Open Data Explorer

Chokki, Abiola Paterne; Frénay, Benoît; Vanderose, Benoit

Publication date:
2022

Document Version
Peer reviewed version

Open Data Explorer: An End-to-end Tool for Data Storytelling using Open Data

Completed Research

Abiola Paterne Chokki
University of Namur
abiola-paterne.chokki@unamur.be

Benoît Frénay
University of Namur
benoit.frenay@unamur.be

Benoît Vanderose
University of Namur
benoit.vanderose@unamur.be

Abstract

Enabling users to better understand government actions is one reason why governments have sought to open their data to the public. Data storytelling tools help to achieve this goal by facilitating users to turn data into stories. However, existing tools are not able to provide the necessary features to overcome the barriers users face at different data storytelling stages. This paper provides findings regarding the features in the design of these data storytelling tools in the open data context and also presents a generic and end-to-end tool called ODE, which helps users through the different data storytelling stages. To achieve the paper's objectives, a literature review was first conducted to collect the features needed for the different data storytelling stages. Then, the identified features were integrated into ODE and its effectiveness in helping users to easily turn data into stories was demonstrated through an evaluation involving 11 users.

Keywords

Open data, features, data storytelling, end-to-end and generic tool.

Introduction

Around the world, many governments at the national, regional and local levels, have published their data on the internet for free use and redistribution by anyone (Attard et al. 2015; Handbook 2015). This movement has been referred to as "open data" and aims, among others things, to help users to better understand governments actions and allow for deeper, and easier monitoring of government work (Brolcháin et al. 2017; Silva and Santos 2019). However, to achieve this goal, users must first turn these published data into stories to generate new knowledge (Matos and Corbett 2019). Data storytelling can be explained as translating data analysis into simple, logical stories that can be understood by a non-technical audience (Brolcháin et al. 2017). It uses graphs that make sense and weave them into compelling, action-inspiring stories[1]. Thus, data storytelling can be considered as a combination of data visualization with addition of analysis or description of the visualization to explain the visualization. In this research, the term "stories" refers to dashboards composed of charts with or without analysis.
We are aware of tools used in the literature to integrate open data as Linked Data and query data on web (Ansari et al. 2022; Böhm et al. 2012; Dadzie and Rowe 2010; Eberhardt and Silveira 2018; Wenzel et al. 2013), however in this study we will focus on tools proposed to facilitate data storytelling. An example of such tools are business intelligence tools (e.g., Tableau[2], Power BI[3]) that are primarily designed for businesses and help them to monitor their performance, but they can be also used in the open data context. However, they require users to download the data before using it and require a steep learning curve (Graves

[1] https://www.storytellingwithdata.com/home
[2] https://www.tableau.com/
[3] https://powerbi.microsoft.com/

and Hendler 2013). Another example are web-based tools (e.g., Datawrapper[4], Google Studio Data[5], The Gamma (Petricek 2017)), which are easier to use, but like their predecessors, do not allow direct connection with open data portals or collection of feedback on stories. On the other hand, generic tools designed for open data, for example, OpenDataVis (Graves and Hendler 2013), SPOD (Cordasco et al. 2017), ChartViz (Pirozzi and Scarano 2016), YDS (Brolcháin et al. 2017) allow for the integration of open data, but have limited features (e.g., in SPOD, there is no feature related to data processing, such as the ability to get a quick overview of columns or data quality or to combine datasets) and then require users to use multiple tools. In summary, despite the fact that many tools have been proposed in the data storytelling process, none of them have implemented the necessary features to process open data across the whole data storytelling process (i.e., from the data collection stage until the deployment and feedback stage) (see table 1). Moreover, to our knowledge, only a few studies (e.g., Brolcháin et al. 2017) have been conducted to propose a list of features in the design of a generic and end-to-end data storytelling tool for open data.
In order to address these gaps, this study aims to identify a list of features (e.g., direct connection to open data portals, data quality estimation) needed in the design of a separate and generic data storytelling tool, implement these features into a usable tool called ODE (Open Data Explorer) and evaluate it. Therefore, our research question is formulated as follows: *"What features are needed in the design of a separate, generic, and end-to-end data storytelling tool in the open data context?"* To answer the research question, we conducted a literature review for the features that need to be integrated into a data storytelling tool to cover all its stages in the open data context. Once the features were identified, we implemented them into an end-to-end tool called ODE. Unlike existing tools, ODE provides additional features to facilitate the data storytelling by users such as: direct connection to open data portals, data quality estimation, data overview, visualization recommendation from selected data, and direct feedback collection on story. We conducted interviews with 11 users to assess whether ODE is easy to use and useful through all stages of data storytelling, but also to gather suggestions for additional features to implement. In this study, we focused on two types of users: users with low data manipulation skills (e.g., citizens and journalists who can at least use basic data manipulation features in Excel such as filtering, sorting, grouping, and changing data types) and users with high data manipulation skills (e.g., researchers, developers).
The remainder of this paper is structured as follows. We present the background related to data storytelling and existing data storytelling tools, explain the methodology used to address the research questions, present the proposed ODE tool and its evaluation, discuss the findings and limitations of this study, and conclude with a summary of the contributions as well as avenues for future work.

Background

In this section, we first clarify the concept of data storytelling. Then, we present its stages in the OGD context. Finally, we present tools used in the literature to facilitate data storytelling.

Data Storytelling

Data storytelling can be explained as a process of translating data analysis into simple, logical stories that can be understood by a non-technical audience (Brolcháin et al. 2017). It can also be seen as a process that consists of using graphs that make sense and weaving them into compelling, action-inspiring stories[1]. A well-known subfield of data storytelling in the literature is the data journalism, where journalists make use of large databases to produce stories (Gray et al. 2012; Kalatzi et al. 2018). In the context of open data, in addition to turning data into stories for data exploration or development of digital services, data storytelling is also about users using datasets published on open data portals for the following purposes: to better understand governments actions and to enable deeper and easier monitoring of government work (Brolcháin et al. 2017).
The data storytelling process is subdivided into 6 stages, as shown in Figure 1 (adapted from (Aanderud et al. 2020; Brolcháin et al. 2017)). The process begins by looking for answers to an identified need or question *(seeking answers* stage*)*. Next, the user tries to identify and collect the datasets needed to answer the question *(data collection* stage*)*, followed by the stage where the user processes the collected data, for example by grouping it or deleting certain rows or columns, to keep only the relevant information *(data processing* stage*)*. Then, the user can create visualizations from the processed data to facilitate

[4] https://www.datawrapper.de/
[5] https://datastudio.google.com/

understanding *(data visualization* stage*)*. Finally, the user can accompany the different visualizations with an interpretation or a small description and share them with other users to present their findings about the studied data *(story creation* stage*)*. Once the story is shared, the user can receive feedback to improve the story or to engage in a discussion with other users *(feedback collection* stage*)*.

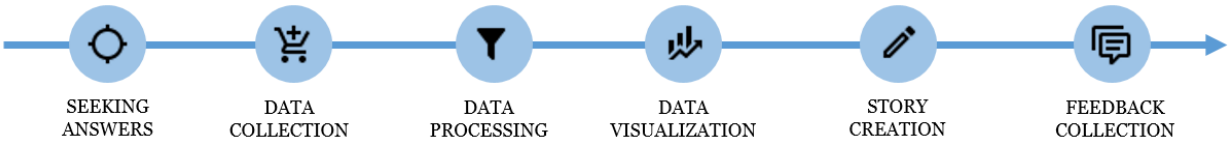

Figure 1. Data storytelling process

Data Storytelling Tools

Many tools have been proposed in the literature to help users to turn data into stories in the context of open data (Ansari et al. 2022; Eberhardt and Silveira 2018). Reviewing each of them would be beyond the scope of this paper. We have focused here only on the popular and generic tools that have been used in the context of open data. We are aware of Linked Open Data (LOD) tools in the literature (Ansari et al. 2022; Böhm et al. 2012; Dadzie and Rowe 2010; Eberhardt and Silveira 2018) but we focused on data storytelling tools for open data in tabular format for two reasons. First, many users, especially those with low technical skills, are more familiar with tabular data than the RDF format which is used in LOD. Second, most of the datasets available on open data portals are published in plain text (CSV) (see e.g., data.gov (US), data.europa.eu (EU)). We grouped these tools into two categories: non-open-data oriented tools (e.g., Tableau, D3) and open-data oriented tools (e.g., ChartViz, SPOD). Non-open-data oriented tools are tools whose primary purpose was not designed for the open data context, but which can be used in that context. The second category includes tools that are initially designed to be used for open data. Table 1 presents the tools reviewed including a short description, their strengths (data storytelling features implemented), and weaknesses (data storytelling features not implemented).

Tools (including a short description)	**Strengths (+) & Weaknesses (-)**
Non-open-data oriented tools	
BI Tools (e.g., Tableau[2], Power BI[3], Qlik[6]) are mainly used by businesses and help them to understand trends and deriving insights from their data so that they can make tactical and strategic business decisions	(+) powerful features from data processing until story creation (+) handle millions of rows of data (-) no direct connection to datasets available on portals (-) no overview of the content of each data column (-) no evaluation of the quality of data and metadata (-) need for technical skills (e.g., SQL) before merging data (-) steep learning curve (-) no information about how a visualization was created and what data was used (-) no customization of others' stories (-) no direct feedback on the story
Web Tools (e.g., Datawrapper[4], Google Data Studio[5]) are mainly designed as web applications and their target users are mainly people with low technical skills	(+) no need to install software before using them (+) basic and easy to use features from data processing until story creation (+) no collaboration on story creation (-) same shortcomings as the BI tools, except that these tools are easier to use than the previous tools
The Gamma (Petricek 2017) provides a scripting language for working with data and produce reproducible source code making the analysis fully transparent and open	(+) same strengths as Web tools (+) information about how a visualization was created (+) customization of others' stories (-) no direct connection to datasets available on portals (-) no overview of the content of each data column (-) no evaluation of the quality of data and metadata (-) need for writing code to create visualizations or perform advanced functions such as merging datasets (-) no collaboration on story creation (-) no recommendation visualizations from selected data (-) no multiple visualizations in a single story (-) no filters on the story (-) no direct feedback on the story
Open-data oriented tools	
Open data portals (e.g., data.gov (US), data.europa.eu (EU)) are mainly used by public organizations to publish their data to the public but also provide visualizations	(+) direct interaction with open datasets on the specific portal (+) easy-to-use features for data visualization (-) no interaction with open datasets from other portals (-) only the data visualization stage is supported

[6] https://www.qlik.com/

YourDataStories (YDS) (Brolcháin et al. 2017) is a European Horizon 2020 project which provides a list of features to integrate data storytelling into open data platforms	(+) predefined stories about public projects and contracts in European countries using maps and graphs (-) as of now, only focused on public projects and contracts in EU countries (-) need to register or log in before accessing the main features (-) no generic tool to allow any users to create their stories using open dataset on portals
OpenDataVis (Graves and Hendler 2013) & (Graves and Bustos-Jiménez 2015) are complementary and allow to interact with the data of any portals by providing a data URL. OpenDataVis provides an easy way to visualize data in less than 5 clicks. On the other hand, (Graves and Bustos-Jiménez 2015) produces a series of visualizations describing the variables of the selected dataset	(+) interaction with a dataset from any portals by providing a data URL (+) provide an easy way to visualize data (+) information about how a visualization was created (+) customization of others' visualizations (+) overview of the data distribution of open dataset (-) no direct connection to open datasets from portals (users need to enter the data URL manually) (-) no use of multiple datasets in a single visualization (-) only handle data overview and data visualization
ChartViz (Pirozzi and Scarano 2016) creates charts from a remote csv open dataset based on a decision tree algorithm	(+) interaction with a dataset from any portals by providing data URL (+) quantitative measure of data quality (homogeneity and completeness) of open dataset (+) provide an easy way to visualize data (-) same shortcoming as OpenDataVis (Graves and Hendler 2013)
DEEP (De Donato et al. 2017) & SPOD (Cordasco et al. 2017) are complementary and enable the creation of interactive, reusable, and shareable visualizations	(+) direct connection to open data portals (+) creation of interactive, reusable and shareable visualizations (+) column completeness information (-) no estimation of data quality (-) no overview of the selected open dataset (-) no use of multiple datasets in a single story (-) no multiple visualizations in a single story (-) no recommendation visualizations from selected data

Table 1. Strengths and weaknesses of data storytelling tools

In summary, although many tools have been proposed for data storytelling, they still lack some features to fully support users to process open data across the different stages of the data storytelling process (see Table 1). To address this gap, we proposed a generic and end-to-end tool called ODE, which will be described in the following sections. ODE differs from existing tools in that it addresses all of their mentioned shortcomings by providing a generic tool to turn any open dataset into stories, providing an end-to-end tool to cover all data storytelling stages, allowing a direct connection to open data, providing a quick overview of the data, providing an estimation of the data quality, allowing multiple datasets to be used in a single story, providing visualization recommendations on the entire data and an easy way to create embeddable and interactive visualization from scratch by simply selecting attributes, integrating up to 17 visualization types, and allowing users to customize existing stories even they are not the owners.

Research Methodology

The research questions of this paper were addressed using the Design Research Science (DSR) methodology (Baskerville 2008; Dresch et al. 2015; Hevner et al. 2004; Hivon and Titah 2017; Peffers et al. 2007; Vaishnavi and Kuechler 2007), as it aims to develop solutions (artefacts, design cycle) that meet defined objectives, contribute to the scientific knowledge base (rigor cycle), and provide utility in the environment (relevance cycle). The methodology of (Hevner et al. 2004) was adjusted in this research because we found that users were less proactive when we simply came to them to ask about their needs rather than presenting them with a prototype built on the basis of a literature review to obtain additional features and feedback.

In the **rigor cycle**, we first conducted a systematic literature review based on the established procedure (Kitchenham, Barbara Brereton et al. 2009) to access existing knowledge on the features needed to be integrated into a data storytelling tool to cover all its stages in the open data context and also on the barriers encountered by users. The literature review was conducted using the databases "Google Scholar" and "Science Direct" with the keywords "open government data" or "open data", + "technical features" or "technical barriers", + "(re)use" or "visualization" or "dashboard" or "data storytelling". Based on the automated search, we obtained 107 articles. We then identified relevant articles in three stages: first, we evaluated the type, domain and title; second, we examined the abstract; and finally, we scanned the content. In the end, we retained 13 articles that were relevant to our research. Features were then collected directly from the retained articles or inferred from the barriers identified in the articles. The literature review along with the feedback collected from users on the *design* and *relevance cycles* will be used to improve the

current knowledge base. This is detailed in *Conclusion Section* by positioning the contributions to the current literature.

In the **design cycle**, we implemented the features gathered in the *rigor* cycle into a generic and end-to-end tool called Open Data Explorer (ODE). ODE is built using three technologies: Python as the programming language, Pandas as the data processing library, and Plotly to create and display the visualizations. An incremental approach based on the agile methodology (Fowler and Highsmith 2001) was used during the implementation of ODE. Once we implemented 2-3 features in ODE, we presented them to 2 users (one with high data manipulation skills and another with low data manipulation skills) to collect additional features, get their feedback and to improve the user interface.

In the **relevance cycle**, we evaluated the overall ease of use and usefulness of ODE as well as each of the implemented features collected during the *rigor* and *design* cycles. The evaluation was conducted through interviews followed by an online survey[7] to assess the usability and usefulness of ODE and to gather additional features for future versions. The survey was pretested with two users to ensure that all kinds of errors associated with survey research were reduced (Grimm 2010). The survey included three types of questions: questions with a 7-point Likert scale (from "Strongly Disagree" to "Strongly Agree") based on the Technology Acceptance Model (TAM) (Davis 1989; Moreno Cegarra et al. 2014) to evaluate two aspects: ease of use and usefulness; free-text questions to collect general opinions and suggestions for additional features for future versions and to justify previous ratings; and 3 additional questions to collect demographic data (level of data manipulation skills, age and education level). To recruit participants, a recruitment survey was sent through UNAMUR mailbox and social media and 11 participants (6 with high data manipulation skills and 5 with low data manipulation skills) were selected from that. Before completing the survey, participants were invited to test ODE with their preferred datasets on the Namur or Liege (Belgium) portals. Their tasks were to analyze their selected datasets and created stories from them. During the ODE test, we adopted an exploratory approach (Rubin and Chisnell 2008), i.e., we let participants did what they considered to be the right action and guided them only when they felt confused and asked for our assistance. Once users completed the survey, they were asked four questions to get their overall opinion of the implemented features: what features should be *kept, improved, removed, or added* to facilitate data storytelling? After collecting user feedback, the median, mean, and standard deviation (SD) were calculated for the questions with a 7-point Likert scale. These statistical measures were chosen because they are the most appropriate for analyzing Likert data and for having a central tendency measure (Boone and Boone 2012). Verbal thoughts and responses collected from the free-text questions were coded using short sentences to retain context and conceptual relations.

Results

In this section, we first presented the features identified from the literature review. Then, we presented how they were implemented. Finally, we reported on the results of the evaluation of ODE.

Features of a Data Storytelling Tool in the Context of Open Data

Table 2 presented the 15 features that we identified by conducting the above literature review. They were either proposed by previous studies or inferred from barriers identified therein. These features are grouped with respect to the different stages of the data storytelling process. No feature has been proposed for the *seeking answers* step, as this step is left to the users to decide if they just want to explore the data or if they already have a specific goal in mind.

<table>
<tr><th colspan="2">Stage/Feature</th></tr>
<tr><th>Data Collection (DC)</th><th>Story Creation (SC)</th></tr>
<tr><td>DC1. Direct access to open datasets from portals (Cordasco et al. 2017; De Donato et al. 2017; Graves and Bustos-Jiménez 2015; Graves and Hendler 2013; Pirozzi and Scarano 2016)</td><td rowspan="2">SC1. Facilitate the creation of story that is easy to understand, use, and trust by potential users (Chokki, Simonofski, Frénay, et al. 2022a; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013)
SC2. Share story (Brolcháin et al. 2017; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013)</td></tr>
<tr><td>Data Processing (DP)</td></tr>
</table>

[7] https://forms.gle/mymVDHNjmoYmkV7i6

DP1. Get a quick overview of the data content (Crusoe et al. 2019; Graves and Bustos-Jiménez 2015) **DP2.** Evaluate data quality (Brugger et al. 2016; Crusoe et al. 2019; Janssen et al. 2012; Pirozzi and Scarano 2016; Zuiderwijk et al. 2012; Zuiderwijk and Janssen 2014) **DP3.** Filter useful data (Brolcháin et al. 2017; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013) **DP4.** Combine multiple data (Brolcháin et al. 2017; Crusoe et al. 2019; Graves and Hendler 2013)	**SC3.** Embed story (Brolcháin et al. 2017; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013) **SC4.** Get information about the story (learning tool) (Chokki, Simonofski, Frénay, et al. 2022a; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013) **SC5.** Customize story (Chokki, Simonofski, Frénay, et al. 2022a; Graves and Hendler 2013)
	Feedback Collection (FC)
	FC1. Give feedback (Chokki, Simonofski, Frénay, et al. 2022a; Cordasco et al. 2017)
Data Visualization (DV)	**Other Features (OF)**
DV1. Facilitate the creation of interactive visualization and provide instant visualizations (Brolcháin et al. 2017; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013; Pirozzi and Scarano 2016) **DV2.** Download or embed visualization (Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013)	**OF1.** Collaborate on story (Cordasco et al. 2017; De Donato et al. 2017) **OF2.** Ease of use and shallow learning curve (Chokki, Simonofski, Clarinval, et al. 2022; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013; Pirozzi and Scarano 2016)

Table 2. Features needed in the design of a data storytelling tool in the open data context

Open Data Explorer (ODE)

ODE is a web application available at https://rb.gy/olmekk. A Video showing the steps performed by the first author to create a story from COVID19 hospitalizations in Belgium collected from the Namur (Belgium) portal[8] is available at https://rb.gy/cor6qt. Screenshots are also available at https://rb.gy/atxowj.

Project Creation. Before moving to the data collection stage, ODE has implemented a workplace concept that helps users to collaborate on a single story *(OF1)*. Thus, before users begin collecting their data, they need to create a project by providing information about its name, description, topic, country/state, and contact information. After filling this information, the system then generates the project with a unique code that can be used to later modify the project or collaborate *(OF1)*.

Data Collection. In this stage, users can search for the data they want that is connected to any CKAN or OpenDataSoft portals *(DC1)*. ODE relies on the APIs provided by these two open data management systems (ODMS) to allow users to collect their data directly from a portal. For now, we have integrated these two ODMS because they are among of the most widely used in the European countries (Berends et al. 2020).

Data Processing. In this stage, once users have selected a dataset, the system presents in table form the content of the selected data, a quick overview of each column of the dataset *(DP1)*, a data quality *(DP2)*, a correlation between the numerical columns and an auto-detection of the data type of each column with the possibility of adjusting the proposed data type. For the data type detection feature, the system relies on regular expressions and data type auto-detection in Pandas library. Currently, ODE covers four data types: numeric (integer, float), categorical (nominal, text, boolean), temporal (date), and geographic (latitude, longitude, geo point). For the quick overview of each column *(DP1)*, ODE proposes graphs with the information about the data type, the percent of missed values in the column and the list of columns that are correlated with the column in case that the column is numeric. ODE offers two types of graphs: a histogram for a numeric column and a word cloud for a categorical column. ODE incorporates four data quality metrics: completeness of cells, completeness of column labels, completeness of column descriptions and completeness of data information such as title or description or timeliness (last modified). The average of these four metrics is then calculated and corresponds to the data quality of the selected dataset. For the correlation feature, ODE shows a heatmap where each cell corresponds to the correlation between two numerical columns. After the users have the important information on the dataset, ODE suggests them four functionalities to process their data before using it *(DP3)*: drop columns, aggregate data, combine data *(DP4)* in case they have selected more than one dataset, search and replace some values.

Data Visualization. In this stage, ODE offers two features: create a visualization or get visualization recommendations *(DV1)*. For the feature of visualization creation *(DV1)*, ODE offers a drag-and-drop option similar to Tableau for users with the difference that in ODE, users do not need to map each selected column

[8] https://data.namur.be/

to the shelves (rows, columns in the case of Tableau) themselves. They simply drag and drop their desired columns into a single shelf and ODE can automatically suggest multiple visualizations that best suits their selected columns based on the decision tree produced by this previous study (Holtz and Conor 2018) and additional rules provided by (Chokki et al. 2021; Munzner 2014; Wilke 2019). For example, if users selected a nominal column and a numerical column with a maximum function applied on the numerical column, then ODE will generate visualizations using the following visualization types: bar chart, pie chart, doughnut, and treemap. Contrary to other tools, the decision tree in ODE can be also improved by integrating the user feedback on the suggested visualizations. Currently, ODE proposed 17 visualization types: boxplot, violin, histogram, scatter, bubble plot, correlogram, bar, line plot, area Plot, connected scatter, bubble map, map, doughnut, pie, sunburst, treemap, and parallel coordinates. For the visualization recommendations feature *(DV1)*, once users have selected their dataset, ODE can suggest several useful visualizations. ODE first detects significant columns. For numerical columns, significant ones are correlated with more numerical columns. For categorical columns, significant ones have few distinct values. Once significant columns are detected, we combine one to two significant categorical columns with one to four numerical columns and apply an average transformation to the numerical columns. After obtaining these combinations, we used the decision tree to generate a visualization for each combination. Each generated visualization is interactive, has a button at the top that allows users to easily embed them into other web pages *(DV2)* and also has a rating option to collect user feedback on the visualization and use it to improve the decision tree.

Story Creation. In this stage, whenever users create the visualization of their choice, they can add it to a story by simply clicking to a button "Add to dashboard" situated on the top of the visualization *(SC1)*. After adding all their desired visualizations to their story, they can then configure some dashboard settings (optional as default settings are already defined) *(OF2)* such as adding filters for the dashboard to allow users to interact with the dashboard, adding the title, description and width of each visualization. Once these parameters are defined, they can save them and the dashboard is generated. The 16 dashboard design principles (eg., use best visualization practices, use the right type of chart, integrate feedback support, allow customization) summarized in (Chokki, Simonofski, Frénay, et al. 2022a) were incorporated into ODE to ensure that the dashboard generated by ODE follows the best practices and thus is easy to use and understandable by the end users. They also help to propose a presentable design of the dashboard without the need for users to do many settings before having an attractive and interactive dashboard *(SC1, OF2)*. In the generated story, users have the following options: view dashboard information (e.g., see open data used in the story, see previous user comments), share dashboard (share the story with others by mail) *(SC2)*, embed dashboard (integrate the dashboard into other web pages), give feedback (submit comments on the story) *(SC3)*, customize dashboard (ODE will automatically duplicate the current story into another story and let users edit it as they wish) *(SC5)*, visualization details (e.g., title, description, visualization marks, and open data used to create the visualization) *(SC4)*.

Feedback Collection. In this stage, users can view or give the feedback on any stories through two other applications: ODEON (Chokki, Simonofski, Clarinval, et al. 2022) and CitizenApps (Chokki, Simonofski, Frénay, et al. 2022b) which are linked to ODE (FC1). Once users generate their story using ODE, their story is automatically published on ODEON and CitizenApps which are platforms that allow users to get feedback of their submitted projects. So, instead of implementing the feedback option in ODE, we simply used the application APIs to directly post the stories published on ODE.

Evaluation Results and Analysis

Through the surveys that participants completed after exploring the ODE prototype, we were able to collect their opinions related about the ease of use and usefulness of ODE and also of each feature implemented in ODE to turn data into stories. A total of 11 participants (5 with low data manipulation skills and 6 with high data manipulation skills) completed the surveys. All participants are between the ages of 18 and 50 and have at least a high school degree. Each participant's evaluation section lasted a maximum of 1 hour.

Table 3 presents the median, mean and standard deviation of the questions with a 7-point Likert scale regarding the 2 aspects (perceived ease of use and perceived usefulness) evaluated for the prototype. The following conclusions can be drawn from the results of Table 3. First, most of the participants with low data manipulation skills agree that the proposed prototype is easy to use, as evidenced by the median and mean of perceived ease of use ≥ 5 and the low standard deviation (SD = 0.93) showing that there is no high significant difference between users' scores. They also agree that the prototype is useful to them to better

explore open data and turn them into stories (median & mean ≥ 5 for perceived usefulness and there is no high significant difference between users' scores (SD = 1.02)). Second, participants with high data manipulation skills find it easier to use the prototype than the participants with low data manipulation skills as evidenced by the median equaling 6. They also agree that the prototype is useful to them in better exploring open data and turning them into stories (median & mean ≥ 6 for perceived usefulness and there is no high significant difference between citizens' scores (SD = 1.02)).

	Participants with low data manipulation skills		**Participants with high data manipulation skills**	
	Median	**Mean (SD)**	**Median**	**Mean (SD)**
Perceived ease of use	5	5.4 (0.93)	6	5.39 (1.10)
Perceived usefulness	6	5.87 (1.02)	6.5 (1.10)	6.1 (1.16)

Table 3. Median, mean and standard deviation (SD) of survey scores

These observations can be justified as follows. First, many participants, even those with high data manipulation skills, found the prototype's interface not too intuitive, but they all agree that they could quickly become proficient if they had more time or if we had done a tutorial at the beginning of the evaluation to show the basic features. Second, all participants were able to create a dashboard from their selected dataset during the evaluation section.

Regarding the ease of use and usefulness of implemented features, all participants found that the implemented features presented in Table 2 are useful in turning their selected dataset into stories, but made the following suggestions. Users with low data manipulation skills suggested that the features related to data processing (*DP1*. Get overview to *DP4*. Combine datasets) to be removed because they are difficult for them to understand. On the other hand, users with high data manipulation skills think that these features are necessary. All of them also found the functionality of recommending visualizations from the selected dataset (*DV1*) useful, as many of them were able to find the visualizations they wanted to create their dashboards from this feature. In addition, they suggested making the interface design more intuitive. However, participants did not suggest any additional features, as many felt that existing features were sufficient and that it was best to avoid making the application more difficult to use.

An important limitation of this research is the representativeness of the participants in the evaluation. The number of participants may be small, but based on previous studies (Faulkner 2003; Nielsen 2000), using at least 5 participants for usability tests is a good baseline. To increase this representativeness, we suggest using other communication channels or collecting data on-site in universities or public places. This was not feasible due to the COVID-19 situation.

Conclusion

The goal of this paper was to identify a list of features needed in the design of a separate and generic data storytelling tool. To achieve this goal, we first conducted a literature review and discussed with potential users to gather features needed in the design of a data storytelling tool in the open data context. Next, we implemented the ODE prototype based on the collected features, and then examined through an evaluation conducted with 11 participants, whether the prototype and each of its features were easy to use and useful in helping users with low or high data manipulation skills to turn their data into stories.

This research contributes to theory in the following aspects. First, this research builds on the previous studies and tools (Brolcháin et al. 2017; Cordasco et al. 2017; De Donato et al. 2017; Graves and Hendler 2013; Pirozzi and Scarano 2016) to propose a list of 15 features needed in the design of a data storytelling tool in the open data context (see Table 2 and Section Open Data Explorer (ODE)). Second, unlike the project YDS (Brolcháin et al. 2017) that focused on 4 aspects (discovery, assistance, insight and leverage) to extend open data platforms with data storytelling features, this research focuses on the technical features needed by a generic tool to facilitate the data storytelling with open data and also categorizes the features according to the different stages of data storytelling. This categorization can better help designers or developers to know what features to implement to cover the needs of users at a specific stage of data storytelling. Third, evaluation results show that all the 15 features implemented in ODE were useful for users except that users with low data manipulation skills suggested to remove the features related to data processing (*DP1*. Get overview to *DP4*. Combine datasets). Furthermore, the results show that the prototype

was useful in turning open data into stories, as most of participants agreed that the proposed prototype met their expectations and they were able to create their stories from their selected data.

This research also contributes to practice in the following aspects. First, unlike previous tools presented in Section Background (see Table 1), ODE addresses each of their shortcomings and thus provides users with an end-to-end tool to turn their data into stories without the need to use separated tools, e.g., to process data or to get overview of data content (see Section Open Data Explorer (ODE)). Second, unless other tools that provide static rules to generate visualizations, ODE allows users to give their feedback on the visualizations and later leverages to improve the initial visualizations rules. Third, we provide access to the ODE source code. This can be used as a starting point for developers to create their tool to facilitate open data storytelling or to improve the prototype.

Future work will investigate whether the identified features can be used or extended to help open data publishers to transform their data into transparency portals. We will also assess whether there is a gap between the features identified for transparency portals from the perspective of publishers and citizens.

REFERENCES

Aanderud, T., Scientific, B., and Agord, J. D. 2020. 'My Sharky Secrets for Telling Fabulous Data Stories', *SAS Global Forum*.

Ansari, B., Barati, M., and Martin, E. G. 2022. 'Enhancing the Usability and Usefulness of Open Government Data: A Comprehensive Review of the State of Open Government Data Visualization Research', *Government Information Quarterly* (39:1), Elsevier Inc.

Attard, J., Orlandi, F., Scerri, S., and Auer, S. 2015. 'A Systematic Review of Open Government Data Initiatives', *Government Information Quarterly* (32:4), Elsevier Ltd, pp. 399–418.

Baskerville, R. 2008. 'What Design Science Is Not', *EJIS* (17), pp. 441–443.

Berends, J., Carrara, W., Engbers, W., and Vollers, H. 2020. 'Recommendations for Open Data Portals: From Setup to Sustainability'.

Böhm, C., Freitag, M., Heise, A., Lehmann, C., Mascher, A., Naumann, F., Ercegovac, V., Hernandez, M., Haase, P., and Schmidt, M. 2012. 'GovWILD: Integrating Open Government Data for Transparency', *Proceedings of the 21st Annual Conference on World Wide Web Companion*, pp. 321–324.

Boone, H. N., and Boone, D. A. 2012. 'Analyzing Likert Data.', *Journal of Extension* (50:2).

Brolcháin, N., Porwol, L., Ojo, A., Wagner, T., Lopez, E. T., and Karstens, E. 2017. 'Extending Open Data Platforms with Storytelling Features', Dg.o '17 , pp. 48–53.

Brugger, J., Fraefel, M., Riedl, R., Fehr, H., Schöneck, D., and Weissbrod, C. S. 2016. 'Current Barriers to Open Government Data Use and Visualization by Political Intermediaries', *CeDEM*, pp. 219–229.

Chokki, A. P., Simonofski, A., Clarinval, A., Frénay, B., and Vanderose, B. 2022. 'Fostering Interaction between Open Government Data Stakeholders: An Exchange Platform for Citizens, Developers and Publishers', *Electronic Government*.

Chokki, A. P., Simonofski, A., Frénay, B., and Vanderose, B. 2021. 'Open Government Data for Non-Expert Citizens: Understanding Content and Visualizations' Expectations', *RCIS*, pp. 602–608.

Chokki, A. P., Simonofski, A., Frénay, B., and Vanderose, B. 2022a. 'Engaging Citizens with Open Government Data: The Value of Dashboards Compared to Individual Visualizations', *DGOV*.

Chokki, A. P., Simonofski, A., Frénay, B., and Vanderose, B. 2022b. 'CitizenApps: Increasing Awareness and Usefulness of Open Government Data via a Mobile Application', *TGPPP*.

Cordasco, G., De Donato, R., Malandrino, D., Palmieri, G., Petta, A., Pirozzi, D., Santangelo, G., Scarano, V., Serra, L., Spagnuolo, C., and Vicidomini, L. 2017. 'Engaging Citizens with a Social Platform for Open Data', *Dg.o '17*, pp. 242–249.

Crusoe, J., Simonofski, A., Clarinval, A., and Gebka, E. 2019. 'The Impact of Impediments on Open Government Data Use: Insights from Users', *RCIS*, pp. 1–12.

Dadzie, A.-S., and Rowe, M. 2010. 'Approaches to Visualising Linked Data: A Survey', (Vol. 1), IOS Press.

Davis, F. D. 1989. 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology', *Management Information Systems* (13:3), pp. 319–339.

De Donato, R., Malandrino, D., Palmieri, G., Petta, A., Pirozzi, D., Scarano, V., Serra, L., Spagnuolo, C., Vicidomini, L., and Cordasco, G. 2017. 'DatalEt-Ecosystem Provider (DEEP): Scalable Architecture for Reusable, Portable and User-Friendly Visualizations of Open Data', *CEDEM*, pp. 92–101.

Dresch, A., Lacerda, D. P., and Antunes, J. A. V. 2015. 'Design Science Research: A Method for Science and

Technology Advancement', *Springer International Publishing*.
Eberhardt, A., and Silveira, M. S. 2018. 'Show Me the Data! A Systematic Mapping on Open Government Data Visualization', *Dg.o'18*, pp. 1–10.
Faulkner, L. 2003. 'Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing', *Behavior Research Methods, Instruments, & Computers 35*, pp. 379–383.
Fowler, M., and Highsmith, J. 2001. 'The Agile Manifesto', *Software Development* (9:8), pp. 28–35.
Graves, A., and Bustos-Jiménez, J. 2015. 'Co-Creating Visual Overviews for Open Government Data', *Dg.o'15*, pp. 37–42.
Graves, A., and Hendler, J. 2013. 'Visualization Tools for Open Government Data', *Dg.o'13*, pp. 136–145.
Gray, J., Chambers, L., and Bounegru, L. 2012. *The Data Journalism Handbook: How Journalists Can Use Data to Improve the News*, O'Reilly Media, Inc.
Grimm, P. 2010. 'Pretesting a Questionnaire', *Wiley International Encyclopedia of Marketing*, John Wiley & Sons, Ltd.
Handbook, O. D. 2015. 'What Is Open Data?', *Open Data Handbook*. (https://opendatahandbook.org/, accessed March 23, 2021).
Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. 'Design Science in Information Systems Research', *MIS Quarterly*, pp. 75–105.
Hivon, J., and Titah, R. 2017. 'Conceptualizing Citizen Participation in Open Data Use at the City Level', *TGPPP*, pp. 99–118.
Holtz, Y., and Conor, H. 2018. 'From Data to Viz'. (https://www.data-to-viz.com/, accessed May 17, 2021).
Janssen, M., Charalabidis, Y., and Zuiderwijk, A. 2012. 'Benefits, Adoption Barriers and Myths of Open Data and Open Government', *Information Systems Management* (29:4), pp. 258–268.
Kalatzi, O., Bratsas, C., and Veglis, A. 2018. 'The Principles, Features and Techniques of Data Journalism', *Studies in Media and Communication* (6:2), pp. 36–44.
Kitchenham, Barbara Brereton, P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. 2009. 'Systematic Literature Reviews in Software Engineering-A Systematic Literature Review', *Information and Software Technology* (51(1)), pp. 7–15.
Matos, U. C., and Corbett, J. 2019. 'Creating Knowledge for Value Creation in Open Government Data Ecosystems', in *AMCIS 2019 Proceedings*.
Moreno Cegarra, J. L., Cegarra Navarro, J. G., and Córdoba Pachón, J. R. 2014. 'Applying the Technology Acceptance Model to a Spanish City Hall', *IJIM* (34:4), Elsevier Ltd, pp. 437–445.
Munzner, T. 2014. 'Visualization Analysis and Design', *Visualization Analysis and Design*.
Nielsen, J. 2000. 'Why You Only Need to Test with 5 Users', *Nielsen Norman Group*. (https://rb.gy/c8jbx4, accessed June 17, 2021).
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. 'A Design Science Research Methodology for Information Systems Research', *JMIS* (24:3), pp. 45–77.
Petricek, T. 2017. 'Tools for Open, Transparent and Engaging Storytelling', in *Companion to the First International Conference on the Art, Science and Engineering of Programming*, pp. 1–2.
Pirozzi, D., and Scarano, V. 2016. 'Support Citizens in Visualising Open Data', in *Proceedings of the 20th International Conference on Information Visualisation*, pp. 271–276.
Rubin, J., and Chisnell, D. 2008. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, Wiley Publishing, Inc.
Silva, D. A., and Santos, C. 2019. 'Open Government Data: From Transparency to Social Participation', in *AMCIS 2019 Proceedings. 94*.
Vaishnavi, V., and Kuechler, W. 2007. 'Design Science Research Methods and Patterns: Innovating Information and Communication Technology', *Auerbach Publications*.
Wenzel, F., Köppl, D., and Kieβling, W. 2013. 'Interactive Toolbox for Spatial-Textual Preference Queries', *In International Symposium on Advances in Spatial and Temporal Databases*, pp. 462–466.
Wilke, C. O. 2019. 'Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures', *O'Reilly Media*.
Zuiderwijk, A., and Janssen, M. 2014. 'Barriers and Development Directions for the Publication and Usage of Open Data: A Socio-Technical View', *Open Government*, pp. 115–135.
Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., and Alibaks, R. S. 2012. 'Socio-technical Impediments of Open Data', *Electronic Journal of Electronic Government* (10:2), pp. 156–172.