Supporting the Generation of Data Narratives

Faten El Outa[1], Matteo Francia[3], Patrick Marcel[1], Veronika Peralta[1], and Panos Vassiliadis[2]

[1] University of Tours, Blois, France firstname.lastname@univ-tours.fr
[2] University of Ioannina, Ioannina, Greece pvassil@cs.uoi.gr
[3] University of Bologna, Cesena, Italy m.francia@unibo.it

Abstract. Data narration has received increasing interest in several communities while lacking models and tools for handling, building and structuring data narratives. We present a simple prototype for supporting data narrative, based on a conceptual model defined in [4]. It guides a data narrator from scratch: fetch and explore data, abstract important messages based on an intentional goal, structure the contents of the data story, and render it in a visual manner. This prototype is implemented in Java as a web application using Spring, d3.js, JFreeChart and Apache PDFBox.

Keywords: Data Narrative · Visual Narrative· Data Storytelling · Data Analysis

1 Introduction

Data narration has received increasing interest in several communities (e.g. journalism, business, e-government). It is defined as the activity of producing narratives supported by facts extracted from data analysis, using interactive visualizations [1].

This paper describes a *prototype* implementing a novel conceptual model of data narrative [4], guiding an author (data narrator) in structuring a data narrative while exploring a database from scratch: fetch and explore data, abstract important messages based on an intentional goal, structure the contents of the data story, and render it in a visual manner. This *prototype* implements the four layers defined in the model (see Figure 1) and based on Chatman's terminology [2], who defines narrative as a couple of story (content of the narrative) and discourse (expression of it). In the data story, the factual layer handles the *exploration* of facts (i.e., the underlying data), via a set of *collectors* that allow for manipulating facts with varied tools in an objective way and the *intentional* layer models the subjective substance of the story, identifying the *messages*, *characters* and *measures* the author intends to communicate and tracing how they are obtained through *analytical questions*, according to an *analysis goal*. In the discourse, the *structural* layer concerns the structure of the data narrative, organizing its *plot* in terms of *acts* and *episodes* and the *presentational* layer serves the rendering of the data narrative, i.e., a *visual narrative*, that is

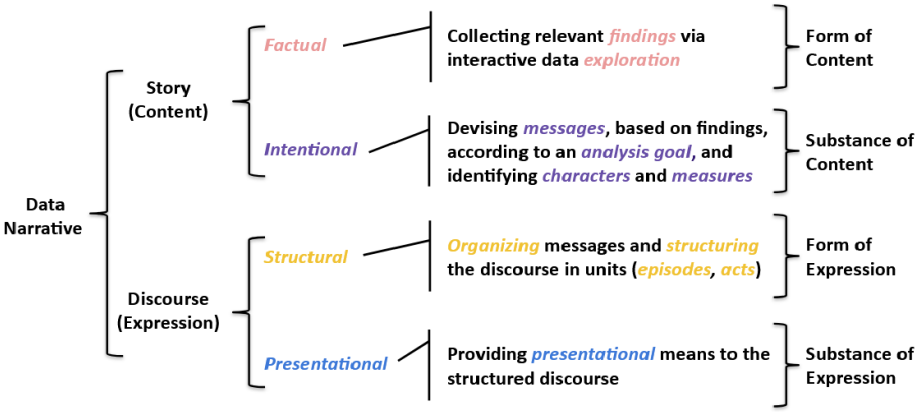

Fig. 1. Data narrative model (adapted from Chatman [2])

communicated to the reader through visual artifacts (*dashboards* and *dashboard components*). While a specific methodology describing how to use the proposed model is yet to come, the prototype supports the generation of a data narrative by organizing the different steps induced by the four layers.

2 Implementing the data narrative model

We propose an interactive interface that gives authors a simple, intuitive, and powerful way to generate data narratives when exploring data with absolutely no code required. Each entity of the model presented in [4] is implemented as an interface. Concrete classes allow to design simple visual narrative based on (i) a factual layer that implements collectors over a relational database, and (ii) a presentational layer that renders stories as a PDF document. The user interface essentially consists of text areas where the author can declare goal, analytical questions, messages, characters, measures, episodes and acts. The application logic controls that these inputs are compliant with the model. Precisely, the author starts a new narrative with a goal, and then expresses some analytical questions. For each question, the author can try the different collectors, and inspect their answers. If the findings brought by a collector are found worth adding to the narrative, they are turned into messages, for which the author must declare at least one character and one measure. Then, an episode can be created only if it can be attached to an act and a message that must have been declared beforehand. The current prototype implements two types of collectors over a relational database. The first collector type allows to send plain SQL queries over the database and obtain the answer as a set of records. The second type implements the Describe operator presented in [3], which allows to enter intentional queries [5], augment the result with automatic model extraction (e.g., clustering), and render the result to appropriate charts. As to the rendering of

the narrative, the current prototype implements two types of visual narrative downloadable as a PDF file. Both use dashboard components that write the texts of acts and episodes, and the image of a chart brought by the Describe collector or produced from the result of a SQL query. Finally, all SQL collectors can be documented and returned in a SQL notebook, using the Franchise SQL notebook application[4]. While for now it can only be used to craft simple narratives, this prototype can be the basis for the creation of more sophisticated ones, once more collectors, dashboard components, and dashboards are implemented.

3 Data narrative in action

This section presents a functional description for supporting data narrative generation with a web application containing a set of text fields titled to understand the story generation path, and a *log and console* to keep track of narrative details. To start crafting a story, the user (or data narrator) clicks *start new story* button, defines the story goal by filling up *analysis goal* and clicking on *define analysis goal* to log the goal into the story's logs. They then pose an analytical question and click on *add new analytical question* to log the question. To answer the analytical question, the user tries different collectors to fetch the data stored in a database by choosing either *create SQL query* or *create describe collector*. The user writes the collector's query and gets the result as a set of tuples and simple charts by clicking on *evaluating this query*. They look over the facts retrieved by the collector, choose the important findings to turn into a message by clicking *validating collector finding*. Important findings are copied into the *message* text area to allow the data narrator to edit it, before logging it. For each message, the user is responsible to fill up its *measure(s)* and *character(s)*. These measures and characters can be recalled later while writing new episodes. When a message is created, the user is allowed to organize the story structure by creating different acts and episodes. The user can create, add and attach different episodes to a specific act, while each episode narrates only one message. The manner of creating and organizing acts and episodes is left to data narrator. At the end, the user can download the story as a PDF document by clicking on *PDF of narrative*. Also, a *notebook SQL* can be generated to document the SQL data exploration.

4 Demonstration Scenario

This section presents the experience to showcase the production of data narratives using interactive querying and visualizations. The demonstration is guided by a generic case study such as "As a journalist, you are investigating how COVID-19 spreads around the world." Authors are asked to extract relevant

[4] `https://github.com/hvf/franchise`

facts from a COVID dataset and to produce a data narrative enriched with visualizations (see Figure 2). The scenario mimics the data narrative published by the European Centre for Disease Prevention and Control (ECDC)[5].

We detailed each layer and component arrangement to reprint this scenario using our prototype for generating a complete data narrative. Figure 2 represents some screenshots of two data narrative versions: initial one in the left part and reprinted version in the right using our prototype. The code, screenshots and a PDF of the reprinted data narrative, generated with the prototype, are available on Github[6].

Intentional layer. Data narrator starts a story by specifying the *analysis goal* of the intended data narrative: report worldwide covid-19 situation as of May 21st, 2020. This goal brings out several *characters* to play a key role in episodes narrated as "worldwide", "covid-19", "cases" and "deaths". A set of *analytical questions* is posed splitting different aspects of the goal: Which is the current covid-19 situation? How daily epidemiological curves evolve? Which is the geographic distribution of cases and deaths? These questions are answered by a set of *messages* (based on findings, see Factual layer below) such as "5 776 934 cases and 360 089 deaths were reported as of 21 May 2020". This message brings out new *characters* and *measures*, for example, "21 May 2020" and "5 776 934", which are narrated in the first episode.

Factual layer. As described in the ECDC web site, every day between 6:00 and 10:00 CET, a team of epidemiologists screens up to 500 relevant sources to collect the latest figures. The data screening is followed by ECDC's standard epidemic intelligence process for which every single data entry is validated and documented in an ECDC database, available from the web site in XLS format. We downloaded the XLS file and inserted data in a relational table (keeping the same structure) for recreating the data exploration. The simplicity of the file structure allowed to produce all the *findings* reported by the data narrative using simple SQL queries as *simple collectors*. For instance, the daily curve of Episode 1 of Act 2 is generated with the following SQL query: SELECT daterep, continentexp, sum(cases) FROM covid19 GROUP BY daterep, continentexp ORDER BY daterep; It is subsequently rendered with a bar chart. Similarly, Episodes 2 and 3 of Act 1 are produced with group by and top queries, while the last episode of Act 3 is produced by joining two group by queries. In other words, the exploration solving this narrative's analysis goal is a sequence of SQL queries over the ECDC database. These queries are available on the project Github.

Structural layer. The plot is organized in 3 *acts*, devoted respectively to narrate: a summary of the situation per continent (Act 1), daily epidemiological curves (Act 2), and geographic distribution of cases (Act 3). Act 1 includes 3 *episodes*, narrating respectively: the worldwide summary of the pandemic, the cases reported per continent (highlighting countries reporting most cases) and the deaths reported per continent (also highlighting countries reporting more deaths). Act 2 includes 2 episodes, narrating respectively: the daily evolution

[5] https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
[6] https://github.com/OLAP3/pocdatastorytelling

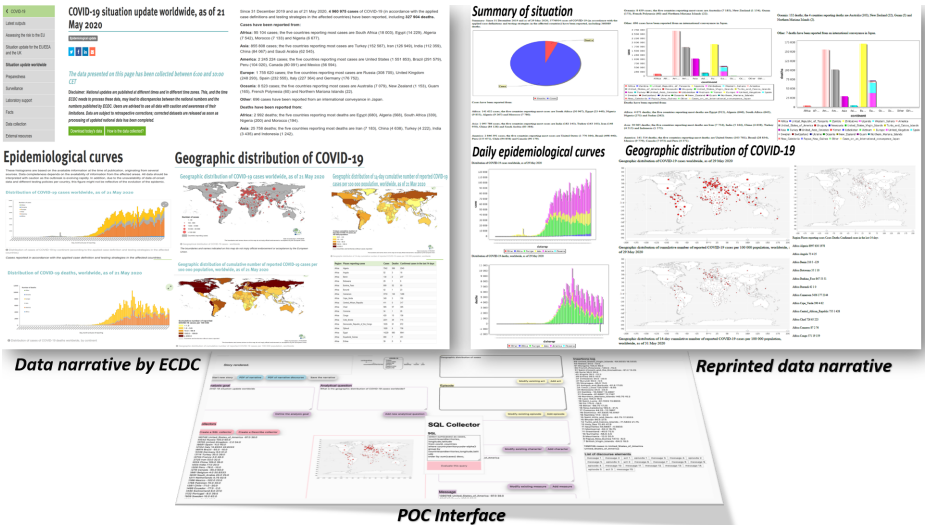

Fig. 2. Some screenshots of two versions of covid data narrative available at `https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases` & `https://github.com/OLAP3/pocdatastorytelling`

of new cases per continent, and the daily evolution of deaths per continent. Act 3 includes 4 episodes. The first 3 narrate the geographic distribution of, respectively, cumulative number of cases, cumulative number of cases per 100 000 population, and 14-days cumulative number of cases per 100 000 population. The last episode details the number of cases, deaths and 14-days cases per country.

Presentational layer. The visual narrative is published as a web page. It contains three dashboards for rendering the 3 acts of the plot. Subtitles are chosen for delimiting dashboards. Dashboard components are responsible for rendering episodes with several visual artifacts: formatted text (episodes of Act 1), bar charts with textual explanations (episodes of Act 2), maps (3 following episodes) and a table (last episode).

References

1. Carpendale, S., Diakopoulos, N., Riche, N.H., Hurter, C.: Data-driven storytelling (dagstuhl seminar 16061). Dagstuhl Reports **6**(2), 1–27 (2016)
2. Chatman, S.: Story and Discourse: Narrative Structure in Fiction and Film. Cornell paperbacks, Cornell University Press (1980)
3. Chédin, A., Francia, M., Marcel, P., Peralta, V., Rizzi, S.: The tell-tale cube. In: ADBIS (2020)
4. El Outa, F., Francia, M., Marcel, P., Peralta, V., Vassiliadis, P.: A conceptual model of data narrative for exploratory data analysis. In: ER (2020)
5. Vassiliadis, P., Marcel, P., Rizzi, S.: Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. Inf. Syst. **85**, 68–91 (2019)