Towards a conceptual model for data narratives

Faten El Outa[1][0000−0002−8153−4252], Matteo Francia[3][0000−0002−0805−1051], Patrick Marcel[1][0000−0003−3171−1174], Veronika Peralta[1][0000−0002−9236−9088], and Panos Vassiliadis[2][0000−0003−0085−6776]

[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 is the activity of producing stories supported by facts extracted from data analysis, possibly using interactive visualizations. In spite of the increasing interest in data narration in several communities (e.g. journalism, business, e-government), there is no consensual definition of data narrative, let alone a conceptual or logical model of it. In this paper, we propose a conceptual model of data narrative for exploratory data analysis. It is based on four layers that reflect the transition from raw data to the visual rendering of the data story: factual, intentional, structural and presentational. This model aims to support the entire lifecycle of building a data narrative, starting from an intentional goal: fetch and explore data, bring out highlights, derive important messages, structure the plot of the data narrative, and render it in a visual manner. Our contributions include a description of the model and its instantiation for several real examples showing that it covers data narration needs.

Keywords: Data narrative, Visual narrative, Data storytelling, Data exploration

1 Introduction

Narrating a story is considered as one of the oldest activities in the world, and a pillar of information communication as a mean of education. Often mistaken with storytelling, which describes the social and cultural activity of sharing stories [13], narration is the use of techniques to convey a story to an audience [12]. Recently, data narration, i.e., narrating with data visualizations [7], 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 [2]. More concretely, such data narratives can be viewed as ordered sequences of steps, each of which can contain words, images, visualizations, audio, video, or any combination thereof, and which are based on data [8].

Apart from these general considerations, and to the best of our knowledge, there is no consensual definition of data narrative, let alone a conceptual or logical model of it. While data narrative essentially attracted attention from the

visualization community [2, 8], we believe that (i) a more global approach to it is needed from domains including visualization, data management, data exploration and machine learning, and (ii) conceptual modeling of the domain should drive further researches, to help the understanding, standardization, reuse and sharing of data narratives. Such a clear foundation of the aspects and design choices in the domain of data narration provided to system builders and algorithm designers will allow to facilitate rapid exploration with automation, to support iterative and collaborative workflows, etc. [14].

The main contribution of this paper is a conceptualization of the domain of data narrative. We propose a novel conceptual model that provides a structured, principled definition of the key concepts of the domain, along with their relationships, and clarifies their role and usage. This model aims to guide an author to build a data narrative 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. Note that automatic decisions for producing a narrative (e.g., automatically deciding the visualization technique given a specific message) or methodological aspects (e.g., guidelines for using the model) are out of the scope of our work.

The paper is organized as follows: Section 2 describes background concepts, presents related work and introduces our definition of data narrative. The proposed model is presented in Section 3. Finally, Section 4 concludes and presents future research directions.

2 From narrative to data narrative

While various models of narratives have been proposed (see [6] for a survey), none of them qualifies for data narrative. However, some aspects of classical narration theory, as described e.g., by Chatman [4], should be reviewed to understand the fundamental structure of narration. This is the topic of Subsection 2.1. In addition, to understand what is particular to the process of communicating the result of a data analysis through visual artifacts, we review recent works on visual narration from the data visualization community. Without offering a precise model nor a consensual definition of data narration, these works shed a light on important aspects of it. This is the topic of Subsection 2.2. Finally, reviewing these aspects helps us to clarify the terminology and to propose a definition of data narrative that will structure the proposed model.

2.1 Narratives

Narrative theoreticians agree that there are at least two levels in any narration: some events happen (what is told) and these events are presented and transmitted to an audience in a certain way (how is it told). In the most widely used structuralist terminology, the answer to the "what" question is called a **story** and the answer to the "how" question is called a **discourse** [1]. Chatman [4] distinguishes narration's elements based on the what and how questions, defining

narrative as a couple of story (content of the narrative) and discourse (expression of it). The story has a form, that is the set of possible objects, events, etc., and a substance, which is a **composition** of story elements (i.e., events, settings, behaviors, characters) as pre-processed by the author's cultural code. A discourse has a form of expression, which is a translation of the story content to a **structured combination** of the story elements. In other words, this means that, out of the entire story as it actually happened, when constructing a discourse, the author picks an interesting subset to present. The discourse also has a substance that includes the set of all **media** used to show structured elements, like text, pictures, tables or charts. In summary, the story can be seen as the logical form of the narrative, while the discourse is its presentable manifestation, obtained through author's editions: prunes unimportant parts out, magnifies some others deemed interesting, rearranges the order of presentation to make it more interesting, etc.

2.2 Visual data narration

While using many terms (e.g., narrative visualization, visual storytelling, data driven storytelling), the data visualization community has recently brought much attention to visual data narration [2]. Kosara and McKinley [8] intuitively define a data story as an ordered sequence of steps, each of which primarily consists of visualization, which can include text and images but essentially are based on data. The authors note that journalists work with a model of story construction where the order of events is consistent and clear, for the story to be comprehensible. Journalists **collect information, which gives them the key facts**, and then they tie those facts together into a story. The authors note that the goals, tasks and tools used during the research phase differ from those in the writing phase, and that **only some of the material from the research phase end up in the final story**, most of the source material only serving as raw background information. Segel and Heer [10] insist that the notion of a chain of causally related events is central to the definition. One typical difference between traditional storytelling and data narration, highlighted in [10], concerns the potential for interactivity in the latter. In an effort to understand what makes a good sequence of visualizations, Hullman et al. [7] estimate the cognitive cost of transiting from one visualization to another. More recently, Chen et al. [5] distinguishes (a) visual analytics, which requires to see all aspects of complex data, explore their interrelationships, and is supported by multiple coordinated views and sophisticated interaction techniques, from (b) storytelling, which is meant to convey only interesting and/or important information extracted through the analysis, presented in a simple and easily understandable way. The two processes differ in their purposes, target users, kind of information dealt with, and methods of presenting the information and interacting with it. To support telling stories of visual analytics findings, there should be an intermediate step between analysis and storytelling, in which **the analyst assembles and organizes information pieces to be communicated**.

2.3 Our definition of data narrative

Inspired by Chatman [4] and Chen et al. [5], we propose the following definition for data narrative: *A data narrative is a structured composition of messages that (a) convey findings over the data, and, (b) are typically delivered via visual means in order to facilitate their reception by an intended audience.*

We borrow Chatman's terminology and extend his structure of narrative considering that data narrative must describe how the content of the story (Chatman's events and existents) is derived from data. This is done by distinguishing 4 layers in our model of data narrative: the first two layers represent the *story* and the last two represent the *discourse*. In the story, a *factual* layer represents the story form while an *intentional* layer represents the story substance. In the discourse, a *structural* layer represents the discourse form and a *presentational* layer represents the discourse substance. Specifically, and originally compared to classical models of narration, the factual layer includes an entity for *findings* based on facts and models collected from data, and the intentional layer includes entities for *messages* derived from findings, where the narrative *characters* (e.g., important business objects) demonstrate interesting measurements. The factual layer can be thought of as the "objective" one, describing the work around data exploration and model construction, while the intentional layer reflects the "subjective" editorial work of pre-processing findings to turn them into messages. Note that our model of data narrative is agnostic of a specific data model; all the specific details on how data and facts are produced to serve the information goal and support the extraction of findings are encapsulated in a *collector* entity. The structural layer includes entities modeling the arrangement of the messages into a structured combination of presentable discourse elements, and the presentational layer includes entities for the assignment of a presentable set of media to each of the narrative's discourse structure.

3 The model

This section presents the conceptual model for data narrative depicted in Figure 2, using UML class diagram notation, but omitting class properties for readability purposes. Subsection 3.1 presents an intuitive introduction to model components with a motivating example, while Subsection 3.2 describes the model, organized in 4 layers.

In what follows, we use the terms *author* or *analyst* for the designer of the data narrative. The author is not necessarily a business analyst, she can be a data journalist, or a plain data enthusiast, aiming to produce a report of findings. We also assume an *audience* for the produced outcome, which includes the people that will see, read or hear the story. Both author and audience can represent several persons, or be confounded into one person.

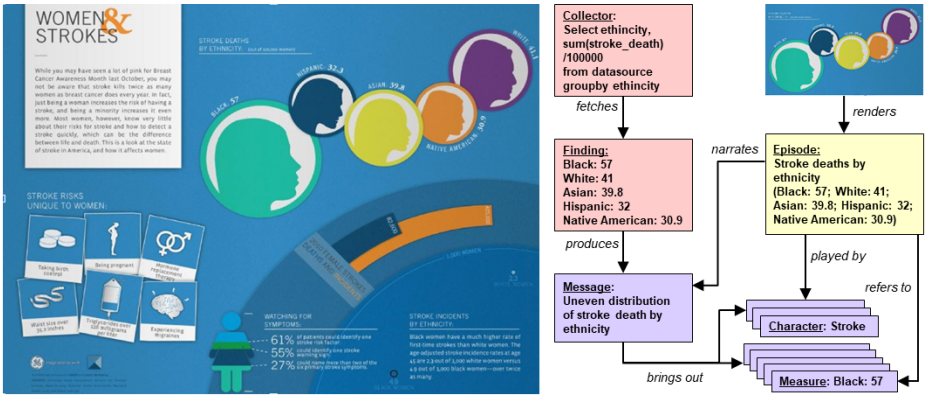

Fig. 1. Example of data narrative, available at `https://www.good.is/infographics/facts-about-women-and-strokes` (left); and a partial object diagram for a particular message (right).

3.1 Motivating example

This subsection illustrates the components of our model (signaled in italics) using a simple visual data narrative about women and strokes, published by GOOD[4]. For illustration purpose, we describe a plausible process for defining analytical questions and collecting data, which is not precised by the author.

The final result, a *visual narrative* is depicted in Figure 1 (left side), taking the form of an infographic. The *plot* warns women about stroke risks by combining diverse information about risks, symptoms and incidents. The plot structures the discourse by arranging messages in coherent pieces of discourse: *acts* narrating a major piece of information and a major part of the narration with a significant communication, and *episodes*, subparts of lesser importance on their own, narrating specific messages. In this example, there is a unique act and six episodes. This act is rendered with a *dashboard* displaying complementary visual information. Six *dashboard components* render the six episodes. For instance, the top right corner of Figure 1 displays stroke deaths by ethnicity. Visual artifacts (in this case, circle sizes) are used for carrying the message (here, putting in evidence that black women are the most impacted by stroke deaths).

We summarize the *messages* in the example, from top-left to bottom-right: (m_1) the overall situation of women's stroke in the USA, (m_2) the uneven distribution of stroke death by ethnicity, (m_3) the risks unique to women, (m_4) the rates of women stroke deaths and incidents, (m_5) the poor ability of patients to identify symptoms, and (m_6) the impact of ethnicity in stroke incidents.

Typically, a data narrative starts with an *analysis goal* and a set of *analytical questions*, reflecting the author's intention. Here, the author's analysis goal is to narrate facts about women and strokes in the USA. An example of analytical question is: Which characteristics of women (age, ethnicity, weight, etc.) have

[4] `https://www.good.is/infographics/facts-about-women-and-strokes`

an impact on stroke deaths? Message m_2 answers this question, evidencing that ethnicity is a critical factor. It brings out ethnicity as a *character*, i.e., a relevant entity or concept of the story, in addition to women and stroke, both already pointed as characters by the analysis goal. Analogously, the ratios by ethnicity are brought out as relevant *measures*, i.e., relevant figures in the story. We can note here that characters may appear in several episodes, esp. the main cast (e.g. women, stroke), while others are only supporting in an episode (e.g. symptoms).

A data *exploration* is built by the author, who called several *collectors* for analysing data and collecting *findings* in order to answer analytical questions. For example, a collector queried a dataset of female patients in the USA, asking for stroke deaths by ethnicity. The ratios of stroke deaths by ethnicity constitute a finding that supports message m_2, stating the uneven distribution of stroke deaths by ethnicity (black women being the most impacted).

Figure 1 (right side) illustrates a partial object diagram concerning message m_2, from the collection of findings to the rendering of an episode.

3.2 Model description

The model we propose for data narrative is depicted in Figure 2. As introduced in Section 2, the organization in 4 layers, adapted from Chatman [4], reflects the transition from raw facts to the visuals communicated to the audience of the data narrative. On their way to the reader, the facts traverse:

1. **Factual layer**. The factual layer models the *exploration* of facts (i.e., the underlying data), via a set of *collectors* that allow for manipulating facts with varied tools. *Findings* emerged from explored facts are candidates for participating in the story.
2. **Intentional layer**. The intentional layer models the 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*.
3. **Structural layer**. The structural layer models the structure of the data narrative, organizing its *plot* in terms of *acts* and *episodes*.
4. **Presentational layer**. The presentational layer models the rendering of the data narrative, i.e., a *visual narrative*, that is communicated to the reader through visual artifacts (*dashboards* and *dashboard components*).

To understand the organization of the model, one should note that the concept of *message* is the model's corner stone, which is clearly evidenced by the way we have related message to the other concepts. Essentially, a specific message is rooted in the facts analyzed, conveying essential findings in the data that answer a, and may raise new, analytical question(s). This specific message is then the discourse structural building block: episodes narrate specific messages, and acts, built as sets of episodes, narrate a broader message. A message is also indirectly connected to the presentational layer: a global message is visually conveyed by one dashboard, each of the dashboard components illustrating one specific message. We now detail each layer by describing its entities and their relationships.

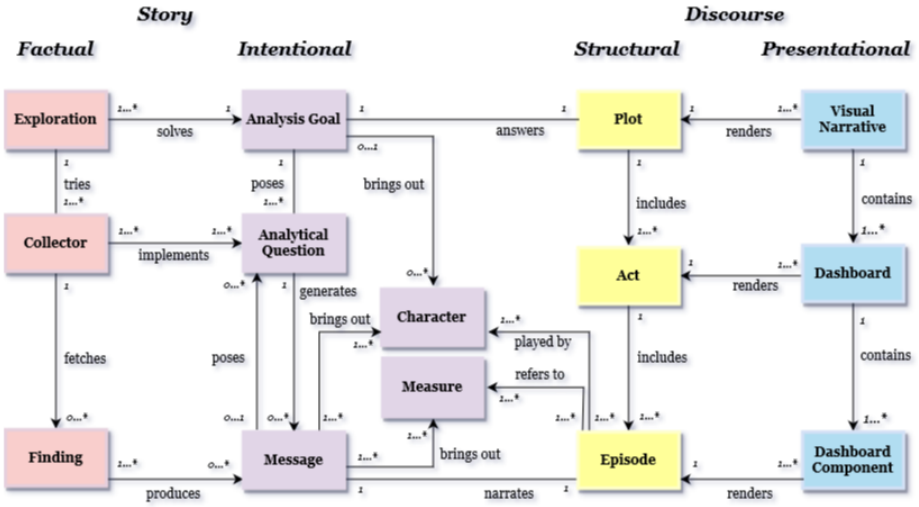

Fig. 2. The data narrative model, organized in layers

Factual layer. Data narratives need data. Data represent the facts that support a story. The factual layer concerns the processes of looking for data in a set of data sources, analyzing them, and obtaining added value and findings. Data are collected via a set of *collectors*, that can be queries in a query language or interface supported by the data sources, extraction tools (e.g. wrappers or loaders), or more generally, all kind of programs able to interact and retrieve data from sources. They may just retrieve data from sources or include functionalities for filtering, checking, building models and reasoning with data. For example, a collector may cluster data, compute correlations, detect outliers or emphasize contradictory data in order to produce insights. For the sake of generality, we do not assume a particular structuring of data (e.g., databases or unstructured files) nor a particular way of collecting findings (e.g., via queries, data analytics or other algorithms), but use collector as an abstraction of data access and manipulation. *Findings* are, among the facts retrieved by collectors, those that are more striking, surprising or relevant and worth narrating in the story. Findings are more than just data returned by collectors, they result from the analysis of facts. A collector may *fetch* one or several findings, or conversely, it may evidence no narrating-worthy finding. The set of actions conducted to collect findings is called a data *exploration*. It aims at keeping trace of the set of collectors tried for addressing an information need. Each collector is part of an exploration while an exploration typically *tries* many collectors.

Intentional layer. The intentional layer models the devising of the substance of the story based upon the author's analysis goal. An *analysis goal* represents the main objective of a story, i.e., the intended information that should be ex-

plained, detailed and transmitted to the reader. A goal is carved up into a set of *analytical questions*, each one concerning specific aspects of the goal, and a set of *messages* are raised in order to answer these questions. Indeed, a goal *poses* a set of analytical questions, and an analytical question *generates* a set of messages, possibly none. Goals are deeply related to explorations. Sometimes, an exploration is built for solving a clear goal, other times, the goal is progressively shaped while exploring data, but frequently, there is an interactive process tying a goal and an exploration. The underlying idea is that an exploration *solves* a goal, while several explorations can be devised for solving a goal. In the same way, an analytical question can be solved by one or more collectors, each collector providing *implementation* means to one or more analytical questions.

As the author explores the data and new findings are collected, progressively the author distills and structures them in her mind. A message is, at the same time, (a) a partial answer to the information need of the author, and (b) the distilling, merging, and translation of a set of related findings into information that is to be conveyed to the audience. The findings raised during the exploration *produce* messages for building the story. A finding may produce many messages and a message results from one or several findings. Possibly, new analytical questions can be *posed* based on the message found, inviting for more exploration. To further structure messages, we introduce two important components of them, *characters* and *measures*, that capture important data values that characterize a message and their fact-based quantification. Both characters and measures belong to the universe understandable by the audience. Messages *bring out* characters (e.g., a set of products causing a drop in sales) and the related measures (e.g. amount of sales for those products). In addition, some characters are previously known and *brought out* directly by the analysis goal (e.g., sales in France). Noticeably, messages serve as a baseline for structuring the story: an episode *narrates* a message. In this way, the intentional layer acts as a bridge between the exploration of facts (factual layer) and the structuring of the story (structural layer). In particular, a message, based on a finding, is the base for building an episode.

Structural layer This layer concerns the form of expression of the data narrative. While previous layers deal with the contents of the narrative, this layer focuses on its discourse. It is important to stress, that there is a design part served here. The idea is that after deciding to address a goal via analytical questions and exploring data, the analysis has resulted in a set of messages to be conveyed to the audience. As reported in Section 2, the literature suggests that, before presenting the messages, there is a synthesis of a story as a coherent composition, where messages must be conveyed to the audience in an organized way [4, 8, 5]. Plots, acts and episodes model the parts of the synthesis of the narrative's content into this composition. A *plot* is the arrangement of messages in a way easily understandable by the audience. To achieve this, the author must put in order a part of the audience-intended report with the messages that conveys an interesting piece of information. Following the terminology of traditional nar-

ration, we introduce an *act* as a constituent part of the plot, which is the mean to convey a specific piece of information. Practically, this means that an act corresponds to a major piece of information and a major part of the plot. Each act is composed of several subparts of lesser importance on their own, which we call episodes. An *episode* is the granular piece of the narrative that conveys a message. A plot *includes* one or several acts and an act *includes* one or several episodes. A plot *answers* a goal and an episode *narrates* a message, the episode text being the shaping of the message. Accordingly, characters and measures brought out by the message appear in the episode text, possibly starring or being highlighted according to author's narration style. One or many characters can *play* in one or many episodes. Analogously, one or many measures can *be referred* in one or many episodes.

Presentational layer. This layer focuses on how the structured story is presented to the audience. Acts and episodes are represented and organized in order to be understandable by the audience. Visualization aspects are the focus of this layer. A *visual narrative renders* a plot. It can be a slideshow, a notebook, a blog, or any other visual art allowing for the visual representation of a story. A visual narrative *contains* a set of *dashboards*, each one *rendering* an act. We use the term dashboard since it is general enough to accommodate varied types of visualizations (e.g. a Business Intelligence dashboard, an infographics, a section in a python notebook, a section in a blog or web page). In the same way, a dashboard contains a set of *dashboard components*, each one *rendering* an episode. Dashboard components include text, images, charts, maps, animations, etc. We remark that a story can be rendered in several ways or formats (e.g., an infographics and a video). In the same way, acts and episodes can be rendered by several dashboards and dashboard components.

4 Conclusion

This paper introduced a conceptual model for data narrative, by extending a classical model of narration [4] to reflect the transition from raw data to the visual rendering of information derived from data analysis. Our model translates fundamental concepts of narration to their respective counterparts when it comes to data narrations and involves the collection of data, the extraction of key findings and the corresponding messages to the audience, the structuring of a presentation of these findings and the ultimate presentation via visual -or other- means via a set of dashboards. To showcase the model, we implemented a proof of concept web application helping an author structuring a data narrative while interactively exploring a database. The code is available on Github[5].

Among the possible refinements of the model, we identified the following as the most desirable: adding a support for transitions between episodes/acts, supporting different semantic structuring (e.g. linear, causality [8], Martini glass

[5] `https://github.com/OLAP3/pocdatastorytelling`

[10]), adding support for plausibilization and coherence of messages and arguments, distinguishing main cast from supporting cast among the narrative characters and adding a layer for interactivity [9] balancing between author- and reader-driven stories. We also aim at improving the web application through tests with data journalists. On the longer run, we plan to tightly integrate the model with our past works on data explorations [3, 11].

References

1. Akleman, E., Franchi, S., Kaleci, D., Mandell, L., Yamauchi, T., Akleman, D.: A theoretical framework to represent narrative structures for visual storytelling. In: Bridges. pp. 129–136 (2015)
2. Carpendale, S., Diakopoulos, N., Riche, N.H., Hurter, C.: Data-driven storytelling (dagstuhl seminar 16061). Dagstuhl Reports **6**(2), 1–27 (2016)
3. Chanson, A., Crulis, B., Labroche, N., Marcel, P., Peralta, V., Rizzi, S., Vassiliadis, P.: The traveling analyst problem. In: DOLAP (2020)
4. Chatman, S.: Story and Discourse: Narrative Structure in Fiction and Film. Cornell paperbacks, Cornell University Press (1980)
5. Chen, S., Li, J., Andrienko, G., Andrienko, N., Wang, Y., Nguyen, P.H., Turkay, C.: Supporting story synthesis: Bridging the gap between visual analytics and storytelling. TVCG pp. 1–1 (2018)
6. Elson, D.K.: Modeling narrative discourse. Ph.D. thesis, Columbia Univ. (2012)
7. Hullman, J., Drucker, S.M., Riche, N.H., Lee, B., Fisher, D., Adar, E.: A deeper understanding of sequence in narrative visualization. IEEE TVCG **19**(12), 2406–2415 (2013)
8. Kosara, R., Mackinlay, J.: Storytelling: The next step for visualization. IEEE Computer **46** (2013)
9. Sarikaya, A., Correll, M., Bartram, L., Tory, M., Fisher, D.: What do we talk about when we talk about dashboards? IEEE TVCG **25**(1), 682–692 (2019)
10. Segel, E., Heer, J.: Narrative visualization: Telling stories with data. IEEE TVCG **16**(6), 1139–1148 (2010)
11. 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)
12. Wikipedia: Narration, `https://en.wikipedia.org/wiki/Narration`
13. Wikipedia: Storytelling, `https://en.wikipedia.org/wiki/Storytelling`
14. Wongsuphasawat, K., Liu, Y., Heer, J.: Goals, process, and challenges of exploratory data analysis: An interview study. CoRR **abs/1911.00568** (2019)