From Data to Narrative: Automating and Engineering the Art of Data Storytelling

Mohammad Daradkeh
College of Engineering and Information Technology
University of Dubai
Dubai, United Arab Emirates
mdaradkehc@ud.ac.ae

Shadi Atalla
College of Engineering and Information Technology
University of Dubai
Dubai, United Arab Emirates
satalla@ud.ac.ae

***Abstract*— This paper addresses the importance of defining data storytelling and its characteristics, distinguishing it from literary narratives. It explores the research landscape in data storytelling and proposes the essence and characteristics based on the findings. The paper then introduces a data science and data engineering approach for automated generation of data stories. Furthermore, it presents a reference architecture for engineering development of data stories from a software engineering perspective. This article contributes by providing a definition of data storytelling, highlighting its unique features, and offering insights for academic research. It also explores automatic generation methods and introduces core concepts such as atoms, operators, and rules. Lastly, an engineering development approach, including a reference framework, is presented to drive the growth of data storytelling tools and industry ecosystem.**

Keywords—Data storytelling, automated generation, engineering development, research landscape, characteristics, reference architecture.

I. INTRODUCTION

Currently, data storytelling has become one of the hot topics in the field of big data as the final mile problem in data science. From the perspective of data perception, perception is the prerequisite for cognition, and cognition is the extension of perception. Data visualization and data storytelling focus on the perception and cognition of data, respectively [1]. Therefore, data storytelling is expected to be widely applied in scenarios where data analysis results need to be explained to non-experts, in order to gain trust from non-experts in data-driven products and services. However, compared to data visualization, there is a lack of substantial research on the essence of data storytelling, and breakthrough progress is urgently needed.

At the same time, compared to the raw data itself, data stories have the characteristics of being more memorable, cognitively appealing, and experiential. Firstly, unlike statistical numbers, data stories are more memorable. A research experiment conducted by Lotfi et al. [2] showed that only 5% of people could remember specific statistical data, but when the statistical data was transformed into a story, 63% of people could remember it. Secondly, data stories align with innate human cognitive characteristics [3]. Finally, data stories offer a higher level of experiential engagement. Throughout the narrative exploration of a story, there is communication and interaction between the storyteller and the audience, with continuous feedback and revision of the storyline based on further data collection [4]. Therefore, research on data storytelling holds significant importance for data presentation and the development of data products.

This paper aims to investigate the concept of data storytelling, elucidate its defining characteristics, and differentiate it from literary narratives. Data storytelling entails the effective communication of insights and narratives through the utilization of data visualization and analysis techniques. Given the exponential growth in data volume, understanding data storytelling is becoming increasingly crucial due to the inherent complexity involved in extracting meaningful information from data, often resulting in its underutilization.

The primary contribution of this paper can be categorized into four key aspects:

- Definition and Characterization: The paper offers a precise definition of data storytelling and provides a comprehensive outline of its distinctive features.
- Research Landscape Exploration: The research examines the existing body of knowledge in the field of data storytelling, effectively summarizing its current state and identifying gaps and opportunities that warrant further investigation.
- Automated Generation Methods: The paper introduces an approach rooted in data science and data engineering for the automated generation of data stories.
- Engineering Development Approach: The article presents a reference architecture that outlines a systematic framework from a software engineering perspective for the engineering development of data stories.

II. DATA STORY AND ITS CHARACTERISTICS

A data story is a type of data product or service that aims to fulfill specific business needs, using data as the raw material and employing data analysis and modeling methods to discover valuable insights from the data and present them in the form of a story to a target audience. The process of transforming data into a data story is referred to as data storytelling. A data story is the final product of the data storytelling process, combining the objectivity of data and the subjectivity of storytelling [5]. Fig. 1 illustrates the pyramid model of data stories. From Fig. 1, it can be observed that a data story starts with the business needs, relies on data as evidence, and involves the sequential stages of analyzing and gaining insights from the raw data, constructing a story model, formalizing the story model, and presenting the story to the target audience to influence their behavior and ultimately achieve the business objectives.

2023 International Conference on Information Technology (ICIT) | 979-8-3503-2006-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIT58056.2023.10225929

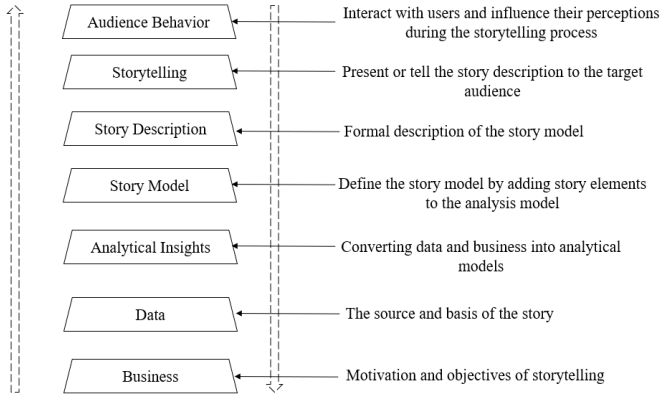

Fig. 1. Pyramid Model of Data Storytelling.

TABLE I. CONCEPTS AND DIFFERENCES OF DATA STORYTELLING

Terminology	Focus
Data-Driven Storytelling	A specific storytelling approach that differs from goal-driven or model-driven methods, emphasizing the dynamism, agility, and personalization of data storytelling.
Visual Storytelling	A primarily visual storytelling approach that emphasizes the importance of visualization techniques in data storytelling.
Analytical Storytelling	Data-driven data storytelling, emphasizing the importance of data analysis in data storytelling.
Interactive Storytelling	A narrative process that emphasizes the importance of user experience and audience feedback through interaction between the storyteller and the audience in data storytelling.
Storytelling with Data	Opposition to data storytelling content being overly fictional, subjective, or detached from business, emphasizing objectivity, quantification, and empiricism in data narration.
Digital Storytelling	Emphasis on the application of digital technology in data storytelling, highlighting the digitization and diversification of data storytelling mediums.

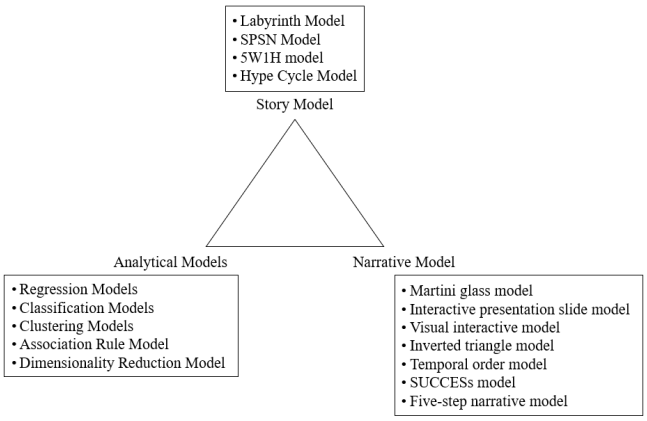

Fig. 2. Model Chain of Data Storytelling.

Currently, there are various related concepts regarding data stories and data storytelling, such as data-driven storytelling, visual storytelling, analytics storytelling, interactive storytelling, and data-driven narrative, which discuss different aspects or dimensions of data storytelling from different perspectives or levels [6], as shown in Table 1.

A. Relevant Models in Data Storytelling

Models related to data storytelling can be categorized into analysis models, story models, and narrative models, representing the three core activities of data storytelling: data analysis and insights, storytelling modeling, and story narration, as shown in Fig. 2.

- Analysis Models: These models primarily describe the business requirements, which involve statistical analysis or machine learning models to transform data into stories.
- Story Models: These models primarily describe the story elements and the structural relationships among them, including story content, plot, and context. For example, the Maze model proposed by Daradkeh [7] emphasizes the contextual, exploratory, intricate, thematic, and simplicity aspects of the data storytelling context. The 5W1H model suggests that a data story should include the elements of What, Why, When, Where, Who, and How. The SPSN framework [8] provides a structural model for stories, including Situation, Problem, Solution, and Next Steps. The Gartner Hype Cycle model [2] offers a developmental model for the plot of a story. Story models serve as a bridge between analysis models and narrative models.
- Narrative Models: Narrative models are the models involved in narrating the story model to the target audience. A single story model can have multiple narrative models to achieve personalized storytelling purposes. For instance, the Martini Glass model [9], interactive presentation slide models, visual interactive models, inverted triangle models, and chronological models provide strategies for story narration. The SUCCESs model [10] presents requirements for effective storytelling, emphasizing simplicity, unexpectedness, concreteness, credibility, emotions, and stories in data storytelling. Dunkleberger's five-step narrative model [11] provides a reference model for story creation and writing.

B. Primary Applications of Data Storytelling

Currently, the theoretical research on data storytelling can be categorized into two main types of applications: direct applications that directly provide data storytelling products and indirect applications that involve the development of data storytelling software products and services.

- Direct Generation of Data Stories: This type involves providing data products in the form of stories to the end audience, such as data journalism, product showcase dashboards, and academic presentation slides. For example, Echeverria, Martinez-Maldonado [12] published a data story in The Wall Street Journal titled "The Fight against Infectious Diseases since the 20th Century: The Impact of Vaccines."
- Development of Data Storytelling Software Products: This type involves providing methods and technical support for platforms or tools used to generate data stories. Currently, data storytelling platforms or tools can be classified into two categories: those that incorporate data storytelling features into existing data visualization tools or business intelligence software, such as Tableau, D3.js, TIBCO Spotfire, Flexdashboard, Qlik Sense, ShortHand, and Microsoft Power BI; and those that are specifically developed for data storytelling, such as Banjo and Narratives for Tableau. Generally, visualization software that supports data storytelling is the

mainstream approach in the development of data storytelling software products, while the number of specialized tools specifically designed for data storytelling generation is relatively limited [9].

III. Automated Generation of Data Stories

From the perspective of automated generation and engineered development, the process of generating data stories can be decomposed into six main activities: data understanding and business understanding, data analysis and insights, storytelling modeling, formalized storytelling description, storytelling narration and interaction, and audience interaction and reflection behaviors, as shown in Fig. 3.

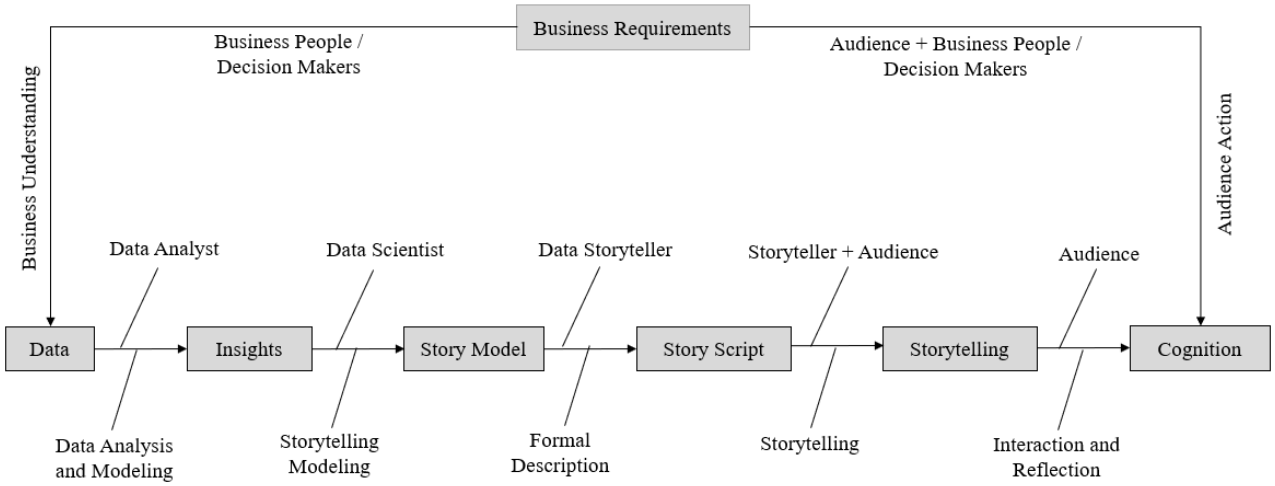

Fig. 3. Key Activities of Data Storytelling.

A. Data Analysis and Insights

Data analysis and insights play a crucial role in data storytelling. On one hand, data analysis and insights facilitate the discovery and understanding of analytical models for data storytelling. On the other hand, data analysis and insights serve as the primary means to understand and characterize business requirements, acting as a bridge between the business demands addressed by data storytelling and the storytelling models. The methods for data analysis and insights in data storytelling can be classified into descriptive analysis, diagnostic analysis, predictive analysis, and prescriptive analysis. Table 3 illustrates the corresponding relationship between the eight business requirements and the four categories of data analysis methods in data storytelling.

TABLE II. Business Requirements, Objectives of Data Storytelling, and Nature of Data Analysis.

Types of business needs	Descriptions of business needs	Types of data analysis methods
Description	To describe data or information to the audience	Descriptive analytics
Recommendation	To recommend products or services to the audience	Predictive analytics s
Explanation	To explain viewpoints and conclusions to the audience	Diagnostic analytics
Investigation	To gather user data or opinions from the audience	Descriptive analytics
Exploration	To explore business or product innovation and optimization through audience participation	Prescriptive analytics
Persuasion	To persuade the audience to believe a certain viewpoint or idea	Diagnostic analytics
Education	To educate the audience in a certain knowledge or skill, or to change their behavior	Descriptive/diagnostic/ prescriptive analytics
Entertainment	To encourage the audience to use and consume a specific product or service	Predictive/prescriptive analytics

IV. Storytelling Modeling

Storytelling modeling is the key to the automated generation of data stories. It relies on data analysis as a foundation and transforms the results of data analysis into a storytelling model. The data storytelling model consists of six core elements: requirements, characters, context, plot, conflict, and resolution, as shown in Fig. 4.

- Requirements: The driving force, goals, and problems that data storytelling aims to address. Data storytelling is a business-oriented analytical modeling activity, and fulfilling business requirements is the ultimate goal of data storytelling.
- Characters: The people and entities involved in the data story. Characters in a data story are not limited to protagonists or human beings.
- Context: The business context, surrounding environment, and initial state in which the story takes place.
- Plot: The plot of a data story should have twists and turns, be filled with major conflicts and contradictions, and generally undergo four stages of intensity: rising action, climax, falling action, and denouement.
- Conflict: The conflicts and contradictions faced by the characters in the data story. Conflict and contradiction are at the core of data storytelling, and the selection of characters and the design of the plot revolve around conflicts or contradictions.
- Resolution: The ultimate resolution to the conflicts and contradictions in the data story.

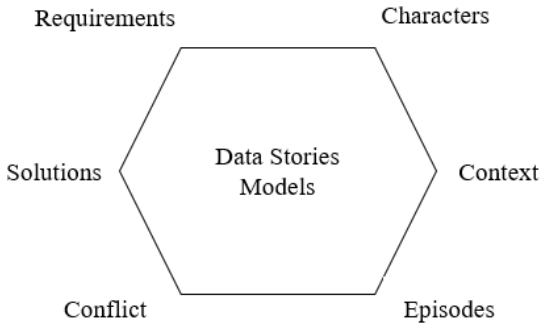

Fig. 4. Components of Data Model.

A. *Formal Description of Storytelling*

Formal description of data stories refers to recording the storytelling model in a way that is understandable, inferable, and verifiable by computers. The formal description of a data story involves three key issues:

- Formalization of story atoms and connections: Story atoms refer to the indivisible and smallest units of a story and can be represented as triples (x, P, y), where x, P, and y represent the subject, predicate, and object, respectively, as shown in Fig. 5. The subject and object of the story primarily describe the five elements of the data storytelling model: requirements, characters, context, plot, conflict, and resolution. RDF or OWL languages can be used for formal representation. Table 4 provides descriptions of the three components.
- Storytelling operators and their formalization: Storytelling operators are operational symbols defined based on story atoms. Unlike the predicates in story atoms, storytelling operators do not represent semantic relationships between elements but rather actions and operations required for storytelling modeling. Common storytelling operators and their inverse operators are shown in Table 5.
- Formalization of rules and inference: The formal representation of rules and inference captures the logical reasoning knowledge in data storytelling and is an indispensable part of the automated generation of data stories.

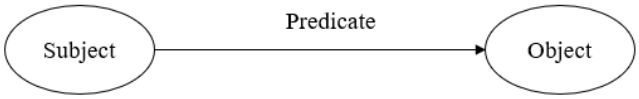

Fig. 5. Triplet Structure of Story Atoms.

TABLE III. STORY ATOMS AND THEIR TRIPLETS.

Atomic	**Possible objects that can be represented**
Subject	Needs, characters, situations, plots, conflicts, and solutions
Object	Needs, people, situations, plots, conflicts, and solutions
Predicate	Different kinds of subject and object words that are expressed in specific business contexts and semantic relationships

TABLE IV. STORYTELLING OPERATORS AND THEIR MEANINGS.

Operator	**Meaning**	**Reverse operator**
Join	Establish connections between two story atoms.	Isolation
Drill Up	Jump from lower-level story atoms to higher-level story atoms.	Drill down
Aggregation	Compute attributes of same-level story atoms, calculating the attributes of upper-level story atoms.	Deduction
Slicing	Extract a story surface at a specific value point on a dimension.	Merge
Chunking	Extract a story block within a specific value range on a dimension.	Integrate
Filtering	Apply conditional filtering to story atoms.	Extract
Drill Down	Traverse or replay story atoms in sequential order from present to past on the time dimension.	Drill up
Crossing	Perform cross-operations on two or more story atoms in a specific dimension.	Parallel
Anonymization	Anonymize the subject, verb, and object of story atoms.	Restore
Pivoting	Alter the analytical dimension or coordinates of story atoms.	Solidify
Swapping	Swap the subject, verb, or object of different story atoms.	Recover
Zoom in	Zoom in on quantitative metrics in story atoms	Narrow
Spotlight	Focus on a particular story atom or class of stories	Generalize

B. *Storytelling and Interaction*

Storytelling and interaction refer to the process of transforming the formalized description of a data story into narrative storytelling. From the perspective of automated implementation and engineering research and development of data stories, the data narrative process should be divided into two stages: data narration and audience interaction, which should be treated differently. For the same story model, different storytelling and interaction strategies are adopted based on different audiences and narrative contexts [13]

V. ENGINEERING DEVELOPMENT OF DATA STORYTELLING

The engineering development of data storytelling should adhere to the principles of hierarchical implementation and component-based development. The data storytelling engineering project is divided into different levels, with each level focusing on a specific set of problems or tasks, maintaining a high level of independence between different levels. Within the same level, the problems to be addressed are divided into multiple sub-functions, which are resolved by different components. Fig. 6 illustrates a reference architecture for the engineering development of data storytelling, where component-based development methods are employed for research and development in all layers except the data layer and user layer.

A. *Insights Layer*

The Insights Layer primarily aims to meet business requirements by applying machine learning and statistical methods to analyze data and derive insight models such as regression models, classification models, clustering models, association rules models, and dimensionality reduction models [14]. The Insights Layer requires four key components:

- Algorithm Selector: Selects appropriate algorithms based on business requirements and data characteristics.
- Algorithm Invoker: Calls the selected algorithm based on the results from the selector and obtains insights.
- Algorithm Evaluator: Assess the credibility and validity of the results returned by the algorithm invoker.

- Algorithm Optimizer: Adjusts hyperparameters in algorithms or selects alternative algorithms based on the results from the algorithm evaluator.

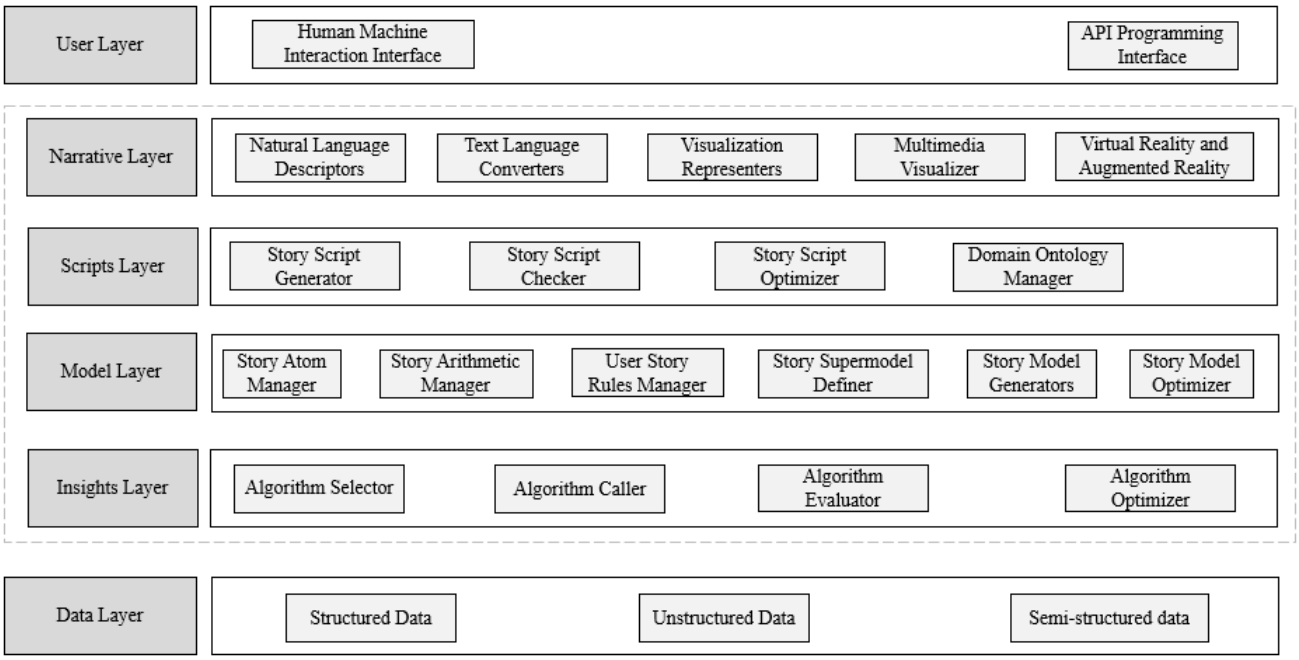

Fig. 6. Reference Architecture for Engineering Development of Data Storytelling.

B. *Story Model Layer*

The Story Model Layer's main function is to transform the insights obtained from the Insights Layer into story models. The story model consists of six core elements: requirement, character, situation, plot, conflict, and resolution [15]. The Story Model Layer requires the following key components:

- Story Atom Manager: Responsible for defining, modifying, and deleting story atoms.
- Story Operator Manager: Defines specific operators and their operational rules for story atoms.
- Story Rule Manager: Defines specific rules and reasoning methods for story atoms and story operators.
- Story Supermodel Definer: Defines the template for story models, where the story supermodel serves as the benchmark for defining specific story models.
- Story Model Generator: Generates concrete story models based on the definition of the story supermodel.
- Story Model Validator: Verifies the syntax and semantic consistency within the story model.
- Story Model Optimizer: Automatically corrects and optimizes issues identified by the story model validator.

C. *Scripting Layer*

The main function of the Scripting Layer is to provide a formal description of the story model. The formal description of the story script serves as the foundation for the processing in the Narrative Layer, enabling support for different storytelling techniques using the same story script [16]. The key components to be implemented in the Scripting Layer include:

- Story Script Generator: Used to formalize the description of the story model and define the story script code.
- Story Script Validator: Performs syntax and semantic consistency checks on the results generated by the story script generator.
- Story Script Optimizer: Optimizes the story script based on issues identified by the story script validator.
- Story Script Inference: Performs semantic inference and narrative reasoning on the story script code to expand the breadth and depth of the story content and generate new story elements.
- Domain Ontology Manager: Manages domain ontologies to support the formal representation and rule inference of story scripts.

D. *Narrative Layer*

The main function of the Narrative Layer is to transform the story script into a narrative. There are various media that can be used for storytelling, such as natural language descriptions, text-to-speech conversion, visual representations, multimedia presentations, virtual reality, and augmented reality.

- Natural Language Description: Converts the story script into a narrative in natural language.
- Text-to-Speech Conversion: Converts the narrative in natural language into speech representation.
- Visual Representation: Converts the story script into visual graphical elements.
- Multimedia Presentation: Uses multimedia technologies to present the story content.
- Virtual Reality and Augmented Reality: Introduces virtual reality and augmented reality technologies into data storytelling to enhance user engagement and experience.

It is worth noting that the engineering development of data storytelling does not require the redundant implementation of existing algorithms and technologies. Existing general-purpose algorithms and technology implementations can be directly called upon [17]. The focus of engineering development in data storytelling lies in the selection, invocation, evaluation, and automatic optimization of existing algorithms or technologies to meet the requirements of data storytelling [18].

VI. Conclusion

Storytelling is an ancient topic, while business-oriented data storytelling is a novel field. This article provides a definition of data storytelling and highlights its key features that differentiate it from literary storytelling. It offers valuable insights for academic research in the field of data storytelling. Another major contribution of this article is approaching data storytelling from the perspective of engineering development. It explores the automatic generation methods of data storytelling from three different levels: analysis models, story models, and narrative models. Additionally, it introduces three core concepts of data storytelling: atoms, operators, and rules, laying a theoretical foundation for the automated generation of data storytelling. Building upon this, the paper presents an engineering development approach for data storytelling, which includes a reference framework for layered implementation and component-based research and development. Moving forward, we will continue our research on several key issues regarding the automatic generation and engineering development of data storytelling.

References

[1] Daradkeh, M., An Empirical Examination of the Relationship Between Data Storytelling Competency and Business Performance: The Mediating Role of Decision-Making Quality. Journal of Organizational and End User Computing (JOEUC), 2021. 33(5): pp. 42-73.

[2] Eckert, H., From Data to Story, in Storytelling With Data: Gaining Insights, Developing Strategy and taking Corporate Communications to a new level, H.-W. Eckert, Editor. 2022, Springer Fachmedien Wiesbaden: Wiesbaden. pp. 69-104.

[3] Lotfi, F., et al., Storytelling with Image Data: A Systematic Review and Comparative Analysis of Methods and Tools. Algorithms, 2023. 16(3): pp. 135.

[4] Shin, M., et al., Roslingifier: Semi-Automated Storytelling for Animated Scatterplots. IEEE Transactions on Visualization and Computer Graphics, 2023. 29(6): pp. 2980-2995.

[5] Santana, B., et al., A survey on narrative extraction from textual data. Artificial Intelligence Review, 2023.

[6] 6. Pozdniakov, S., et al., How Do Teachers Use Dashboards Enhanced with Data Storytelling Elements According to their Data Visualisation Literacy Skills?, in LAK23: 13th International Learning Analytics and Knowledge Conference. 2023, Association for Computing Machinery: Arlington, TX, USA. pp. 89–99.

[7] Oberascher, L., et al., Data - driven Storytelling to communicate Big Data internally – a Guide for Practical Usage European Journal of Management Issues, 2023. 31(1).

[8] Lopezosa, C., M. Pérez-Montoro, and J. Guallar, Data Visualization in the News Media: Trends and Challenges, in Technology, Business, Innovation, and Entrepreneurship in Industry 4.0, T. Guarda, C. Fernandes, and M.F. Augusto, Editors. 2023, Springer International Publishing: Cham. pp. 315-334.

[9] Ghodoosi, B., et al., A systematic literature review of data literacy education. Journal of Business & Finance Librarianship, 2023: pp. 1-16.

[10] Halperin, B. and S. Lukin, Envisioning Narrative Intelligence: A Creative Visual Storytelling Anthology, in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023, Association for Computing Machinery: Hamburg, Germany. pp. Article 240.

[11] Zhang, Y., et al., A Visual Data Storytelling Framework. Informatics, 2022. 9(4): pp. 73.

[12] Zdanovic, D., T. Lembcke, and T. Bogers, The Influence of Data Storytelling on the Ability to Recall Information, in ACM SIGIR Conference on Human Information Interaction and Retrieval. 2022, Association for Computing Machinery: Regensburg, Germany. pp. 67–77.

[13] Echeverria, V., et al., HuCETA: A Framework for Human-Centered Embodied Teamwork Analytics. IEEE Pervasive Computing, 2022: pp. 1-11.

[14] Cheng, H., et al., Investigating the Role and Interplay of Narrations and Animations in Data Videos. Computer Graphics Forum, 2022. 41(3): pp. 527-539.

[15] Calegari, D. Computational narratives using Model-Driven Engineering. in 2022 XVLIII Latin American Computer Conference (CLEI). 2022.

[16] Sharda, R., D. Delen, and E. Turban, Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support (11th ed.). 2021: Pearson.

[17] Nieto, G., Kitto, K., Buckingham, S., & Martinez-Maldonado, R. (2022). Beyond the Learning Analytics Dashboard: Alternative Ways to Communicate Student Data Insights Combining Visualisation, Narrative and Storytelling. Paper presented at the LAK22: 12th International Learning Analytics and Knowledge Conference, Online, USA.

[18] Southekal, P., Analytics Best Practices: A Business-Driven Playbook for Creating Value Through Data Analytics. 2020, NJ, USA: Technics Publications, LLC.

[19] Boldosova, V., Telling stories that sell: The role of storytelling and big data analytics in smart service sales. Industrial Marketing Management, 2020. 86(1): pp. 122-134.

[20] Golnaz, A., et al., Bop or Flop?: Integrating Music and Data Science in an Elementary Classroom. The Journal of Experimental Education, 2023: pp. 1-25.

[21] Mathisen, A., et al., InsideInsights: Integrating Data-Driven Reporting in Collaborative Visual Analytics. Computer Graphics Forum, 2019. 38(3): pp. 649-661.

[22] Jones, P. and D. Comfort, Storytelling and corporate social responsibility reporting: A case study commentary on U.K. food retailers. Journal of Public Affairs, 2019. 19(4): pp. 1-8.

[23] Hair, J., et al., When to use and how to report the results of PLS-SEM. European Business Review, 2019. 31(1): pp. 2-24.

[24] Ciancarini, P., et al., Software as storytelling: A systematic literature review. Computer Science Review, 2023. 47: pp. 100517.

[25] Chen, Q., et al., How Does Automation Shape the Process of Narrative Visualization: A Survey of Tools. IEEE Transactions on Visualization and Computer Graphics, 2023: pp. 1-20.