Information Systems Frontiers (2024) 26:1785–1801
https://doi.org/10.1007/s10796-023-10418-1

Highlighting the Importance of Intentional Aspects in Data Narrative Crafting Processes

Faten El Outa[1] · **Patrick Marcel**[1] · **Veronika Peralta**[1] · **Panos Vassiliadis**[2]

Accepted: 18 June 2023 / Published online: 22 August 2023

Abstract
Data narration is the activity of crafting narratives supported by facts extracted from data exploration and analysis, using interactive visualizations. While data narration has recently attracted much attention, the process of crafting data narratives is loosely documented and has not yet been formally described. In this article, we propose a comprehensive and well-founded process to fill this need. It aims at (i) supporting the complete cycle of data narration, from the exploration of data to the visual rendering of the narrative, (ii) being flexible enough to cover a wide range of crafting practices, and (iii) being well founded upon a conceptual model of the domain. In addition, we investigate several crafting scenarios that represent typical situations and detail the workflow of one particular phase, which reflects the intentional aspects.

Keywords Data narrative crafting · Data journalism · Process

1 Introduction

Data narratives are receiving increasing interest from several research communities (e.g., visualization, data management, computer-human interfaces) Carpendale et al. (2016) and many application domains (e.g. journalism, business, e-government, health). They are largely used by journalists, scientists, and other communicators, to convey striking messages to a given audience. A data narrative could take the form of a data video, an infographics, a news article, etc. Any sort of narration that is constructed based on data can be considered a data narrative. In addition, the crafting of a data narrative includes a variety of activities, including the analysis of data, the drawing of relevant messages from data, the structuring of messages into a coherent story and its visual rendering. Despite this diversity of activities, sometimes even conducted by different people with varied professions and skills, there is no framework, workflow, or tool for supporting the crafting of data narratives.

In an effort to clarify the concepts of data narratives, we recently defined a data narrative *as 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*, and we proposed a conceptual model describing and structuring the key concepts around data narratives Outa et al. (2020). This model (described in Section 2) is organized in 4 layers: factual, intentional, structural and presentational, which reflect the transition from raw data to the visual rendering of the story. With this definition and model in mind, our aim in this paper is to contribute with a study of the dynamic aspects of data narrative crafting. Like many works in the literature (e.g., Kosara et al. (2017); Lee (2015); Chen et al. (2018)), we postulate that the different forms of data narration can be described by a comprehensive process encompassing the various activities ranging from data exploration to the rendering of the data narrative. A formal description of this process will benefit novice data narrators, like e.g., non technical data journalists, and will be instrumental to the development of tools for supporting advanced data narrators.

Accordingly, we reviewed the literature around the process of crafting data narratives, and we conducted a survey with data journalists in order to understand how they craft a data narrative. As an outcome of the former, we found

✉ Patrick Marcel
patrick.marcel@univ-tours.fr

Faten El Outa
faten.elouta@univ-tours.Fr

Veronika Peralta
veronika.peralta@univ-tours.fr

Panos Vassiliadis
panos.vassiliadis@cs.uoi.gr

[1] University of Tours, Blois, France

[2] University of Ioannina, Ioannina, Greece

that the research communities globally agrees in the fact that the crafting process includes three main phases: (i) the *analyzing* phase that handles the activities of exploring data, retrieving findings and formulating messages learned from data, (ii) the *structuring* phase that includes the activities to organize the plot of the narrative in an understandable way and, (iii) the *presenting* phase that covers the activities to convey the structured messages visually. At the same time, our bibliographical study revealed the absence of a comprehensive and well-founded process that covers the main activities of the crafting process, specially those dealing with user intentions and their tight relation to data analysis. Apart from the bibliographical study, the conducted survey allowed us to observe the crafting workflows regularly followed by 18 data journalists, and we contrasted them to the literature. It turned out that journalists follow the same three phases, mostly in a linear way, attaching less attention to the structuring phase, while spending more time in the analyzing phase.

These considerations from the literature study and the survey with data journalists enabled us to identify the activities (and their chaining) for crafting data narratives. Based on those, we propose a comprehensive and well-founded process that (i) covers the whole cycle of data narrative crafting, from exploration of the data to the visual presentation of the narrative, (ii) accommodates a wide range of practices observed on the field, and, (iii) is founded on a conceptual model of the domain that clarifies the concepts involved in the process Outa et al. (2020).

This paper is a follow-up to Outa et al. (2022), where the process was originally motivated and introduced. In particular, we enhance the process description by investigating several crafting scenarios that represent typical situations, and we detail the workflow of the key, and often overlooked, *answer question* phase. Specifically, we have improved over Outa et al. (2022) on the following aspects:

1. We enhance the description of the process by further detailing activities, adding a motivating example and presenting several scenarios of typical crafting situations.
2. We specify the internals of the *answer question* phase. Concretely, we propose activity diagrams for describing the workflow between activities. This workflow covers the activities and paths reported by several practitioners, while also being founded upon and coherent with the conceptual model.
3. The state of the art section is enriched with recent proposals for the automatic crafting of data narratives, as well as works highlighting the importance of intentional aspects of data narrative crafting.
4. We report various experiments conducted to understand to what extent the proposed process (i) covers all necessary activities performed by observed data narrators; (ii) contributes to the improvement of the quality of handcrafted data narratives; and (iii) is consistent with processes documented by one data journalist and one data scientist.

The scope of our process targets the population of data journalists or any other data enthusiast that craft data narratives out of existing data. The reason for proposing the process is exactly the observed discrepancy between literature and practice, with omissions of important parts from both sides. *Thus one significant contribution of our work is the explicit treatment of all the steps that should be involved in the process, as well as providing detailed descriptions of the activities that accurately reflect the intention of the data narrator*. Secondly, apart from providing a methodological guidance, *our process* can enable the support of the process via tooling. Indeed, there is a lack of integrated tools covering the whole crafting process and recommending actions to less-experienced narrators. In particular, an application that would automatically document the data exploration and narration crafting is desperately needed by data workers, who spend hours to document their work. This is important for reproducibility, transparency, and linkage, and requires a conceptual model and a process that are both consensual.

The paper is organized as follows: Section 2 recalls the key concepts of the conceptual model proposed in Outa et al. (2020). Section 3 reviews the related work concerning processes for crafting data narratives, and Section 4 depicts a brief introduction to understand the process with a relevant motivating example and discusses the survey we conducted with data journalists. The proposed process is described in Section 5 and the question answering phase is detailed in Section 6. Section 7 describes the experiments and Section 8 concludes and draw research directions.

2 A Conceptual Model for Data Narratives

We recently proposed in Outa et al. (2020) a conceptual model of data narratives providing a principled definition of the key concepts of the domain, along with their relationships, and clarifying their role and usage (see Fig. 1).

This model is based on 4 layers following Chatman's organisation Chatman et al. (1980), who defined narrative as a pair of (a) *story* (content of the narrative), and, (b) *discourse* (expression of it). In our model, 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 and fetching *findings*, in an objective way, while the *intentional* layer models the subjective substance of the story, identifying the *messages*, *characters* and *measures* the narrator intends to communicate, and tracing how they

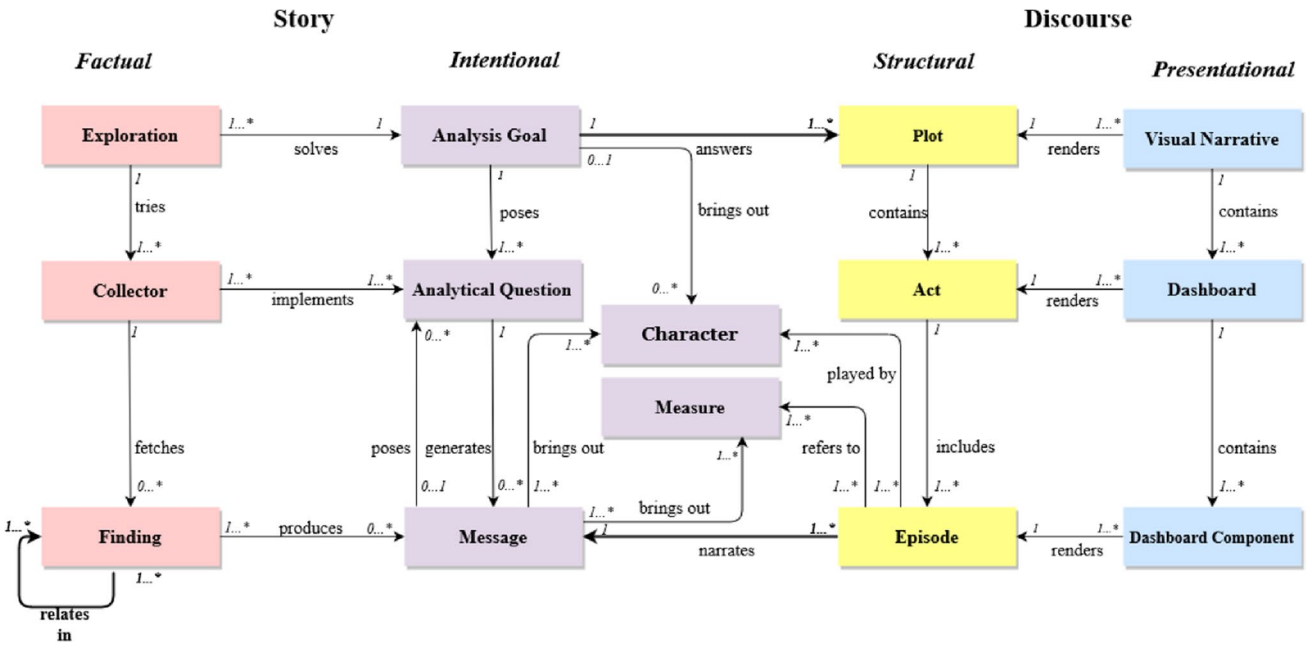


Fig. 1 The conceptual model for data narratives (relations in bold were extended w.r.t. the original version in Outa et al. (2020))

are obtained through *analytical questions*, according to an *analysis goal*. As to the discourse, the *structural* layer models the structure of the data narrative, its *plot* being organized in terms of *acts* and *episodes*, while the *presentational* layer deals with its rendering, that is communicated to the audience through visual artifacts (*dashboards*[1] and *dashboard components*).

The interested reader is redirected to Outa et al. (2020) for a deeper presentation of the model. Here, we will highlight the main decisions behind the model that are necessary for grasping its essence. Importantly, it should be noted 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. A specific message is rooted in the facts analyzed, conveying essential findings, potentially raising new analytical questions.

While a finding can be a pattern like an association rule, or a path in a decision tree, or the verification or rejection of a hypothesis over the data, a message, on the other hand, is the answer to the intentional question that exploits a finding to label a character with respect to other characters or a measure. The following examples illustrate the difference between these two concepts:

- By comparing *Daily Infection* in *France* to *EU Average*, we *find* that they are similar. A message corresponds to the labeling of measure *Daily Infections* of character *France* with respect to another peer character, *EU Average*.
- The message that a *Media outlet* cannot determine the existence of fake news answers the following hypothesis (analytical question): can the outlet solely determine fake news? This message follows from the finding that, by correlating the concept *News Authenticity* to the concept *Media outlet*, we find a non-significant correlation.

The message allows introducing episodes, the building blocks of the discourse. Each episode of the discourse is specifically tied to a message which it aims to convey. The relationship between messages and episodes is the basis for structuring stories that address analysis goals, narrated by structured discourses (with cohesive acts being the backbone of the narrative structure) and dashboards their presentational counterpart.

3 Related Work

In this section, we review the works describing the internals of the data narration process, as well as the tools that automate (part of) the crafting process.

3.1 Global data narration processes

Data narration is a complex process, at the crossroads of several domains: data exploration, data visualization, data

[1] We use the term dashboard since it is general enough to accommodate various 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).

management, etc. Despite the many contributions in each of these areas, few works offer comprehensive workflows describing the entire data narration process. The first attempt to model data narration processes come from the visualization community. For example, Kosara and Mackinlay (2013) proposed a two-phases process: First, narrators collect information and *explore* their interrelationships, pointing to key facts, and then, they *tie* those facts together into a story. Chen et al. (2018) surveyed early proposals and concluded that their crafting processes are composed of two main phases: (a) *visual analytics*, which requires seeing all aspects of complex data, explore their interrelationships, and is supported by multiple coordinated views and sophisticated interaction techniques, and (b) *storytelling*, which is meant to convey only interesting or important information (i.e., findings) extracted through the analysis, presented in a simple and easily understandable way.

To bridge the gap between these two phases, Chen et al. proposed an intermediate one, called *data synthesis* Chen et al. (2018). In this phase, the narrator assembles and organizes the findings to be communicated, to represent explicitly the essential relationships between them, building a compelling narrative. Lee et al. (2015) also identified three main phases: *explore data* to retrieve findings, *make a story* to turn findings into a sequence of narrative pieces to build the plot of the narrative, and *tell a story* to materialize the plot in a visual manner. The authors stress the importance for the data narrator to go back and forth between the exploration and the story-making phases. Duangphummet et al. (2021) proposed a protocol consisting of the following phases: *conceptualization* of the data narrative domain, targeted audience and distribution channel, *data preparation* to deliver data that is relevant to the use, *realization* to deliver a storyline with detailed content and an initial form of key visualizations, *visualization design* to redesign the visualizations and create visualization prototypes, and finally, the *visualization development* where technical requirements are defined, and the key visualizations for target devices are developed and deployed.

In addition to Lee (2015), many works underline the importance of moving between the data narrative crafting phases. For instance, Wang et al. (2019) ran a workshop on data comics, organized by an interdisciplinary team with expertise in data visualization, graphic design, data comics, and illustration. They observed that to create stories, students require to *move back and forth between the story, visualizations, and the data*.

Besides the previously described works proposing global crafting processes, some works describe subprocesses, focusing on the necessary activities to be conducted. Without being exhaustive, we mention here some major contributions.

Battle and Heer (2019) identified three ways to start a data narrative: having a precise idea in mind, having a vague idea refined during data exploration, or having no idea before exploring the data. Weber et al. also point that the crafting process starts by either an idea, a problem or a question Weber et al. (2018).

Notably, many works underline the importance of different story structures and different kinds of interactivity in data narration Segel and Heer (2010); Weber et al. (2018). In particular, Weber et al. (2018) encourage to use non-linear structures and set up interactivity. Many works specifically deal with the phase of structuring the narrative Wang (2020); Shi et al. (2021b).

Finally, very few works highlight the importance of intentional aspects. Bach et al. (2018a) found 18 narrative patterns to provide guidance on how to achieve five general storytelling intents (i.e., argumentation, flow, framing, emotion, and engagement). Similarly, design patterns have been proposed for data comics Bach et al. (2018b) and dashboards Bach et al. (2023).

Thudt et al. (2017) stress that subjective perspectives can be introduced at every step of visualization creation: during data collection and processing, visual encoding, and when refining the presentation. In the context of OLAP cube exploration, Vassiliadis et al. (2019) propose a set of intentional operators to express high-level analytical intentions and automate their translation to database queries.

3.2 Automated data narration

Many recent works addressed the automatic generation of data narratives, providing another source of insights on how this process is perceived.

Wang et al. (2020) conducted a qualitative analysis of 245 infographics examples to explore the infographics design space in terms of structures, sheet layouts, fact,[2] types, and visualization styles. Based on those, the authors propose a system for supporting a fact sheet generation pipeline consisting of three phases: (i) fact extraction, (ii) fact composition, and (iii) presentation synthesis. Shi et al. (2021a, b) proposed Calliope, a system that can automatically generate visual data stories with facts arranged into a logical sequence. It consists of two main modules: (i) the story generation engine, for generating, choosing and organizing the facts that will participate in the narrative, and (ii) the story editor, that visualizes the data story (generated as a series of visualization charts) and allows the users to change it based on their preferences. Sun et al. (2023) proposed Erato, a human–machine cooperative data story editing system that allows users to generate insightful and fluent data stories. It consists of three major modules: (i) a fact embedder that

[2] We worth noting that the term *fact* used by many authors of the visualization community, corresponds to the concept of *finding* discussed in the conceptual model Outa et al. (2020).

takes a fact's specification string as the input and converts it into a vector representation, (ii) an interpolator that generates new story content by interpolating between two data facts, and (iii) a story editor that enables user to verify, refine, and incorporate data facts to make a more smooth and compelling story. Park et al. (2022) proposed Storyfacets, a communication-minded visualization system that maintains the provenance of all data exploration and provides multiple, linked visual formats for analysis and presentation. The workflow typically begins with an analyst exploring data in a full-fledged analysis view. The system automatically maintains multiple distinct views of the same data-driven narrative, corresponding to the user's expertise level.

Shi et al. (2021a, b) described the workflow for crafting data videos, consisting of 4 phases: (i) *collecting a series of data facts* around a certain topic, (ii) *constructing a storyline* as an assembly of these data facts into a sequence, (iii) *choosing data visualizations* for the data facts and deciding how to animate them by drawing a storyboard, and finally, (iv) *realizing the storyboard* via a design software in which the narrator edits and combines the animated visualizations until a coherent data video is accomplished.

In CineCubes Gkesoulis et al. (2015), Gkesoulis et al. detail the process of crafting a data movie in the form of a powerpoint presentation, to answer a specific user's need described by a query. First, an introductory act is built with the initial query, and two subsequent acts are used to put context. These acts contain visualizations highlighting important facts, as well as text and audio describing these facts. A summary act concludes with all the important highlights of the previous acts.

In all these works, the proposed phases are consistent with those described in the previous subsection. Being a mostly automatic generation, the construction is linear in the sense that there is no back and forth movements between phases. In addition, they target a specific domain or data format and organize stories accordingly to pre-established patterns. In particular, we highlight the absence of intentions, that are, at best, modeled via an initial query or a topic.

Lessons learned. Most of the works describing the data narration process agree on the 3 general phases of: exploration (to retrieve findings), structuring (organizing the information gathered into narrative pieces) and presentation (crafting visual artifacts). Automated data narration is still in its infancy, mainly applying rigid patterns and lacking the necessary flexibility of moving between the 3 phases. One of the key findings is that the intentional layer of the model presented in Fig. 1 is largely absent from the works reviewed. This means the substance of the story, i.e., the **composition** of story elements (analytical questions and hypothesis, messages, etc.) as pre-processed by the author's cultural code Chatman et al. (1980) is ignored. We claim that this absence is regrettable; if data narrations are to be shared, reused, and have their crafting process documented, then this intentional layer deserves more attention.

4 Data Journalist Practices

In this section, we explore the professional practices of data journalists. We start with an example of a real use case of data narrative crafted by a data journalist about COVID deaths in a French region. We next report the results of a survey conducted with 18 french data journalists.

4.1 Example of a data narrative handcrafted by a data journalist

This subsection describes the steps a data journalist took to craft a data narrative about the COVID pandemic in a French region. Activities leading to explore the data and reflect the intention of the data journalist were deduced from the analysis documented in a notebook,[3] while the activities leading to structure and present the plot of the data narrative were deduced from the visual narrative published on Rue89 Strasbourg.[4] To ease the reading, we describe these steps in four main parts, following the four layers of the conceptual model presented in Section 2. Note that this does not reflect the order of activities taken by the journalist, who, in that case, started with a clear goal in mind before even collecting and exploring the data.

1. *Explore:* The data journalist collected the datasets of deceased people on the "data.gouv.fr" (open data) portal. The exploration of the data via several collectors written in Python revealed a number of key **findings**. For example, one finding is: "Between 2010 and 2019, there were an average of 1100 deaths between the months of March and April in the Upper Rhine area, while In 2020 (the year COVID19 struck) there were 2347 deaths between the months of March and April in the same area". The data journalist tried multiple visualizations, to aid the understanding of the data and retrieval of the findings. All such findings obtained from the data are the output of this explore phase.
2. *Answer questions*: The data journalist aimed to "analyze mortality figures in a specific area (the French Alsatian departments)", and this goal was clear from the start. Several analytical questions were posed that were answered through data exploration. One of the questions was: " Compare the mortality figures in 2020 with those of previous years in the Alsace departments between March 1st

[3] https://tinyurl.com/yc5chu57(in french)

[4] https://tinyurl.com/24ubaanu(in french)

and April 24th, 2020." On the basis of the finding which is retrieved from the analysis, the data journalist formulated one or more **messages** that provide an answer to the analytical question. One message is: "In the Upper Rhine region, it can be said that the difference of number of deaths is undoubtedly largely related to COVID19". This message was validated by an expert (Dean of the faculty of Medicine of the University of Strasbourg). The messages formulated are used in the next phase.

3. *Structure answers*: The data journalist addressed the general public to communicate the mortality figures in the Alsatian departments. He structured the plot of the data narrative into six acts, each composed of one or multiple **episodes**. Each episode narrates a message formulated in answering a question. One such episode narrates the message stated above that COVID19 is largely responsible of the increase in mortality in March-April 2020, including the validation and an explanation by the Dean of the Faculty of Medicine. The episodes and the plot produced during this phase feed the last phase.
4. *Present*: Once the plot of the data narrative was created, the data journalist conveys the plot by creating a visual narrative in the form of a post published on Rue89 Strasbourg (see footnote 4). Each act of the narrative is represented by a different section, while each episode is further turned into a **dashboard component** in the form of texts and often a visualization. In the case of the episode about the increase in mortality in March-April 2020, an interactive line chart plots, for the period and for years 2010 to 2020, the number of deaths by day and its moving average. The output of this phase is the dashboard components forming the data narrative.

4.2 Surveying data journalist practices

We report the results of a survey Chagnoux et al. (2020) (in French), aiming at investigating the professional practices of data journalists.

The survey consisted of 32 questions[5] (in French). Note that for some questions more than one answer was possible, and that journalists could leave the questions unanswered. The survey was answered by 18 data journalists from 14 French regions, who have worked on a big variety of topics, including elections, environment, cinema, terrorism, paradise-papers, real estate.

For nearly 50% of them, data narration is at the core of their professional activity, and is occasional or marginal for the others. Concerning training, 56.3% studied social sciences, 18.8% studied sciences and 24.9% graduated from law or journalist schools. One of the journalists works for the International Consortium of Investigative Journalists (ICIJ), 5 of them work for the national press, and the 12 remaining work for the regional press.

Regarding their general working habits, 75% of them work alone. They usually work on open data (72.2%) and more specifically on data from public institutions (44.4%). They consume from minutes to months during the data narration and use different tools during data exploration, such as spreadsheets (93.8%), scripts (50%), notebooks (18.8%), powerBI-like tools (31.3%) and some machine learning tools (28.6%).

Two main questions were asked on their data narration practices. For the first one, "How does a data story's subject emerge?", multiple answers were possible. The answers showed that the goal, or subject, of an article emerges from: an idea to be confirmed by data (68%), a dataset which needs exploration to reveal important facts (68%), a refinement of the subject while exploring the dataset (48%).

The second, open question was: "What is the general workflow you apply for data narrative crafting?". Figure 2 sketches the answers provided by 14 of the 18 journalists, where activity names summarize journalists' descriptions of their main activities,[6] rows correspond to journalists and column numbers reflect the sequence of activities.

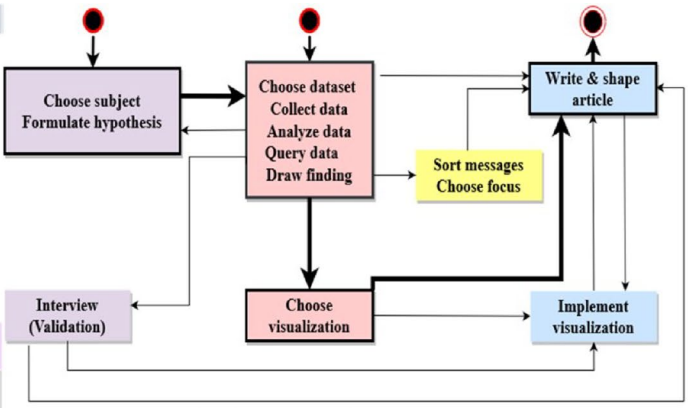
We color these activities according to the layers of the conceptual model: factual (pink), intentional (purple), structural (yellow) and presentational (blue). Gray-colored cells indicate that the activity may overlap structural and (more probably) presentational tasks. In addition, activities concerning the checking of findings and the validation of messages (namely interviews, validation or cross-checking), aiming at transforming a factual object into an intentional one, are in between the factual and intentional layer. Similarly, visualizations are used both in the factual layer, to understand data and retrieve findings, and in the presentational layer, to choose the most suitable one for communicating findings to the audience in a visual manner. We have abstracted these sequences in the form of an activity diagram (top-right corner of Fig. 2). Most frequent paths are highlighted by larger arrows.

Lessons learned. Fig. 2 shows that many activities under different names aim towards the same action, and that different paths can be followed by journalists when crafting a data narrative. *The figure also shows a preponderance of activities from the factual and the intentional layer*. The activity diagram shows that journalists enter the workflow either in the factual layer, i.e., by exploring a dataset, or by the intentional layer, i.e., having at least a vague idea of the subject. After this, the workflow becomes mostly linear, with some movements between the factual and intentional layers.

[5] https://tinyurl.com/ynjzjs63

[6] Since the question was open, we homogenized the answers and grouped them into few categories.

Journalists	1	2	3	4	5							
1	Choose datasets	Query data	Implement visualization									
2	Choose datasets	Analyze data	Write article									
3	Analyze data	Validation	Write and shape the article									
4	Analyze data	Sort messages	Choose focus	Shape the article								
5	Analyze data	Choose visualization	Implement visualization	Test visualization								
6	Formulate hypothesis	Choose datasets	Verify hypothesis (Analysis)	Write and shape the article								
7	Formulate hypothesis	Collect data	Query data	Draw patterns	Interviews (validation)							
8	Formulate question	Choose datasets	Analyze data	Interviews (validation)	Write article							
9	Overview of data	Choose subject	Prepare data	Analyze data	Implement visualization	Write article						
10	Choose subject	Formulate hypothesis	Verify hypothesis (Analysis)	Design visualization	Implement visualization	Write article						
11	Idea of subject	Choose datasets	Prepare data	Analyze data	Choose visualization	Write article						
12	Choose datasets	Prepare data	Query data	Draw findings	Choose visualization	Write article	Choose final visualization					
13	Formulate hypothesis	Choose datasets	Collect data	Prepare data	Analyze data	Validation (cross-checking)	Implement visualization					
14	Choose subject	Choose dataset	Collect data	Analyze data	Refine subject	Discussion (Check finding)	Search further datasets	Analyze data	Interviews (validation)	Prepare visualization	Choose form of article (paper/web)	Write article



Activity diagram for activities reported by data journalists

Fig. 2 Sequence of activities reported by journalists

Usually, data journalists start writing their articles once the analysis phase is over, and there is no backtrack once the presentational layer is entered.

Notably, the journalists attached less importance to the activities of the structural layer. At the exception of one of them, structuring activities are either hidden in writing activities or even not mentioned explicitly. Precisely, many of them agree that while data exploration usually takes long, visual storytelling can be extremely fast, potentially done on the fly, with some of them actually not even involved in the writing of the article. For those that mention it, the activity "write article" includes several hidden details concerning the organization of messages that should be communicated, the visual presentation and communication of the analysis results.

Overall, we can say that there is a chasm between what practitioners do and what literature suggests – and in fact, there are deficits in both sides. On the one hand, compared to what is reported in the literature, the work of the data journalists is over-emphasizing the intentional part and under-investing on the structural and the presentational part. On the other hand, when it comes to the literature, the presented methodologies overemphasize presentation and (to some extent) structuring, and pay much less attention to the intentional part. A process that gracefully hosts all aspects of narrative construction would facilitate narratives that are more complete and intuitive.

5 A Process for Crafting Data Narratives

From the literature review and the survey with journalists, we synthesize a set of requirements for a comprehensive data narration process, and we propose a process that fulfills the requirements.

5.1 Requirements

First of all, we note the absence in the literature of a whole workflow for crafting data narratives, including all the activities identified in Section 3 and Section 4. Figure 3 depicts the activities as phrased in the literature (in gray boxes) and by journalists (in green boxes). We group those referring to the same task and propose new names (the bold ones in Fig. 3) which are consistent with the conceptual model of Fig. 1.

In more details, most authors Chen et al. (2018); Lee (2015); Wang (2020); Shi (2021a, b); Duangphummet et al. (2021); Shi et al., 2021a, b; Kosara & Mackinlay (2013); Wang et al. (2019) agree that data narration process includes three main phases: (i) *analyze*, (ii) *structure answers*, and (iii) *present*.

The survey reveals that the data journalists agreed with the literature, especially on the phases (i) and (iii). In Fig. 3, activities are grouped according to these phases. We remark

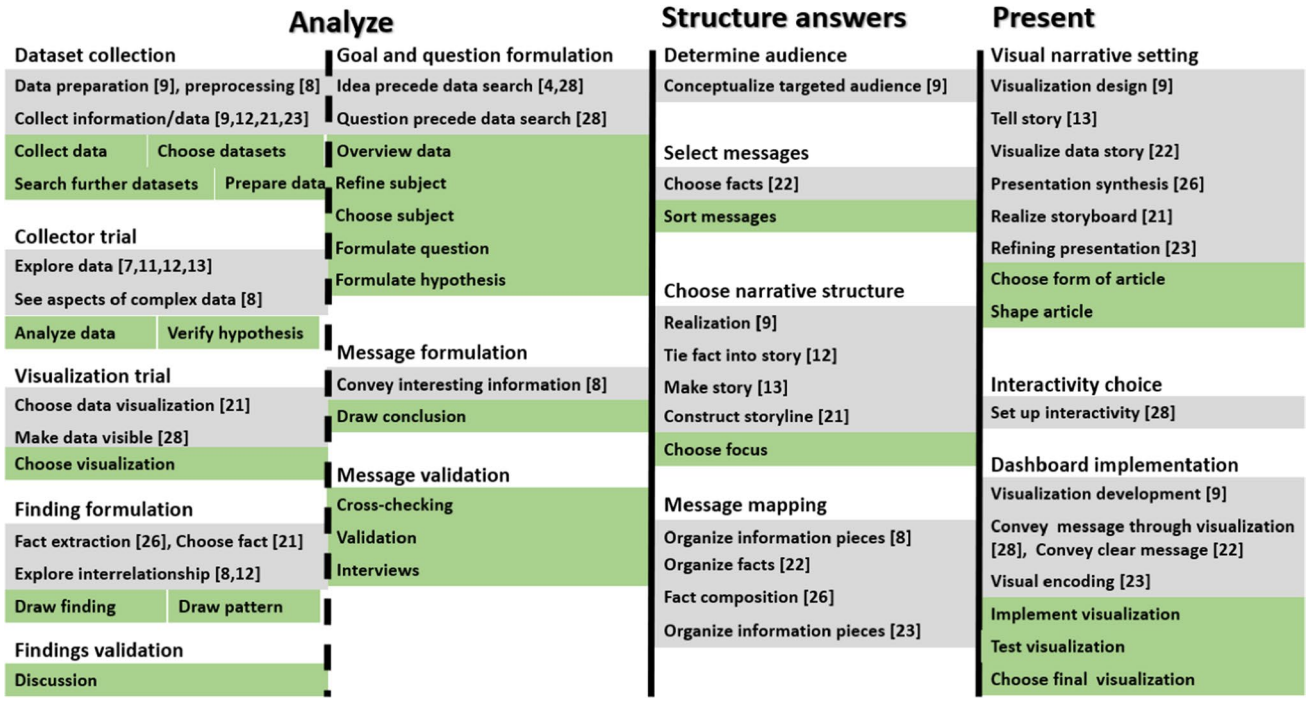


Fig. 3 The main activities for crafting data narratives identified from the literature (in gray boxes) and a survey with data journalists (in green boxes)

that activities pertaining to the factual and intentional layers of the conceptual model

are mixed in phase (i). In addition, while the literature rarely mentions the activities pertaining to the intentional layer, these activities are often pointed by data journalists. Furthermore, as we explained in Outa et al. (2020), the substance of a story, representing the narrator's intention in reporting the story, is a constituent of the data narrative Chatman et al. (1980). Conversely, while the journalists did not attach much importance to the activities of the structure answers phase, this aspect is emphasized in the literature. Finally, as noted in Wang et al. (2019); Lee (2015), the narrator should have the possibility to move freely back and forth between the different phases of data narration. However, this movement should not prevent that different groups of activities could be conducted by different persons with different profiles.

These groups of activities, identified by layers in the conceptual model Outa et al. (2020), should be as isolated as possible.

To summarize, a comprehensive workflow for crafting data narratives should satisfy the following requirements:

- (R_1) cover the activities and the paths identified by the survey with data journalists, reflecting the intention of the data narrator, which are depicted in Fig. 2,
- (R_2) cover the activities of the three phases identified from the literature,
- (R_3) allow the free back-and-forth transition between phases,
- (R_4) clearly delineate the different layers of the conceptual model Outa et al. (2020) within its activities.

5.2 The process of crafting data narratives

In this subsection, we propose a comprehensive process for the crafting of data narratives that covers the activities and paths proposed in the literature and reported by journalists (requirements R_1 and R_2), while also being founded upon and coherent with the conceptual model (R_4) and allowing the back and forth movement between its phases (R_3).

The phases of the process are illustrated in Fig. 4. All phases are accompanied by the resulting outcomes, which are exactly the basic constituents of our conceptual model (R_4). Note that, the incomes of the *structure answers* and *present* phases are more than just the basic constituents; rather, they are the organization of episodes and dashboard components (see Chatman et al. (1980)). We retain the same coloring (pink for factual exploration, purple for intentional question-answering, yellow for the structuring of the answers of the intentional questions into a plot, and blue for presentation).

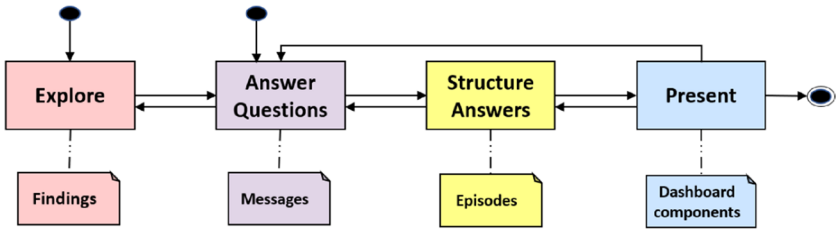


Fig. 4 The process of data narrative crafting

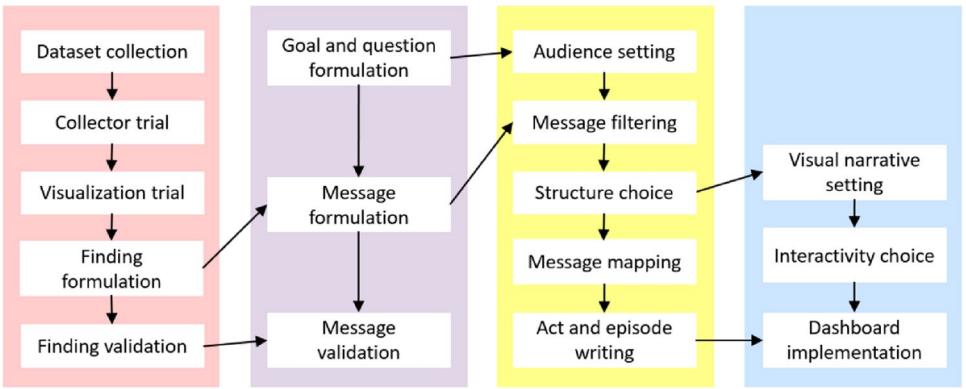


Fig. 5 Activities for data narrative crafting (→ indicates a depends on relationship)

Observe that the factual and intentional layers of the conceptual model are well differentiated here, contrarily to the literature that mix them into one phase.

Consistently with Fig. 2, the process flexibly starts either with the existence of a data set, which is to be explored for findings, or with the emergence of an initiating question to be answered. This flexibility is important in the sense that prescribing a specific starting point for the collection of findings from the data is not what practitioners typically do. The internals allow the flexibility of exploring several paths, that can be chained according to narrator's habits and specificities of the task on hand, alternating the exploration of data, answering questions by deriving messages, structuring the answers and presenting visually the structured answers.

In any case, the identification of the answer questions in terms of messages and their formulation is a task that is practically absent from the related literature, significantly present in the everyday work of practitioners, and structured in our model for the first time.

The following paragraphs present the activities associated with each phase. These activities are abstracted from the literature and survey results (shown in Fig. 3). A new activity *act and episode writing* is added in order to explicitly state the task of conceiving, naming and eventually writing some notes about episodes and acts. In this way, the plot of the data narrative is produced. This activity materializes the concepts of acts and episodes depicted in the conceptual model for data narratives Outa et al. (2020) , which are implicit both in the survey and the literature.

Note that such activities should not be considered as steps to be executed sequentially. Conversely, many activities can be initiated and executed in parallel, and many activities are frequently performed asynchronously. The arrows in Fig. 5 indicate a depends on relationship. For example, message validation depends on message formulation, as it is necessary to formulate messages before validating them. In addition, at any time, it is possible to come back to previously executed activities (e.g. to rewrite messages or formulate new ones). Backtrack arrows are omitted for clarity.

Exploration The exploration phase, handling the factual layer, concerns several activities: (i) dataset collection, concerning source selection, data extraction, integration and preprocessing, (ii) trial and reuse of several collectors (i.e. querying, profiling and mining tools) and (iii) trial of diverse visualizations (crosstabs, graphics, clusters, etc.) for collecting findings, then, (iv) finding formulation, concerning the expression of findings and their relationships, and (v) finding validation, which is typically done via statistical tests, but also by discussing and crosschecking with additional data sources (as done by data journalists) and confronting with the state of the art (as done by data scientists Mbenga et al. (2022)). Note that some findings may lead to additional analysis, triggering more collectors and visualisations, or even the collection of more datasets. The exploration phase is time-consuming (data journalists measure it in days or even in months).

Question-answering This phase, neglected in the literature, handles the intentional layer and concerns activities for (i) formulating goals and questions, (ii) drawing

Scenarios	Regular expressions for crafting data narratives
Exploratory	
Pre-canned	
Question-by-question	
Delegated-presentation	

Fig. 6 Regular expressions representing the unfolding of phases in different scenarios for crafting data narratives. Colored boxes represent phases, respectively pink for Explore, purple for Answer questions, yellow for Structure answers and blue for Present

messages from findings, and (iii) validating messages. It supports explicitly the data narrator intention, as its proposed activities help in formulating an analysis goal and a set of analytical questions that reflect their intention.

Furthermore, to cope with literature lacks (evidenced in Fig. 3), we propose a message formulation activity, concerning the derivation of messages from findings, and the identification of characters and measures (the relevant constituents of messages Chatman et al. (1980)) to be highlighted to the audience.

Structuring The structuring phase, the most overlooked part in the data journalists practices, handles the structural layer, describing activities for organizing the plot of the narrative in terms of acts and episodes Gkesoulis et al. (2015). Plot setting starts by (i) determining the audience, (ii) eventually selecting a subset of messages for such audience, and (iii) choosing an appropriate narrative structure. Then, (iv) messages are mapped to acts and episodes. In more details, these activities allow the arrangement of the messages into different layers: an act which is a major piece of information, and is composed of several episodes that are of lesser importance on their own Outa et al. (2020).

Remember that the result of the structuring is an *episode*, which is the annotation of a message (which has a simple structure and labeling) with comments on the context, significance, essence, etc., in other words with the content that makes the message interpretable by human beings.

Also, observe in Fig. 5, the existence of a specific activity to make the actions of writing acts and episodes explicit. Such activities can be performed before or at the same time as choosing visual means.

Presentation Finally, the presentation phase handles the presentational layer, and includes activities for (i) setting the type of visual narratives, ii) setting the interactivity mode, and (ii) implementing dashboards for conveying acts and episodes to the audience. Such activities carry on the visualization level and build for each act an associated dashboard and present the narration in a complete visual narrative. Remember that *dashboard components* are representations of episodes in (typically) a visual form of communication, including text, figures, charts, data plots, or any other means to convey the message.

5.3 Scenarios for crafting data narratives

The proposed process allows the free back and forth transition between phases (requirement R_3), some paths being more typical in specific situations. This subsection presents several examples of such situations, representing some common unfolding scenarios described by practitioners or observed. Scenarios are identified based on the following: the study about data journalist practices described in Section 4, the analysis of several data narratives and their associated processes published by data journalists, and the observation of several practitioners (as will be detailed in Section 7). The scenarios are sketched in Fig. 6 by means of regular expressions.

An *exploratory* scenario is commonly observed when the analyst does not have in-depth knowledge of the datasets. It represents situations where the narrator only has a vague idea of the analysis goal (or no goal at all), where many iterations of questions-explorations are necessary to formulate and answer clear questions. This scenario contains many activities and transitions between the phases of *explore* and *answer questions*. Once the exploration is completed and messages are validated, next activities can be linearly performed to structure and then present the data narrative. A good example of this scenario is a data journalist's notebook[3] describing the process followed to build a data narrative[4] about covid pandemic in a French region.

In this notebook, the data journalist shows the effort put in the many iterations to collect, clean and explore the data and highlights the formulation and validation of messages.

A *pre-canned* scenario corresponds to crafting processes where goals and questions are well defined from the beginning. It is typically observed for periodic or repeated studies, looking for well-known patterns, for example, reporting the results of an election. In this scenario, phases are chained quite straightforwardly, with no need to come back to precise questions or refine collectors. The structure and presentation are typically reused.

A *question-by-question* scenario consists in chaining all phases one question at a time. In a loop, for each question, an exploration is launched in order to find one or several messages that answer this question. Then, these messages are structured and presented in the rendered data narrative before proceeding with a new question. This scenario

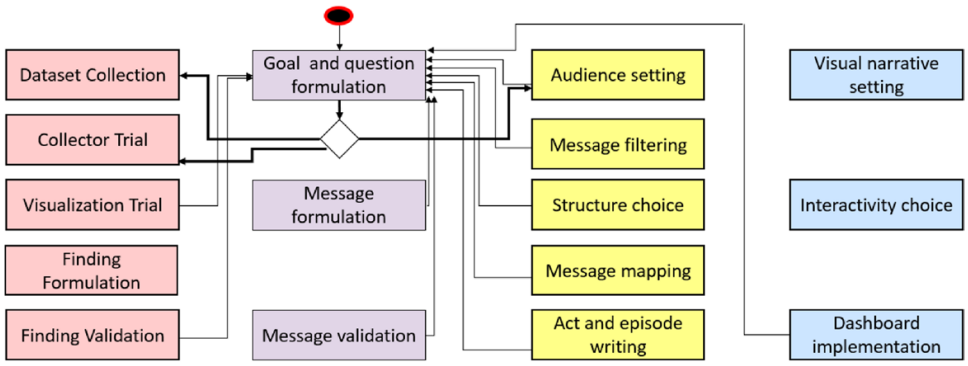


Fig. 7 Goal and question formulation flow

concerns more back and forth transitions among all phases. We observed this scenario with beginners, who tried to order and present messages just after their formulation before posing new questions. Students can even go message by message. On the contrary, professionals tend to express most analytical questions at an early stage.

A *delegated-presentation* scenario corresponds to professional environments where the presentation phase is delegated to a specific team at the end of the process. There can be (or not) among the previous phases, preparing the plot. This scenario was reported by several interviewed data journalists Chagnoux et al. (2020).

6 A Focus on The Answer Questions Phase

The *Answer questions* phase, which has previously been neglected in the literature, covers the intentional layer and its interdependencies with the factual, structural and presentational layers.

In this section, we detail the *workflow* for the answer questions phase that covers the activities and paths reported by data journalists, while also being founded upon and coherent with the conceptual model. Several paths are added based on discussions with data journalists and observations of many data narrators.

The workflow is modeled by three activity diagrams, respectively in Subsections 6.1, 6.2, and 6.3. For readability purpose, each diagram highlights one of the 3 activities of this phase (goal and question formulation, message formulation and message validation), by detailing the incoming and outgoing arcs to facilitate the understanding of the succession of activities. The paths from the activity are depicted with bold arcs, while the paths to the activity are depicted with regular arcs.

6.1 Goal and question formulation

The *Goal and question formulation* activity reflects the high-level intentions of the data narrator, and therefore helps identifying potential data for exploration and influences the structure of the data narrative. Figure 7 depicts its flow.

Incoming arcs. This activity can be the first of the process (top incoming arc), when a goal, and eventually some questions, are initially formulated.

Conversely, it can be entered after some data exploration (incoming arcs from the left). Indeed, while exploring and visualizing a collector's output in order to gain a deeper understanding of the data, possibly, new analytical questions can be posed, seeking for different information and inviting for further exploration. The same may arise after a finding is formulated and validated.

Furthermore, the selection of a previously asked question that was not (completely) answered is also possible.

Later, when data exploration has lead to the formulation and validation of messages, new questions may appear or old questions may be reformulated, so this activity can also be reached after message formulation or message validation (bottom incoming arcs). Many iterations can follow in this way, specially in exploratory and question-by-question scenarios, allowing the expression of new questions after some exploration and some messages, which in turn, will trigger more exploration.

Finally, after partially structuring the plot or even implementing visualizations, the data narrator can come back to formulate new questions (incoming arcs from the right). For example, while mapping messages to acts or writing acts and episodes, a new question can be asked if some missing aspects are detected, in order to complete the plot data narrative. In a question-by-question scenario, this come back to the intentional phase, looking for the following question to deal with, is particularly frequent.

Outgoing arcs. After formulating goal and questions, the natural sequel is to explore data, either by collecting datasets (especially the first time) or just trying or reusing collectors (outgoing arcs to the left). If formulating goal and questions was the first activity in the crafting process, dataset collection is necessary to start exploration. In addition, even after some exploration, a question may require the collection of additional datasets if the available ones lack some information. Conversely, this activity may directly be followed by collector trial to directly explore existing datasets.

Once a goal and some questions have been formulated (and typically some messages have been validated), the data narrator can set the appropriate audience for the data narrative (outgoing arc to the right) and eventually start structuring.

6.2 Message formulation

The *Message formulation* activity concerns the derivation of messages from findings, intended to answer analytical questions. Its flows are depicted in Fig. 8.

Incoming arcs. Message formulation naturally takes place after finding validation. Indeed, during data exploration, some findings arise, which are in turn validated via cross-checking or statistical testing. These findings are then compiled into messages to the audience, highlighting characters and measures. This is the unique incoming arc to this activity.

Outgoing arcs. After formulating a message, there are many options to continue the data narrative crafting. Message validation (outgoing arc to the bottom) is the more natural, allowing for a direct verification of the validity of the message.

But note that the validation of messages is not required to be done immediately after they are formulated. Indeed, some narrators prefer to validate all messages together, especially when such validation involves external experts. This can help save time, allowing the data narrator to explore more data before sending a set of messages to be validated.

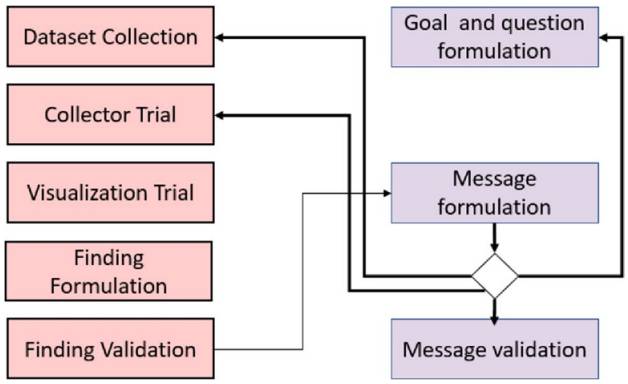


Fig. 8 Message formulation flow

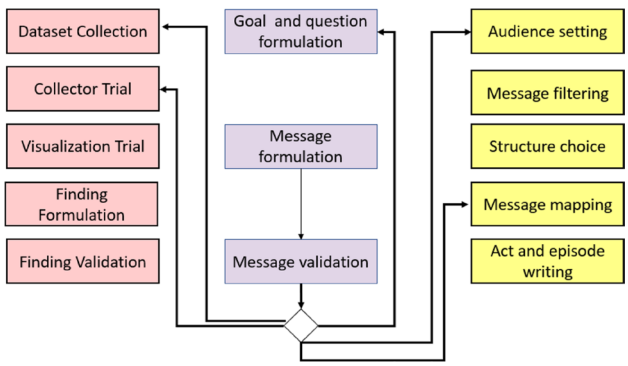


Fig. 9 Message validation flow

The data narrator may prefer to continue exploring data, in order to find additional substance to answer the analytical question at hand. This can be done by collecting a new dataset or trying a collector to further explore an existing dataset (outgoing arcs to the left).

In turn, the narrator can choose to express a new question, or select another existing question to treat (outgoing arc to the top). This latter flow was already explained in the previous subsection.

6.3 Message validation

The *Message validation* activity ensures the validity of messages, typically asking for expert's advice or comparing to the state of the art. Figure 9 illustrates its flows.

Incoming arcs. Message validation comes after message formulation (unique incoming arc). As explained in the previous subsection, this can occur either one by one, immediately after formulating each message, or all messages at a time, after formulating several messages.

Outgoing arcs. After validating a message, there are many options to continue the data narrative crafting. A data narrator may, for instance, pose a new analytical question or continue answering the same question by collecting new datasets or applying a collector to explore an existing dataset.

Also, the data narrator can pass to the *Structure Answers* phase, by setting the audience of the data narrative, or mapping the validated message to acts. The former is done in the first passing to the structure answers phase, typically when enough messages are validated and the analytical questions are reasonably answered, having a good idea of story to be tell. The latter is done when the plot is already initiated and some additional messages arrive, in order to map them to acts and start writing.

7 Experiments

To validate experimentally the proposed process, we organized two challenges and analyzed several publications describing some crafting processes followed by data journalists and data scientists.

The challenges aim at answering two questions: (i) Does the process *cover* all necessary activities performed by data

Table 1 Characteristics of the crafted narratives observed during the workshop

Teams	A	B	C
Topics	Migration	Vegetation	Woman-named streets
Visual narratives	Video	Notebook	Interactive book
Starting from	Vague idea	Precise idea	Vague idea

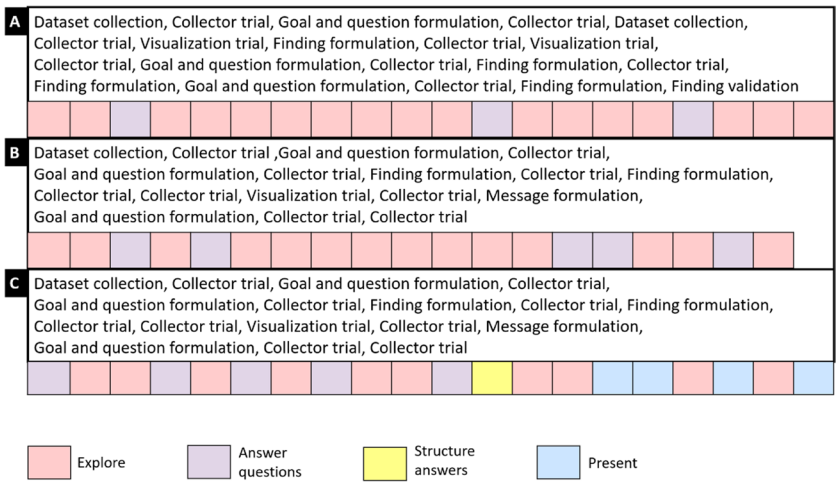


Fig. 10 Activities for crafting data narratives observed during the workshop

narrators? and (ii) To what extent do the process *phases contribute* to the quality of the data narratives? The analysis of published processes aim at answering the question: (iii) Is the proposed process consistent with the reported ones?

Subsection 7.1 addresses the first question. During a challenge in a workshop, we observed several narrators with various profiles while they crafted data narratives for answering the challenge. In particular, we observed whether their actions corresponded to the activities defined in our process. The second question is addressed in Subsection 7.2. An experienced data journalist assessed the quality of data narratives crafted by Master students, and judged the completion of each process phase. Concretely, we investigate the correlation among phase completion and narrative quality.

Subsection 7.3 describes our analysis of some published narratives and the associated processes followed by their narrators. Concretely, we investigate whether the proposed process is coherent with the documented ones, highlighting the scenarios that better represent them.

7.1 Coverage

We organized a one day challenge called "Narrating Rennes by the data"[7] with data enthusiasts and data scientists, aiming at producing data narratives using the open data of the French city of Rennes.[8] Three teams (A, B, C) were constituted, mixing one or two data enthusiasts (among which journalists, students, social workers) and a data scientist. An external observer (lecturer or PhD student in Computer Science) annotated the crafting process followed by each team. In particular, they wrote down the sequences of activities that were performed.

It should be noted that the teams were allowed to continue their crafting work during 3 additional days. During the annotation period (only the initial day, during the workshop), all teams mainly performed exploration and question answering activities; only one team (C) started the structuring and presentation of the data narrative. Importantly, the teams were not asked to follow the process proposed in this paper; only the observers were aware of it.

The details of our analysis and the narratives produced (in French) are publicly available.[9] A prize was awarded to the best one.

Table 1 reports, for each team, the topic of the data narrative, the style of visual narrative, and its starting point. Fig. 10 depicts, for each team, the sequence of activities observed during the workshop. Activities are colored according to the phases of our process.

The main observation one can make from Fig. 10 is that the proposed process covers the data narrators activities and their chaining, whatever their initial idea, the topic chosen, or the style of visual narrative. In more details, we found that each group struggled at the beginning with the choice of the analysis goal and the datasets to use. In all cases, the first explorations did not return any findings (finding formulation activity arrived a bit later after the trial of several collectors). This did not prevent the groups to continue with the narrative crafting, and more importantly, the observers noted that no activity conducted by the group was absent from those listed in Fig. 5.

Interestingly, we remarked that all teams started with a vague idea of the topic they wanted to treat, which was refined after many iterations among data collection, data analysis and question formulation. This clearly correspond to an exploratory scenario. Furthermore, we identified some repeated sequences of activities, e.g. goal and question

[7] Sponsored by CNRS https://www.madics.fr/event/titre1617704707-3351/#madona.

[8] https://data.rennesmetropole.fr/

[9] https://drive.google.com/drive/folders/1zDzP_ndSlQUJCbtFMVzJDnIbyXK1D2_l?usp=sharing (in French)

Table 2 Assessed quality (informativity, comprehensibility, expertise, and average quality) and perceived completion (of answer questions, structure answers and present phases) of data narratives of Master students. We report minimum, maximum, average, and standard deviation for each criteria

		Assessed quality				Perceived completion		
		Info	**Comp**	**Expe**	**Avg_Q**	**C_ans**	**C_str**	**C_pre**
All	**Min**	1	1	1	1	1	1	1
	Max	5	6	5	5.33	6	6	7
	Avg	3.38	3.63	3.21	3.43	3.00	3.67	4.17
	Stddev	1.13	1.50	1.18	1.14	1.44	1.37	1.52
M1	**Min**	1	1	1	1.00	1	1	1
	Max	5	6	5	5.33	5	5	7
	Avg	3.00	3.43	2.86	3.10	2.57	3.29	4.00
	Stddev	1.41	2.07	1.35	1.58	1.27	1.50	2.00
M2	**Min**	2	1	1	1.89	1	2	2
	Max	5	6	5	5.33	6	6	7
	Avg	3.53	3.71	3.35	3.57	3.18	3.82	4.24
	Stddev	1.01	1.26	1.11	0.92	1.51	1.33	1.35

formulation followed by collector trial, which also illustrate the tight link between explore and answer questions phases.

All of them used a unique timeline for structuring their narratives, which were rendered with varied styles.

We can also note that our proposed process remains tailored for the task at hand. Indeed, the activities reported in Fig. 10 cover almost all the activities of our process. Activities that were not reported in Fig. 10, particularly those related to structure answers and present phases, were likely completed after the workshop.

7.2 Phases contribution to narrative quality

For assessing the relationship between process phases and narrative quality, we asked an experienced data journalist to evaluate a set of data narratives, assessing both their quality and the perceived phase completion. Narrative quality was assessed on a scale from 1 (lowest) to 7 (highest), using 3 criteria (previously proposed in El et al. (2020)): (1) Informativity – How informative the narrative is, and how well does it capture dataset highlights? (2) Comprehensibility – To what degree is the narrative comprehensible and easy to follow? (3) Expertise – What is the level of expertise of the narrator? The level of completion of each phase (answer questions, structure answers and present), was deduced from the narrative, as the data journalist was not present during the crafting. The data journalist was asked to assess how much of the *answer questions* phase had been completed, based on how well the data narratives translate the expression of the intention of the data narrator and how much the subject was investigated.

To this end, we implemented a challenge for Master students in Computer Science, specialized in data analysis. 44 students participated in the challenge, 14 of the first year of master (hereafter called M1), 30 of the second year (called M2). Obviously, M2 students have more experience with data analysis and visualization tools, however, all students were familiar with the dataset (they previously did some data cleaning tasks in class) and none of them had previous experience with data narratives. Students were asked to craft a data narrative about fatal encounters in the USA, using an open dataset.[10] They received a one-hour tutorial on data narratives, presenting definitions and examples, and introducing typical crafting activities. Students worked by pairs or alone. We received 7 data narratives from M1 students and 17 from M2 students.

The results of the evaluation are reported in Table 2. Both M1 and M2 students produced narratives graded from 1 to 6, with similar average quality (with less deviation for M2 students), despite their background differences. Students were observed during crafting, and some of them, especially the M1, were asked to indicate the sequence in which they completed the activities depicted in Fig. 5. This helped them to start, particularly having to write down the analytical questions that guided the data analysis, and to write down messages and initially consider how to structure them.

As to the different phases, the present phase was better completed than the two others. In addition, we measured the correlation (using Pearson correlation coefficient) between the average quality (Avg-Q in Table 2) and the completion of the three phases. The correlations were, respectively, 0.7 for answer question completion (C_ans), 0.85 for structure answers completion (C_str), and 0.87 for present completion (C_pre). Interestingly, the completion of the three phases was correlated to the overall narrative quality.

We also measured the correlations between the level of expertise and the completion of three phases, the results being slightly higher for the answer question phase (0.79 for answer question completion, 0.77 for structure answers completion, and 0.73 for present completion). These correlations evidence that the answer question phase influence narrative

[10] https://fatalencounters.org/

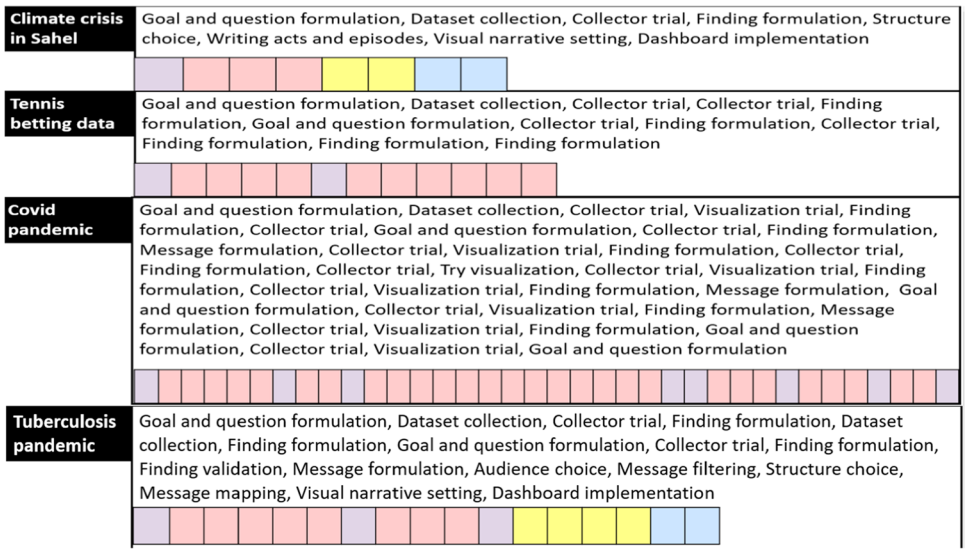


Fig. 11 The activities of documented processes created by various data narrators with specialized skills

quality at least as much as the other phases, which confirms our claim about its importance for data narrative crafting.

7.3 Comparison to documented processes

In this subsection, we study four works that documented (at least some portions) of the crafting process followed to produce a data narrative, namely, (i) a data narrative[11] about *the climate crisis in the Sahel* documented by a data journalist in the form of a blog,[12] (ii) a data narrative[13] about *tennis betting data* documented by an investigative data reporter for BuzzFeed News in the form of a sport news,[14] (iii) a data narrative[4] about *the COVID pandemic in a French region* documented by a data journalist in the form of a notebook[5], and (iv) a data narrative about the *tuberculosis pandemic in Gabon*[15] documented by a data scientist in the form of a research article Mbenga et al. (2022).

Some of the works merely described the main activities accomplished, without detailing every iteration adopted during the crafting process. Other ones only detailed the early phases of the process.

For three of the narratives, namely those about Climate crisis in Sahel, Tennis betting data and Tuberculosis pandemic, the process was clearly described. For analysing them, we just needed to match the activities listed by data narrators to those of our process, highlighting the flow of activities.

However, for the narrative about Covid pandemic, the process is reported by a Python notebook, mostly detailing the data exploration, with references to goals and questions, but few explicit references to messages. Therefore, we also analysed the visual data narrative for matching messages. We proceed as follows: We began by identifying the goal and analytical questions, which were explicitly stated at the beginning of the notebook. Collectors were implemented as python code, results being commented. We identified findings within the data journalist's comments. Then we attempted to locate these findings within the data narrative, looking both for text explaining the finding and a visualization similar to the collector output. When we succeed to match some textual or visual artefact, we take them as the formulation of a message. The activities of structuring and presenting weren't mentioned explicitly by the data journalist.

Figure 11 lists the activities performed in the analyzed processes, which are also sketched as a sequence of boxes, colored as the phases of our process.

As first remark, all the reported processes and activities could be matched to those of our process and the flow between activities is also congruent with our process. In addition, all processes describe many iterations among the initial phases, even if some of them just illustrate some examples of questions and collectors. All of them follow the exploratory scenario. Furthermore, we remark that intentional activities (those of the answer questions phase) are present in all the reported processes.

8 Conclusion

In this paper, we proposed a process for crafting data narratives, that covers the whole cycle of data narration, from data exploration to the visual presentation of the narrative. Importantly, the process reflects the intention of the data narrator by incorporating activities covering the formulation of

[11] https://data.humdata.org/visualization/climate-crisis-sahel/
[12] https://tinyurl.com/ynjzjs63
[13] https://tinyurl.com/zxwf34xt
[14] https://tinyurl.com/wa4jaenj
[15] https://www.youtube.com/watch?v=u_KoBWc_qJU (in french)

their goals, questions, and messages. Backed by a literature review and a survey with data journalists, it accommodates a wide range of practices observed on the field, via clearly delineated activities, while being well founded upon a conceptual model of the domain Outa et al. (2020).

We believe that these two models, static and dynamic, can serve as a stepping stone for future research in the area of data narration. Implementing tools for guiding the narrator all along the process as well as automating tedious or complex tasks is a clear path for future work. We indeed believe that holistic approaches to data narration (from exploration to visual presentation) should be adopted, and we particularly insist on the importance of the intentional phase of the crafting activities. Activities in this phase (e.g., message formulation, message validation) are likely to be the most difficult to automate. This a clear first step to the development of approaches for data narrative manipulation and sharing. Finally, benchmarking data narrative development, not only the final data narration but all steps pertaining to its construction, is another challenge.

References

Bach, B., et al. (2018a) *Narrative Design Patterns for Data-Driven Storytelling*. CRC Press (Taylor & Francis).

Bach, B., Wang, Z., Farinella, M., Murray-Rust, D., Riche, N.H. (2018b) *Design patterns for data comics*. In: CHI. ACM.

Bach, B., Freeman, E., Abdul-Rahman, A., Turkay, C., Khan, S., Fan, Y., & Chen, M. (2023). Dashboard design patterns. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 342–352.

Battle, L., Heer, J. (2019) Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Computer Graphics Forum 38*(3)

Carpendale, S., Diakopoulos, N., Riche, N.H., Hurter, C. (2016) *Data-driven storytelling (dagstuhl seminar 16061)*. Dagstuhl Reports.

Chagnoux, M. (2020) La datavisualisation, double point d'entrée du data-journalisme dans la PQR (in french). *Interfaces numériques* 9(3).

Chatman, S. (1980) *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press.

Chen, S., et al. (2018) Supporting story synthesis: Bridging the gap between visual analytics and storytelling. *TVCG*.

Duangphummet, A., et al. (2021) Visual data story protocol: Internal communications from domain expertise to narrative visualization implementation. In: *VISIGRAPP*.

El, O.B., Milo, T., Somech, A. (2020) Automatically generating data exploration sessions using deep reinforcement learning. In: *SIGMOD*. pp. 1527–1537. ACM.

Gkesoulis, D., Vassiliadis, P., & Manousis, P. (2015). Cinecubes: Aiding data workers gain insights from OLAP queries. *Information Systems, 53*, 60–86.

Kosara, R. (2017) An argument structure for data stories. In: Kozlíková, B., Schreck, T., Wischgoll, T. (eds.) EuroVis.

Kosara, R., & Mackinlay, J. D. (2013). Storytelling: The next step for visualization. *Computer, 46*(5), 44–50.

Lee, B., et al. (2015). More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications, 35*(5), 84–90.

Mbenga, R.O., et al. (2022) A data narrative about tuberculosis pandemic in gabon. In: *Proceedings of the Workshops of the EDBT/ICDT*. vol. 3135.

Outa, F.E., Francia, M., Marcel, P., Peralta, V., Vassiliadis, P. (2020) Towards a conceptual model for data narratives. In: ER. pp. 261–270.

Outa, F.E., Marcel, P., Peralta, V., da Silva, R., Chagnoux, M., Vassiliadis, P. (2022) Data narrative crafting via a comprehensive and well-founded process. In: ADBIS.

Park, D.G., et al. (2022) Storyfacets: A design study on storytelling with visualizations for collaborative data analysis. *Information Visualization 21*.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *TVCG, 16*(6), 1139–1148.

Shi, D., et al. (2021a). Calliope: Automatic visual data story generation from a spreadsheet. *TVCG*, *27*(2), 453–463.

Shi, D., Sun, F., Xu, X., Lan, X., Gotz, D., & Cao, N. (2021b). Autoclips: An automatic approach to video generation from data facts. *Computer Graphics Forum 40*(3), 495–505.

Sun, M., Cai, L., Cui, W., Wu, Y., Shi, Y., & Cao, N. (2023). Erato: Cooperative data story editing via fact interpolation. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 983–993.

Thudt, A. F., Perin, C., Willett, W., & Carpendale, S. (2017). Subjectivity in personal storytelling with visualization. *Information Design Journal, 23*(1), 48–64.

Vassiliadis, P., Marcel, P., & Rizzi, S. (2019). *Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP*. Syst: Inf.

Wang, Z., Dingwall, H., Bach, B. (2019) Teaching data visualization and storytelling with data comic workshops. In: CHI. ACM.

Wang, Y., et al. (2020). Datashot: Automatic generation of fact sheets from tabular data. *TVCG, 26*(1), 895–905.

Weber, W., Engebretsen, M., & Kennedy, H. (2018). Data stories: Rethinking journalistic storytelling in the context of data journalism. *Studies Community Science, 18*, 191–206.

Faten El Outa a doctoral candidate at the University of Tours, is currently engaged in developing a framework for data narratives. Her research revolves around the modeling of concepts and workflows to present a comprehensive approach to data narratives.

Patrick Marcel is an Associate Professor at the University of Tours, France. His current research focuses on OLAP and data warehousing, recommender systems, exploratory data analysis and data narration. Patrick served as program committee member in top tier international conferences, including ER, VLDB, EDBT. He is a member of the steering committee of DOLAP and a member of the regular editorial board of DKE.

Veronika Peralta is an Associate Professor at the University of Tours (France) where she is head of Computer Science department. Her current research interests include data and information quality, exploratory data analysis, business intelligence and data narration. She served as program committee member and guest editor in many international conferences and journals.

Panos Vassiliadis is a professor at the University of Ioannina, Greece. His research focuses on the rigorous modeling of data, software, and their interdependence. Currently he works in the areas of business intelligence and schema evolution. He is a senior member of both ACM and IEEE. More information is available at https://www.cs.uoi.gr/~pvassil.