

Tracing Personal Data Using Comics

Andreas Schreiber¹(✉) and Regina Struminski²

¹ German Aerospace Center (DLR), Intelligent and Distributed Systems,
Linder Höhe, 51147 Cologne, Germany

Andreas.Schreiber@dlr.de

² Faculty of Media, University of Applied Sciences Düsseldorf,
Münsterstraße 156, 40476 Düsseldorf, Germany

regina.struminski@study.hs-duesseldorf.de

<http://www.dlr.de/sc>, <https://medien.hs-duesseldorf.de>

Abstract. Personal health data is acquired, processed, stored, and accessed using a variety of different devices, apps, and services. These are often complex and highly connected. Therefore, privacy violations and other use or misuse of the data are hard to detect for many people, because they are not able to understand the trace (i.e., the provenance) of that data. We present a visualization technique for personal health data provenance using comics strips. Each strip of the comic represents a certain activity, such as entering data using an app, storing or retrieving data on a cloud service, or generating a diagram from the data. The comic strips are generated automatically using recorded provenance graphs. The easy-to-understand comics enable all people to realize crucial points regarding their data.

Keywords: Provenance · Quantified Self · Personal informatics · Visualization · Comics

1 Introduction

Understanding how a piece of data was produced, where it was stored, and by whom it was accessed, is crucial information in many processes. Insights into the data flow are important for gaining trust in the data; for example, trust in its quality, its integrity, or trust that it has not unwantedly been accessed by organizations. Especially, detecting and investigating privacy violations of personal data is a relevant issue for many people and companies.

A specific area where integrity and privacy of data is crucial is health and fitness. For example, personal health data should not be manipulated, if doctors base a medical diagnosis on that data. Health-related data and personal data from self-tracking (Quantified Self; QS) [6, 8] should not be available to other people or companies, as this might lead to commercial exploitation or even disadvantages for people. In this field, data is often generated by medical sensors or wearable devices, then processed and transmitted by smartphone and desktop applications, and finally stored and analyzed using services (e.g., web or cloud

services operated by commercial vendors). Following the trace of data through the various distributed devices, apps, and services is not easy. Especially, people who are not familiar with software or computer science are often not able to understand where their data is stored and accessed.

To understand the trace of data, the provenance [15] of that data can be recorded and analyzed. Provenance information is represented by a directed acyclic property graph, which is recorded during generation, manipulation, and transmission of data. The provenance can be analyzed using a variety of graph analytics and visualization methods [11]. Presenting provenance to non-experts is an ongoing research topic (“*Provenance for people*”). As a new presentation and visualization technique for provenance, we introduce *provenance comics*:

- We explain the general idea of *provenance comics* for provenance compliant with the PROV standard [16] (Sect. 3).
- We describe a specific visual mapping between the provenance of Quantified Self applications [19,20] and their graphical representations in comic strips (Sect. 4).
- We briefly describe our prototype for *automatically generating provenance comics* (Sect. 5).
- We give details and results of a qualitative user study (Sect. 6).

2 Motivation

The provenance of data is usually represented as a directed acyclic graph. In many visualizations the graph is sorted topologically from left to right or top to bottom. Much like in a family tree, the “oldest” data can then be seen at the left or top and the “youngest,” most recent data at the right or bottom.

While these graphs may, to some extent, seem quite self-explaining to scientists, they can be rather hard to understand for laymen who are not usually concerned with graphs at all and have not been trained to read them.

Furthermore, provenance graphs can sometimes grow to enormous sizes, becoming so huge that even experts will have a hard time reading them. Since the span of immediate memory is limited to 7 ± 2 entities at a time [14], graphs containing more than five to nine items will become gradually harder to interpret with every new item being added. However, 7 ± 2 is a value that is easily reached and exceeded by even simple examples of provenance graphs (see Fig. 1). The larger the graphs become, the more difficult it is to draw conclusions and derive new findings from the provenance data.

The possibility to view their own provenance data is of no value to end users, if the visualization of that provenance is unintelligible to them. It cannot be expected that they learn how to read an abstract, possibly complex graph. Instead, the visualization should be simple, self-explaining, and familiar in such a way that end users can read and understand it almost effortlessly.

4 Visual Mapping






To generate the provenance comics, we defined a consistent visual language [21]. This visual language allows to “translate” the provenance data into corresponding drawings. Generally speaking, we mapped elements of the PROV standard (*Entity, Activity, Agent*) onto three distinctive features: *shapes, colors, and icons or texts*.

4.1 Shapes

We designed and selected shapes according to several criteria. Most importantly, we created shapes that do not show much detail. Instead, they have a “flat” look without any textures, decorations, shadows, or three-dimensional elements.

Table 1 gives an overview of the shapes we selected to reflect the different types of elements in the Quantified-Self PROV model [19]. Activities are not directly listed here. Unlike agents or entities, activities are actions that take place over time, as described in Sect. 3. Thus they are not depicted as a single graphic; instead, they represent a temporal progress and only become visible through the sequence of events in the next three to five panels of the comic.

Table 1. Shapes defined for different types of PROV elements.

Element type	Shape	Example
Agent type: Person	human silhouette	
Agent type: SoftwareAgent	smartphone, computer, ... (depending on the agents "device" attribute)	
Agent type: Organization	office building	
Entity	file folder, document, chart, ... (depending on the entity's type attribute)	
Activity-related objects	button, icon, ... (depending on the activity's name or "role" attribute)	

4.2 Icons, Letters, and Labels

As a second distinctive feature, all main actors in the comics carry some kind of symbol on them, whether it be an icon, a single letter, or a whole word (Fig. 2).

- **Person** agents always wear the first letter of their name on the chest.
- **Organization** agents display their name at the top of the office building.
- **SoftwareAgents** show an application name on the screen.
- Entities are marked by an icon representing the type of data they contain. A few icons have been defined for some types of data that are common in the Quantified-Self domain (Table 2).

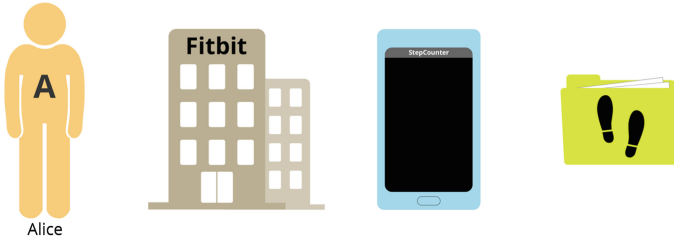


Fig. 2. Agents and entities using three distinctive features (shape, color, icons/text).

Table 2. Icons for some typical Quantified Self data types.

Data type	Icon	Description
Blood pressure		a heart outline with a pressure indicator
Heart rate		a heart containing an ECG wave
Sleep		a crescent moon with stars
Steps		a pair of footprints
Weight		a weight with the abbreviation “kg” cut out

4.3 Colors

We defined colors for entities as well as the different types of agents. For example, *Person* agents use a light orange color, while *SoftwareAgents* have a light blue and *Organization* agents a tan color. Entities are always colored in a bright yellowy green.

Alternative color shades have been defined for both agents and entities in case that two or three objects of the same type ever need to appear at once. We took care that colors are well-distinguishable even for people suffering from color vision deficiencies (pronatopia, deuteranopia, tritanopia, and achromatopsy). In the few cases where they are not, discriminability is still granted through the other two distinctive features, namely shape and icons or labels.

4.4 Captions and Text

We aimed to include as little text as possible in the comics. Most of the information should be conveyed by the graphics to provide an effortless “reading” experience. However, in certain cases, a few words are useful to support the interpretation of symbols. For example, when up- or downloading data, the words “Uploading...” or “Downloading...” are added below the cloud icon. These short annotations take only little cognitive capacity to read, but may greatly help understand certain icons.

Buttons also use textual labels, as it is very difficult to convey the actions they represent in the form of graphics. The labels are only very short though, mostly consisting of only one or two words (e.g., “View graph” or “Export CSV”).

Captions are used to expose the date and time when activities took place. Every comic strip begins with such a caption in the very first panel to give the reader temporal orientation. If a relevant amount of time has passed between two activities, a caption may be used again to communicate this to the reader.

The comic depicted in Fig. 3 contains examples of these textual annotations, button labels, and captions.

4.5 Level of Detail

The comics are characterized by extreme simplicity and reduction to the essentials. The reader should never have to look for the important parts of the image. Thus, only relevant items are pictured; no purely decorative graphics are used. This includes the background, which is plain white at all times. No surroundings or other possible distractions are ever shown. By eliminating details, reducing images to their essential meaning, and focusing on specific elements, the emphasis is put on the actual information.

4.6 Commonly Known Symbols

Some of the graphics used in the comics rely on the reader’s experience. For example, “sheet of paper” and “document folder” icons have been used for decades to symbolize data and collections of data, and in recent years, the “cloud” icon has become a widely known symbol for external data storage space.

Conventions like these are useful when it comes to depicting rather abstract items. Concrete objects, such as a person, a smartphone, or a computer, can easily be drawn as a simplified graphic, but it is not as easy with more abstract

notions like “data.” The graphics representing exported files, collections of Quantified Self data, but also data transmission and synchronization build upon icons that have been adopted into many peoples’ “visual vocabulary.”

4.7 Example

Figure 3 shows an example of two comic strips that correspond to the provenance graph from Sect. 2 (Fig. 1). The example contains the consecutive strips for two user actions: downloading steps count data from a cloud service to the user’s smart phone, and visualizing the steps data in a line chart.

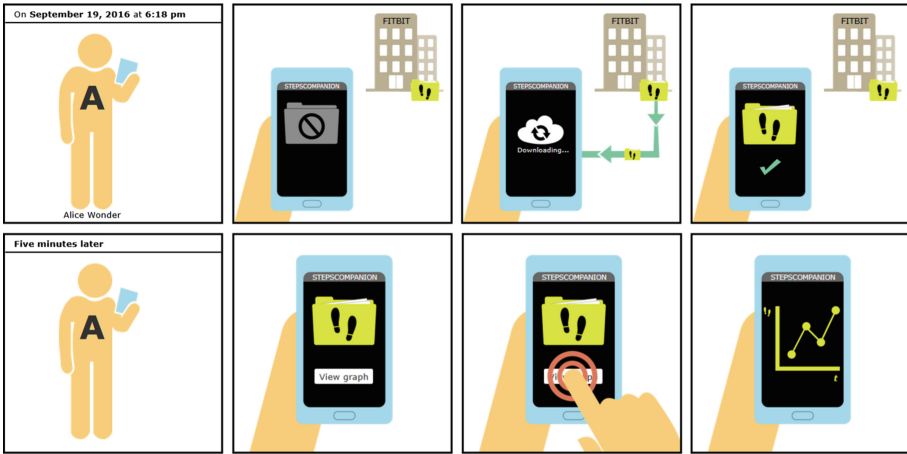


Fig. 3. Generated provenance comics strip for two consecutive user actions.

5 Implementation

For generating the comic strips, we developed a web application in JavaScript¹. This web application fetches the provenance directly from a provenance store (we support ProvStore [10]).

The script first looks for activities in the provenance document to determine what kinds of panels need to be displayed. If there is more than one activity, the correct order can be derived from the activities’ timestamps.

After that, the script reads the attributes of involved agents, entities, and relations to decide which graphics to include in these panels. For example, the attributes indicate whether to display a smartphone or a computer, a folder or a single document, a steps icon or a weight icon, etc.

¹ <https://github.com/DLR-SC/prov-comics>.

6 Qualitative User Study

We conducted a qualitative user study to evaluate the clarity and comprehensibility of the provenance comics. Ten test subjects were shown a number of test comics and asked to re-narrate the story as they understood it.

6.1 Study Design

Research Question. The general research question that was to be answered by the study is whether the comics are comprehensible to average end users: *Are the selected graphics and the visual language they form understandable?* and *Do users understand the history of their own data (i.e., when and how their data originated, what conversions and transformations it underwent, and who had access to or control over it in the course of time)?* The study was also to reveal misunderstandings that may arise from a lack of technical knowledge on the reader's part and help determine passages where the images are not explanatory enough and need to be improved or extended.

Test Comics. We selected five different scenarios as test comics to be included in the user study [21]. The first three test comics each depicted a combination of two activities (e.g., *Input* and *Visualize*). The fourth and fifth comics are a little longer, combining three to four activities.

Questions. We decided to have test readers speak freely about the comics and do a qualitative analysis afterwards. However, to make the test readers' answers accessible to statistics and comparison, we created a list for each of the comics, containing 10 to 23 findings that participants might discover and verbalize. It was thus possible to gain quantitative data by calculating the percentage of discovered findings.

Timing. Test readers were interviewed one at a time, and each reader was interviewed only once; there were no repeated interviews with the same persons. All participants were shown the same comics in the same order. The interviews took about thirty minutes each and were conducted over a period of several days.

Selection of Test Subjects. No special background was required of the test persons; on the contrary, it was desired that they have no previous knowledge about data provenance and no special expertise in the Quantified-Self domain. No limitations were set in terms of age, gender, or occupation.

Tasks, Rules and Instruments. For each participant, five different sheets with comic strips were printed out and handed to them on paper. To obtain comparable results, all test subjects were asked to fulfill the exact same tasks

for each of the five comics: first read the comic silently for themselves, and then re-narrate their interpretation of the story. To avoid influencing the process in any way, the examiner did not talk to participants at this stage. A smartphone running a dictaphone app was used to record the participants' re-narrations of the comics.

Debriefing. After all comics had been worked through, any difficult parts were revisited and analyzed in an informal conversation. Participants were encouraged to comment freely on the comics, giving their own opinion and suggestions for improvements.

6.2 User Study Results

The average percentage of findings that participants verbalized over all five comics was 77%. The value was remarkably high for some particular comics, the highest one being 87%. On a side note, women showed a better overall performance than men (84% for women vs. 73% for men).

There were certain difficult parts in some of the comics, which mostly stemmed from a lack of experience with Quantified Self applications or web services. However, even in these cases, the general essence of the story was largely interpreted correctly.

Participants had no difficulties recognizing and interpreting the different icons for concrete elements, like persons, smartphones, computers, and bracelets or smartwatches. But even more abstract notions (e.g., “transmitting data from one device to another,” “synchronizing data with a cloud”) were well-understood, since they relied on icons that are commonly used in software and web applications and were understood by most readers without any confusion.

In summary, all users were able to explain correctly the scenarios depicted in the comic strips. Some users suggested minor changes and improvements to the visual representation.

Current work includes user studies with a much broader set of people, especially with very limited knowledge about the technology behind wearable devices, smartphone apps, and services.

7 Related Work

Usually, visualization in Quantified Self focuses on the *data*, where all kinds of visualization techniques are used [13]. For example, time series visualizations or geographical visualization are very common².

For *provenance* visualization, most tools found in literature visualize provenance graphs using ordinary node-link diagrams, or tree representations similar to node-link diagrams. Provenance Map Orbiter [12], Provenance Browser [1],

² See visualization examples at the “Quantified Self” website: <http://quantifiedself.com/data-visualization/>.

and Provenance Explorer [9] are based upon node-link diagrams. Large provenance graphs are then simplified by combining or collapsing sub-nodes or hiding nodes that are not of interest right now. The user can interactively explore the graph by expanding or zooming into these nodes.

Other tools, such as VisTrails [3], use a tree representation similar to node-link diagrams. Visual clutter is reduced by hiding certain nodes, limiting the depth of the tree, or displaying only the nodes that are related to the selected node.

Probe-It! [7] and Cytoscape [5] basically display provenance as ordinary graphs. However, Probe-It! does not only show the *provenance* of data, but also the *actual* data that resulted from process executions. In Cytoscape, users can create their own visual styles, mapping certain data attributes onto visual properties like color, size, transparency, or font type.

One work that stands out due to its completely different and novel approach is InProv [4]. This tool displays provenance using an interactive radial-based tree layout. It also features time-based grouping of nodes, which allows users to examine a selection of nodes from a certain period of time only.

There are some more related works, even though they are not directly concerned with provenance visualization. A non-visual approach to communicating provenance is natural language generation by Richardson and Moreau [17]. In this case, PROV documents are translated into complete English sentences.

Quite similar to provenance comics are Graph Comics by Bach et al. [2], which are used to visualize and communicate changes in dynamic networks using comic strips.

8 Conclusions and Future Work

The goal of this work was to develop a self-explaining, easy-to-understand visualization of data provenance that can be understood by non-expert end users of Quantified-Self apps.

A detailed concept has been created that defines a consistent visual language. Graphics for PROV elements like different agents and entities were designed, and sequences of comic panels to represent different activities were determined. Symbols, icons, and panel sequences were specified in an exact and uniform manner to enable the automatic generation of comics.

As proof of concept, a prototypical website has been developed which is able to automatically generate comics from PROV documents compliant with the existing Quantified-Self data model. The documents are loaded from the ProvStore website.

A reading study involving ten test readers has shown that a non-expert audience is mostly able to understand the provenance of Quantified-Self data through provenance comics without any prior instruction or training. The overall percentage of 77% for findings verbalized by participants is deemed a good result, given that the checklists were very detailed and contained findings that some readers probably omitted, because they seemed too obvious and self-evident to them.

Future work will focus on graphical improvements. This includes suggested improvement measures that resulted from the reading study. A major step will be quantitative comics, which also show actual measured values. For example, diagrams on depicted devices could show real plots of health data, and single comic panels may include real geographical information. Another improvement could be the use of glyph-based depiction [18], where the body shape of depicted humans represent real values such as weight.

A useful improvement of the provenance comics would be to make them application-generic to some extent, (i.e., not restricted to the Quantified Self domain). We plan to explore whether provenance comics might be useful for other application domains, such as electronic laboratory notebooks or writing news stories in journalism.

References

1. Anand, M.K., Bowers, S., Altintas, I., Ludäscher, B.: Approaches for exploring and querying scientific workflow provenance graphs. In: McGuinness, D.L., Michaelis, J.R., Moreau, L. (eds.) IPAW 2010. LNCS, vol. 6378, pp. 17–26. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-17819-1_3](https://doi.org/10.1007/978-3-642-17819-1_3)
2. Bach, B., Kerracher, N., Hall, K.W., Carpendale, S., Kennedy, J., Henry Riche, N.: Telling stories about dynamic networks with graph comics. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 3670–3682. CHI 2016. ACM, New York (2016). <http://doi.acm.org/10.1145/2858036.2858387>
3. Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Vo, H.T.: VisTrails: enabling interactive multiple-view visualizations, pp. 135–142. IEEE (2005)
4. Borkin, M.A., Yeh, C.S., Boyd, M., Macko, P., Gajos, K.Z., Seltzer, M., Pfister, H.: Evaluation of filesystem provenance visualization tools. *IEEE Trans. Vis. Comput. Graph.* **19**(12), 2476–2485 (2013). <https://doi.org/10.1109/TVCG.2013.155>
5. Chen, P., Plale, B., Cheah, Y.W., Ghoshal, D., Jensen, S., Luo, Y.: Visualization of network data provenance. In: 2012 19th International Conference on High Performance Computing, pp. 1–9, December 2012. <https://doi.org/10.1109/HiPC.2012.6507517>
6. Choe, E.K., Lee, N.B., Lee, B., Pratt, W., Kientz, J.A.: Understanding quantified-selfers’ practices in collecting and exploring personal data. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 1143–1152. ACM (2014)
7. Rio, N., Silva, P.P.: Probe-It! visualization support for provenance. In: Bebis, G., et al. (eds.) ISVC 2007. LNCS, vol. 4842, pp. 732–741. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-76856-2_72](https://doi.org/10.1007/978-3-540-76856-2_72)
8. Hoy, M.B.: Personal activity trackers and the quantified self. *Med. Ref. Serv. Q.* **35**(1), 94–100 (2016)
9. Hunter, J., Cheung, K.: Provenance explorer—a graphical interface for constructing scientific publication packages from provenance trails. *Int. J. Digit. Libr.* **7**(1–2), 99–107 (2007). <https://doi.org/10.1007/s00799-007-0018-5>
10. Huynh, T.D., Moreau, L.: ProvStore: a public provenance repository. In: Ludäscher, B., Plale, B. (eds.) IPAW 2014. LNCS, vol. 8628, pp. 275–277. Springer, Cham (2015). doi:[10.1007/978-3-319-16462-5_32](https://doi.org/10.1007/978-3-319-16462-5_32)

11. Kunde, M., Bergmeyer, H., Schreiber, A.: Requirements for a provenance visualization component. In: Freire, J., Koop, D., Moreau, L. (eds.) IPAW 2008. LNCS, vol. 5272, pp. 241–252. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-89965-5_25](https://doi.org/10.1007/978-3-540-89965-5_25)
12. Macko, P., Seltzer, M.: Provenance map orbiter: interactive exploration of large provenance graphs. In: Proceedings of the 3rd Workshop on the Theory and Practice of Provenance (TaPP). USENIX Association (2011)
13. Marcengo, A., Rapp, A.: Visualization of human behavior data: the quantified self. In: Innovative approaches of data visualization and visual analytics, pp. 236–265. IGI Global (2014)
14. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81–97 (1956)
15. Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V., Varga, L.: The provenance of electronic data. *Commun. ACM* **51**(4), 52–58 (2008)
16. Moreau, L., Missier, P., Belhajjame, K., B’Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.: PROV-DM: The PROV data model 30 April 2013. <http://www.w3.org/TR/2013/REC-prov-dm-20130430/>
17. Richardson, D.P., Moreau, L.: Towards the domain agnostic generation of natural language explanations from provenance graphs for casual users. In: Mattoso, M., Glavic, B. (eds.) IPAW 2016. LNCS, vol. 9672, pp. 95–106. Springer, Cham (2016). doi:[10.1007/978-3-319-40593-3_8](https://doi.org/10.1007/978-3-319-40593-3_8)
18. Riehmann, P., Möbus, W., Froehlich, B.: Visualizing food ingredients for children by utilizing glyph-based characters. In: Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, AVI 2014. pp. 133–136. ACM, New York (2014). <http://doi.acm.org/10.1145/2598153.2598203>
19. Schreiber, A.: A provenance model for quantified self data. In: Antona, M., Stephanidis, C. (eds.) UAHCI 2016. LNCS, vol. 9737, pp. 382–393. Springer, Cham (2016). doi:[10.1007/978-3-319-40250-5_37](https://doi.org/10.1007/978-3-319-40250-5_37)
20. Schreiber, A., Seider, D.: Towards provenance capturing of quantified self data. In: Mattoso, M., Glavic, B. (eds.) IPAW 2016. LNCS, vol. 9672, pp. 218–221. Springer, Cham (2016). doi:[10.1007/978-3-319-40593-3_25](https://doi.org/10.1007/978-3-319-40593-3_25)
21. Struminski, R.: Visualization of the provenance of quantified self data. Master thesis, Hochschule Düsseldorf (2017), <http://elib.dlr.de/110996/>