Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

Article

Visualizing the Provenance of Personal Data using Comics

Andreas Schreiber [1,*] and Regina Struminski [1,2]

[1] Intelligent and Distributed Systems Department, Simulation and Software Technology, German Aerospace Center (DLR), Cologne, Germany; andreas.schreiber@dlr.de

[2] Faculty of Media, University of Applied Sciences Düsseldorf, Düsseldorf, Germany; regina.struminski@study.hs-duesseldorf.de

* Correspondence: andreas.schreiber@dlr.de; Tel.: +49-2203-601-2485

Abstract: Personal health data is acquired, processed, stored, and accessed using a variety of different devices, applications, and services. These are often complex and highly connected. Therefore, use or misuse of the data is hard to detect for people, if they are not capable 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 a smartphone application, 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 notice crucial points regarding their data such as, for example, privacy violations.

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 been accessed by organizations unwantedly. Especially, detecting and investigating privacy violations of personal data is a relevant issue for many people and companies. 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) [1,2] should not be available to other people or companies, as this might lead to commercial exploitation or even disadvantages for people, such as higher health insurance contributions.

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, applications, 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* [3] 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 [4]. Presenting provenance to non-experts is an ongoing research topic (*"Provenance for people"*). As a new visualization technique for provenance, we present *provenance comics* that we introduced and applied to trace personal data [5].

The remaining article is organized as follows:

- We shortly give an overview about provenance and our provenance model for Quantified Self data and self-tracking workflows [6,7] (Section 2).

© 2017 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

- We explain the general idea of *provenance comics* for provenance compliant with the PROV standard [8] (Section 3).
- We describe a visual mapping between the provenance of Quantified Self data and their graphical representations in comic strips (Section 4).
- We briefly describe our prototype for *automatically generating provenance comics* (Section 5).
- We give details and results of a qualitative user study (Section 6).

2. Provenance of Quantified Self Data

2.1. Provenance of Electronic Data

The definition of *provenance* is: *"Provenance is a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. In particular, the provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it. In an open and inclusive environment such as the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgments [8]"*.

With the previous definition, World Wide Web Consortium (W3C) started in 2011 and finalized in 2013 the generic provenance model PROV, which has specifications for a data model PROV-DM [8] and an ontology PROV-O [9], among others. PROV was inspired by various different approaches [10], that is adaptable to any domain. The general provenance model can be seen as a property graph with three different types of nodes: *Entities*, *Activities*, and *Agents*. Entities represent physical (e.g., sensors or medical devices), digital (e.g., data sets), conceptual (e.g., a workflow description), or any other kinds of objects. An activity is a process that uses or generates entities and that can be associated with an agent, meaning that the agent is responsible for the activity.

Provenance is being recorded during runtime of a process. To make Quantified Self workflows provenance-aware requires to gather information that is required by the provenance model (see [7] for some possible approaches). This information is stored in a provenance database or provenance store. For example, PROVSTORE [11] is publicly available provenance store. Large provenance graphs of long running real world workflows are stored in scalable databases more efficiently (e.g., using graph databases such as NEO4J [12]).

2.2. Provenance Visualization

For analyzing data provenance, visualization is a feasible method. Several solutions to visualize provenance exist, for example, publicly available web-based tools such as PROV-O-VIZ [13], desktop tools such as VISTRAILS [14], or numerous other graph visualization tools.

Provenance is usually represented as a directed acyclic graph (DAG). 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 [15], 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. 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 the provenance of their own 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

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.

2.3. *Quantified Self Provenance Model*

Based on a requirements study of Quantified Self workflows and analysis of documentation from breakout sessions at Quantified Self Conferences (such as the QSEU14 Breakout session on Mapping Data Access [16]), we developed a provenance model for Quantified Self workflows [6].

The possible activities in Quantified Self workflows are categorized into six abstract functionalities: *Input*, *Sensing*, *Export*, *Request*, *Aggregation*, and *Visualization*. We defined a provenance sub model for each of these abstract functionalities[1].

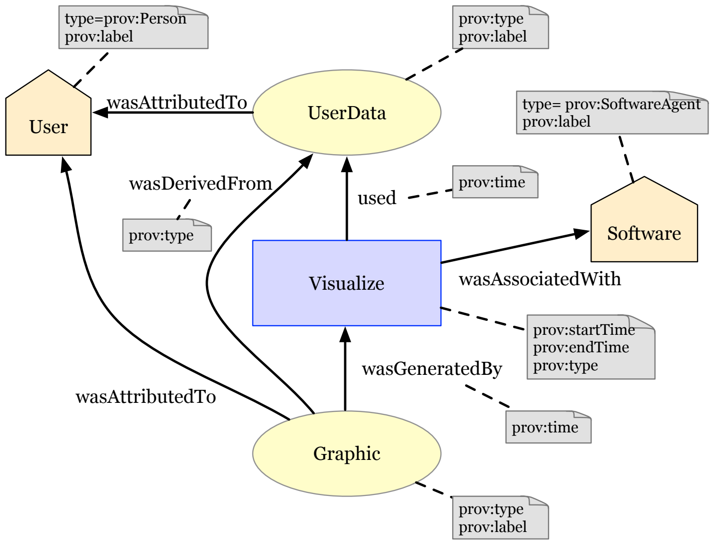


Figure 1. Provenance model for the Quantified Self activity *Visualize*.

As an example, Figure 1 show the provenance model for the *Visualize* activity where data (PROV entity "UserData") that belongs to a human (PROV agent "User") is visualized by method (PROV activity "Visualize") from a certain software (PROV agent "Software") which results in a graphic (PROV entity "Graphic"). The respective PROV elements can contain attributes, which specify meta information such as time of creation, names, or data types.

While the basic Quantified Self activities and the provenance of these activities are easy to understand conceptually, the representation of that provenance can be difficult to understand as explained in Section 2.2. For example, the two most common representations of provenance are a graphical representation as a graph (Figure 2) and a textual representation in PROV-N notation (Figure 3).

3. Provenance Comics

The basic idea of *provenance comics* is to present the provenance information of data processes in a visual representation, which people can understand without prior instruction or training. A general advantage of comics over conventional visualizations, like node-link diagrams, is their familiarity: Almost anyone has probably seen some comics in their life. No training is required to read them, and they can transport meaning with minimal textual annotation. They are easy to interpret and not as strenuous to read as, for example, a graph or a long paragraph of continuous text.

[1] https://github.com/onyame/quantified-self-prov

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

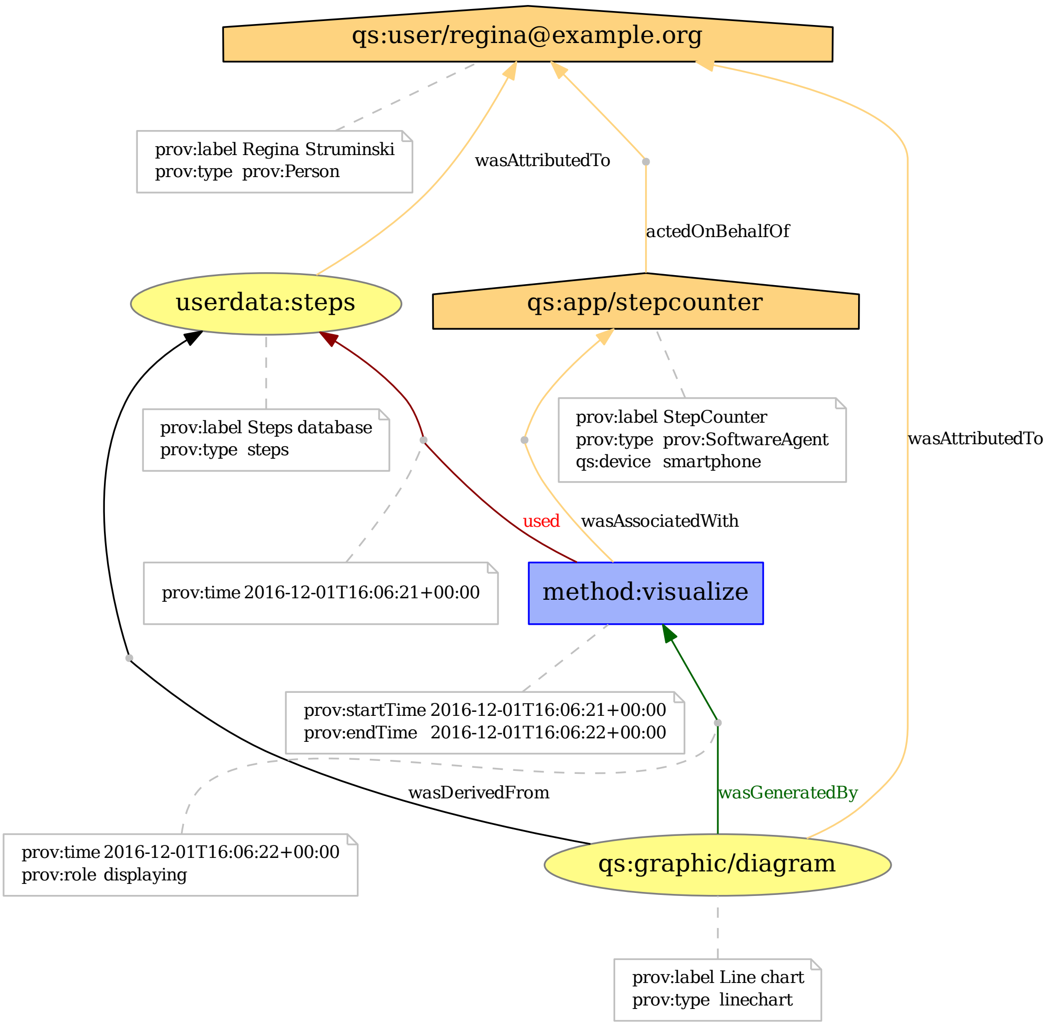


Figure 2. Graphical representation of the provenance for the QS activity *Visualize* as a directed acyclic graph (https://provenance.ecs.soton.ac.uk/store/documents/115521/).

Data provenance has a temporal aspect: origin, manipulation, transformation, and other activities happen sequentially over time. The directed, acyclic provenance graph guarantees that, while moving through its nodes, one always moves linearly forward or backward in time. It is therefore possible to derive a temporal sequence of happenings from the graph that can be narrated like a story.

We generate a comic strip for each basic activity in the provenance data (e.g., for the activity "Visualize" in Figures 1 or 2). Each strip consists of a varying number of panels, which are small drawings that provide further details about the activity. The comic strip for the earliest activity in the provenance document is at the top, while the strip for the newest, most recent activity is at the bottom. The complete set of comic strips shows the "story" of the data. Of course, when there are many activities, the collection of comic strips could become quite large. In this case, one could choose a subset of the provenance, containing only those activities that are relevant in real use cases.

Some questions that the provenance comics should answer and explain are *When was data generated or changed?*, *Where was the user?*, or *Where was the user's data stored?* At this time, the comics do not contain the actual data. They only represent information contained in the provenance of the user's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1


```
document
  prefix userdata <http://software.dlr.de/qs/userdata/>
  prefix qs <http://software.dlr.de/qs/>
  prefix graphic <http://software.dlr.de/qs/graphic/>
  prefix app <http://software.dlr.de/qs/app/>
  prefix user <http://software.dlr.de/qs/user/>
  prefix device <http://software.dlr.de/qs/device/>
  prefix method <http://www.java.com>

  wasGeneratedBy(qs:graphic/diagram, method:visualize, 2016-12-01T16:06:22+00:00, [prov:role="displaying"])
  activity(method:visualize, 2016-12-01T16:06:21+00:00, 2016-12-01T16:06:22+00:00)
  entity(qs:graphic/diagram, [prov:type="linechart", prov:label="Line chart"])
  entity(userdata:steps, [prov:type="steps", prov:label="Steps database"])
  agent(qs:user/regina@example.org, [prov:type="prov:Person", prov:label="Regina Struminski"])
  agent(qs:app/stepcounter, [prov:type="prov:SoftwareAgent", qs:device="smartphone", prov:label="StepCounter"])
  wasAttributedTo(qs:graphic/diagram, qs:user/regina@example.org)
  wasAttributedTo(userdata:steps, qs:user/regina@example.org)
  actedOnBehalfOf(qs:app/stepcounter, qs:user/regina@example.org, -)
  used(method:visualize, userdata:steps, 2016-12-01T16:06:21+00:00)
  wasDerivedFrom(qs:graphic/diagram, userdata:steps, -, -, -)
  wasAssociatedWith(method:visualize, qs:app/stepcounter, -)
endDocument
```

Figure 3. Textual representation of the provenance for the Quantified Self activity *Visualize* in PROV-N (https://provenance.ecs.soton.ac.uk/store/documents/115521/).

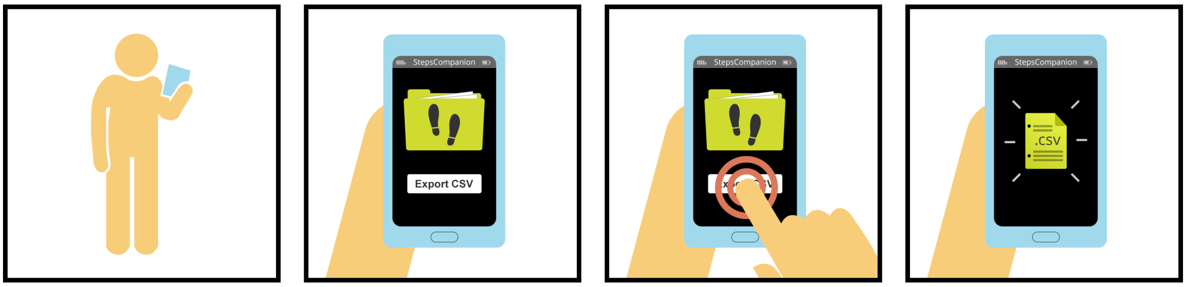


Figure 4. Generated provenance comic strip depicting the export of step data into a file in CSV format.

data. This might be extended in the future by using (parts of the) data for representing the real measurements, geographical coordinates, etc.

4. Visual Mapping

To generate the provenance comics, we defined a consistent visual language [17]. 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 graphical features: *shapes*, *colors*, and *icons or texts*.

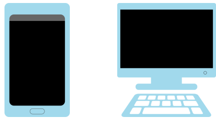
4.1. Shapes

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. Flat design became popular in mobile UI and icon design [18] and despite of the fact that study results shows a higher cognitive load for searching flat icons [19], we stick to flat design in the first appraoch since we have use cases in mind, where the comics are incoporated into mobile applications.

Table 1 gives an overview of the shapes we selected to reflect the different types of elements in the Quantified Self PROV model [6]. Activities are not directly listed here. Unlike agents or entities, activities are actions that take place over time, as described in Section 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

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 agent's "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 (Figure 5).

- `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).

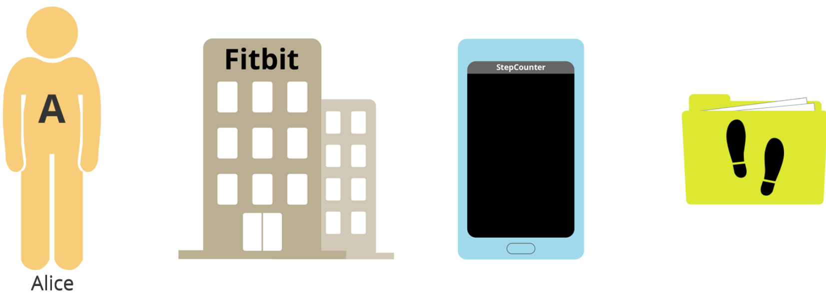


Figure 5. Agents and entities using three distinctive features (shape, color, icons/text).

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. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

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

4.3.1. Colors for objects of the same type

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.

The first alternative was determined by reducing the main color's lightness (in the HSL color space) by 60%, the second alternative by reducing the lightness by 30%–45%. Figures 6, 7, and 8 examplarily simulate the effect of different types of color blindness on agent and entity colors[2].

(a) Normal **(b)** Protanopia **(c)** Deuteranopia **(d)** Tritanopia **(e)** Achromatopsia

Figure 6. `Person` agent color shades and how they are seen by colorblind people

(a) Normal **(b)** Protanopia **(c)** Deuteranopia **(d)** Tritanopia **(e)** Achromatopsia

Figure 7. `SoftwareAgent` color shades and how they are seen by colorblind people

(a) Normal **(b)** Protanopia **(c)** Deuteranopia **(d)** Tritanopia **(e)** Achromatopsia

Figure 8. Entity color shades and how they are seen by colorblind people

In a previous approach, colors had been rotated by 180°, 90°, and 270° to obtain well-matched second, third and even fourth colors. However, two problems arose: First of all, the whole comic would generally have become very colorful, which would possibly have led to confusion. Depending on the situation, there might, for example, have been a blue person that owns a blue phone and a pink

[2] Simulations generated by http://www.color-blindness.com/coblis-color-blindness-simulator/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

entity, while at the same time a pink person is present owning a blue entity. Some similar items would have had very dissimilar colors, while some dissimilar items would have had very similar colors. Apart from causing a certain visual inconsistency, this might also have suggested to the reader that there were some deeper meaning to the colors, other than discriminability. For example, the reader might have thought that similar colors indicate a grouping of some kind (e.g. that a pink entity belongs to a pink person).

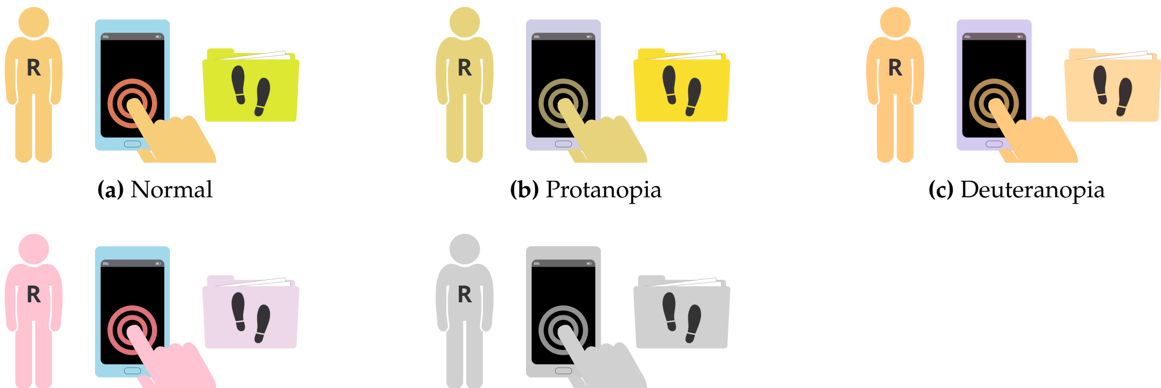
4.3.2. Colors for objects of different types

The distinctiveness between the colors of different object types is not as important as that between colors of the same types of objects. That is to say: Color is more important for distinguishing two items that have the same shape than it is for two items with different shapes. Thus the selection and discriminability of colors need not be handled as strictly for different types of actors.

Figure 9 shows that especially the default colors of `Person` agents and entities are not well distinguishable by readers suffering from color vision deficiencies. However, since shape and icon or text will be different, the weak color difference is neglectable. Figure 10 shows that items are still well distinguishable due to their shapes and icons.

(a) Normal **(b)** Protanopia **(c)** Deuteranopia **(d)** Tritanopia **(e)** Achromatopsia

Figure 9. Default colors for `Persons`, `SoftwareAgents`, entities, and a "button press" effect and how they are seen by colorblind people



(a) Normal **(b)** Protanopia **(c)** Deuteranopia **(d)** Tritanopia **(e)** Achromatopsia

Figure 10. Default colors and shapes for different objects and how they are seen by colorblind people.

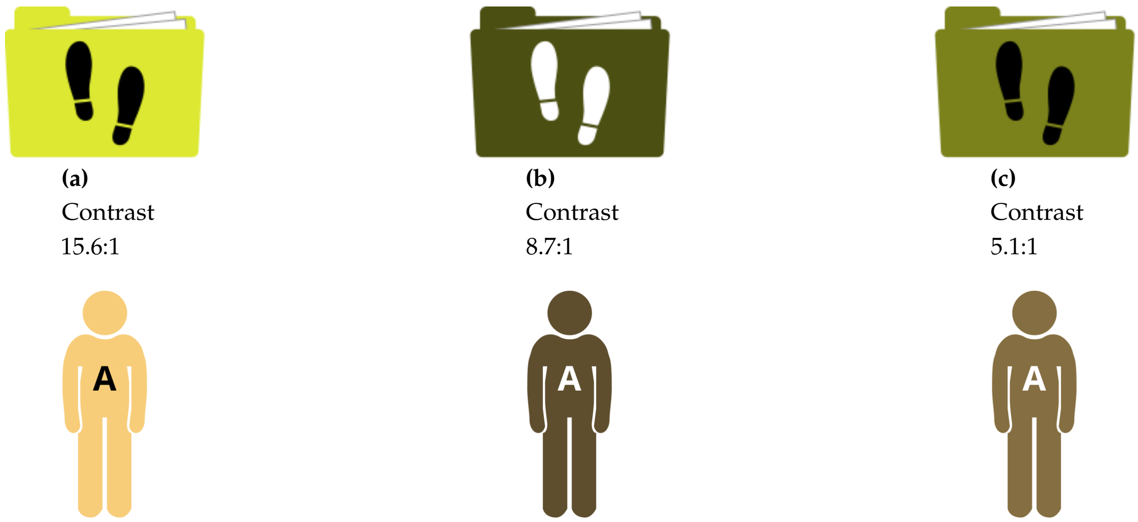
4.3.3. Text and icon colors

In a number of cases, agents and entities will be labeled with texts, letters or icons. To keep those recognizable on different background colors, a simple rule of thumb has been established using the colors' equivalents in the Lab color space:

- If a color's L (lightness) value is between **0 and 49**, the text or icon color is **white**.
- If a color's L value is between **50 and 100**, the text or icon color is **black**.

By choosing the font color this way, a contrast ratio of at least 3:1 (often a lot higher) is achieved, which is "the minimum level recommended by ISO-9241-3 and ANSI-HFES-100-1988 for standard text and vision" [20]. The WCAG's SC 1.4.3 (MINIMUM CONTRAST) requires a ratio of 4.5:1 for standard text, and 3:1 for "large-scale text and images of large-scale text", with "large-scale text" having a size of at least 18 point, or 14 point and bold style. The even stricter SC 1.4.6 (ENHANCED CONTRAST) requires a ratio of 4.5:1 for large-scale text and 7:1 for standard text [20].

The majority of icons and letters used in the PROV COMICS qualify as large-scale text. By choosing the font or icon color according to the simple "black or white" rule proposed here, it is guaranteed that a contrast ratio of at least 3:1 is always achieved. In fact, when combined with the previously defined agent and entity colors, this rule yields a contrast ratio of at least 4.5:1 for all graphics containing text or icons. Thus, they even fulfill the stricter SC 1.4.6 (ENHANCED CONTRAST) for large-scale text. Figure 11 shows some example graphics with high-contrast icons or letters[3].



(a) Contrast 15.6:1

(b) Contrast 8.7:1

(c) Contrast 5.1:1

(d) Contrast 14:1

(e) Contrast 8.1:1

(f) Contrast 4.9:1

Figure 11. Examples of entities and agents with icons passing the WCAG SC 1.4.6 (ENHANCED CONTRAST).

4.4. Panels and layout

All panels are perfect squares. Horizontally, they are separated from each other by a whitespace of 10% of the panel size, while the vertical distance between rows of panel is 20% of the panel size. For example, 600x600 pixel panels have 60 pixels of white space between them horizontally, and 120 pixels of white space vertically. By arranging them this way, panels are grouped into rows, helping the reader determine the correct reading direction. This is explained by the gestalt law of proximity: Objects that are close to each other are perceived as a group [21].

However, no requirements are made as to how many panels each row should contain. Due to the fact that the comics are to be viewed on different devices the layout needs to be scalable. While a row may consist of four or five panels on a desktop or tablet computer, there might only be enough space for one panel per row on a smartphone.

[3] Contrast ratios calculated by http://leaverou.github.io/contrast-ratio/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

The panels have black borders, the width of which should amount to 1% of the panel size. For example, a 600x600 pixel panel should use a 6 pixel border. In case a caption or introductory text is added to the top of a panel, it is separated from the rest of the panel by a bottom border with the same properties. Borders group the different graphics inside a panel together, so they are perceived as one large image. This is an application of the law of closure, which states that objects in a framed area are perceived as one unit [21].

4.5. 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 Figure 14 contains examples of these textual annotations, button labels, and captions.

4.6. 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.

4.7. Recurring image structures

Activities will not be represented by a single graphic, but by a sequence of three to five comic panels. Similar activities should be illustrated by similar sets of panels, making use of recurring image compositions. For example, the activities of the data sub-models *Export*, *Aggregate*, and *Visualize* are comparable in that they take one kind of data and create a different kind of data from it. They can thus be visualized in a very similar manner (see Figures 4, 12, and 14).

Using recurring image structures whenever possible adds to the comics' consistency, comprehensibility and learnability: Once readers have understood the *Export* panels, for example, they will easily be able to understand *Aggregate* and *Visualize* panels, too.

4.8. 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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

Figure 12. Comic depicting the aggregation of step count and heart rate data into a new set of data.

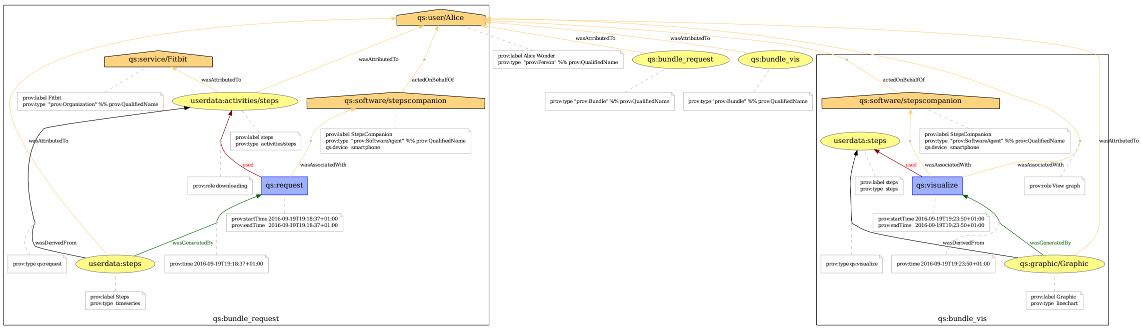


Figure 13. Provenance graph of two user actions (https://provenance.ecs.soton.ac.uk/store/documents/115642/)

collections of Quantified Self data, but also data transmission and synchronization build upon icons that have been adopted into many peoples' "visual vocabulary."

4.9. Example

Figure 14 shows an example of two comic strips that correspond to the provenance graph in Figure 13. The example contains the consecutive strips for two user actions: *downloading steps count data from a cloud service to the user's smart phone* (PROV activity "request"), and *visualizing the steps data in a line chart* (PROV activity "visualize").

5. Implementation

For generating the comic strips, we developed the web application PROV COMICS in JavaScript [22] (Figure 15). This web application fetches provenance documents directly from a provenance store. The current prototype supports the publicly available provenance store PROVSTORE [11] using the PROVSTORE JQUERY API to retrieve public documents from the PROVSTORE for a certain user.

Within the provenance document, the script first looks for activities to determine what kinds of panels need to be displayed. If there is more than one activity, the correct order is derived from the activities' timestamps. As mentioned earlier in Section 4.7, activities will not be represented by a single graphic, but by a sequence of three to five comic panels. Similar activities are illustrated by similar sets of panels.

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.

For generating the comics, the `ProvComics.js` script defines three JavaScript prototypes ("classes"):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

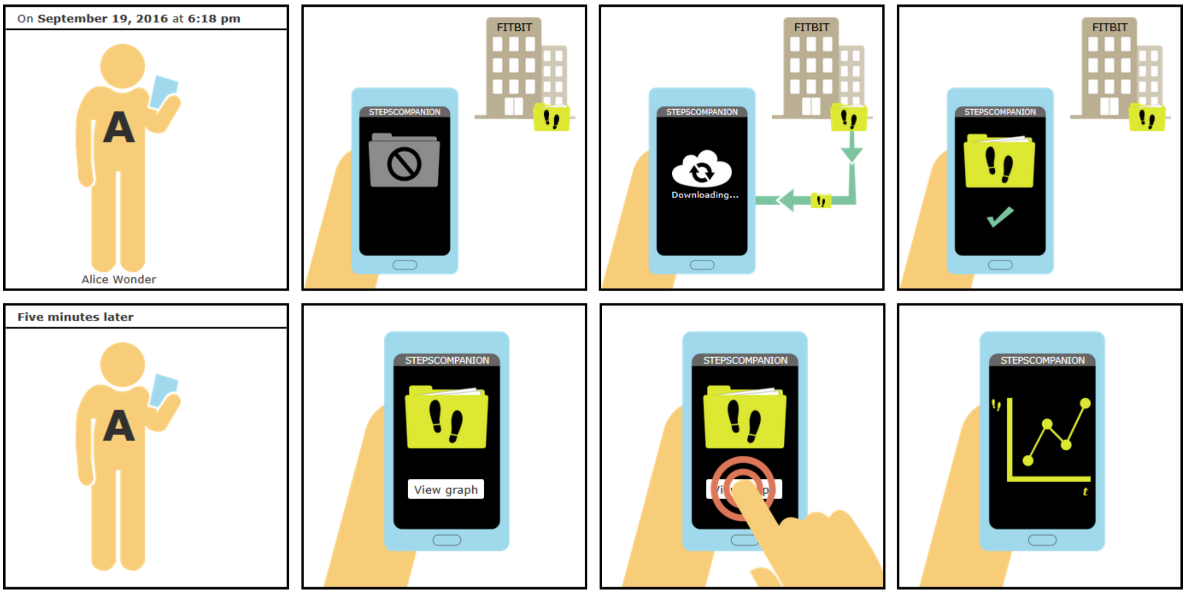


Figure 14. Generated provenance comics strip for two consecutive user actions.

ProvComic serves as a frame to contain all comic panels. It is also the general starting point for creating a *provenance comic* inside a given HTML element. For example, if there is a `<div id=''comic''>` tag in the HTML, a new *provenance comic* may be started within the `div` element by declaring `var comic = new ProvComic(''#comic'')`.

Panel represents a single comic panel and has all necessary abilities to create any of the panels described in the concept. For example, it provides functions to add captions, `Persons`, `SoftwareAgents`, `Organizations`, different types of entities, etc.

PanelGroup represents a predefined sequence of panels. They make it easier to insert recurring panel sequences. For example, it provides a function to add all panels depicting a download *Request* at once.

6. Qualitative User Study

We conducted a 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

We decided that a *qualitative study* was the better choice—in contrast to a quantitative study—in order to find out whether or not the PROV COMICS are comprehensible. Different people may understand the comics in different ways, or have different problems when reading them. These can hardly be compared or measured in numbers, and creating a standardized questionnaire with closed questions would have been very difficult. Moreover, it would probably have led to further problems; for example, if asking about certain features of the comics using single or multiple choice questions, the question itself as well as the available answers might have provided hints and suggested something to the participants that they actually did not understand by themselves when they first read the comics.

Due to these considerations, we let test readers speak freely about the comics and performed 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

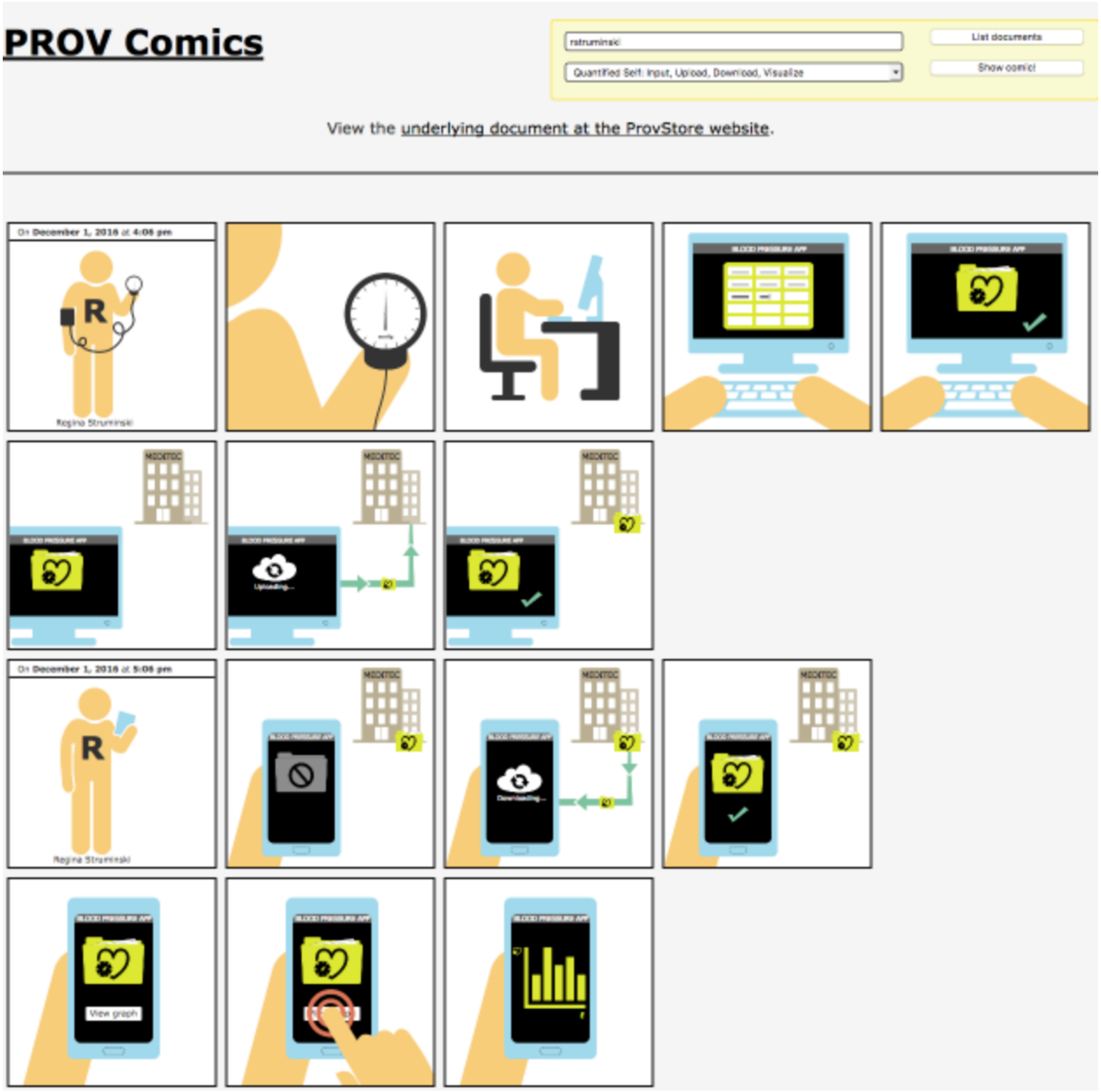


Figure 15. Screenshot of the PROV COMICS web application (http://provcomics.de).

306 6.1.1. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

6.1.2. Test comics

We selected five different scenarios as test comics to be included in the user study [17]. 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.

6.1.3. 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.

6.1.4. 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.

6.1.5. 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. Table 3 gives an overview about the selected participants.

Table 3. Study participants.

Test subject	Gender	Age	Technical expertise (0 = none, 3 = expert)	# QS applications used	Profession
on	f	28	2	4	Cook's mate / waitress
er	f	63	1	4	Senior executive in aged care
mm	m	25	2	4	Student (computer science)
42	m	25	3	4	Student (computer science)
ab	m	26	3	4	Student (computer science)
nn	f	43	2	3	Primary school teacher
al	m	49	1	1	Commercial clerk
ud	f	40	2	1	Optometrist
te	m	49	2	0	Soldier
xe	m	29	2	1	Computer scientist / programmer
Average	n/a	37.7	2	2.6	n/a
Median	n/a	34.5	2	3.5	n/a

6.1.6. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

6.1.7. 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 %. Women showed a better overall performance than men (84 % for women vs. 73 % for men). Figure 16 shows results for all test comics. However, the number of test subjects in this small study is too low to draw any general conclusions from that.

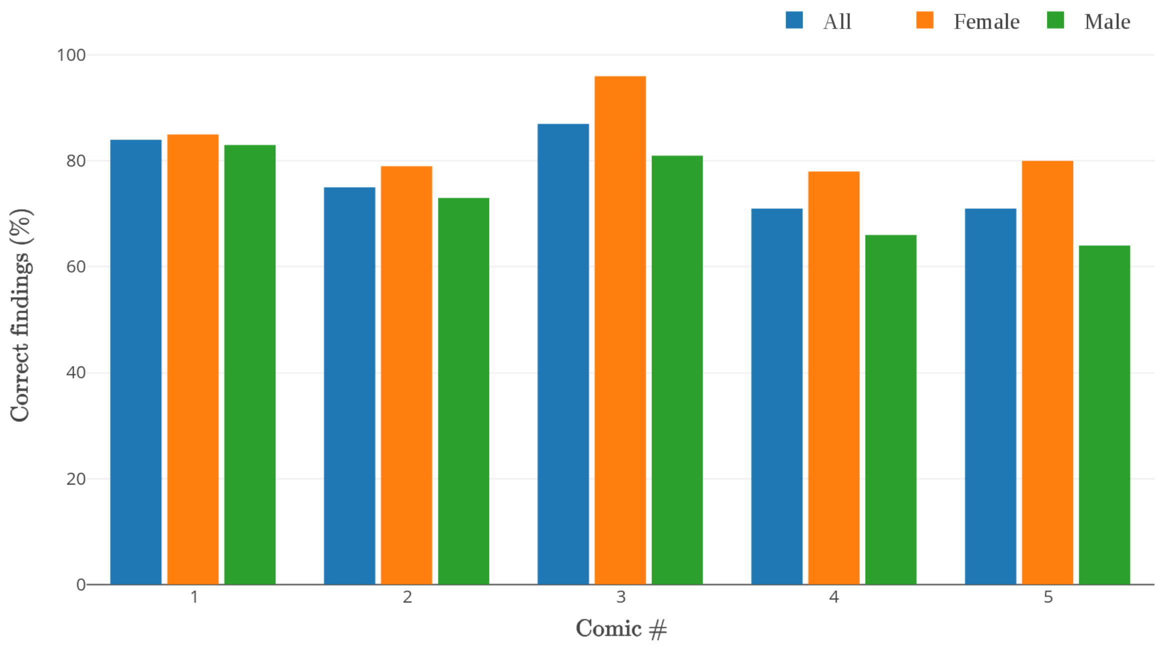


Figure 16. Evaluation of results: Percentage of correct findings for all participants as well as for women and men only (https://plot.ly/~onyame/50/).

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.

Readers also had no problem identifying themselves with the comic figure (human silhouette). Almost every re-narration was told from a first-person point of view, using sentences like "*I* was walking", "*I* was wearing a bracelet", "*I* clicked the button", etc.

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 applications, and services.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

7. Related Work

Usually, visualization in Quantified Self focuses on the *data*, where all kinds of visualization techniques are used [23]. For example, time series visualizations or geographical visualization are very common[4].

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 [24], PROVENANCE BROWSER [25], and PROVENANCE EXPLORER [26] 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 [14], 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! [27] and CYTOSCAPE [28] 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 [29]. 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 [30]. In this case, PROV documents are translated into complete English sentences.

Quite similar to provenance comics are *Graph Comics* by Bach et Al. [31], 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 applications.

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

[4] See visualization examples at the"Quantified Self" website: http://quantifiedself.com/data-visualization/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

health data, and single comic panels may include real geographical information. Another improvement could be the use of glyph-based depiction [32], where the body shape of depicted humans represent real values such as weight. A more technical improvement will be the consequent use of *provenance templates* [33,34], which will help to standardize the recorded provenance with templates provided to tool developers and which then helps tools for generating comic strips based on these standardized provenance.

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, writing news stories in journalism, or security breaches in Internet-of-Things environments. For example, using provenance comics seem to be a feasible approach to communicate hacking attempts in smart home systems, if provenance of such attacks is available (such as by the recent works of Wang et Al. [35]).

References

1. Choe, E.K.; Lee, N.B.; Lee, B.; Pratt, W.; Kientz, J.A. Understanding quantified-selfers' practices in collecting and exploring personal data. Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM, 2014, pp. 1143–1152.
2. Hoy, M.B. Personal Activity Trackers and the Quantified Self. *Med Ref Serv Q* **2016**, *35*, 94–100.
3. 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. *Communications of the Acm* **2008**, *51*, 52–58.
4. Kunde, M.; Bergmeyer, H.; Schreiber, A. Requirements for a Provenance Visualization Component. Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers; Freire, J.; Koop, D.; Moreau, L., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp. 241–252.
5. Schreiber, A.; Struminski, R., Tracing Personal Data Using Comics. In *Universal Access in Human–Computer Interaction. Design and Development Approaches and Methods: 11th International Conference, UAHCI 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I*; Antona, M.; Stephanidis, C., Eds.; Springer International Publishing: Cham, 2017; pp. 444–455.
6. Schreiber, A. A Provenance Model for Quantified Self Data. Universal Access in Human-Computer Interaction. Methods, Techniques, and Best Practices: 10th International Conference, UAHCI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part I; Antona, M.; Stephanidis, C., Eds.; Springer International Publishing: Cham, 2016; pp. 382–393.
7. Schreiber, A.; Seider, D. Towards Provenance Capturing of Quantified Self Data. Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings; Mattoso, M.; Glavic, B., Eds.; Springer International Publishing: Cham, 2016; pp. 218–221.
8. 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, 2013.
9. Lebo, T.; Sahoo, S.; McGuinness, D.; Belhajjame, K.; Cheney, J.; Corsar, D.; Garijo, D.; Soiland-Reyes, S.; Zednik, S.; Zhao, J. PROV-O: The PROV Ontology, 2013.
10. Moreau, L.; Groth, P.; Cheney, J.; Lebo, T.; Miles, S. The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web* **2015**, *35, Part 4*, 235–257.
11. Huynh, T.D.; Moreau, L. ProvStore: A Public Provenance Repository. Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers; Ludäscher, B.; Plale, B., Eds.; Springer International Publishing: Cham, 2015; pp. 275–277.
12. Schreiber, A.; Ney, M.; Wendel, H. The Provenance Store prOOst for the Open Provenance Model. Provenance and Annotation of Data and Processes: 4th International Provenance and Annotation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

Workshop, IPAW 2012, Santa Barbara, CA, USA, June 19-21, 2012, Revised Selected Papers; Groth, P.; Frew, J., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp. 240–242.

13. Hoekstra, R.; Groth, P. PROV-O-Viz – Understanding the Role of Activities in Provenance. Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers; Ludäscher, B.; Plale, B., Eds.; Springer International Publishing: Cham, 2015; pp. 215–220.
14. Bavoil, L.; Callahan, S.P.; Crossno, P.J.; Freire, J.; Vo, H.T., VisTrails: enabling interactive multiple-view visualizations. In *Visualization, 2005. VIS 05*; IEEE, 2005; pp. 135–142.
15. Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **1956**, *63*, 81–97.
16. QSEU14. Breakout: Mapping Data Access, 2014.
17. Struminski, R. Visualization of the Provenance of Quantified Self Data. Master thesis, Hochschule Düsseldorf, 2017.
18. Stickel, C.; Pohl, H.M.; Milde, J.T., Cutting Edge Design or a Beginner's Mistake? – A Semiotic Inspection of iOS7 Icon Design Changes. In *Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II*; Marcus, A., Ed.; Springer International Publishing: Cham, 2014; pp. 358–369.
19. Burmistrov, I.; Zlokazova, T.; Izmalkova, A.; Leonova, A., Flat Design vs Traditional Design: Comparative Experimental Study. In *Human-Computer Interaction – INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II*; Abascal, J.; Barbosa, S.; Fetter, M.; Gross, T.; Palanque, P.; Winckler, M., Eds.; Springer International Publishing: Cham, 2015; pp. 106–114.
20. Cooper, M.; Kirkpatrick, A.; Connor, J.O. Understanding WCAG 2.0: Contrast (Minimum), 2016.
21. Böhringer, J.; Bühler, P.; Schlaich, P. *Kompendium der Mediengestaltung für Digital- und Printmedien*; Number Bd. 1 in Kompendium der Mediengestaltung für Digital- und Printmedien, Springer, 2008.
22. Struminski, R.; Bieliauskas, S.; Schreiber, A. DLR-SC/prov-comics: QS PROV Comics Prototype - Big fixes [Data set]. Zenodo, 2017.
23. Marcengo, A.; Rapp, A., Innovative Approaches of Data Visualization and Visual Analytics; IGI Global, 2014; chapter Visualization of Human Behavior Data: The Quantified Self, pp. 236–265.
24. Macko, P.; Seltzer, M. Provenance map orbiter: Interactive exploration of large provenance graphs. Proceedings of the 3rd Workshop on the Theory and Practice of Provenance (TaPP), USENIX Association, 2011.
25. Anand, M.K.; Bowers, S.; Altintas, I.; Ludäscher, B. Approaches for Exploring and Querying Scientific Workflow Provenance Graphs. Provenance and Annotation of Data and Processes: Third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010. Revised Selected Papers; McGuinness, D.L.; Michaelis, J.R.; Moreau, L., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2010; pp. 17–26.
26. Hunter, J.; Cheung, K. Provenance Explorer-a graphical interface for constructing scientific publication packages from provenance trails. *International Journal on Digital Libraries* **2007**, *7*, 99–107.
27. Del Rio, N.; da Silva, P.P., Probe-It! Visualization Support for Provenance. In *Advances in Visual Computing: Third International Symposium, ISVC 2007, Lake Tahoe, NV, USA, November 26-28, 2007, Proceedings, Part II*; Bebis, G.; Boyle, R.; Parvin, B.; Koracin, D.; Paragios, N.; Tanveer, S.M.; Ju, T.; Liu, Z.; Coquillart, S.; Cruz-Neira, C.; Müller, T.; Malzbender, T., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007; pp. 732–741.
28. Chen, P.; Plale, B.; Cheah, Y.W.; Ghoshal, D.; Jensen, S.; Luo, Y. Visualization of network data provenance. 2012 19th International Conference on High Performance Computing, 2012, pp. 1–9.
29. Borkin, M.A.; Yeh, C.S.; Boyd, M.; Macko, P.; Gajos, K.Z.; Seltzer, M.; Pfister, H. Evaluation of Filesystem Provenance Visualization Tools. *IEEE Transactions on Visualization and Computer Graphics* **2013**, *19*, 2476–2485.
30. Richardson, D.P.; Moreau, L. Towards the Domain Agnostic Generation of Natural Language Explanations from Provenance Graphs for Casual Users. Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings; Mattoso, M.; Glavic, B., Eds.; Springer International Publishing: Cham, 2016; pp. 95–106.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2017 doi:10.20944/preprints201712.0153.v1

31. Bach, B.; Kerracher, N.; Hall, K.W.; Carpendale, S.; Kennedy, J.; Henry Riche, N. Telling Stories About Dynamic Networks with Graph Comics. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; ACM: New York, NY, USA, 2016; CHI '16, pp. 3670–3682.
32. Riehmann, P.; Möbus, W.; Froehlich, B. Visualizing Food Ingredients for Children by Utilizing Glyph-based Characters. Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces; ACM: New York, NY, USA, 2014; AVI '14, pp. 133–136.
33. Curcin, V.; Fairweather, E.; Danger, R.; Corrigan, D. Templates as a method for implementing data provenance in decision support systems. *Journal of Biomedical Informatics* **2017**, *65*, 1–21.
34. Moreau, L.; Batlajery, B.; Huynh, T.D.; Michaelides, D.; Packer, H. A Templating System to Generate Provenance. *IEEE Transactions on Software Engineering* **2017**, *PP*, 1–1.
35. Wang, Q.; Hassan, W.U.; Bates, A.; Gunter, C. Provenance Tracing in the Internet of Things; USENIX Association: Seattle, WA, 2017.