Journal of Visualization and Interaction

Provenance Visualization as an Entry Point to the History and Curation of Information Collections

Tomas Vancisin, Mary Orr, Loraine Clarke, Uta Hinrichs

A peer-reviewed publication in the Journal of Visualization and Interaction (JoVI).

This article is an extended version of a publication from the Information+ Conference 2023.

DOI: 10.54337/jovi.v1i1.8436 (the public peer reviews can also be found at this link)

- Submitted: 2024-05-03
- Accepted: 2025-03-07
- Published: 2025-04-22 (version 1)

Publisher

Aalborg University Open Publishing
Kroghstræde 1-3
9220 Aalborg Øst
DENMARK
ISSN: 2794-5502

Provenance Visualization as an Entry Point to the History and Curation of Information Collections

Tomas Vancisin, Mary Orr, Loraine Clarke, Uta Hinrichs

Abstract—Provenance—the origin and construction methods of an artifact or piece of information—is essential across all areas of knowledge production. Its disclosure ensures authenticity, reproducibility, and transparency. While digital tools can automate provenance tracking and disclosure, the amount and complexity of provenance information present a challenge, particularly within the context of cultural collections, where 'physically-born' artifacts are transformed into digital space. This process introduces several methodological and curatorial decisions that, in turn, can greatly influence how the—once-physical—collection is represented and how it will be interpreted. In previous work, we have started to address this issue by introducing *provenance-driven visualization* as an approach to provenance disclosure that (1) traces and categorizes the physical and digital provenance of information collections (e.g., transcriptions, modifications of content and structure, ex/inclusion of information items) and (2) uses visualization to disclose and make provenance explorable in interactive ways. While this approach has shown potential, there are challenges in designing provenance-driven visualizations because they can be perceived as complex and abstract and, ultimately, a distraction from the information collection's content. *How can visualization design navigate tensions between making visible provenance information and underlying curatorial decisions in a holistic and compact way, while enabling easy entry points to and promoting the critical interpretation of the collection's content?* In this paper, we present a new design approach to provenance-driven visualization that combines abstract visualization, textual descriptions, and representations of artifactual form with storytelling techniques and animation to portray provenance information. Our findings from a qualitative study demonstrate the success of this approach in (1) providing a visual entry point into the collection's provenance, (2) promoting an in-depth understanding of the transitions and underlying curatorial decisions the collection has gone through, all the while (3) positively influencing the collection's content exploration and its critical interpretation. Our work contributes new perspectives on how visualization can be applied to add transparency and raise awareness of the constructed and situated nature of data, especially in the context of cultural collections, but also beyond.

Index Terms—Provenance, Visualization, Provenance-Driven Visualization, Historical University Records, Digital Humanities

1 Introduction

Provenance, i.e., the origin of an object and its transformation over time, is an important aspect that underpins research across disciplines. While some research fields (mostly humanities) focus on provenance as a way to promote authenticity and originality of artifacts [16], others (mostly sciences) study provenance to provide validity of data collection and as a way to enable reproducibility [19]. Provenance not only provides the foundation for a sensible interpretation of and discussion about the presented information items (be it a piece of art, the results of an experiment, or the interactive visualization of a cultural collection), it also ensures transparency of the underlying methods applied. This can shed light on the methods' soundness and validity but also on the ethical dimensions of these methods (e.g., how were information items obtained, who was involved in the labor, was consent provided, etc.).

In the digital age, the amount and nature of provenance-related information further emphasizes the importance of provenance disclosure, but also adds to its complexity. Provenance information includes (a) curatorial decisions related to in/exclusion of certain information, (b) changes to artifactual representation (e.g., from physical to digital format), (c) differences in information classification and imposing high-level structures (e.g., indexing), (d) 'cleaning' processes to address missing information or uncertainty, and (e) design decisions that determine the (visual) representation of the information collection as part of or at the end of the process. Traditionally, none of these manipulations and inherent situated curatorial decisions are visible in the resulting datasets or their visual representation. The issue of tracing and making *data provenance* transparent has started to become a focus of various data analysis tools [8, 35, 44]. However, these tools are mostly applicable to digital, machine-processable data collections that are 'born-digital' and more or less structured. It is difficult to apply these tools to information collections that have undergone both qualitative and quantitative changes in terms of content and structure, e.g., as part of a digitization process. Furthermore, making provenance as well as related transformations of a collection visible is still a challenge. While selected visualization approaches have very recently been explored [9], it is still most common to describe provenance via text or diagrammatic flowcharts (e.g., [24]) that provide a high-level view on the applied transformations but do not portray their complexity and potential implications.

To enable a holistic perspective on provenance that considers manipulations of the information collection as a result of curatorial decisions which informed transitions from physical to digital form, as well as transformations applied to the collection after its digitization, we have previously proposed *provenance-driven visualization* [46]. The aim of provenance-driven visualization is to capture and make visible the transformation steps that have been applied to an information collection—be it a scientific data set or collection of historical documents. Rather than providing a high-level generic overview of the transformations applied, provenance-driven visualization makes visible each transformation step at a much more granular level. As we have demonstrated previously in the context of a historical record collection, a provenance-driven visualization could make each transformation of this collection visible at the level of individual records within the collection [46]. We believe that provenance-driven visualization could be applied to research across a variety of fields, but it is particularly relevant in the field of digital humanities, where the information collections often span the physical and digital realms.

- *Tomas Vancisin and Uta Hinrichs are with the University of Edinburgh. E-mail: {tvancisi | uhinrichs}@ed.ac.uk.*
- *Mary Orr and Loraine Clarke are with the University of St Andrews. E-mail: {mmo | lec24}@st-andrews.ac.uk.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Using our first provenance-driven visualization prototype, we have already conducted a qualitative study where we employed a range of experts and non-experts in archiving, history, and visualization [46]. The study revealed the potential of provenance-driven visualization to promote a more in-depth understanding of the situatedness of information collections and related curatorial processes. However, it also highlighted the challenges of this approach in terms of design and overall impact. While the participants acknowledged the importance of making provenance visible, they also found our initial design approaches too abstract and complex, distancing them from the actual information and stories inherent in the collection. This raises the question: ***How can visualization design navigate tensions between making visible provenance information and underlying curatorial decisions in a holistic and compact way, while enabling easy entry points to and promoting the critical interpretation of the collection's content?***

Based on the feedback we gathered from the first study, we developed a completely new provenance-driven visualization prototype aimed at addressing the question above from a design and qualitative research perspective. Our contributions are as follows:

- Presentation of a new design approach to provenance-driven visualization that combines abstract visualization, textual descriptions, and representations of artifactual form with storytelling techniques and animation to portray provenance information.
- Findings from a qualitative study that demonstrate the success of this approach in (1) providing a visual entry point into the collection's provenance, (2) promoting an in-depth understanding of the transformations and underlying curatorial decisions the collection has gone through, all the while (3) positively influencing the collection's content exploration and its critical interpretation.
- A discussion of new perspectives and open questions on how provenance-driven visualization can be applied to add transparency and to raise awareness of the constructed and situated nature of data, particularly in the context of cultural collections, but also beyond.

2 Related Work

Our research draws from current practices on provenance research and its disclosure. In the following paragraphs, we outline how provenance is discussed in different fields of knowledge production, focusing in particular on methods of provenance tracking and disclosure.

2.1 Provenance across Disciplines

Provenance tracking and disclosure play a crucial role across disciplines, although the nature of objects in focus as well as the goals of provenance disclosure may differ. To emphasize the extent to which the study of provenance may differ, we focus on two examples. One is applied mainly in the humanities, and another is commonly used in the sciences. It is crucial to highlight that variations and overlaps between the two approaches may exist, but for simplicity and clarity, we only focus on those two, and we refer to them as 'humanities provenance' and 'science provenance'.

'Humanities provenance' can be defined as *"the history of ownership of a valued object or work of art or literature"* [34] which ensures **authenticity** and **originality** of a given artifact [16]. Researchers gather data about the artifact's creator, names of current and past owners, how these owners acquired the artifact, geospatial and temporal information about the artifact's travels, as well as gaps indicating where provenance is not documented. The extent to which this type of provenance is important can be seen in the standards GLAM institutions follow [1] [2] and also in the provenance convention adopted internationally to prevent illegal trade of cultural property [29, 45].

'Science provenance' can be described as *"information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness"* [36] which enables **reproducibility** and ensures **validity** of the research outcome [19]. Researchers try to capture a great variety of provenance data, and for this purpose, they develop provenance-tracing tools [4, 6, 27, 32]. These tools are often specialized for a particular kind of provenance they capture [39]. Whether it is the provenance of data and its changes over time [23], visualization and its states [15], history of interactions with the given tool [33], insights from the analysis [9, 21], or the history of decisions and reasons behind these decisions [28]. The applicability and far-reaching impact of these tools can be seen in the variety of disciplines where they are used: quantum physics [18], psychiatry [2], cosmology [1], ecography [37], biomedical research [43] etc.

Apart from the fact that one focuses mainly on the originality of the artifact while the other on the reproducibility of an experiment, there are further differences between how the two approaches treat and disclose provenance. In 'humanities provenance', tracing the history of an artifact always tries to go back to when the artifact was created, and all the subsequent steps are taken into consideration, up to and including its current state. In contrast, even though in exquisite detail, 'science provenance' traces the transformation steps only within the 'microcosm' of a given project or experiment. The disclosure of provenance in these two approaches differs also. Textual descriptions are by far the most common way to show the 'journey' of an artwork over time, while diagrammatical approaches and scientific workflow charts are popular in empirical sciences.

2.2 Provenanance in Information Visualization and Digital Humanities

In information visualization and digital humanities, the 'artifact' for which provenance is traced and disclosed is often an interactive visualization and its underlying dataset. Even though not in such depths as in the approaches discussed above, the potential benefits of revealing the history of this 'artifact' have also gained attention.

In the context of information visualization, Dörk et al. highlight **trust** between the visualization researcher and the visualization user as an outcome of disclosing provenance [13]. Hullman & Diakopoulos [26] emphasize that disclosing provenance leads to **transparent research**, whereas **replicability** was pointed out by Fekete & Freire [17] as yet another benefit. Further still, Correll [10] stresses that provenance disclosure ensures that the research outcome can be **critically evaluated**.

In digital humanities, D'Ignazio & Klein emphasize that the labor behind a particular research outcome, in this case, a visualization, deserves attention as a means for more **ethical research** [11, 12]. Lamqaddam et al. highlight **user trust** and **authenticity** as two crucial benefits of disclosing provenance [30].

Current practices of disclosing provenance in these two fields resemble 'science provenance' in terms of how far back in history researchers go. Provenance is studied and shown only within the 'microcosm' of the given visualization project, and what happened before is disregarded. When it comes to the form in which the history behind the visualization is portrayed, information visualization and digital humanities use diagrammatical as well as textual approaches, both of which are common for provenance disclosure across a variety of fields. Our approach to some extent complements the current practices in information visualization and digital humanities but its primary aim is to push the entire area of provenance disclosure in these two fields much further by taking inspiration both from 'humanities' as well as 'science provenance' approaches.

[1] https://www.ifar.org/provenance.php

[2] https://museum.cornell.edu/provenance-guidelines-resources

3 Our Previous Work on Provenance-Driven Visualization

In order to address the challenges described above, we have previously introduced provenance-driven visualization as a concept that aims at (1) integrating provenance information directly into the visualization of information collections (such as historical document collections) and (2) making provenance visible and traceable at the level of individual information items within the collection.

Provenance-driven visualization borrows from and expands 'humanities provenance' and 'science provenance' practices. Drawing from the former, provenance-driven visualization aims at taking a holistic approach to provenance tracing and disclosure, i.e., it includes transformations applied to information collections both in their physical and digital form. At the same time, provenance-driven visualization borrows from 'science provenance' practices in that it traces and makes visible transformation steps applied to a collection at the item level. This includes transformations such as translations, edits and expansion of items' content, exclusion of items, modifications to the structure and organization of the information collection (e.g., the introduction of categories or transitions, from a temporal to an alphabetical index), and changes to artifactual form (e.g., from handwritten parchments to printed paper, to digital database), in addition to temporal data related to the duration of individual transformation steps, when these transformations took place, and who was involved in these transformations.

The key to provenance-driven visualization is then the visual portrayal of this qualitative and quantitative provenance data that can be gathered directly from the collection, from reports and writings about the collection, as well as from direct discussions with people involved in working and transforming the collection in question. In contrast to visualizing the data in its final stage, common in information visualization practice and research, provenance-driven visualization also focuses on representing the provenance data that has led to the final dataset into which the initial information collection has been transformed. Below, we provide a brief overview of a case study we have presented in our previous work [46] to illustrate provenance-driven visualization on a practical example and highlight the questions and challenges this example brought to the fore.

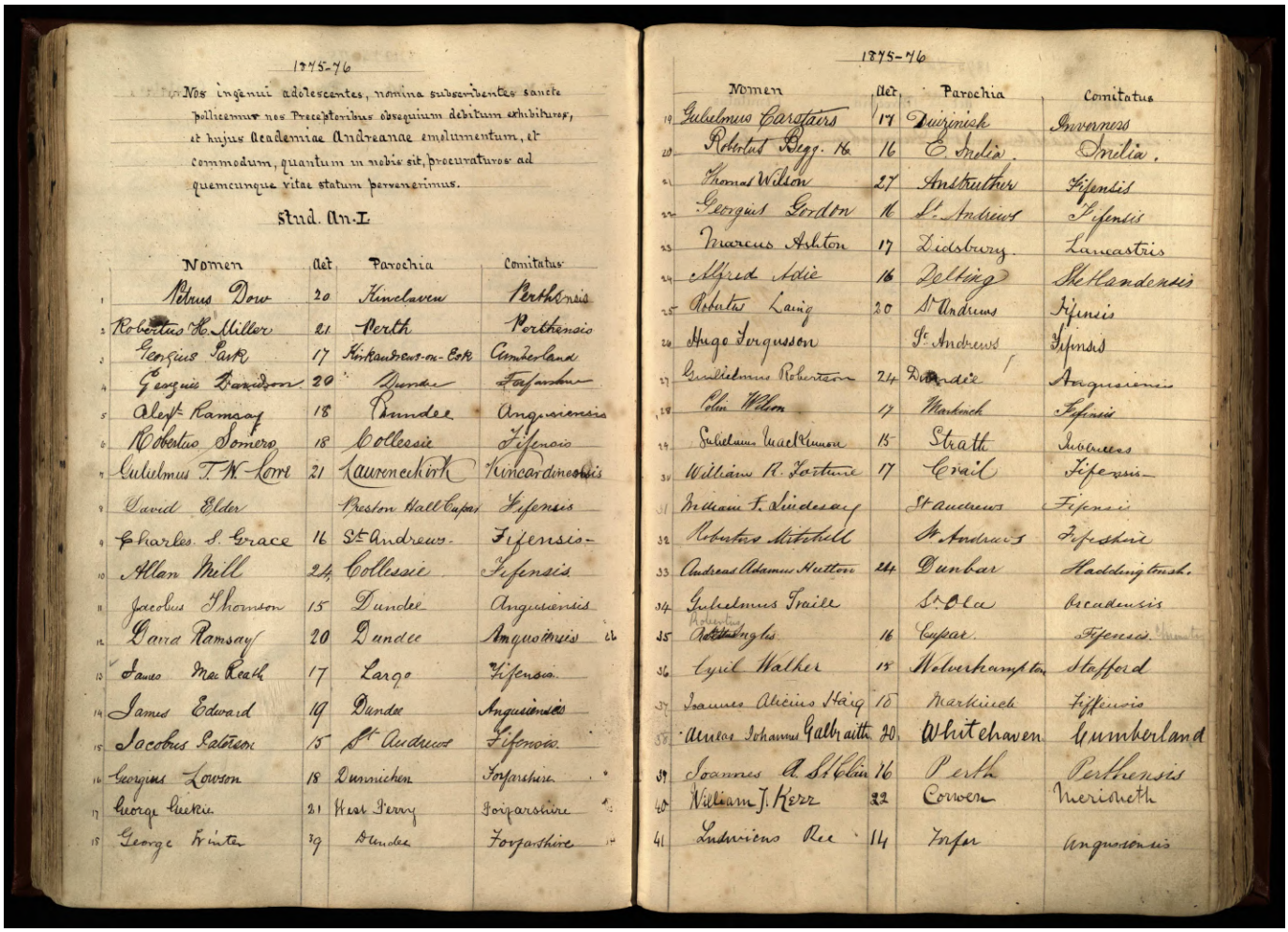

Fig. 1: Matriculation/Graduation Roll (1875–1876). Image Courtesy of the University of St Andrews Library [UYUY309].

3.1 Case Study

The provenance-driven visualization prototype we have presented in previous work [46] is based on a collection of historical university student/staff records—the Biographical Records of the University of St Andrews (1579–1897). These records (see Fig 1 for an example) contain information about student and staff dates/locations of birth/death, parentage, when they studied at the University of St Andrews, what subjects they took, as well as their career after the University. There is and always has been an interest in these records from the perspective of the University's history, but also considering the development of knowledge institutions within the UK and beyond and who had access to them. As a result, the records have gone through a variety of transformations that have had an impact on their content, structure, organization, and artifactual form. We provide a more comprehensive overview of these transformations in [47], but in a nutshell, the records have been transcribed, expanded to include more information about individual students and staff members, re-structured from a temporal to an alphabetical index, digitized and further categorized

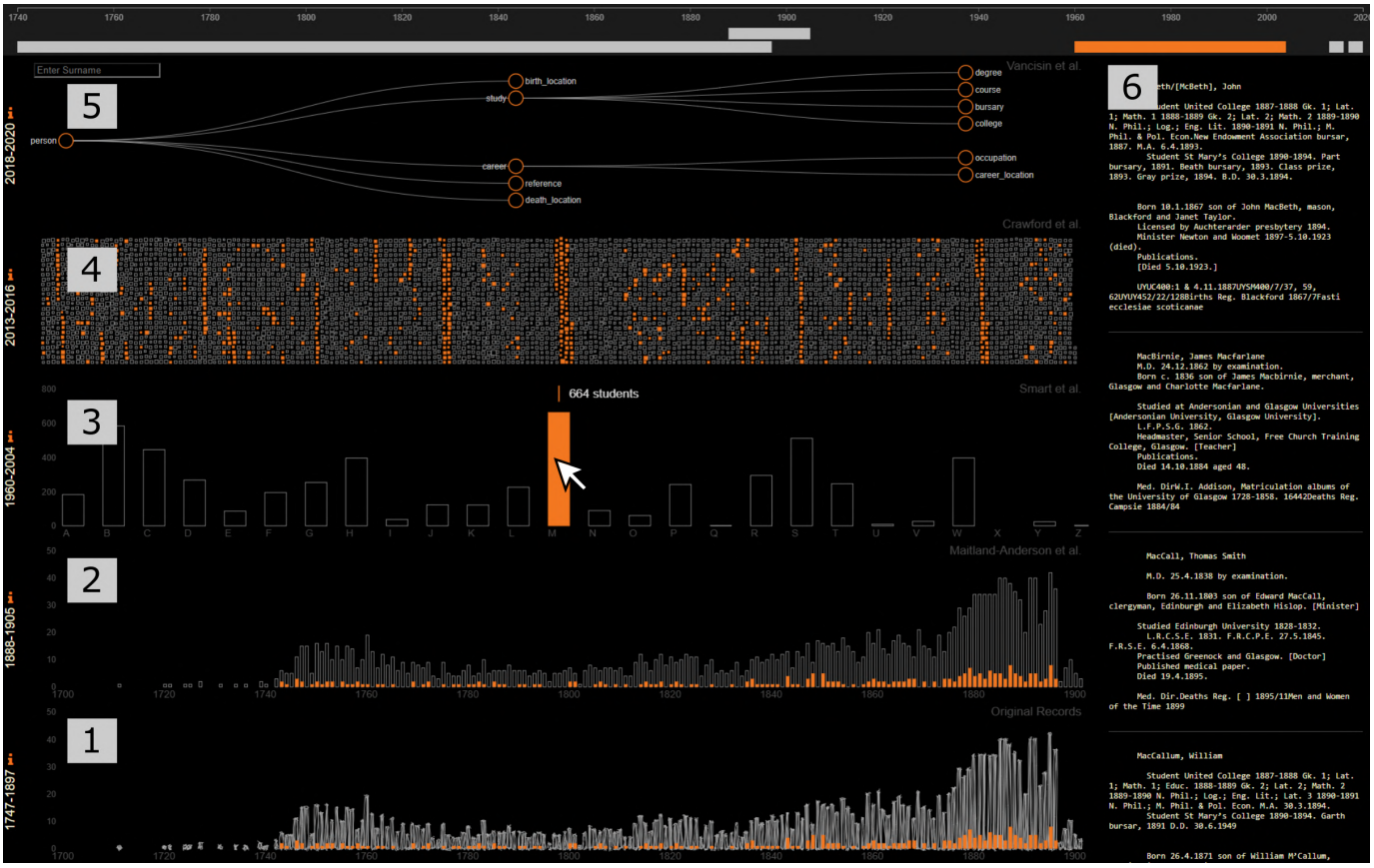

Fig. 2: Provenance-driven visualization prototype showing sequential layers of transformation steps. Layer 1: Original records. Layer 2: Manual transcription. Layer 3: Expansion & restructuring. Layer 4: Reorganization & restructuring into XML:TEI. Layer 5: Relational database structure. 'Record View' (6) shows the artifactual form of records in each layer. Selecting records in one layer, in this case Layer 3, highlights these records in all the other layers. The prototype was developed entirely using the JavaScript library for dynamic web-based visualizations, D3.js.

to enable the development of a relational database for flexible querying, and then visualized. Our initial visualization work on making visible geographic, temporal, and gender-related trends within these records [47] started to bring to the fore provenance-related questions about their constructed and curated nature. This motivated us to collect both qualitative and quantitative provenance data (e.g., content expansion in terms of word units, changes in the structure and order of these records, information describing their artifactual form) about these transformations through interviews with people who were directly involved in transforming the records, by studying library catalogs, prefaces to transcribed books and alphabetical indexes, as well as 'About' pages for the digitized register [48]. We then experimented with ways of visualizing the transformation steps which led to a first visualization prototype that illustrates the idea of provenance-driven visualization (see Fig. 2).

The visualization prototype (described in [46] in more detail) provides an all-in-one overview of five key transformation stages of the historical records. Each visualization layer 1–5 represents the content and structure of the records after the transformation, with Layer 1 representing their original temporal structure. Within each layer, records can be explored to reveal their representational form and content after each transformation step (see Fig. 2.6).

3.2 Challenges and Open Questions

We evaluated this prototype by conducting a study with people from various backgrounds (visualization, archival work, history, and the general public) which revealed that making provenance-related information visible through interactive visualizations promotes discussions about the records beyond their content and around *transparency*, *authenticity*, and *ethical considerations*, as well as their collection, curation, and representation (see [46] for more details). At the same time, however, the participants' interest and focus was often on the records' content as they found the provenance-driven visualization (due to the visual abstraction of transformation processes) too complex.

In parallel, our study of the historical student/staff records revealed that the initial collection we worked with was missing thousands of records of female students who obtained a Lady Literate in Arts degree from the University of St Andrews between 1877 and 1897, which was one of the first university programs in the UK where women were allowed to enroll [42]. These records were deliberately excluded by archivists from the initial University biographical records because the LLA was not considered a 'proper' degree. Our first experiments with provenance-driven visualization did not consider these types of curatorial decisions. Still, our initial explorations raised the following questions and challenges:

- How to navigate the tension between the compact and abstract representation of record transformations while still foregrounding the records' content and artifactual form?
- How to visualize provenance-related transformations in ways that can provide easier entry points into provenance exploration?

It is these questions that we address in this article. In the following sections, we first describe how we explored these challenges from a design perspective and then describe the findings of a qualitative study conducted with visualization experts, archivists, and historians.

4 Conveying Provenance through Visualization + Storytelling (New Prototype)

To address the questions above, we experimented with different visualization approaches. First, we combined text, visualization, and artifactual representations as a means to bridge abstract visualization with concrete artifactual representations of the records at different stages of the process. Second, we used storytelling techniques and animation to gradually introduce users to the transformation processes the historical University records underwent. Using narrative visualization and storytelling techniques has shown to enhance comprehension [41] and memorability [3], and combining it with animated transitions can help with user engagement and understanding [20]. The resulting prototype follows the 'martini glass' structure which first takes the user through the different transformation processes in a linear and interactive way, before they arrive at an interactive visualization where the St Andrews student and staff records can be explored in more detail from a geographic perspective. Our prototype is available online for exploration [3] and below we provide an outline of its design and features.

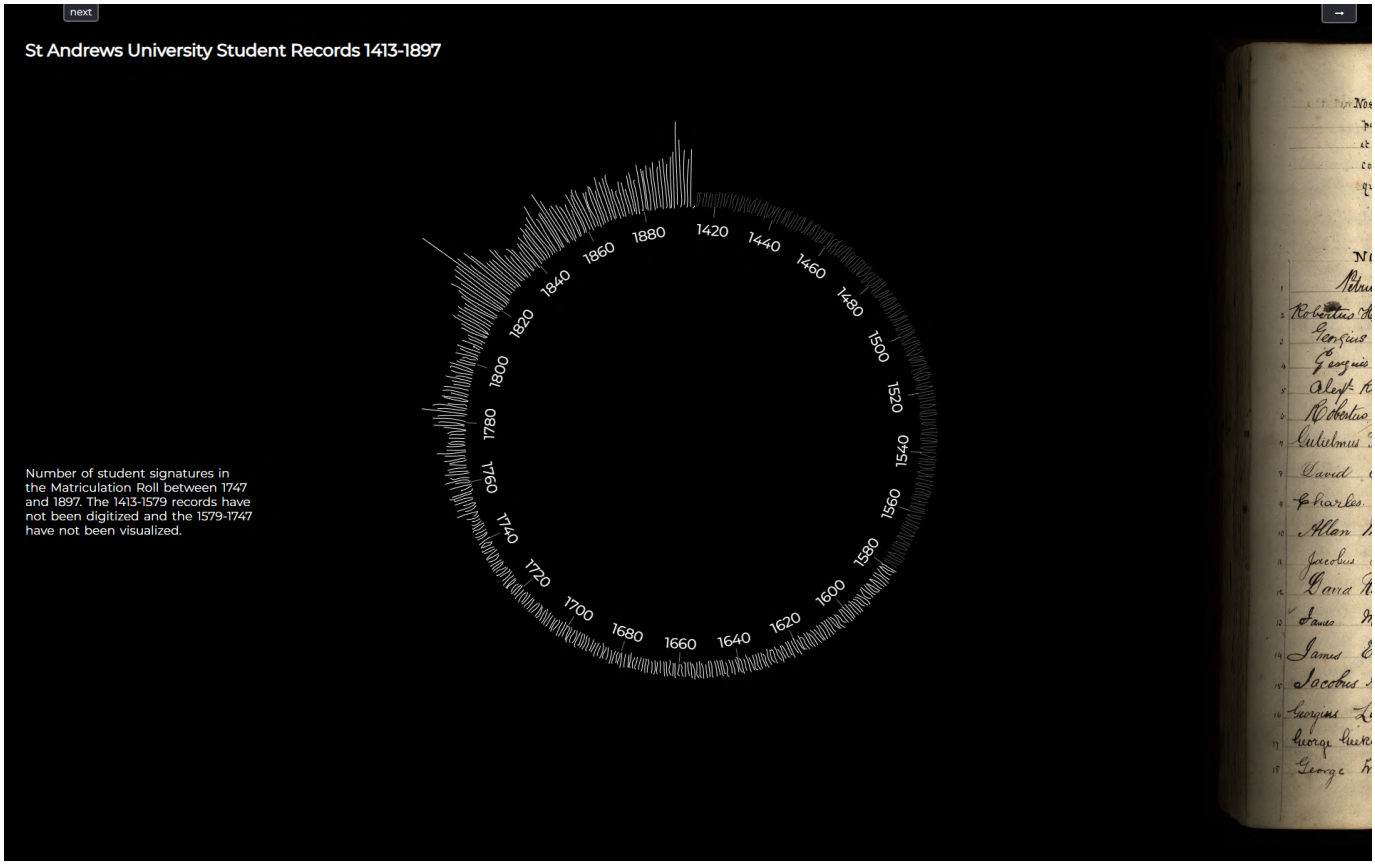

Fig. 3: Original Records: Initial screen shows the number of signatures in the Matriculation Roll using sketchy line rendering to mimic the handwritten form of the records. 1415-1579 records in lower opacity indicating their current manual transcription. 1579-1747 records not ready for computational aggregation. 1747-1897 records' line length indicating number of students per year.

4.1 Illustration of Transformation Steps

As shown in Figure 3, the visualization prototype consists of three elements: a visualization at the center, a textual description of the visualization to the left, and a visual hint at the artifactual representation of the historical records to the right. The visualization consists of a circular timeline that presents the creation years of the records (1413—1897) in clockwise order. Each line represents a year of matriculation. The timeline is divided into three periods 1413–1579, 1580–1746, and 1747–1897, which is based on the way the archivists divided the records: the medieval University: 1413–1579, the University between the Reformation and the creation of the United College 1580–1747, and the University after the creation of the United College 1747–1897. The lines representing the time period from 1747–1897 vary in length to indicate the number of students who started at the University that year. In contrast, the records from earlier years are represented as lines of the same length. These records are in the process of being manually transcribed (1413–1579; as indicated by the light grey color of lines) and digitized (1580–1746; as indicated by the white lines); these records were not ready for computational quantitative aggregation when we developed our visualization prototype, but we wanted to include them in our visualization to make the early years of the University visible. Clicking the top right button expands the scan of the matriculation roll page to contextualize what the original records looked like.

From this initial view of the original records and their timeline, the user can walk through and explore the different transformations the records have gone through over time step-by-step (See Fig 4). The 'Next' button at the top left takes the user through each step. This transforms the circular timeline into a dual timeline where the original creation years of the records remain visible in a circular fashion, but the years of key transformations (1888–2022) appear as a radial outer axis. With every transformation step, the timeline gradually expands radially, making visible the nature of this step and how it has affected the records. At the bottom left, little figurine icons indicate the number of people involved in each transformation step. For example, Figure 4.1 shows the manual transcription of the records from 1747–1897; the straight lines indicate the transcription from the handwritten text to print, and line length indicates the duration of this transformation (17 years). Figure 4.2 shows

[3] https://provenance.tomasvancisin.co.uk/ [Source Code: https://github.com/tvancisin/provenance_storytelling]

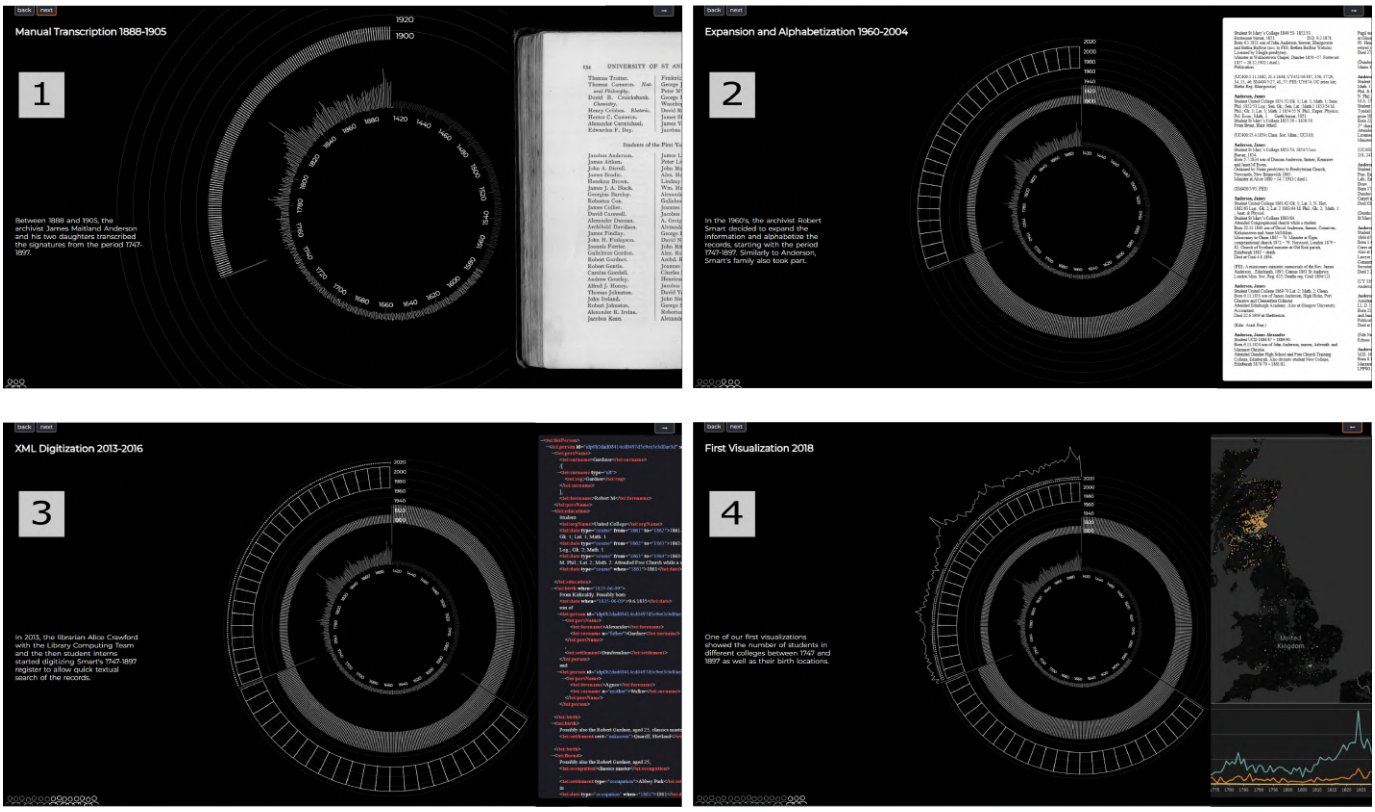

Fig. 4: Transformations of BRUSA records. 1: manual transcription by the Andersons; 2: content expansion by the Smarts; 3: XML:TEI digitization by the University Library Research Computing Team; 4: our first Tableau Desktop visualization.

the expansion of records which added information about students and staff, as well as the transformation of the records into an alphabetical index (1960–2012). This was followed by a full digitization and re-structuring of the records into XML:TEI (2013–2016; see Fig. 4.3). Finally, Figure 4.4 shows a further digital transformation of the records into a database which enabled first visualizations of the geographic distribution of birthplaces of St Andrews students and staff as well as the temporal distribution of the records.

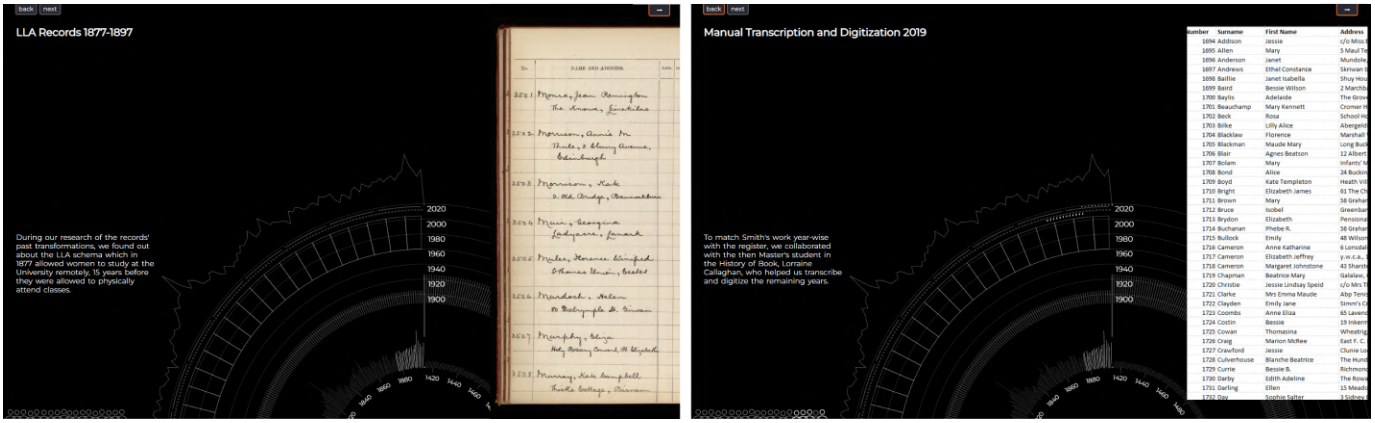

Fig. 5: LLA records. Highlighted sketchy lines at the bottom indicate the period 1877–1897 when the LLA's were originally recorded. Left: original handwritten records; Right: our digitization of the LLA records.

4.2 Inclusion of Missing Records

As mentioned above, our initial work with the student/staff records revealed that the records of women who were enrolled in the University's LLA program (1877–1897) were excluded from the official biographical register of the University because the LLA was not considered a 'proper' University degree. Hence, these records of female students at the University did not undergo the same extensive transformation processes as the original records and were not visible in our first visualizations [46,47]. The last three transformation steps in our visualization prototype therefore focus on the nature of these records, their late transcription and digitization process (see Fig. 5), and the impact our inclusion of these records had in terms of student numbers (see Fig. 6).

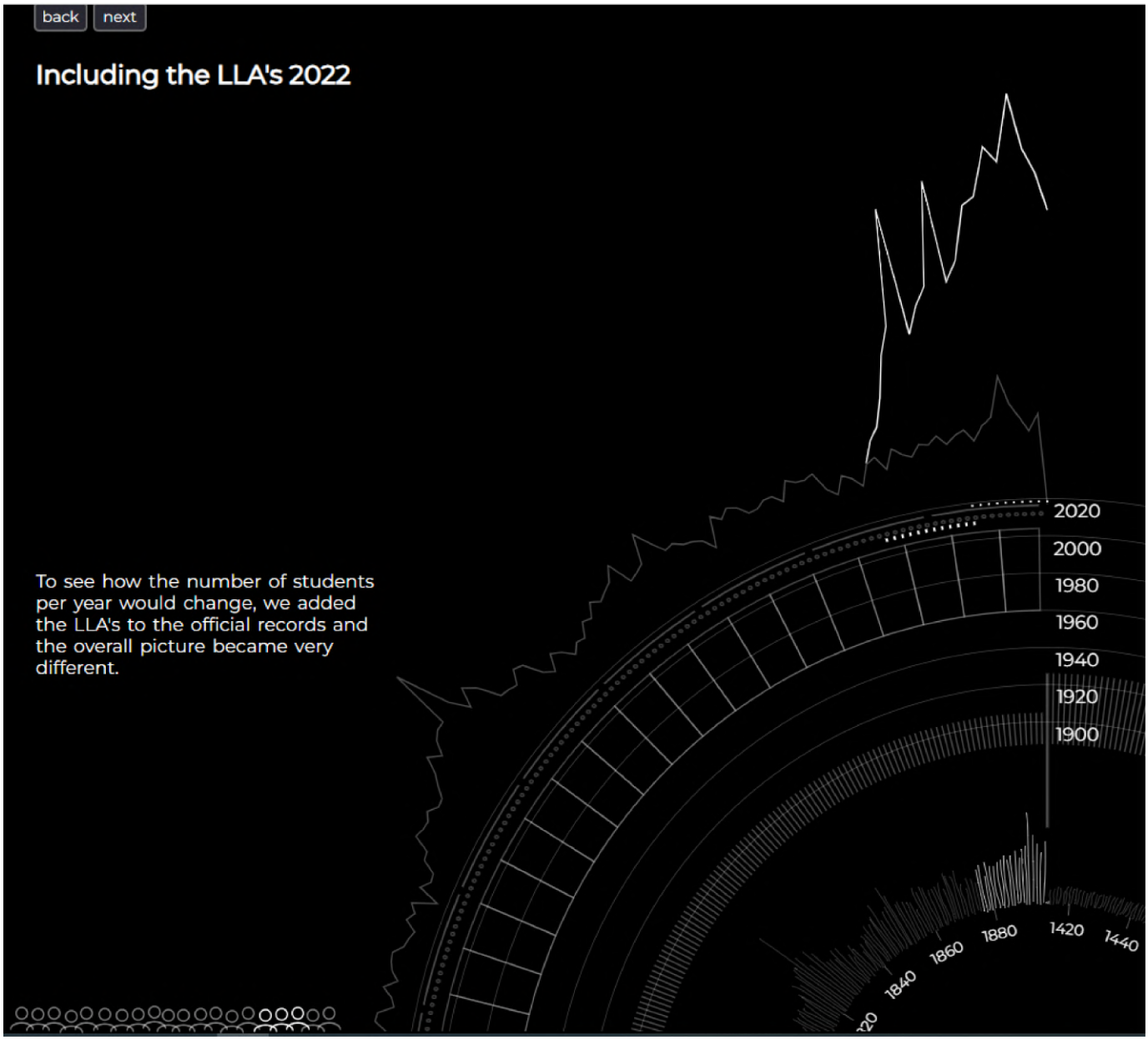

Fig. 6: Contrasting the number of LLA records (white line) with the official register (grey line) shows the scale and popularity of the LLA schema.

4.3 Geographic Visualization of University Records

The step-by-step visualization of record transformations described above finally leads to a geographic visualization of student/staff records which allows the exploration of where they came from (birthplaces; see Fig. 7), where they pursued additional degrees after leaving St Andrews, and where their subsequent careers took them after University. We decided to focus on the period of 1877–1897 to allow the comparison of female and male students who entered the University at the time because this includes the LLA students. Some extracts of these geographic visualizations are shown in Figure 7. The map views can be explored interactively, revealing individual student records by location (see Fig. 7.right). Filtering options allow the user to explore all records at once, or to filter by male and female students. This filtering reveals the discrepancy between male and female student records in terms of numbers but also when it comes to content: while we only have a few records about women whose biographical information was expanded with additional information about, e.g., their careers after leaving the University, this information does not exist for the LLA students because their records were never expanded.

To summarize, this prototype explores how to gradually reveal and narrate the transformation steps involved in transitioning the historical University records into digital and visual form by combining textual elements, visualization, and representations of artifactual form. As illustrated, the prototype follows a 'martini glass' structure where the user is gradually guided through these transformation processes in a linear way, leading to a final geographic visualization that allows the exploration of what is actually within student and staff records in a more open-ended way. With this new prototype, we were interested in investigating (1) how the visualization would be used and interpreted by people with different backgrounds, and (2) if and how it would trigger critical discussions of provenance and curatorial decision-making in the context of historical collections, also in comparison to comments we received about our previous provenance-driven visualization design approach [46]. We therefore conducted a qualitative study of a similar nature, as described below.

Fig. 7: Visualizing geographical contents of the records: Paths from birth locations to St Andrews. Left: male students; Middle: female students; Right: remote LLA (female) students.

5 Study

The visualization prototype described above can be considered a departure from and an expansion of our previous approach to provenance-driven visualization [46] in that it combines visualization and representations of artifactual form with text and storytelling elements to communicate provenance. To explore if and how this approach can bring across and, potentially, act as an entry point to understanding the provenance of the historical University records and its impact, we conducted a study with people from different professional backgrounds and knowledge about the records. We were particularly interested in the following questions:

- What types of insights can participants extract from our visualization prototype?
- If and how does the provenance-driven visualization influence their understanding of the collection's transformation steps and linked provenance?
- How do participants experience and make sense of the abstract visualization of transformation steps on the one hand, and the records' content and artifactual form on the other?
- If and how can the visualization highlight and promote discussions of curatorial decisions that took place during the transformation of records into digital form?

5.1 Participants

12 participants (8 females and 4 males) took part in our study; six had a background in computer science, specializing in human-computer interaction and visualization, and six specialized in history and/or archival practice in the context of interpreting historical document collections. Three of the HCI/visualization specialists and three of the specialists in history/archival work had participated in our previous study [46] described earlier. The six participants who did not take part in our previous study had no or very limited knowledge about our research. Participants were recruited from the School of Computer Science (St Andrews and Edinburgh University), the University Library Research Team (St Andrews), the Special Collections Department (St Andrews), and the School of History (St Andrews) through official University channels (email circulars and print advertisements).

5.2 Study Design & Procedure

Our study was of qualitative nature and included observations of participants interacting with our prototype, followed by semi-structured interviews. Study sessions took approximately one hour and were conducted online with individual participants. Each session started with an introduction to the visualization prototype and its functionality. We then invited participants to explore the visualization in an open-ended way, based on their own interests and curiosities, without giving them particular pointers or tasks. During this exploration period, participants were encouraged to comment on their experiences and ask any questions that came up during these explorations. This was followed by in-depth interviews, which focused on participants' understanding of the visualization and the insights they gained from it. In particular, we focused on insights and questions related to the provenance of the historical University records, but also on the information within the records, and any questions the visualization had raised.

5.3 Data Collection & Analysis

All interactions with the visualizations were recorded via screen captures and all interviews were audio recorded. Our qualitative data analysis mostly focused on the audio recordings which were transcribed word-by-word (using an automatic pass first, followed by a manually refined pass). The transcripts were then analyzed using a thematic analysis approach [7, 22] where we iteratively developed and refined a coding scheme to capture the range of participants' reactions and comments on the visualization prototype. The qualitative coding and development of the coding scheme were conducted by two members of the research team and refined multiple times. Codes were informed partly by the interview questions but also emerged from participant statements. The emerging coding scheme was then further refined through discussions among all four members of the research team. Occasionally, findings derived from the thematic analysis were verified and enhanced by observations of selected sequences of participants' interactions with the visualization prototype.

5.4 Researcher Reflexivity

Because of the qualitative nature of our study, we found it important to reflect on aspects that may have influenced our research process. Therefore, we summarize our background and positionality as researchers and authors of this article, our relationship to study participants, and our biases and assumptions below.

Personal Background & Positionality: We, the researchers on this project and authors of this paper, have an interdisciplinary background spanning computer science (specializing in HCI/information visualization; 1st, 3rd, and 4th author) and the humanities (modern languages and literature; 1st and 2nd author). While our individual background has influenced our approach to this research project in general, and possibly, the interpretation of participant responses, we tried to mitigate this by discussing and reflecting on our research approaches and methods as part of regular meetings (typically, weekly meetings of team diads and monthly meetings with the whole team). These discussions became crucial during the thematic analysis of the participant interviews. The interdisciplinary nature of our team allowed us to critically reflect and balance our perspectives on our research.

Biases and Assumptions: We decided to make the inclusion of the LLA students the focus point of this project and our resulting visualization, as it (we believe) nicely exemplifies the far-reaching impact of studying provenance. While this influenced the design of the visualization, we did not, however, communicate this intention to participants during our study, for example, as part of our interview questions. Instead, as we will outline in our findings below, participants observed the importance of the LLA inclusion in the visualization and highlighted this during the interviews. The LLA inclusion, however, was only one observation that participants made, among other things (as further discussed below).

Relationship to Participants: We deliberately recruited four participants for our study who were deeply familiar with the historical University records, in order to get an expert perspective on our visualization. The lead author had worked with these four participants (three from the University of St Andrews Special Collections and one from the School of History) throughout the project and had already established a professional relationship with them. The six participants with an HCI/visualization background can be considered as a convenience sample; four of them were members of the University of St Andrews HCI group, of which the 1st, 3rd, and 4th authors are also members; two participants were recruited from outside the group.

It is therefore clear that the responses of our study participants represent specific perspectives on the presented visualization. However, we would like to emphasize that there were no power dynamics between participants and our research team, and we deciphered from their responses that participants felt free to speak their opinions and also bring forward critique. While it is clear that widening the participant sample to include more and different perspectives on our visualization approach would be beneficial, our findings from our observations and interviews already present interesting insights into the impact of making provenance visible through a combination of visualization, text, and artifactual form.

Below, we describe the core findings of our study. Individual results are illustrated by direct participant quotes, which are coded as follows: [CS] and [H] indicate that the participant has a background in HCI/information visualization or history, respectively. An asterisk (i.e., [CS*] or [H*]) indicates that the participant took part in our previous study and, hence, has some familiarity with the record collection and visualization approach. Furthermore, each participant was assigned a unique number.

6 Findings

Our findings are structured as follows: we first provide an overview of the types of insights our participants were able to extract from our visualization, both in terms of higher-level impressions of the University records and the transformations they have gone through, as well as in terms of in-depth interpretations of the individual transformation steps. We then describe how individual design elements of the visualization prototype—the data visualization itself, the accompanying text, the representation of artifactual form, and the storytelling elements—have contributed to participants' understanding and interpretation of the records' transformation processes. Finally, we describe how the visualization-based disclosure of record transformations influenced participants' interpretation of the geographic visualization of student records. Finally, we describe the limitations that participants mentioned concerning our visualization prototype.

6.1 General Impressions & Insights

Participants' general impressions and insights they were able to extract from our visualization within the relatively short period of open-ended exploration time revolve around three themes: (1) insights from the student and staff records and their geographic distribution, (2) the extensive amount of time and labor that was involved in processing the University records, and (3) the importance of being aware of these transformations when working with and interpreting such records.

University Students and Staff Geographies The final visualization of student and staff geographies before and after entering the University led to a number of insights and raised curiosity among participants. For example, participants highlighted differences between the birth locations of different student cohorts—those represented by the original records (mostly male students) vs. the female LLA students: "*How much more varied their [the LLAs'] places of origin are. That's right down in the South of Ireland. And there's the sisters... That's interesting. We do tend to have, you know, the association of being the family University.*" [H2]; "*There's bloody a lot of them [LLA's]. They're not all in the UK as well [...] way out here, and I think Austria, Switzerland.*" [CS*6].

Participants also inquired about aspects in the geographic visualization directly linked to the exclusion of the LLAs. They were curious about the lines connecting geographical locations (i.e., students/staff places of origin before coming to St Andrews and places they went to after University) which were not present for the LLA student cohort, partially because of the remote nature of the degree, but also because the LLA records have not been expanded with additional information. In this context, 3/12 participants noticed that a single LLA student whose career path was represented on the map, even though most LLA records do not include career information: "*I had a lot of questions about this [single LLA career] that you answered. I was very curious as to why only one person had a career after the LLA, but it turned out to be a record-keeping issue.*" [CS2]. This particular student started 'proper' studies at St Andrews in 1895, after the 1892 Ordinance[4] which allowed women around Scotland to physically attend university classes. She did not gain a 'proper' degree, but later on started studying for her LLA and obtained it in 1898. Because of her unfinished 'proper' studies, she was included in the official records and, thus, her student record was expanded with additional information.

Time & Labor on the University Records. In terms of provenance, one of the main messages that 7/12 participants took away from the visualization was the extensive time and labor that went into processing and transforming the University records over time. Representatives from both participant groups—those with a background in history and those with a background in HCI/information visualization—emphasized how these aspects are often invisible in the resulting dataset and/or its visualization, and yet how important they are: "*It's the story of the manual labor*

[4] https://api.parliament.uk/historic-hansard/commons/1892/may/09/universities-scotland-ordinances

and how long it takes to produce these outputs. The number of different people involved. Working over successive generations, often bringing in family members and a mixture of students and archivists. The job is really a sort of team effort over a century or more, which is something that I don't think is often apparent to the end user." [H*1]. "*The work and the effort involved in the transformation processes, I think, the storytelling aspect of that, really drives how much work goes into these different processes of digitization. And as someone that works with data in data visualization, I sometimes kind of take it for granted that data is readily available. [...] even a simple visualization would take months of human work to create the transformations required. So that's really impactful for me.*" [CS3]. Another participant, who was involved in the digitization of the records, stressed the importance of the work done on the records pre-digitization: "*The fact that this job of just transcribing handwritten stuff and then gathering further information about these people takes so much more time and so many people, compared to the modern effort, which admittedly, probably has more people attached to it, but for shorter periods of time. I was only there for a year at most. And none of the work that I did and that you've done would have been possible without those [previous transformations]. What is it, 60 years of preparatory work? Without Smart and without the Andersons, none of this would have been possible*" [H*5].

Awareness of Record Curation & Transformation. When describing the insights gained from the visualization prototype, participants often commented on the LLA student records which were—as highlighted in our visualization—missing from the original Biographical Register of the University and only later included by us. In this context, 4/12 participants stressed the importance of being aware of record curation and transformation and the inherent biases for the analysis and interpretation of records—something that they felt our visualization has achieved to a certain extent: "*I'm also doing work on gender bias, so I'm quite intrigued by the kind of existence of this [...] set of records that was found during your work and the story that kind of comes out of that—of all these other women. Like this spike [in the timeline visualization] is just really quite impressive to me. There has been [...] a lot of women that were very excited to be able to enroll in a university, even if they couldn't go in person. And so that to me just opens up a lot of future research directions and says a lot about what isn't being included, and the kind of information that we just don't have, that limits our analysis and limits the ability to be certain about the kind of analysis we can do on historical records.*" [CS5]. "*And it's funny how these women are taken out of it [the initial Biographical Register of the University]. But the doctors who did exactly the same thing; they didn't study in St Andrews, right? They just sent off paperwork that their mentor wrote. He's a good doctor, basically. And, they get their papers. It's the exact same thing. Right? But because they're doctors, they get to be included. Or the honorary degrees. They didn't even do anything, and they get to be included, but then these women don't get to be included because they don't fit the bias of what a student should or what a university graduate should look like. And so I think the metadata and the context of the source is as important as the source to make a complete analysis of something.*" [H*5].

One participant emphasized that making the records' past visible can help portray "*how it changed over time, how many different sources of different types of data there are, and how they could all be combined together to make a more complete set.*" [CS*1]. Another participant even highlighted that awareness and the ability to show the records' transformations can help justify funding applications: "*You also had to know how the source got to you, and this is knowing how the source [...] came into existence. I think that's a very important thing to keep track of and to show, and also, you know, for funding reasons be like 'look this took 60 years to gather. We're gonna need more money'.*" [H*5].

These statements illustrate participants' general impressions and insights and reveal that the visualization prototype was successful in highlighting and creating an awareness of the extent of transformation processes that the University records underwent—both in terms of time and labor, as well as curatorial decisions. Furthermore, the analysis of the interviews revealed no significant differences between participants who took part in the previous study and those with no prior knowledge of our research. Below, we highlight the depth of participants' understanding of individual transformation steps and how different visualization elements contributed to this.

6.2 Understanding of Individual Transformations

The way in which participants were able to describe the individual transformation steps and speculate about their—potentially positive or negative—impact shows that our visualization facilitated an in-depth understanding of the records' transformations.

Transformations to Increase Research Potential & Accessibility. 6/12 participants mentioned that to them the records' transformations mean an increase in usability, accessibility, and research potential of the University records: "*The intention [of each transformation step], I think, is to make the data more usable, more accessible, and to allow different kinds of interrogation of the data by different audiences. Making use of the forms of communication of the day, whether that's a printed catalog, a database, an image on the screen, or a visualization.*" [H*1]; "*[...] it seems like all these transformations are about improving access to information, making it easier for people to discover the records and do further research on them.*" [CS5]. The controversy about the importance of certain transformation steps (potentially due to participants' background) revealed participants' in-depth understanding of these as made visible by our visualization. For example, one of the historian participants highlighted the transcription of records as the most important transformation: "*So manual transcription [...] is hopefully getting [the records] into a form where it's kind of readable. [..] It goes from being one document held in the archive. [...] you've got something that can be distributed publicly where it can be consulted without damage to the original materials, and where it's just much easier to read. It's much more approachable. So that is a massive, massive step forward. [...] and I think the moment you've made that step, you've got a resource that will be used by scholars and that was probably the most important step.*" [H3]. In contrast, a participant with a background in HCI/information visualization stated: "*Obviously XML makes it possible to be parsed via computer. It's not obviously as useful as the database which came later, but it allowed it to be computer-readable and can be understood by computers and because it can be understood by computers, you can get information out of that. Make databases out of it, which they did later on. And this means you can do anything at that point. That transformation to the database, that was the most valuable transformation because at that point you can analyze so much. [...] so like the percentage of students that were female versus male and things like that, which you couldn't do with the original data or any of the previous versions very easily.*" [CS*1].

Transformation Downsides. Participants' statements also reveal an awareness of the potential downsides of transformation processes. For example, 5/12 participants emphasized that with each transformation, contextual information gets lost: "*In terms of that book particularly [the original Matriculation Roll], the way that records are kept is by class [year-by-year]. [...] So what you've got there is the ability to see: Who was in class with each other? Who are the peers? Who was actually sitting next to 'Andrew Lang', when he was a student in the 1860s? That kind of thing. And you don't necessarily easily manage to get that information when you're doing it through your database.*" [H2].

3/12 participants highlighted a loss of context in terms of the human aspect which was present in the original records through individual handwritten student and staff signatures (see Fig. 1): "*I guess, the context sort of disappears a bit as you go through the transformations. You know, when you look at this beautiful page with whatever lovely Latin is going on here, very nice slightly flex nib pen, it has a personality that is missing from this lovely block of JSON.*" [CS*6].

3/12 participants mentioned the potential introduction of errors as part of the record transformations: "*I think it's very, very important to recognize that is the thing that is most essential [original Matriculation Roll]. That's what everything is based on, and that is the thing that we*

have to go back to and check because we've got to reckon. It's possible that at every stage we've added an additional error." [H2]; "*There's always a possibility when you go from one format to the next that, I suppose, records are gonna be missed and some information might be left out accidentally.*" [CS5].

2/12 participants took it even further and emphasized that even the original records might have gaps we do not know about, hinting at the missingness of the LLA records: "*I've been working with some AI. We've had discussions about it, not actual AI, but like talking about the issues with AI and the whole thing with biases and how that skews your data and skews your AI is relevant here as well because like without the LLA, you've skewed the whole thing and you don't even know it. Because it [original Matriculation Roll] looks like a complete register. But we're missing these women.*" [H*5]; "*I mean, because these are physical artifacts, that message about the data representing women's presence in academia, it's very easy to think that when you have a nice, neat database that you have all of the data and that there aren't things missing and that when you're presented with this, it's as it is. But having that book [original Matriculation Roll] and being able to see that there are two separate books [original LLA records] and that there are holes in the information or differences in the way it was recorded gives you more of a feel for just what was left out, and what might not be accurate, and what might be like, you know, the more human aspects of the data.*" [CS3].

We discuss the impact of the different visual elements of our prototype in more detail below.

6.3 The Role of Different Visualization Elements

When asked about the role the different visualization elements—text, visualization, and artifactual representation—played in portraying the nature of the transformations, 6/12 participants stated that they worked well in combination: "*I guess, the text I'm sort of getting who did the work, and more precisely, what it involved, and perhaps why it happened. From the visualization I'm getting when the work happened, how long it took, and what section of the records is being worked on, and perhaps how it relates to previous work. And from the image, I'm getting some idea of what it sort of might have looked like. The sort of manifestation of the transformation.*" [H*1]. However, participants also commented on the role of the individual elements.

The Role of the Visualization. 3/12 participants stated that the visualization element was the most useful for ***portraying the process of transformations***: "*I think they [the different elements] worked nicely together. I think, having the visualization as the kind of focus and having the images of different materials or the data as something that you kind of saw hints of, had to click to see [...] that put emphasis on the process rather than on the output of the transformation, which seems like that's what you wanted. At least that's what I'm getting from this whole visual and the way that you have the captions and stuff on the side explaining the process. But I think at the same time having those photos is super helpful because I don't think the process would be as easy to understand the significance of without seeing that very first format of the records all the way to the end. But I think the way the timeline is centered puts emphasis on the process of the transformations.*" [CS5]. Even participants with prior knowledge of the University records and their background found the visualization to be successful in portraying the background of the records, and were able to extract new information: "*I think the visualized version is more powerful than just a paragraph or text would have been. [...] I knew about 3/4 of this story and, I think, [...] seeing it in the visualized version made me see things that I haven't really picked up despite actually having quite a lot of the background information.*" [H3].

That being said, walking the user through these transformations step-by-step; ***the storytelling element of the visualization***, seemed to be a key element in conveying this process. 4/12 participants explicitly commented on this: *So in terms of the transformation of the data, the circular visualization was definitely most useful. Well, given that I clicked through. I think if you just showed me the end results, I wouldn't have any idea what was going on, but they're kind of building up very well*" [CS2]. "*I like the layering of the history of the time. And then the way that it's kind of combined with this storytelling [...] you're not just presented with all of the information at once. It's really cool how it builds up [...]. It tells a really interesting picture on the levels of details of the various pieces of history. But also it shows just how much work and how many processes have gone into creating this data so that it's usable in the digital format now. And that's super cool. I don't think you would get the same appreciation of the work and the gravity without this build-up. Without the story.*" [CS3]. "*The way that you work through that gradually and then steps to show it [the visualization] building up was really helpful. I think it might be quite difficult to interpret all those different things. Coming at that full-diagram.*" [H2]; "*I think this graph [visualization element] shows it well and as you watch the graph being constructed, you gain the ability to read the graph. I mean, if you were given this graph at the beginning, you have no idea what it meant. But the fact that you see the graph being constructed, I think that it's a valuable way to represent the data.* [CS*1].

3/12 participants highlighted that the visualization element allowed for ***the comparison of transformations***: "*This one [the visualization] was nice because I can compare the Andersons to the Smart to, you know, the work that was done. I don't have to leave one behind to get to the next.*" [H*5]; "*I quite like seeing the narrative of how things have progressed over time. I don't think you necessarily see the connections. I think it's helpful to have a visual representation of the things and how they're connected. Having, you know, the script here from the XML next to just an entry from a calendar [Smart's records] or something, you wouldn't necessarily make those connections. I think having this here [visualization] helps you compare them.*' [H*6].

This comparison also included ***temporal aspects*** as highlighted by 3/12 participants: "*I think it's interesting how much more quickly you've been able to do the recent digitizing. I think honestly, coming at it from a historian's eye view, I think that indicates on the one hand how computers are quite quick at processing data. [...] I am not at all surprised that it took Robert Smart [expansion], the amount of time it did to do that work. [...] I think what you've done is a really, really interesting visualization of how long it actually took for each project, and I think that's a really interesting story in itself. And I think the point you're making that there's all these layers of work to get to the ultimate digital resource that you've got, I think it's a really interesting point and I don't think we spend enough time either as the public or as historians talking about the layers of work that we are reliant on.*" [H3].

Other participants positively commented on our ***visual interpretation of the transformation steps***, as highlighted in the visualization: "*The move from the scratchy to the more ordered, I think, is a really nice reflection of what's going on over here. And I do like this sort of way of trying to visually categorize what is going on between the different transformations, where the value added is in a particular transformation.*" [CS*6]; "*I like the idea of seeing the continuity of the record creation visualized as a circle. That's really helpful because I think that gives a sense of the fact that this is an ongoing task which covers a whole set of years of the University's history, but also is tackled in various times by different people. I think this gives a sense of continuity and effort and endeavor. So I really like seeing that portrayed in a circular fashion and being able to visualize which aspects of the work on the collections were done at what particular time.*" [H2].

Nevertheless, 3/12 participants (from both history and HCI/information visualization backgrounds) also admitted that the ***abstract nature of the visualization element can cause confusion and requires more guidance and explanation***: "*I think with the central visualization, I would need your assistance to understand it.*" [H*4]; "*[The visualization is] a bit too abstract, but not everything lends itself so easily to visualization.*" [CS*4].

The Role of Text. Participants found the textual description alongside of the visualization to be useful, even essential, especially for users unfamiliar with the historical records or the project: "*I think it's super useful information. It explains a lot.*" [CS3]; "*I think they [the descriptions] were essential in that you need to have them to be able to understand what's going on.*" [H2]. Indeed, participants often used the text captions to better understand the visualization and artifactual fom elements, even though one participant did highlight: "*I think that a lot of this visualization does speak for itself.*" [CS*6].

7/12 participants, both those who were new to the project and those who participated in the previous study, found the text and the amount of information to be helpful and sufficient: "*I think as it is, it's probably exactly what you need.*" [CS*1]. One participant even felt the brevity of text to be a feature: "*I am quite happy with the snippets. Because it [the text] can get in the way of everything else [...] You get bogged down with how they did it [the transformations], but it's enough who did it, when they did it, and what they did is perfectly fine.*" [H4].

4/12 participants would have preferred more information, also to better understand what is presented, if nobody was around to ask questions: "*I think I asked you a few questions, I guess, maybe I need a little bit more information because I wasn't clear on a couple of points.*" [CS2]. "*I am a words' person. I will always want more words. [...] I always take more information, but I understand that not everyone likes words.*" [H*5];

The Role of Representations of Artifactual Form. Our visualization prototype presents the impact of each transformation on the records' artifactual form/presentation (see Fig. 4). 3/12 participants stated that these photographs and screenshots—at least for some of the transformations—were useful for providing context. One participant even described them as a *"visual hook"* that provides a better idea about the transformation itself.

Much like the textual descriptions, participants thought the photographs and screenshots complemented the visualization: "*This [visualization] gives you a bit of a hint about what's going on, but it's kind of useful to actually see the difference in the structure of the data.*" [CS*6]; "*With pictures like this, where you can actually see in practice what we start with; [...] that we start with this handwritten thing [the matriculation roll]. And then, we get the transcription by the Andersons and a couple of layers [...], and then you get the Smarts which looks different again. That will help people who haven't worked with it [the historical records] as in depth to kind of understand what these lines [in the visualization] mean in practice. [...] You go from pure kind of old fashioned research into technology and digital tools. And I think without the images for a lot of people, that wouldn't make any sense.*" [H*5]; "*I think the visualization puts the emphasis on their process, but that process becomes more meaningful and the significance of that process becomes easier to grasp when you have the digitized photos and screenshots on the right side.*" [CS5].

In contrast, 3/12 participants considered the photographs and screenshots as non-essential: "*It wasn't really as information-rich because it's just a single screenshot.*" [CS*1]; "*I think probably this image, these screenshots added less than the other two parts. And, of course, they're just screenshots, so you're not really getting a sense of how they were created or used. So I'm not sure, I got that much from the image in the context of the other two parts.*" [H*1].

6.4 Provenance-Driven Visualizations as a Backdrop to Visualizations of Historical Records

As stated earlier, the provenance-driven visualizations (i.e., the gradually building circular timeline) can be considered as a backdrop to the actual visualization of the historical records (i.e., the geographic visualizations) in that the user has to go through the provenance-driven visualization before they can explore the records through the geographical visualization in a more self-driven and open-ended way. Participants appreciated this approach that borrows from the 'martini glass narrative' visualization concept because it highlighted how the historical records shown in the final visualization actually came into place: "*[The provenance-driven visualization is] showing you all this information, all the transformations, how the data went through all this process, but that last step, being able to move across to here [the geographic visualization] and then explore the data... Even beyond just allowing you to explore the data yourself, it emphasizes the actual value of the whole process and what this whole process has enabled. We couldn't have done this with the original data or any of the earlier stages, but because the data is now transcribed and digitized in a database, because of all this, just being able to go across to here [the geographic visualization] and say 'This is the data, now explore it!'... It emphasizes why this whole process was so valuable.*" [CS*1]. Another participant stated: "*I think [the provenance-driven visualization is] a particular lens to look at the final visualization. And it's a lens that doesn't exist unless you do something like this. It's very easy for that stuff to get missed, and it is all part of the history.*" [CS*6].

6.5 Limitations & Open Questions

Our way of combining the two visualization approaches also brought limitations and questions to the fore. For example, 3/12 participants pointed out ***the lack of interactivity*** in the provenance-driven visualization; one cannot interactively explore the records in the circlar timeline. This was supported in our previous provenace-driven visualization work [46] and it certainly is an aspect that should be incorporated in the future.

Our previous work revealed that while people seem to see the importance and merit of provenance-driven visualization, their mindset and focus was still on the exploration of the information within the University records [46]. In this study, participants made similar comments, although the positive impact of the provenance-driven visualization clearly came across. For example, one participant suggested: "*I think you could probably also do a different presentation which went straight into the visualization of the data. I mean, the record-keeping stuff might not be as important to some people. [...] The record-keeping story is interesting, but I know from experience some people just want the answers, and people's questions might be still related to the individuals rather than to the trends. And so you have to choose your tool according to the question that you're asking, I guess.*" [H2]. Along similar lines, another participant stated that the exploration of the provenance-driven visualization would be more of a one-off thing: "*I think looking at this [the provenance-driven visualization] has made a very powerful point. [...] But I suspect I wouldn't consult that very often. I think, having seen it once or twice, it's like, that's a cool thing that made the point. But I've got the point now, if that makes sense. So I would be interested in going back and looking at 'OK, that data on the LLAs and where they came from and all that". I'd go back and look at that. [...] Most of the time I wouldn't be regularly consulting the story of the transcription of the records and the publication of the records, unless, I suppose, you had a lot of information on the methodology, in which case I might possibly consult it for my own, as a model for my own research projects.*" [H3].

The same participant also pointed out that they were unsure about the ***target audience*** of the provenance-driven visualization approach: "*I think it's a really interesting resource which I enjoyed using, and I learned stuff from it. I think I suppose I'd have a question as to what the intended audience was. Because I suspect the public would be much more interested in exploring the data and having a gateway into how to research individuals within the university [...] I think there's a lot of people who would be interested in the history of their families, and I think you'd also find scholars who wanted to research things like, shall we say, women education or like ministers, relationships between people who go on into various careers and their relationship with St Andrews [...]. I said my question is: 'Who is the audience? Who is to be interested in the story of what happens to the records beyond people working in the archives and museums sector and, to an extent, maybe university management who have to listen to presentations from the archives and museums?' I think that's an advertising tool for what the archives do. It is really good.*" [H3].

7 Discussion

Our study findings show that the combination of visualization, textual descriptions, and representations of artifactual form, paired with storytelling techniques and animation, can make provenance not only visible but also a central element of discussion. Interview findings reveal that our visualization approach brought across the time and labor involved in transforming the University records, and there is evidence that participants gained a clear awareness of record curation and its potential impact on the interpretation of these records— all of these aspects are important prerequisites for historical and archival research, critical humanistic interpretation, and knowledge production.

Our previous design approach [46] presented the provenance-driven visualization as a standalone interface that allowed for the detailed exploration of transformation steps from a 'birds-eye' and individual record perspective. In contrast, our new approach combines different representational elements with storytelling, where the provenance-driven visualization gradually unfolds and leads toward a more traditional visualization of historical records. Comparing our study findings with those from the previous study shows that this new approach works better in conveying an understanding of the records' provenance. In our previous study, participants agreed on the importance of disclosing provenance, but the provenance-driven visualization with its abstract nature was found to be unnecessary to those who preferred to simply focus on the exploration of the University records' content. In addition, while critical questions and reflections on the records' transformations and curation were brought up by participants in our previous study, they were much more concrete and pronounced in the study findings we have presented here. Our findings show that participants thought critically and in-depth about the transformations, which was reflected in the variety of responses regarding the increase of usability and research potential, the de/re-contextualization each transformation introduces, the decline of the human aspects visible in the original records, the potential of error introduction, as well as the far-reaching implications of the decisions made as part of the transformation steps. That is not to say that participants did not have any inquiries regarding the content of the records, but in comparison to the previous study, again, their understanding of the records' provenance (labor involved, decisions made, biases, time taken, etc.) was much more pronounced and influenced their exploration of and comments on the geographic visualization of the records. Below, we discuss aspects that contributed to this in more detail, address our research questions, and highlight questions our work raises for future research directions.

7.1 Visualization as an Entry Point to Provenance

Our previous work has highlighted that provenance-driven visualizations run the risk of introducing a level of visual abstraction and complexity that can be difficult for participants to decipher—particularly if they are not familiar with the underlying information collection. While providing context in the form of textual descriptions can help, these are often ignored if disconnected from the visualization itself. Therefore, in this paper, we studied ***how to visualize provenance-related transformations in ways that can provide easier entry points into provenance exploration***.

Our study findings indicate the following: (1) The combination of visualization, brief textual descriptions, and representations of artifactual form helped participants gain an in-depth understanding of the nature of transformations (e.g., the changes applied and their duration, as well as the people involved) and their (potentially positive or negative) impact; (2) Based on participants' interactions and comments, the storytelling and gradual build-up of the timeline using animation played a crucial role in helping to understand the nature of the transformations in a more engaging way, as it emphasized the notion of a 'process' that led to the final, digitized version of the historical records we have visualized.

That being said, our approach still raises questions. The provenance-driven visualization we present in this paper is still quite complex (as emphasized by at least two participants) and the level of visual abstractions requires some explanation. Future design research with other case studies and different information collections is required to explore how to visualize the types of qualitative transformation steps we encountered in the context of the historical University records. Furthermore, the exploration of storytelling elements that allow the interactive exploration of transformation steps at a data level is another interesting avenue to explore in the future.

7.2 Linking Provenance Visualization with Visualizations of Collection Items' Content

In our previous work, we created two different visualizations—one provenance-driven and one geographic visualization of the University records' content—that were presented independently to study participants' engagement and questions [46]. In a way, this resembles traditional approaches where provenance-related information is presented separately (e.g., in textual or diagrammatic form) from the actual visualization of the information collection. In the work presented here, we explored ***how to navigate the tension between the compact and abstract representation of record transformations while still foregrounding the records' content and artifactual form***.

We decided to use the 'martini-glass structure' as a common storytelling technique in visualization to create a link between the provenance-driven visualization and geographic visualizations of the historical University records. The provenance-driven visualization acted as a lead or an entry point into the geographic record visualization, which, ultimately, meant that participants had to go through the provenance-driven visualization before they could explore the records from a geographic perspective. This approach had a highly positive effect because by the time participants explored the geographic visualization, they had acquired a strong sense and awareness of the character of transformations that preceded this visualization—also reflected in their in-depth understanding of the transformations processes, time, and labor involved. The positive effect of the provenance-driven visualization was also reflected in the ways participants critically interrogated the geographic visualization's content, which was clearly informed by the insights they gained from the provenance-driven visualization. While this form of narrative structure can be applied widely across different case studies, it assumes a rather linear storytelling approach, which, as some participants stated, may not be suitable for every use case. In fact, some participants were asking for more open-ended interaction approaches to interrogate the provenance-driven visualization. These findings point to interesting future directions, where different storytelling and narrative patterns could be explored to link provenance-driven and content-focused visualizations.

7.3 Applicability of Provenance-Driven Visualization

Both the previous and our current work point to the potential that lies in making provenance visible through visualization. As a visualization community, engaging with questions around the constructed nature of data [14] and how we can represent, explore, and interrogate processes that preceded their visualization, and even their digital form, is vital for critical discussions about visualizations and the data they represent. We believe that provenance-driven visualization can be applied to most information collections and not only in the area of historical research. Future studies can explore provenance-driven visualization by applying it to much more recent data and making it more versatile as a result.

However, collecting provenance-related information and making transformation processes of information collections visible can be a non-trivial and challenging process, especially if the provenance information is qualitative, goes far into the past, and cannot be computationally gathered. While we were fortunate enough that all the work on the historical University records had happened within a single institution, and we were even able to interview some of the early archivists who had worked with the University records [48], it is not always possible. This raises the question of ***how to construct provenance-related information in ways that enable their visualization***? There is no easy answer to this question as it strongly depends on the information collection in focus.

Interdisciplinary Collaborations. Our research would not have been possible without working closely and discussing our approaches with archivists, librarians, historians, and visualization experts—invaluable questions, perspectives, and pointers derived from these conversations that, for example, enabled us to include the missing LLA students into the record collection.

Interrogating Use Cases & Existing Publications. Other ways of gathering provenance information systematically include the ***exploration of previous uses of and publications around the given collection***, i.e., the contexts in which they were used and described and/or the purposes and audiences for which they were reconstructed. Furthermore, seeking out and studying possible metadata related to the previous iterations of the information collection in question can provide invaluable provenance information. Such information can, for example, be found in the prefaces of publications or 'About' sections of web pages around the information collection in focus.

Explorative Visualization Sketches. Another vital approach that helped us discover the past of the historical University collection was the ***continuous sketching of visualizations to make the records and their metadata visible***. Sketching involved the use of visualization tools such as Tableau Desktop that allowed quick explorations of emerging patterns. It was one of these quickly built visualization sketches that first brought our attention to the inconsistencies in the number of female students. The visualization sketch also facilitated discussions with historians and archivists in our interdisciplinary network. Visualization can therefore not only provide invaluable pointers early on in the exploration process but also act as a facilitator of interdisciplinary discussions, as highlighted in previous work [25].

Missing and Incomplete Provenance Information. There may be cases where it is clear that the provenance information contains gaps or cannot be constructed at all. What is crucial here is to visually acknowledge these gaps, rather than dismiss them. This opens up the possibility of using **uncertainty visualization**, which is a research area of high interest in visualization, and many approaches exist already [5, 31, 38, 40]. Similarly, there is an increasing discussion about how to visualize missing information, which would apply to the idea of provenance-driven visualization. With time, more resources can become available, reveal previously unknown transformations, and fill knowledge gaps.

Target Audiences to Provenance Visualization. A few participants in our previous and current studies raised the question of who is the target audience for provenance-driven visualization, assuming that this may be relevant only for academics in certain fields (e.g., archivists or historians). However, vast majority of participants in both studies found provenance-driven visualization insightful regardless of their background and with this in mind, we argue that provenance is relevant to everybody who engages with data and related visualizations. Considering the provenance of any information collection is crucial and should be part of any interpretation process, no matter in what context it is conducted. Just as we teach media literacy from a relatively early age, we need a means of raising awareness that a visualization can be only as good as its underlying data, which, in turn, is directly connected to the data construction, transformation processes, and underlying curatorial decisions. Our findings show the potential of provenance-driven visualization as an approach that can provide perspectives and put people in an informed position to critically discuss what they see in a visualization. From this perspective, provenance-driven visualization is relevant and applicable across all sorts of information collections, not only in the context of history or the humanities at large. That being said, design approaches to provenance-driven visualization may vary and we hope to see more design ideas and visualization prototypes emerge in the future.

8 Conclusion

To conclude, this paper presents a novel approach for studying and disclosing provenance in information visualization and digital humanities using interactive data-driven visualization. Borrowing from well-established 'humanities provenance' and 'science provenance' approaches, our provenance-driven visualization explores provenance from the point when information was first captured and studies (and gathers data about) all subsequent transformations of this information, up to and including the visualization. Building on our previous provenance research [46], this paper describes a follow-up provenance-driven visualization/study that solidifies the importance of not only looking back at the history behind a visualization and its underlying dataset but also finding ways to show the changes, decisions, biases, and in/ex-clusions that accompany the lifetime of any piece of information in a visualization/dataset. Our provenance research has shown far-reaching implications of this approach for ethical, transparent, and critical digital research in information visualization and digital humanities but at the same time, it emphasized the gaps these two disciplines, we argue, need to face. As a visualization community, we should ask difficult questions about the quality and trustworthiness of the information we visualize and dig deeper into the past of what we show. By collaborating with experts who know the visualized data from completely different perspectives and by finding ways to represent often very difficult-to-visualize qualitative data the past has to offer, provenance can become a crucial source of critical insights and new research questions.

9 Information+ Presentation

An early version of this paper was presented by the first author at the 2023 Information+conference under the title: *Provenance in Information Visualization and Digital Humanities*

10 Open Data & Research Materials

Our qualitative study was approved by the School of Computer Science Ethics board at the University of St Andrews in Scotland. Due to ethics restrictions, we cannot share the raw study data, but provide access to all anonymized interview transcripts.

The visualization prototype described in this paper is available here: https://provenance.tomasvancisin.co.uk/

The underlying data of the visualization prototype is currently embargoed but will be publicly available in February 2026. A preliminary version of parts of the data included in our visualization is already available here: https://arts.st-andrews.ac.uk/biographical-register/

11 Authorship

Tomas Vancisin: Investigation, Visualization Development, Study, Data Collection & Curation, Writing – Original Draft, Review & Editing; Uta Hinrichs: Investigation, Writing – Review & Editing; Loraine Clarke: Writing – Review & Editing; Mary Orr: Writing – Review.

12 License

13 Conflicts of Interest

The authors declare that there are no competing interests.

REFERENCES

[1] E. W. Anderson, J. P. Ahrens, K. Heitmann, S. Habib, and C. T. Silva. Provenance in comparative analysis: A study in cosmology. *Computing in Science & Engineering*, 10(3):30–37, 2008. doi: 10.1109/MCSE.2008.80 2

[2] E. W. Anderson, G. A. Preston, and C. T. Silva. Towards Development of a Circuit Based Treatment for Impaired Memory: A Multidisciplinary Approach. In *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pp. 302–305. IEEE, 2007. doi: 10.1109/CNE.2007.369671 2

[3] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, 10 pages, p. 2573–2582. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753716 5

[4] O. Biton, S. Cohen-Boulakia, S. B. Davidson, and C. S. Hara. Querying and managing provenance through user views in scientific workflows. In *2008 IEEE 24th International Conference on Data Engineering*, pp. 1072–1081. IEEE, 2008. doi: 10.1109/ICDE.2008.4497516 2

[5] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. *Overview and State-of-the-Art of Uncertainty Visualization*, pp. 3–27. Springer London, London, 2014. doi: 10.1007/978-1-4471-6497-5_1 14

[6] M. A. Borkin, C. S. Yeh, M. Boyd, P. Macko, K. Z. Gajos, M. Seltzer, and H. Pfister. Evaluation of filesystem provenance visualization tools. *IEEE transactions on visualization and computer graphics*, 19(12):2476–2485, 2013. doi: 10.1109/TVCG.2013.155 2

[7] R. E. Boyatzis. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Sage Publications, 1998. 8

[8] C. Chabot, C. Stotle, A. Beers, and P. Hanrahan. Tableau prep, 2018. 1

[9] M. Chakhchoukh, N. Boukhelifa, and A. Bezerianos. Understanding how in-visualization provenance can support trade-off analysis. *IEEE Transactions on Visualization and Computer Graphics*, 29(9):3758–3774, 2023. doi: 10.1109/TVCG.2022.3171074 1, 2

[10] M. Correll. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 13 pages, p. 1–13. Association for Computing Machinery, 2019. doi: 10.1145/3290605.3300418 2

[11] C. D'Ignazio and L. F. Klein. Feminist Data Visualization. In *Workshop on Visualization for the Digital Humanities (VIS4DH), IEEE*, 2016. 2

[12] C. D'Ignazio and L. F. Klein. *Data Feminism*. MIT press, 2023. 2

[13] M. Dörk, P. Feng, C. Collins, and S. Carpendale. Critical infovis: Exploring the politics of visualization. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 2189–2198. Association for Computing Machinery, 2013. doi: 10.1145/2468356.2468739 2

[14] J. Drucker. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 5(1):1–21, 2011. 13

[15] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson. Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1663–1672. Association for Computing Machinery, 2012. doi: 10.1145/2207676.2208293 2

[16] G. Feigenbaum and I. Reist. *Provenance: An Alternate History of Art*. Getty Publications, 2012. doi: 10.11588/jfk.2016.2.69492 1, 2

[17] J.-D. Fekete and J. Freire. Exploring reproducibility in visualization. *IEEE Computer Graphics and Applications*, 40(5):108–119, 2020. doi: 10.1109/MCG.2020.3006412 2

[18] M. H. Freedman, J. Gukelberger, M. B. Hastings, S. Trebst, M. Troyer, and Z. Wang. Galois conjugates of topological phases. *Physical Review B*, 85(4):045414, 2012. doi: 10.1103/PhysRevB.85.045414 2

[19] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for Computational Tasks: A Survey. *Computing in Science & Engineering*, 10(3):11–21, 2008. doi: 10.1109/MCSE.2008.79 1, 2

[20] C. Gonzalez. Does animation in user interfaces improve decision making? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 27–34. Association for Computing Machinery, 1996. doi: 10.1145/238386.238396 5

[21] D. P. Groth and K. Streefkerk. Provenance and annotation for visual exploration systems. *IEEE transactions on visualization and computer graphics*, 12(6):1500–1510, 2006. doi: 10.1109/TVCG.2006.101 2

[22] G. Guest, K. M. MacQueen, and E. E. Namey. *Introduction to Applied Thematic Analysis*. Sage Publications, 2012. doi: 10.4135/9781483384436 8

[23] Z. Hensley, J. Sanyal, and J. New. Provenance in sensor data management. *Communications of the ACM*, 57(2):55–62, 2014. doi: 10.1145/2556647.2556657 2

[24] U. Hinrichs, B. Alex, J. Clifford, A. Watson, A. Quigley, E. Klein, and C. M. Coates. Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*, 30(suppl_1):i50–i75, 2015. doi: 10.1093/llc/fqv046 1

[25] U. Hinrichs, S. Forlini, and B. Moynihan. In defense of sandcastles: Research thinking through visualization in digital humanities. *Digital Scholarship in the Humanities (DSH)*, 34(Issue Supplement_1):i80—i99, 2019. doi: 10.1093/llc/fqy051 14

[26] J. Hullman and N. Diakopoulos. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231–2240, 2011. doi: 10.1109/TVCG.2011.255 2

[27] D. Huynh, D. R. Karger, D. Quan, et al. Haystack: A platform for creating, organizing and visualizing information using rdf. In *Semantic Web Workshop*, vol. 52, 2002. 2

[28] W. Javed and N. Elmqvist. Explates: spatializing interactive analysis to scaffold visual exploration. In *Computer Graphics Forum*, vol. 32, pp. 441–450. Wiley Online Library, 2013. doi: 10.1111/cgf.12131 2

[29] M. P. Kouroupas. U.s. efforts to protect cultural property: Implementation of the 1970 unesco convention. *African Arts*, 28(4):32–41, 1995. doi: 10.2307/3337290 2

[30] H. Lamqaddam, A. V. Moere, V. V. Abeele, K. Brosens, and K. Verbert. Introducing layers of meaning (lom): A framework to reduce semantic distance of visualization in humanistic research. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1084–1094, 2021. doi: 10.1109/TVCG.2020.3030426 2

[31] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE transactions on visualization and computer graphics*, 18(12):2496–2505, 2012. doi: 10.1109/TVCG.2012.279 14

[32] E. Maguire, P. Rocca-Serra, S.-A. Sansone, J. Davies, and M. Chen. Visual compression of workflow visualizations with automated detection of macro motifs. *IEEE transactions on visualization and computer graphics*, 19(12):2576–2585, 2013. doi: 10.1109/TVCG.2013.225 2

[33] J. Matejka, T. Grossman, and G. Fitzmaurice. Patina: Dynamic heatmaps for visualizing application usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3227–3236. Association for Computing Machinery, 2013. doi: 10.1145/2470654.2466442 2

[34] Merriam-Webster. Provenance. 2

[35] Microsoft. Power query, 2010. 1

[36] P. Missier, K. Belhajjame, and J. Cheney. The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, 4 pages, p. 773–776. Association for Computing Machinery, 2013. doi: 10.1145/2452376.2452478 2

[37] J. T. Morisette, C. S. Jarnevich, T. R. Holcombe, C. B. Talbert, D. Ignizio, M. K. Talbert, C. Silva, D. Koop, A. Swanson, and N. E. Young. VisTrails SAHM: Visualization and workflow management for species habitat modeling. *Ecography*, 36(2):129–135, 2013. doi: 10.1111/j.1600-0587.2012.07815.x 2

[38] G. Panagiotidou, H. Lamqaddam, J. Poblome, K. Brosens, K. Verbert, and A. V. Moere. Communicating uncertainty in digital humanities visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):635–645, 2022. doi: 10.1109/TVCG.2022.3209436 14

[39] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2016. doi: 10.1109/TVCG.2015.2467551 2

[40] C. Schulz, A. Nocaj, J. Goertler, O. Deussen, U. Brandes, and D. Weiskopf. Probabilistic graph layout for uncertain network visualization. *IEEE transactions on visualization and computer graphics*, 23(1):531–540, 2016. doi: 10.1109/TVCG.2016.2598919 14

[41] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, 2010. doi: 10.1109/TVCG.2010.179 5

[42] R. N. Smart. Literate Ladies - A Fifty Year Experiment. *St Andrews University Alumnus Chronicle*, 59:21–31, 1968. 4

[43] H. Stitz, S. Luger, M. Streit, and N. Gehlenborg. Avocado: Visualization of workflow-derived data provenance for reproducible biomedical research. *Computer Graphics Forum (EuroVis '16)*, 35(3):481–490, 2016. doi: 10.1111/cgf.12924 2

[44] D. Stoecker, O. Duane Adams, and N. Harding. Alteryx, 1997. 1

[45] R. Straughn-Navarro. Provenance in situ: Documenting information of origins across glam contexts. *The International Information & Library Review*, 48(4):287–293, 2016. doi: 10.1080/10572317.2016.1243964 2

[46] T. Vancisin, L. Clarke, M. Orr, and U. Hinrichs. Provenance visualization: Tracing people, processes, and practices through a data-driven approach to provenance. *Digital Scholarship in the Humanities*, 38(3):1322–1339, 2023. doi: 10.1093/llc/fqad020 1, 2, 3, 4, 6, 7, 8, 12, 13, 14

[47] T. Vancisin, A. Crawford, M. Orr, and U. Hinrichs. From people to pixels: Visualizing historical university records. In *Proceedings of the 5th Biennial Transdisciplinary Imaging Conference 2018 (Transimage 2018)*, pp. 41–57, 2018. doi: 10.6084/m9.figshare.6104699 3, 4, 6

[48] T. Vancisin, M. Orr, and U. Hinrichs. Externalizing transformations of historical documents: Opportunities for provenance-driven visualization. In *Proceedings of the 5th Workshop on Visualization for the Digital Humanities (VIS4DH2020)*, 2020. doi: 10.1109/VIS4DH51463.2020.00011 4, 13