies have relied on similar sample sizes (e.g., [21, 24]). Since this study focuses on perception and understanding, its scope is the general public and not exclusively people who have experienced pregnancy. All participants could complete our survey on a laptop or desktop device and reported English language proficiency. They were between 21 and 59 years old ($mean = 29$, $median = 27$). Participants were compensated according to the minimum wage in [country of author's affiliation]. All participants completed the survey in less than 25 minutes.

3.2 Procedure

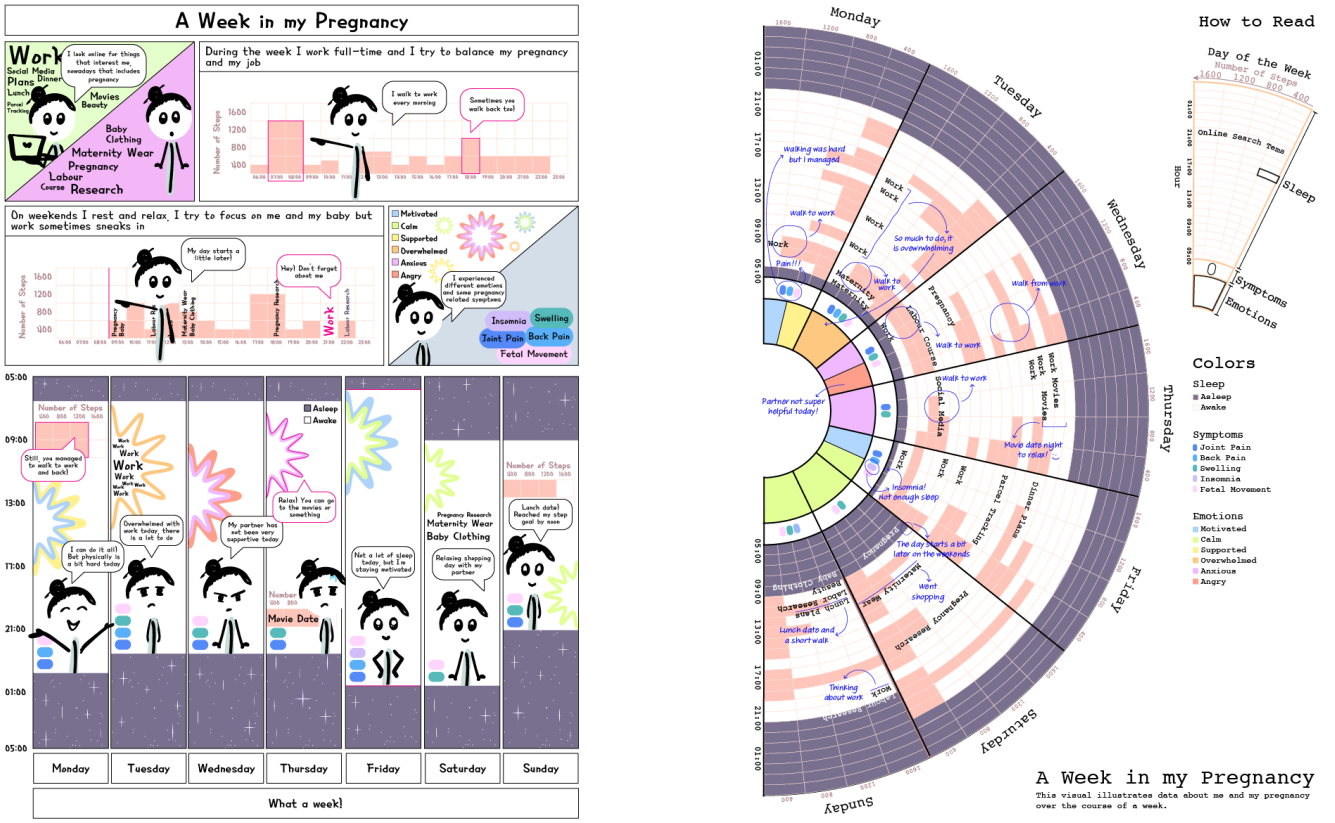
We designed a data comic (Fig. 3a) and an annotated data visualization (Fig. 3b) with a synthetic dataset, derived from the aggregation of anonymized data from the participants in the creation study. The visuals represented a week in the life of a pregnant person through three different types of data: (1) browsing history logs (e.g., search terms and keywords), (2) physical activity (e.g., daily steps and sleep), and (3) pregnancy logs (e.g., symptoms and emotions). We made sure to include the same data, encode it in the same way (i.e., using the same symbols and colors), and integrate the same textual elements either as narrative elements (i.e., captions and speech bubbles) or annotations. We used these visuals as prompts during a between-subject study (i.e., every participant read either the data comic or the annotated data visualization) delivered through an online survey.

Participants were randomly assigned to a visual and were invited to explore it for a few minutes. Then, they were invited to answer six multiple-choice questions measuring their understanding of the visual. Participants were instructed to assume that the visual was about them, hence the questions were formulated in the second person. The questions covered the information types proposed by [21], including (1) *time* (e.g., In what time frame did you take the most steps on weekdays (i.e., from Monday to Friday)?); (2) *distribution* (e.g., What day of the week did you sleep the least hours?); (3) *single-fact* (e.g., What pregnancy-related symptoms did you experience during the week?), (4) *encoding* (e.g., Which color is used to present anger?), (5) *comparison* (e.g., Do you get up earlier or later during the week or on weekends?), and (6) *outliers* (e.g., What was the most prominent keyword in your online search logs on Tuesday?). During this stage, the participants could refer back to the visual. Finally, to measure immediate recall and engagement, participants were invited to answer open-ended questions about the visual, as proposed by Bateman et al. [2], and the short-form user engagement scale [17]. During this stage, participants could not refer back to the visual.

3.3 Analysis and Results

During the evaluation study, we collected different types of data: (1) time spent exploring the visual and answering the multiple-choice questions; (2) error rate from multiple-choice questions to measure accuracy (i.e., correct understanding), as proposed by Wang et al. [21]; (3) visual recall scores, coded as proposed by Bateman et al. [2], and (4) engagement from the short-form user engagement scale [17]. We report on our findings from the statistical analysis of these data.

- **Effectiveness - Efficiency:** It was measured as the average time spent exploring the visual and answering each of the multiple-choice questions. We used the Shapiro-Wilk normality test and found that the average time for the annotated data visualization was not normally distributed. Using the Mann-Whitney test we found no significant difference ($\alpha = .05$) between the average time spent with the data comic ($M = 41.5s$) and the annotated data visualization ($M = 50.9s$).
- **Effectiveness - Accuracy:** It was measured based on whether the answer to each of the six multiple-choice questions was correct (score = 1) or not (score = 0). We computed the



(a) Data Comic. **(b) Annotated Data Visualization.**

Figure 3: Visuals designed for the evaluation study.

mean accuracy score across all the answers as described by Wang et al. [21]. We used the Shapiro-Wilk normality test and found that the accuracy scores were not normally distributed. Using the Mann-Whitney test we found no significant difference ($\alpha = .05$) in mean accuracy scores between the data comic ($M = .93$) and the annotated data visualization ($M = .83$).

- **Effectiveness - Recall:** It was measured following the methodology described by Bateman et al. [2]. The first author computed the immediate recall scores by coding the responses to the open-ended questions according to the following scale: all correct (score = 3), mostly correct (score = 2), mostly incorrect (score = 1), all incorrect (score = 0). We used the Shapiro-Wilk normality test and found that the immediate recall scores for the annotated data visualization were not normally distributed. Using the Mann-Whitney test we found no significant difference ($\alpha = 05$) in immediate recall scores between the data comic ($M = 3.90$) and the annotated data visualization ($M = 3.65$).
- **Engagement:** We used the short form of the user-engagement scale proposed by O'Brien et al. [17] to measure participants' engagement. We calculated the scores for each of the subscales (i.e., Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward) and the overall engagement score. We used the Shapiro-Wilk normality test and found the overall engagement scores to be normally distributed. Using an independent-sample t-test ($\alpha = .05$) we found that there is a significant difference ($t(32) = -2.72, p < .05$) in the overall engagement between the data comic ($M = 3.23$) and the annotated data visualization ($M = 2.75$).

3.4 Summary

Data comics and annotated data visualizations scored similarly in terms of efficiency, accuracy, and recall. Hence, we could not find evidence that data comics are more effective than annotated data visualizations. Nonetheless, both formats were regarded as highly efficient (i.e., average time spent identifying information less than 50 s) and accurate (i.e., average accuracy score greater than 0.8). While there was no significant difference in terms of effectiveness, participants perceived data comics as more engaging than annotated data visualizations. These findings align with previous research that concluded that data comics are generally perceived as engaging, enjoyable, and fun (e.g., [21, 24]) and with the qualitative results from the creation study.

4 DISCUSSION

The independent creation of personal data comics is one of active and playful engagement with personal data. It enabled participants to engage with their data as prompts to build a story, with their personal experiences as the driving plot. It facilitated a two-way dialogue between data and people; data was confirmed and augmented by an experience, event, or activity, which in turn were used to challenge or contest (other) data. In addition, it allowed the smooth integration of different types of data from multiple sources, which were tied to a specific life event or context. Moreover, it was perceived as effective and engaging in delivering what is expected by the reader to be a complex message, mainly because it breaks the information into small *chunks* that are easy to read and digest. Hence, data comics facilitate explanation and support data legibility [15].

The comic format allows complex information to be represented and communicated in an accessible and entertaining way [21, 24], and the comic layout and its linear sequential structure offer a guide for the reader [1]. We found that these attributes provide enough structure and familiarity for non-data experts to create their personal data comics, reducing the need for expert facilitation and guidance besides data processing and visualization.

In our study, participants began creating their comics from visual representations of their data, and not from the raw data. Bach et al. [1] describe the different starting points for data comics, including raw data, a single visualization, and a set of visualizations. We recommend a set of visualizations as a starting point, as it removes the initial barrier of manipulating the raw data, which was perceived as challenging and intimidating by the participants in our study. For this reason, (pre)processing and visually representing the data is an important step in supporting non-data experts in creating (personal) data comics. Tools that support the creation of personal data comics, such as [20] and [24] could integrate data visualization libraries that generate multiple visual representations to choose from. In addition to facilitating the interaction with the data, these tools should stimulate the creative process. Bach et al. [1] describe that there are no clear guidelines or approaches for creating data comics. They propose a series of design patterns that could support data comic creators in their process, such as narrative patterns, (i.e., creating connections between the data visualization, the narrator, and the context of the data), and temporal patterns (i.e., communicating a temporal change in the data). Some of these were used instinctively by the non-data expert participants in our study. Yet, providing an overview of possibilities and layouts could better support people in independently creating personal data comics. In addition, the different modalities of interacting with and relating to data (visualizations) used by our participants (i.e., *cropping*, *contesting*, *highlighting*, *structuring*, *relating to*, and *chatting with* the data) could enrich existing design patterns.

5 CONCLUSION

We proposed personal data comics as a data storytelling approach to support personal data literacy. We conducted two complementary studies in the context of pregnancy, to investigate (1) how independently creating data comics fosters personal data literacy among *non-data experts* and (2) how effective and engaging are data comics when compared to annotated data visualizations. Our results indicate that *non-data experts* can successfully create engaging personal data comics and develop their personal data literacy in the process. Our contribution is threefold, we (1) illustrate how non-data experts create personal data comics, (2) describe the modalities of interaction with personal data that originate from the independent creation of personal data comics, and (3) evaluate how personal data comics are perceived and understood by the general public. They lead to valuable implications for the design of future comics and tools that facilitate independent comic creation.

ACKNOWLEDGMENTS

We sincerely thank the participants in the two studies described in this paper. We especially wish to acknowledge the contribution of the three pregnant women who shared with us a bit of their pregnancies, their data, their time, and their creativity. We thank Roos Teeuwen and Hosana Morales for their helpful feedback on the visuals. In addition, we thank the anonymous reviewers for their valuable feedback.

REFERENCES

[1] Benjamin Bach, Zezhong Wang, Matteo Farinella, Dave Murray-Rust, and Nathalie Henry Riche. 2018. Design patterns for data comics. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–12. https://doi.org/10.1145/3173574.3173612

[2] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. *Conference on Human Factors in Computing Systems - Proceedings* 4 (2010), 2573–2582. https://doi.org/10.1145/1753326.1753716

[3] Sander Bogers, Joep Frens, Janne van Kollenburg, Eva Deckers, and Caroline Hummels. 2016. Connected Baby Bottle. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 301–311. https://doi.org/10.1145/2901790.2901855

[4] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. 2022. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501947

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[6] Virginia Braun and Victoria Clarke. 2013. *Successful Qualitative Research. A Practical Guide for Beginners*. SAGE Publications Ltd, London. 400 pages.

[7] European Parliament. 2018. General Data Protection Regulation (GDPR). https://gdpr.eu/

[8] Sandra Gabriele and Sonia Chiasson. 2020. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376651

[9] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2022. Reconstructing Intimate Contexts through Data Donation: A Case Study in Menstrual Tracking Technologies. In *Nordic Human-Computer Interaction Conference*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3546155.3546646

[10] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2023. What is Sensitive About (Sensitive) Data ? Characterizing Sensitivity and Intimacy with Google Assistant Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581164

[11] Jonathan Gray, Carolin Gerlitz, and Liliana Bounegru. 2018. Data infrastructure literacy. *Big Data and Society* 5, 2 (2018). https://doi.org/10.1177/2053951718786316

[12] Awais Hameed Khan, Stephen Snow, Scott Heiner, and Ben Matthews. 2020. Disconnecting: Towards a semiotic framework for personal data trails. *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), 327–340. https://doi.org/10.1145/3357236.3395580

[13] Albrecht Kurze, Andreas Bischof, Sören Totzauer, Michael Storz, Maximilian Eibl, Margot Brereton, and Arne Berger. 2020. Guess the Data: Data Work to Understand How People Make Sense of and Use Simple Sensor Data from Homes. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376273

[14] Richard Mortier, Hamed Haddadi, Tristan Henderson, Derek McAuley, and Jon Crowcroft. 2014. Human-Data Interaction: The Human Face of the Data-Driven Society. *SSRN Electronic Journal* 10 (2014), 1–11. https://doi.org/10.2139/ssrn.2508051

[15] Richard Mortier and Tristan Henderson. [n.d.]. *Human-Data Interaction: The Human Face of the Data-Driven Society.* Technical Report.

[16] Heather O'Brien and Elaine Toms. 2009. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology* 64, July (2009), 1852–1863. https://doi.org/10.1002/asi

[17] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human Computer Studies* 112, January (2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[18] Bahador Saket, Alex Endert, and John Stasko. 2016. Beyond usability and performance: A review of user experience-focused evaluations in Visualization. *ACM International Conference Proceeding Series* 24-October (2016), 133–142. https://doi.org/10.1145/2993901.2993903

[19] Peter Tolmie, Andy Crabtree, Tom Rodden, James Colley, and Ewa Luger. 2016. "This has to be the cats" - Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16).* ACM, New York, NY, USA, 491–502. https://doi.org/10.1145/2818048.2819992

[20] Zezhong Wang, Hugo Romat, Fanny Chevalier, Nathalie Henry Riche, Dave Murray-Rust, and Benjamin Bach. 2022. Interactive Data Comics. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 944–954. https://doi.org/10.1109/TVCG.2021.3114849

[21] Zezhong Wang, Shunming Wang, Matteo Farinella, Dave Murray-Rust, Nathalie Henry Riche, and Benjamin Bach. 2019. Comparing effectiveness and engagement of data comics and infographics. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–12. https://doi.org/10.1145/3290605.3300483

[22] Annika Wolff, Daniel Gooch, Jose Cavero Montaner, Umar Rashid, and Gerd Kortuem. 2017. Creating an understanding of data literacy for a data-driven society. *Journal of Community Informatics* 12, 3 (2017), (In press). www.ci-journal.net/index.php/ciej/article/view/1286.

[23] Xinglin Zhang, Zheng Yang, Wei Sun, Yunhao Liu, Shaohua Tang, Kai Xing, and Xufei Mao. 2016. Incentives for mobile crowd sensing: A survey. *IEEE Communications Surveys and Tutorials* 18, 1 (2016), 54–67. https://doi.org/10.1109/COMST.2015.2415528

[24] Zhenpeng Zhao, Rachael Marr, and Niklas Elmqvist. 2015. Data Comics: Sequential Art for Data-Driven Storytelling. *HCIL Technical Report* 15, Figure 1 (2015), 12. https://www.semanticscholar.org/paper/Data-Comics-%3A-Sequential-Art-for-Data-Driven-Zhao-Marr/43f6a7a70a9cc3dfdaec99f0c240a04830191827?p2df

[25] Ying Zhu. 2007. Measuring Effective Data Visualization. In *Advances in Visual Computing, Third International Symposium.* Vol. Part II.