Towards Enabling the Analysis of Visual Exploration Processes through Interaction Provenance*

Lyncoln S. de Oliveira[1], **Gustavo Moreira**[2], **Fábio Miranda**[2]
Marcos Lage[1], **Daniel de Oliveira**[1]

[1]Institute of Computing – Universidade Federal Fluminense (UFF)

[2]Department of Computer Science – University of Illinois Chicago (UIC)

lyncolnsousa@id.uff.br, {gmorei3,fabiom}@uic.edu, {mlage,danielcmo}@ic.uff.br

Abstract. *The rapid growth of data has made accessing, integrating, and analyzing information increasingly challenging. While large-scale systems support processing and querying, interactive visualizations are essential for exploring complex datasets. Understanding how users gain insights from these visualizations requires capturing their interactions. Provenance data offers a natural solution, but current methods often fail to capture interaction-level provenance effectively. This paper presents an approach to capture and record user interaction provenance and integrate it with both prospective and retrospective provenance. We implement this approach in the* `Curio` *framework, which builds urban data visualization pipelines. Results demonstrate its effectiveness in capturing user behavior during visual exploration.*

1. Introduction

Over the last decades, the volume and complexity of urban data have grown quickly, driven by computing advances and data sources such as sensor networks [Ang and Seng 2016], street-level imagery [Biljecki and Ito 2021], and building geometries [Miranda et al. 2024]. These data capture different aspects of urban dynamics, allowing us to address key issues in areas such as mobility [Ferreira et al. 2013], accessibility [Miranda et al. 2020], urban climate [Loibl et al. 2021], and extreme events response [Oliveira et al. 2023]. However, their heterogeneity, scale, and dynamic nature become a challenge for analysis, requiring advanced tools to support the decision-making process.

While many solutions are suited for managing large-scale data by providing features for storage, processing, and querying (*e.g.*, the Apache Spark ecosystem), they are not designed to support the exploratory nature of human data analysis. Interactive visualizations [Heer et al. 2010] bridge this gap by presenting complex datasets in intuitive visual interfaces that enable users to interact with the data. These visual interfaces allow users to filter, zoom, compare, and drill down into subsets of the data, identifying insights, anomalies, and hidden relationships in tabular query results. However, building an interactive visualization is far from a trivial process.

In this context, dataflow-based frameworks have emerged to provide flexibility and interoperability [Bavoil et al. 2005]. However, only a few of these frameworks are designed to address the challenges posed by urban data, which often entails complex spatial layers and

*The authors would like to thank the financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, CNPq and FAPERJ.

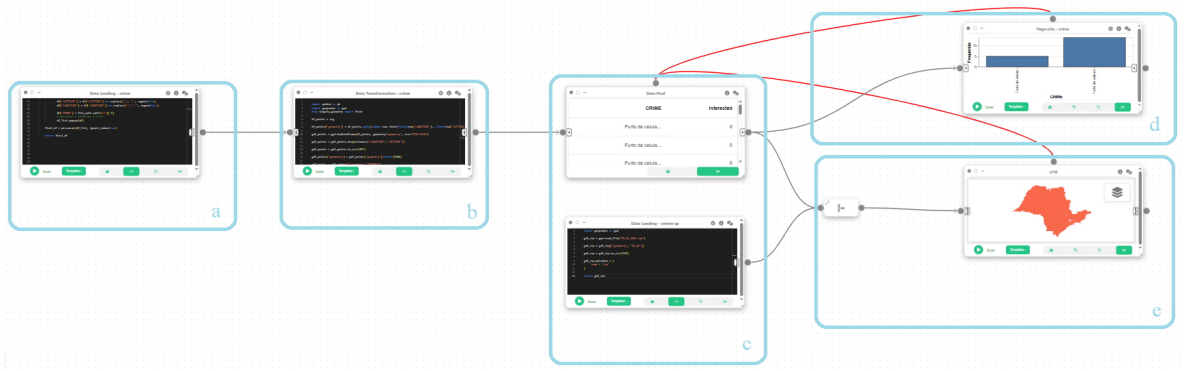

Figure 1. A dataflow in Curio to analyze crime hotspots in São Paulo.

requires interactive visualizations. Furthermore, collaboration among experts from various domains, *e.g.*, urban planners, data scientists, and visualization researchers, remains limited due to the absence of frameworks that integrate their perspectives and dataflows. Curio [Moreira et al. 2025] and VisFlow [Yu and Silva 2017] are examples of frameworks built on a dataflow model aimed at developing interactive visualizations for urban data.

Although the aforementioned frameworks represent a step forward, it is still a challenge to understand the rationale behind how interactive visualizations are constructed within them. While some frameworks already capture provenance data [Herschel et al. 2017] (historical information about the specification and execution of the dataflow) to support reproducibility and traceability, this captured provenance does not account for user interactions. Consider the example in Figure 1, which shows a dataflow modeled in Curio to visualize crime hotspots in São Paulo. In this dataflow, a sequence of activities loads crime datasets, groups them by type (*e.g.*, theft), and then plots the identified hotspots (*i.e.*, areas with high crime intensity) on a map. Additionally, it generates a chart showing the number of events by subtype within a specific crime category (*e.g.*, theft can be subdivided into mobile phone theft or car theft).

Curio already captures two types of provenance data: the specification of the dataflow (*i.e.*, prospective provenance or p-prov) and the execution results, such as the generated maps and charts (*i.e.*, retrospective provenance or r-prov). However, to produce the final visualization, users often interact with the map or chart, *e.g.*, by selecting sub-regions or filtering by specific crime types. We refer to this kind of user interaction as *provenance of interactions (i-prov)*. This type of provenance, which complements both p-prov and r-prov, is important for understanding the rationale behind the final visualization and for ensuring full traceability of the analytical process.

In this paper, we propose an extension to the provenance model of Curio by introducing a mechanism for capturing i-prov in urban visualizations and storing it in a queryable format. This extended model enables representing interactions, *e.g.*, selections made within maps, which are essential for understanding the rationale behind the construction and refinement of a visualization. The i-prov capturing feature is also important in collaborative scenarios to understand the broader context of analytical decisions across team members. To evaluate the proposed approach, we considered a dataflow modeled in Curio to visualize crime hotspots in São Paulo. The results showed the feasibility and potential of the proposed approach to provide traceability and interpretability in interactive visual analytics dataflows.

2. Related Work

The provenance of interactions, particularly concerning visualizations, has received less attention over the last years than other types of provenance data [de Oliveira et al. 2018]. [Fekete and Freire 2020] highlight the challenges of reproducibility in visualizations, claiming that capturing data and results and the user's interactions is important. The authors outline directions for visual tools to capture and store this history of actions, emphasizing the importance of provenance for verifying and sharing analyses. The approach proposed in this paper aims to advance in this direction by proposing mechanisms within the Curio framework to capture and store every action performed during the visual exploration process.

In [Psallidas and Wu 2018], the authors propose capturing the provenance of user interactions within visualizations by introducing operators that track both the source of the data and the effects of user actions. This approach enables the implementation of features such as linked selection, drill-down, and synchronized views, all based on provenance data. The approach presented in this paper shares a similar goal on traceability and reproducibility but advances the concept further by integrating i-prov into a dataflow-based environment. Unlike the approach in [Psallidas and Wu 2018], which is tailored to instrument interactions within a specific visualization, our approach supports the capture of i-prov across a broader range of applications, as Curio is designed to be domain-agnostic.

[Walchshofer et al. 2021] propose a visual approach to the analysis of i-prov, called Provectories, in which user sessions are represented as high-dimensional state graphs based on embeddings. Provectories adopts an exploratory strategy, emphasizing the understanding of user behavior through visual representations. Although this approach represents a step forward, the approach presented in this paper provides a finer-grained capture of provenance, wherein interactions are recorded at the level of attributes and entities. Also, while the approach in [Walchshofer et al. 2021] relies on high-dimensional embeddings of application states, our approach preserves both the structural and semantic aspects of user actions, thereby enhancing explainability. Furthermore, the provenance of each activity of the dataflow modeled to construct the visualization is persistently stored and can be directly queried from a provenance database, easing auditing, retrospective analysis, and the reuse of analytical workflows.

3. Capturing *i-prov* in the Curio Framework

The approach proposed in this paper extends the Curio framework architecture by adding mechanisms to capture, store, and query i-prov. Figure 2 presents the main components of Curio. The architecture consists of a frontend for modeling urban analysis dataflows and visualizing results, along with a backend infrastructure responsible for data management, provenance capture, storage, and access to APIs. Following, we describe the core components of Curio and highlight those that have been extended to support the capture of i-prov.

The frontend of Curio (Figure 1) provides two main modes of operation: (i) workspace mode, where users design urban analysis dataflows, and (ii) visualization mode, which presents a visual analytics interface generated from the modeled dataflow. In workspace mode, users construct dataflows on a canvas by inserting and connecting nodes, which represent the individual activities within the dataflow. Nodes can be resized, repositioned, collapsed, or deleted, and connections are established by linking nodes through interactive handles. Each node is composed of four main elements: (i) a header that displays the node type, (ii) a body that contains the node's code and parameter settings, (iii) a footer

with buttons for toggling between code and parameter views, and (iv) connection handles that enable the creation of edges between nodes. When two nodes are connected, an edge is automatically added to the canvas, visually encoding the dataflow dependencies.

To capture i-prov, we performed several modifications on Curio's frontend. We introduced a monitoring mechanism to detect user actions involving clicks on components that render charts using Vega-Lite [Satyanarayan et al. 2017]. In addition, we extended the visualization component using Vega-Lite specification to include a highlight mechanism triggered by identified user interactions, allowing selected bars within the chart to be marked.

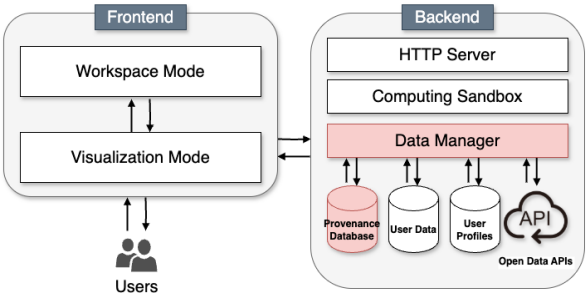

Figure 2. The Architecture of Curio.

Curio's original architecture already supports inter-component interactivity through red connections to a shared datapool (a data table that feeds visual components). With the proposed extensions, when a user interacts with a chart, Curio captures the event, updates the datapool by flagging the relevant entries (*e.g.*, by setting an interactive attribute), and propagates these changes to all connected components. As a result, the affected components, *e.g.*, charts, respond automatically by highlighting the corresponding data points, effectively reflecting the user's interaction in real-time and making it available for provenance tracking. Curio's backend is composed of four main components: (i) the HTTP server, (ii) the data manager, (iii) the computing sandbox, and (iv) the provenance database. The HTTP server handles requests and manages communication between the frontend, data manager, provenance database, and sandbox. The computing sandbox executes the user's Python code within Docker containers.

The Data Manager is responsible for downloading and storing datasets required for processing, maintaining the dataflow state, and collecting provenance data. This component was extended to receive and process user interaction events identified by the frontend, enabling the capture and storage of i-prov in the provenance database. Initially, the provenance database supported two types of provenance: (i) p-prov, which tracks the specification and structure of dataflows, and (ii) r-prov, which records node executions and resulting attribute values. With the extensions proposed in this paper, the provenance database has been enhanced to store i-prov, allowing Curio to trace user interactions with the visualizations. The extended provenance schema used to store these three types of provenance, *i.e.*, p-prov, r-prov, and i-prov, is presented in Figure 3.

The provenance schema consists of several classes that represent different dimensions of provenance within Curio. The classes *dataflow*, *activity*, *relation*, *attribute*, and *attributeRelation* correspond to p-prov, capturing the specification of the dataflow, its activities, and the data dependencies established through relations. On the other hand, r-prov is represented by the classes *dataflowExecution*, *activityExecution*, *relationInstance*, and *attributeValue*, which record execution traces of the dataflow and the computed attribute values. The classes *user*, *version*, and *versionedElement* are responsible for maintaining version control within Curio, enabling the tracking of how dataflow specifications evolve over time. To i-prov, three new classes were introduced: *visualization*, *interaction*, and *attrValueChange*.

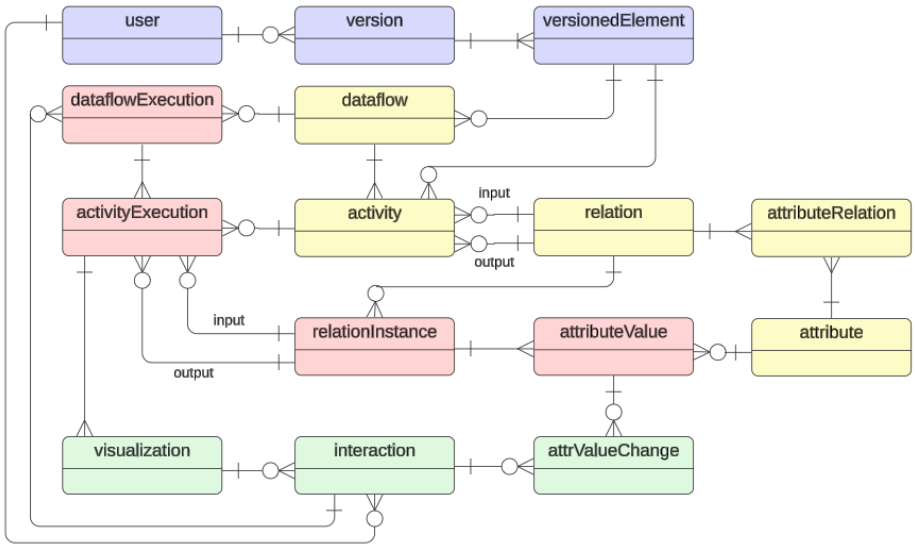

Figure 3. The provenance schema of the proposed approach.

The *visualization* class stores metadata describing the interactive visualizations generated by a dataflow execution. User interactions with these visualizations are captured as instances of the *interaction* class. As previously discussed, such interactions may result in updates to the datapool, which are recorded in the *attrValueChange* class, allowing for detailed tracking of changes to attribute values in response to user actions. The extended provenance schema is designed to be compliant with the W3C PROV recommendation, ensuring interoperability and adherence to established standards for provenance representation. The source code for Curio is available at `https://urbantk.org/curio/`.

4. Experimental Evaluation

This section presents a feasibility analysis in Curio to evaluate the effectiveness of the proposed i-prov capture approach. The dataflow modeled in Figure 1 analyzes crime hotspots in São Paulo, Brazil. It begins with data ingestion using the *Data Loading* component, which imports datasets from SSP-SP [Sá et al. 2021]. The data is then processed in the *Data Transformation* component and passed to the datapool, which manages interactivity between visualizations. Another *Data Loading* component imports a shapefile with São Paulo's geographic boundaries. The dataflow includes two visualizations: a bar chart (using Vega-Lite) and a map (using UTK [Moreira et al. 2024]), both sharing the same dataset. Interactions start with the bar chart, *i.e.*, when a user clicks a bar, corresponding crime locations are highlighted on the map. The datapool captures the interaction, updates its state, and propagates the changes to the map. These interactive links are presented in Figure 1 with red arrows.

We defined four provenance queries to evaluate the effectiveness of the captured i-prov within Curio. Query `Q1` identifies the users involved in a specific dataflow and the activities they interacted with. As shown in Table 1, two users (lyncolnsousa and gmorei3) interacted with the same dataflow. Query `Q2` retrieves which activities within the crime analysis dataflow were affected by interactions with a specific visualization, *i.e.*, the user selects "Mobile Theft" as type of crime of interest, enabling the analysis of how such interactions influence the input data of other visualizations (results presented in Table 2).

activity_name	workflow_id	workflow_name	user
DATA_LOADING-e1c1d0fa-bd51-425f-bdf1-ca17221fec2e	2	CRIMES	lyncolnsousa
DATA_TRANSFORMATION-2a5bdf75-ca49-44fb-ae60-da85b674db62	3	CRIMES	lyncolnsousa
DATA_LOADING-e1c1d0fa-bd51-425f-bdf1-ca17221fec2e	3	CRIMES	lyncolnsousa
DATA_POOL-36979878-c295-415e-bdce-24a6889d885a	4	CRIMES	lyncolnsousa
DATA_TRANSFORMATION-2a5bdf75-ca49-44fb-ae60-da85b674db62	4	CRIMES	lyncolnsousa
DATA_LOADING-e1c1d0fa-bd51-425f-bdf1-ca17221fec2e	4	CRIMES	lyncolnsousa
VIS_VEGA-b1405e01-8b7d-4afc-8a44-06fc977e23cf	5	CRIMES	gmorei3
VIS_UTK-bae88bfc-0a7c-4d7f-9131-7dc440f9be76	5	CRIMES	gmorei3
MERGE_FLOW-a0520f42-9c49-4599-ab18-746de4ec0e3a	5	CRIMES	gmorei3
DATA_LOADING-41c3f954-d8e1-47be-8ddf-eb3725da531f	5	CRIMES	gmorei3
DATA_POOL-36979878-c295-415e-bdce-24a6889d885a	5	CRIMES	gmorei3
DATA_TRANSFORMATION-2a5bdf75-ca49-44fb-ae60-da85b674db62	5	CRIMES	gmorei3
DATA_LOADING-e1c1d0fa-bd51-425f-bdf1-ca17221fec2e	5	CRIMES	gmorei3

Table 1. Results of query `Q1`

activity_name
DATA_POOL-36979878-c295-415e-bdce-24a6889d885a
VIS_UTK-bae88bfc-0a7c-4d7f-9131-7dc440f9be76
MERGE_FLOW-a0520f42-9c49-4599-ab18-746de4ec0e3a

Table 2. Results of query `Q2`

Query `Q3` identifies the users who interacted with visualizations in the crime analysis dataflow (results shown in Table 3). Query `Q4` determines in which executions of the crime analysis dataflow the same interaction, *i.e.*, selecting the "Mobile Theft" crime type, was performed (Table 4). While these represent initial examples of analyses enabled by i-prov, the results show the feasibility of capturing and querying i-prov to better understand the exploration process involved in developing complex urban visualizations.

user_name
gmorei3

Table 3. Results of query `Q3`.

execution_id
12
14
16

Table 4. Results of query `Q4`.

5. Conclusions

In this paper, we address a gap in provenance management for interactive visualizations by extending the Curio framework to capture and store i-prov. While some frameworks already support p-prov and r-prov, they often lack mechanisms to document user interactions that influence analytical outcomes. Our approach aims at enhancing reproducibility and traceability by recording fine-grained user actions, *e.g.*, selections and filters, within visual components. We propose architectural extensions to the Curio framework, including frontend instrumentation to detect user interactions and backend changes to persist them in a provenance database. The resulting W3C PROV-compliant schema supports queries that can be used to reveal how users interact with complex urban datasets. We demonstrate the feasibility of our approach using a dataflow visualization for analyzing crime hotspots in São Paulo. Provenance queries show that i-prov enables users and collaborators to trace back visual decisions, audit workflows, and explore how different interactions shape outcomes. Future work will focus on supporting temporal replay of interactions and extending the approach to other domains and visualization types.

References

Ang, L.-M. and Seng, K. P. (2016). Big sensor data applications in urban environments. *Big Data Research*, 4:1–12.

Bavoil, L., Callahan, S., Crossno, P., Freire, J., et al. (2005). Vistrails: enabling interactive multiple-view visualizations. In *IEEE Vis 2005*, pages 135–142.

Biljecki, F. and Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, 215:104217.

de Oliveira, W. M. et al. (2018). Provenance analytics for workflow-based computational experiments: A survey. *ACM Comp. Surv.*, 51(3):53:1–53:25.

Fekete, J.-D. and Freire, J. (2020). Exploring reproducibility in visualization. *IEEE Computer Graphics and Applications*, 40(5):108–119.

Ferreira, N., Poco, J., et al. (2013). Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE TVCG*, 19(12):2149–2158.

Heer, J., Bostock, M., and Ogievetsky, V. (2010). A tour through the visualization zoo. *ACM Queue*, 8(5):20–30.

Herschel, M., Diestelkämper, R., and Ben Lahmar, H. (2017). A survey on provenance: What for? what form? what from? *VLDB J.*, 26(6):881–906.

Loibl, W., Vuckovic, M., Etminan, G., Ratheiser, M., Tschannett, S., and Österreicher, D. (2021). Effects of densification on urban microclimate—a case study for the city of vienna. *Atmosphere*, 12(4).

Miranda, F., Hosseini, M., et al. (2020). Urban Mosaic: Visual exploration of streetscapes using large-scale image data. CHI '20, page 1–15, New York, NY, USA.

Miranda, F., Ortner, T., Moreira, G., et al. (2024). The state of the art in visual analytics for 3D urban data. *Computer Graphics Forum*, 43(3):e15112.

Moreira, G., Hosseini, M., et al. (2024). The Urban Toolkit: A grammar-based framework for urban visual analytics. *IEEE TVCG*, 30(1):1402–1412.

Moreira, G., Hosseini, M., et al. (2025). Curio: A dataflow-based framework for collaborative urban visual analytics. *IEEE TVCG*, 31(1):1224–1234.

Oliveira, L. F., Oliveira, D., and Frota, Y. (2023). Defining routes for emergency response from climate events: a data-oriented approach. *IEEE Latin Am. Trans.*, 21(10):1064–1072.

Psallidas, F. and Wu, E. (2018). Provenance for interactive visualizations. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, HILDA '18, New York, NY, USA.

Satyanarayan, A., Moritz, D., Wongsuphasawat, K., and Heer, J. (2017). Vega-lite: A grammar of interactive graphics. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Sá, B. et al. (2021). Polroute-ds: um dataset de dados criminais para geração de rotas de patrulhamento policial. In *DSW'21*, pages 117–127. SBC.

Walchshofer, C. et al. (2021). Provectories: Embedding-based analysis of interaction provenance data. *IEEE TVCG)*.

Yu, B. and Silva, C. T. (2017). Visflow - web-based visualization framework for tabular data with a subset flow model. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):251–260.