Digital Scholarship in the Humanities, 2023, **00**, 1–18
https://doi.org/10.1093/llc/fqad020
Full Paper

OXFORD

Provenance visualization: Tracing people, processes, and practices through a data-driven approach to provenance

Tomas Vancisin [1]*, Loraine Clarke[1], Mary Orr[1], Uta Hinrichs[2]

[1]University of St Andrews, St Andrews, UK
[2]University of Edinburgh, Edinburgh, UK

*Correspondence: Tomas Vancisin, School of Computer Science, University of St Andrews, Jack Cole Building, North Haugh, St Andrews KY16 9SX, UK. E-mail: tv8@st-andrews.ac.uk

Abstract

Provenance disclosure—the documentation of an artifact's origin and how it was produced—is an important aspect to consider when working with historical records which undergo multiple transformations in preparation for and during digitization. Provenance in this context is commonly communicated through explanatory text or static diagrams. However, the methodological and curatorial decisions that have influenced the records' data are easily overlooked, in particular when exploring the records through visualization as a result of digitization processes. We propose a data-driven approach to provenance disclosure which (1) traces provenance back to when the records were created, (2) documents and categorizes the records' transformations (transcriptions, content modifications, changes in organization, and representational form), and (3) uses data visualization to disclose provenance in interactive ways. We reflect on how this approach can be practically applied in the context of historical record collections, and we present findings from a qualitative study we conducted to investigate the merits and limitations of provenance-driven visualization. Our findings suggest that data-driven provenance disclosure has the potential to (1) promote transparency and deeper interpretations of historical records, (2) provide rigor in researching historical document collections and underlying production processes, and (3) encourage ethical considerations by making visible labor and implicit bias that influence the production and curation of historical records.

1 Introduction

The digitization of historical documents such as manuscripts, national and university records, letters, or books, and, as part of this, the transformation of such records into computer-readable formats, has given rise to the visualization of such records, to facilitate their interrogation from multiple perspectives (Windhager *et al.*, 2019). Visualization can enable the identification of higher-level patterns across historical record collections and facilitate explorations that are not possible through close-reading techniques (Jänicke *et al.*, 2017; Windhager *et al.*, 2019). In the past few years, visualization has become an important means in various indicative research projects for investigating and interpreting digitized historical record collections (Betti *et al.*, 2014; Edelstein *et al.*, 2017; Hinrichs *et al.*, 2015; Mäkelä *et al.* 2012).

What is less clear in such visualizations of digitized historical record collections, however, is their provenance disclosure, and why it matters quantitatively and qualitatively for users with regard to how they then interrogate the visualized data. Historical records undergo a number of transformation steps, including transcription and (re-)structuring (e.g. via tagging), before interactive visualizations can be developed. These transformations variously change the data's content, organization, and artifactual form, in turn influencing how the records can be interrogated (e.g. what research questions are asked) and interpreted—individually and as a collection. Traditionally, provenance information is either described in textual form (e.g. Betti *et al.*, 2014; Hyvönen *et al.*, 2017) or illustrated through process diagrams (e.g. Capodieci *et al.*, 2015; Hinrichs *et al.*, 2015). Both approaches can only address provenance at a high level, and treat provenance as secondary information that is easily overlooked when exploring the records in question through digital search interfaces or visualization.

This is an issue in particular in the context of historical record collections because their provenance information can be vast and complex, yet just as important for the interpretation and understanding of the records as the history they capture. Ultimately, provenance disclosure is crucial to promote transparency and ethical approaches to research on historical record collections (e.g. disclosing interpretations and curatorial decisions that have accompanied transformation steps, and acknowledging the labor involved in historical records' transformations). Our research, therefore, focuses on the following question: "How can we disclose provenance in ways that capture and make visible transformation steps of historical records in a holistic and engaging way?" Our interest here is not only the fuller disclosure of the actors behind the multiple transformation processes that constitute 'the product'. More ambitiously, our interest is in how provenance visualization can better engage users from multidisciplinary perspectives interactively to ask more probing research questions of the historical dataset.

We have started to address this question in previous work by characterizing the types of transformations historical record collections typically go through and by exploring how such transformation steps could be visualized in an interactive way (see the extended abstract we presented at the 2020 ADHO DH conference (Vancisin *et al.*, 2020a,b) and—in a slightly extended version—at the workshop for Visualization in the Humanities (VIS4DH'20) (Vancisin *et al.*, 2020a,b). In this article, we consolidate and expand on this preliminary work at a theoretical and empirical level.

- We propose a data-driven approach to provenance disclosure which (1) traces provenance all the way back to when the records were first created, (2) systematically documents and categorizes record transformations, including transcriptions, content modifications, and changes of organization and representational form, and, based on this data, (3) uses visualization to disclose provenance in an interactive way. This in turn allows for a representation of provenance at an individual record level as well as across the entire record collection.
- We present findings of a qualitative study which suggest that provenance-driven visualization presents opportunities for (1) promoting transparency and deeper interpretations of historical records, (2) ensuring rigor in researching historical document collections by highlighting modifications and interpretations introduced through transformation processes, and (3) highlighting ethical considerations for working with historical records in issues such as hidden labor and implicit bias that influence the curation and production of data within historical records.
- Based on our findings, we critically reflect on the proposed data-driven approach to provenance disclosure, outlining research questions for future work and potential limitations (required time and resources, and visualization challenges).

Our work contributes to the fields of digital humanities and data visualization.

2 Related work

Recognition of provenance is as important in the Arts and Humanities as it is in Science and Technology research, although its treatment and definition vary respectively. Provenance research in the Arts and Humanities focuses on 'the history of ownership of a valued object or work of art or literature' (Merriam-Webster, 2022). In the Sciences provenance is considered to be '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' (Missier *et al.*, 2013). A brief outline of different disciplinary perspectives on provenance and provenance disclosure demonstrates how these have influenced our work and our main research question.

2.1 Provenance in the arts and humanities

Provenance, in the context of the arts, provides pointers to the given piece's authenticity and originality (Feigenbaum *et al.*, 2012, p. 98)—crucial aspects for collectors and curators who need to be sure about the origin of artworks and items of historical and/or cultural value. Provenance research in the Arts and Humanities, therefore, includes questions of past and present ownership, how the item was acquired, its creator(s), and its geographical and temporal context. Along these lines, museums across the world follow provenance standards,[1,2] further highlighting the importance of provenance in this context. The 1970 UNESCO Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property urges countries and their museums to make sure the artifacts they obtain do not come from any illegal trade. While ratifying this convention might not be a straightforward process (Kouroupas, 1995), the fact that 141 countries have done so by now[3] shows the extent to which provenance research and disclosure are taken seriously in order to maintain professional and ethical integrity in the GLAM sector. Museums around the world investigate the provenance of their artifacts to uncover illegal acquisitions and, where appropriate, try and return the

artifacts in question to their rightful owners (American Alliance of Museums, 2022; The Metropolitan Museum of Art, 2022; Bartrum, 2000).

In the Arts and Humanities, provenance is usually disclosed and presented in textual form. However, researchers have also explored new ways of representing provenance. For example, in collaboration with the Carnegie Museum of Art, the Art Tracks project uses visualization to present provenance of historical paintings in time and space using interactive maps and timelines (Berg-Fulton *et al.*, 2015).

2.2 Provenance in science and technology research

The importance of capturing and disclosing provenance in the context of Science and Technology research is rooted in its aims for reproducibility of scientific experiments and ensuring validity of corresponding data (Freire *et al.*, 2008). The term data provenance has been coined to specifically focus on issues related to the lineage or pedigree of digital information. Buneman *et al.* describe data provenance as 'the origins of a piece of data and the process by which it arrived in a database' (Buneman *et al.*, 2001).

Different systems have been introduced to facilitate the recording and documentation of scientific data to enable the reproduction of experiments and the validation of results. These so-called scientific workflow systems focus on provenance capture and disclosure, data management, analysis, simulation, and visualization (Barker and Van Hemert, 2007). Systems such as Avocado (Stitz *et al.*, 2016), Kepler (Ludäscher *et al.*, 2006), or VisTrails (Bavoil *et al.*, 2005; Freire and Silva, 2012) provide the necessary infrastructures to manage and explore data and its provenance. VisTrails has been used in the contexts of quantum physics (Freedman *et al.*, 2012), psychiatry (Anderson *et al.*, 2007), cosmology (Anderson *et al.*, 2008), and ecography (Morisette *et al.*, 2013), while Avocado has been proposed to support biomedical research (Stitz *et al.*, 2016). These systems capture provenance from data acquisition to the final experimental result and make the steps and decisions during the experimental process visible via graphical workflows (Bavoil *et al.*, 2005; Stitz *et al.*, 2016).

Our research builds on considerations and approaches to provenance capture and disclosure that transect Arts and Humanities and Science and Technology research.

2.3 Discussions of provenance in DH and VIS

The importance of provenance disclosure has also been discussed in the fields of Digital Humanities (DH) and Visualization—both inherently interdisciplinary fields that combine approaches from the Sciences and the Arts and Humanities.

2.3.1 Provenance in the digital humanities

In advocating for more feminist approaches to data analysis, D'Ignazio and Klein argue for a stronger acknowledgment of labor involved in data production, maintenance, representation, and communication as a means to promote more ethical research (D'Ignazio and Klein, 2016; D'Ignazio and Klein, 2020). Disclosing provenance and, with it, the disclosure of how data in the widest sense has been transformed—by, and also for whom—is directly linked to this call. Lamqaddam *et al.* add to the ethical dimension of provenance by highlighting authenticity and user trust as crucial benefits provenance disclosure can promote (Lamqaddam *et al.*, 2021).

In the field of digital humanities, provenance is typically presented through textual descriptions (see Hyvönen *et al.*, 2017; Betti *et al.*, 2014). However, diagrammatic approaches also exist and these provide abstract, high-level overviews of the data production steps, such as data transformation, digitization, and visualization (e.g. Boer *et al.*, 2015; Capodieci *et al.*, 2015; Hinrichs *et al.*, 2015).

2.3.2 VIS perspectives

The importance of provenance disclosure is also being discussed in the field of visualization. Referring to Tufte's Beautiful Evidence (Tufte, 2006), Hullman and Diakopoulos (2011) emphasize the importance of data provenance in visualization for promoting trust and transparency in the data represented. Similarly, Dörk *et al.* advocate for more critical approaches to data visualization, and emphasize disclosure of the decisions made about the data as crucial for developing 'trust between visualization creators and viewers' (Dörk *et al.*, 2013). Michael Correll emphasizes the need to visualize and communicate data provenance as 'a key component of both, affording criticism and supporting transparency, in data-driven decision-making.' (Correll, 2018). Paralleling approaches in the Sciences, Fekete and Freire emphasize the importance of reproducibility and replicability in the visualization process and highlight four techniques that have been used for provenance disclosure in the Visualization community: Algorithmic Reproducibility, Technique Reproducibility & Replicability, System Reproducibility & Replicability, and Application Reproducibility & Replicability (Fekete and Freire, 2020).

3 Challenges of provenance disclosure

Although the above demonstrates that provenance and its disclosure in VIS and DH are given serious

attention, important issues remain to be addressed, especially when working with historical records.

3.1 The starting point of provenance disclosure

Provenance is typically addressed from the point of institutional data recording/acquisition, that is, from the point when the museum, library, or archive provides records/data for a project. The transformations the historical records have undergone before this point are often not reported.

We argue that provenance should be studied in full, including the transformations that preceded record/data acquisition, the decisions made, and people behind them.

3.2 Forms of provenance disclosure

As outlined above, the most common way of disclosing provenance is through text. While it can be precise and detailed in reflecting on transformation processes of historical records, there are limitations. Provenance information is typically provided in the form of a separated preamble (for books) or an 'About' page (for digital projects), in the main as collection-level descriptions. However, not every record in the collection is necessarily transformed in the same way. For example, during content expansion, the archivist may find additional information for one record but not for others. It can therefore be valuable to provide provenance information at a record-level, which can be cumbersome in textual form. Textual descriptions may also omit the wider spectrum of changes (form, structure, etc.) that individual records, or parts of the collection (e.g. in time, scope) have undergone.

Visual approaches to disclosing provenance typically consist of schematic diagrams, which provide an overview of the transformation steps of historical records (see Boer *et al.*, 2015; Capodieci *et al.*, 2015; Hinrichs *et al.*, 2015). However, diagrams can only provide a high-level glimpse into the processes and decisions made. The necessary abstraction of nuanced processes that may affect individual records differently can hide important aspects of transformation and re-interpretations. Moreover, in diagrams, transformations are often represented as homogenous geometric shapes, which might give a false impression of a straight-forward, complete, and objective process, when transformations are in fact rich in interpretation and situated decisions.

We argue for a data-driven approach to provenance disclosure that uses visualization to represent transformations both at collection level and at the level of individual records.

3.3 Placement of provenance information

We typically find detailed provenance information—in both textual and diagrammatic form (Boer *et al.*, 2015; Capodieci *et al.*, 2015; Hinrichs *et al.*, 2015; Hyvönen *et al.*, 2017)—in academic publications or placed alongside the corresponding artifacts and items, be it physically in the form of separate/dedicated information panels (e.g. at museums) or in the form of textual metadata presented alongside an item in digital space. However, on digital web-based platforms, provenance information is often presented separately from the collection items themselves, typically as part of an 'About' page. This may cause important provenance information to be overlooked, and hence open the presented historical data to user misinterpretation (minor and major).

We argue for exploring the question of how to directly integrate provenance information into visualizations of historical record collections.

4 Toward a data-driven approach to provenance disclosure

To address the challenges outlined above, we propose a data-driven approach to provenance disclosure which (1) traces provenance all the way back to when the records were first created, (2) systematically documents and categorizes record transformations, including transcriptions, content modifications, and changes of organization and representational form, and, based on this data, (3) utilizes visualization to disclose provenance in an interactive way. This in turn allows for a representation of provenance at an individual record level as well as across the entire record collection. We argue that making the full spectrum of transformation steps, related curatorial decisions, and the people behind them visible is (1) a step toward more ethical research approaches that acknowledge the labor behind data production from historical records, and (2) allows for a more critical interpretation and contextualization of historical records and how perspectives on these have changed over time.

The idea of a data-driven approach to provenance disclosure is inspired by existing scientific practices where data transformations are carefully documented, as well as by practices in the humanities which study provenance in its entirety. It is also grounded in our previous work which has (1) identified and characterized the transformation processes that historical documents often go though and how these influence the interpretation of digitized historical collections and (2) explored how these transformation processes can be made visible through provenance-driven visualization (Vancisin *et al.*, 2020a,b)

Below, we summarize this previous research to illustrate how a data-driven approach to provenance disclosure can be applied in practice to a collection of historical student records in order to support new ways of disclosing provenance and to enable the study and (re)interpretation of records from different perspectives.

4.1 Provenance of the historical student records

As many other universities around the world (Spörlein, 2014), the University of St Andrews has been keeping records of its students since its foundation (1413) (Maitland Anderson, 1905; Maitland-Anderson, 1926; Smart, 2012). These records provide biographical information such as student name, parentage, time and location of birth and death, courses taken, and even careers after studies. Providing insights about the University's history, links between academic institutions, and information about societal structures of the era at the time, these records are of great historical and public interest. The records initially took the form of a handwritten Matriculation Roll that students signed year-by-year (see Fig. 1.1).

Our initial project focused on applying visualization techniques to study these records' content from new perspectives (Vancisin *et al.*, 2018), but we soon started discovering the complexities of the records' provenance and decided to explore ways of foregrounding this aspect for a more nuanced portrait of this historical record collection.

The St Andrews historical student records have been curated, transformed, and re-presented by a number of experts—historians, archivists, digitization officers, and librarians. Between 1888 and 1905, archivist James Maitland-Anderson transcribed the Matriculation Roll, resulting in a printed version of the records (see Fig. 1.2). Between 1960 and 2004, another archivist, Dr Robert Smart, expanded the individual record content, drawing from more than 1000 additional sources. He transformed the collection into the Biographical Register of the University of St Andrews 1747–1897 (BRUSA) (Smart, 2004), a printed alphabetical index that includes additional information about student demographics as well as academic and subsequent careers (see Fig. 1.3). From 2013 to 2016, the University Library's DH and Research Computing team led by Dr. Alice Crawford transformed BRUSA into a searchable digital format using XML: TEI (see Fig. 1.4), resulting in an online textual search interface.[4] In 2018, we further transformed the XML: TEI files, each representing one person, into a relational database and created a number of interactive visualization sketches in Tableau Desktop (Vancisin *et al.*, 2018) (see Fig. 2).

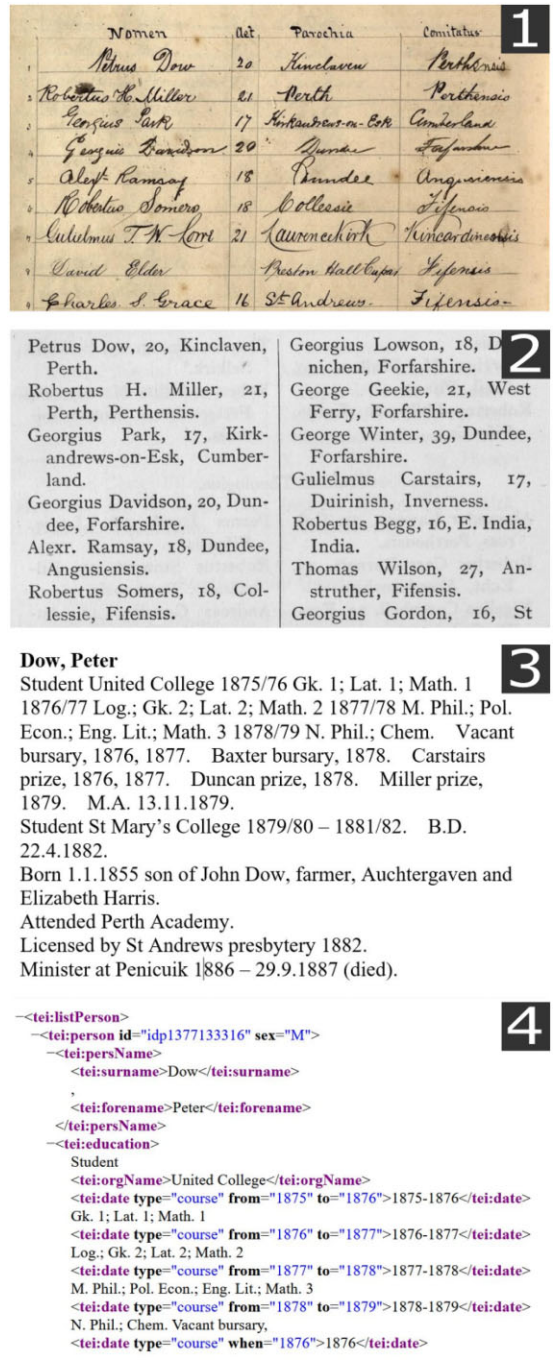

Figure 1. Provenance of the records

We conducted interviews with experts who had worked with the University records as well as experts who work with similar historical collections (Vancisin *et al.*, 2020a,b) to identify and characterize key transformations the University records underwent over all

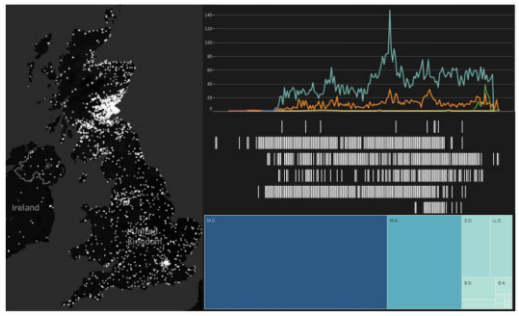

Figure 2. Visualizations of the St Andrews University student records

of these years. We identified four key transformation steps, as summarized below.

4.1.1 Transcription

Like many historical documents that come in handwritten form, our records were transcribed. Transcription is a necessary step toward any computational treatment of historical records, but it often requires an interpretation of the original records (e.g. deciphering the historical spelling of names or determining if different records refer to the same person) and can potentially lead to errors.

4.1.2 Content modification

Not only has content been added to the records from different sources, it has also been reviewed for inconsistencies introduced during the transcription. Individual records were modified to different extents; while some were enhanced significantly, others remained nearly unchanged.

4.1.3 Structural modification

Transformations can also take the form of structural modifications, which influence how the historical records and related data can be interrogated. In the case of our records, structural changes included a shift from temporal to alphabetical order, followed by a removal of an inherent order and structure by slicing up the record contents based on attributes such as name, time, and geolocation.

4.1.4 Artifactual form

Perhaps the most obvious are the changes in artifactual and medial form that the records have seen over the years (see Fig. 1). These changes (e.g. from handwritten form to print text; from text to interactive, visual representation) significantly change the modes of reading and interpretation.

4.2 Provenance disclosure through data visualization

When looking at the historical student records in their digital representation—either as part of a textual search interface or in the form of visualizations, none of the outlined transformations are visible (see Fig. 2). As we have argued previously, this can hamper the interpretation of historical records, but it also raises ethical issues regarding the attribution of labor and transparency of knowledge making (Vancisin *et al.*, 2020a,b). A data-driven approach to provenance disclosure allows to systematically visualize the transformations at an individual record level, which can make aspects of provenance more prominent in digital representations of these historical records. As part of our previous work, we have experimented with and iteratively refined visualizations to highlight the transformation steps that historical records have gone through (Vancisin *et al.*, 2020a,b), leading to the prototype presented below (see Fig. 3).

The bottom layer of this provenance-driven visualization (see Fig. 3.1) represents the original records aggregated and organized according to their temporal distribution into a bar chart, hinting at the chronological order in which student signatures were originally collected. Using a sketch-based stroke for bars[5] we emphasize the unique characteristics of the original, handwritten records. This bar chart is mirrored in the layer above by another temporal bar chart, which emphasizes Maitland-Anderson's → *transcription* of the records by using a smooth stroke (see Fig. 3.2). Layer 3 (see Fig. 3.3) highlights Smart's alphabetization of the records and his expansion of their content. The bar chart represents an aggregation of records by alphabet, and the bar width indicates the amount of content present in these records. Layer 4 (see Fig. 3.4) shows Crawford's work, which revoked both the temporal and alphabetical structure of the records and applied more structure to individual records. Each individual record is represented as a square with no inherent ordering. Square size indicates the length of individual records. Our migration of the records into a database is represented in the top-most layer of the provenance-driven visualization (see Fig. 3.5). The database structure is represented as a hierarchical tree diagram where each circle represents a database table.

The five layers are interactive and interlinked. For example, hovering over and/or clicking on a bar in one of the bar charts, highlights these records in all the other layers (see Fig. 3) and shows the corresponding individual records in the 'Record View' to the right (see Fig. 3.6). The textual representation of individual records directly corresponds to the transformation layer that has been selected (see Fig. 4). In Fig. 3, for instance, a bar in Layer 3 has been selected which represents the records after their indexation and expansion

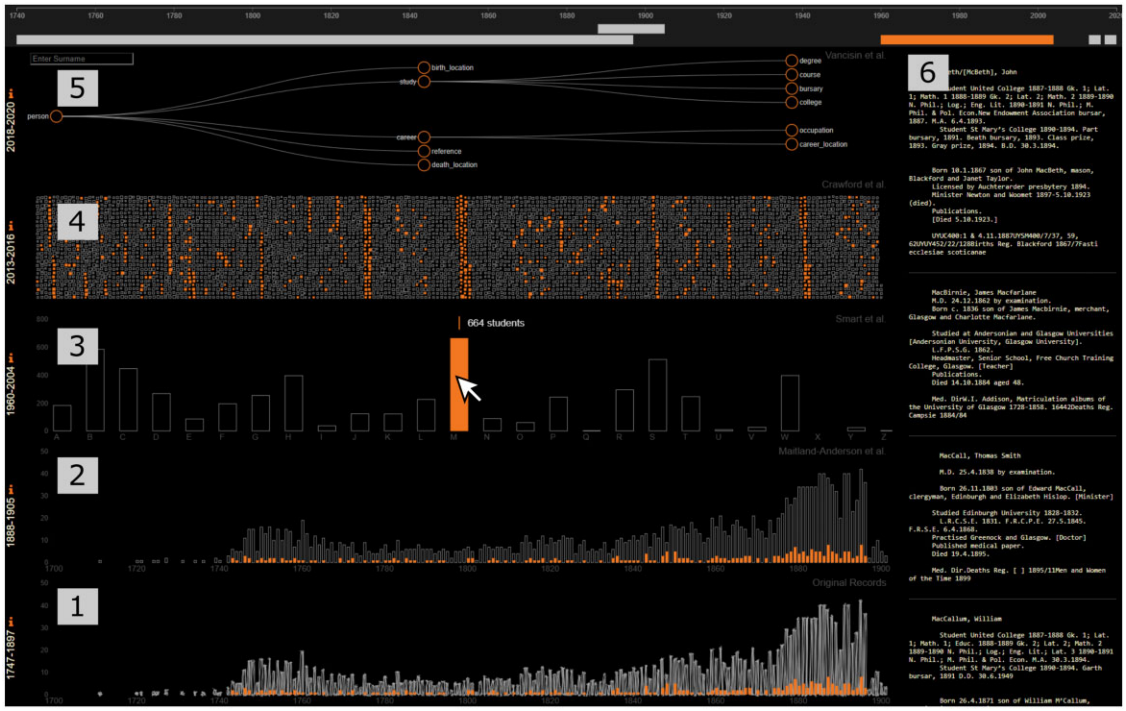

Figure 3. Provenance-driven visualization; Selecting 'M' surnames in the third layer shows their form in the 'Record View' to the right and highlights the structure and organization of those records in the other layers

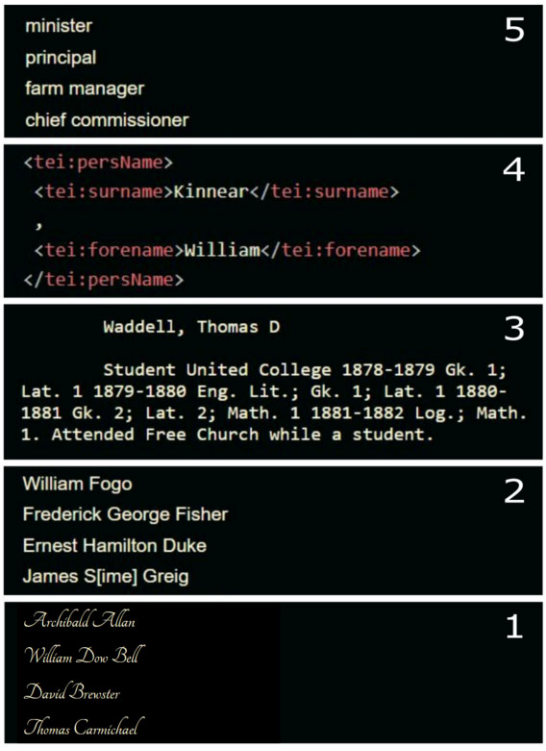

Figure 4. Provenance-driven visualization's different record views

by Smart. The Records View (Fig. 3.6) therefore shows the content of individual records after this transformation in a layout, font, and style that resembles BRUSA as published by Smart (Smart, 2004). A selection of one of the squares in Layer 4 would show the corresponding record in the XML: TEI form (see Fig. 4.4).

This example of a provenance-driven visualization illustrates past forms of the student records and inherent changes to their content and structure in a visual way. The records' content and provenance information are visually linked. Following a data-driven approach to provenance disclosure, we focused on capturing and disclosing aspects of the full spectrum of identified transformation processes, mostly drawing on provenance information that is already implicit in the different artificial forms of the records and that can be directly visualized. For example, we highlight → transcription through the choice of different stroke types (see Fig. 3.1 and 2). We make the → structural modifications introduced by Smart, Crawford and by ourselves explicit through the spatial distribution of elements representing the records (see Fig. 3.3, 4, and 5). We show → content modifications at the record-level and in aggregated ways by modifying the size of visual elements (see Figs 3.3 and 4). Modifications of → artifactual form are illustrated in the Record View directly capturing the original record forms and through typographical representation (see Figs 3.6 and 4).

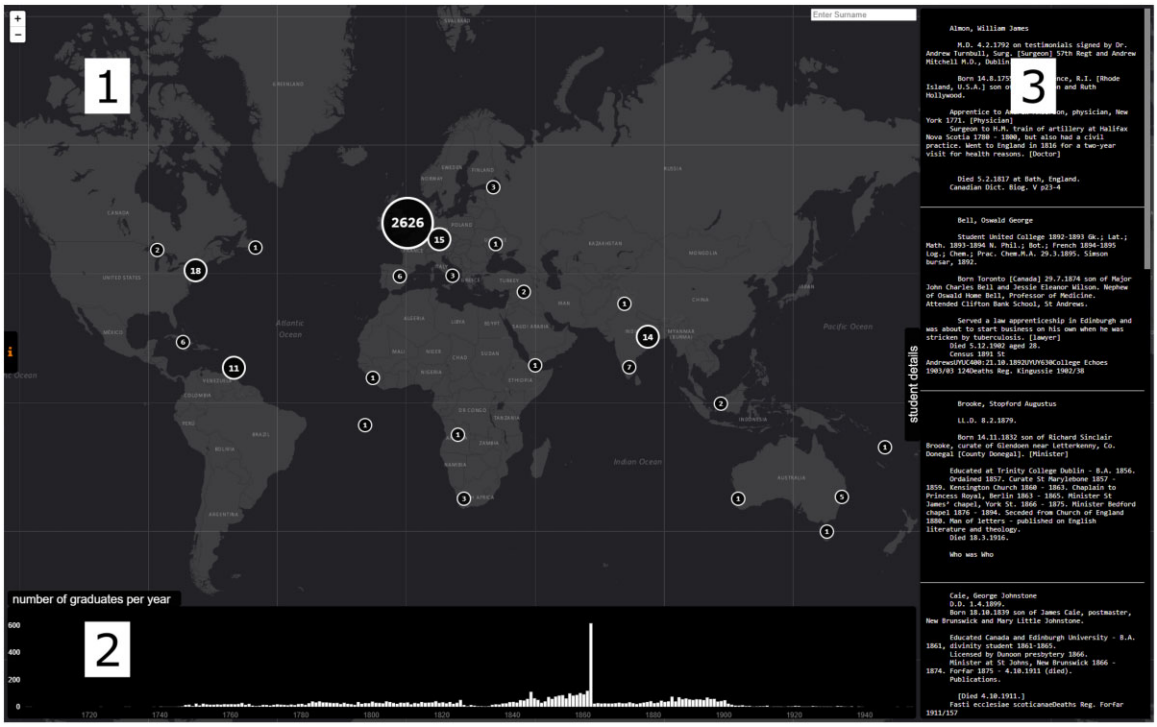

Figure 5. Traditional visualization

5 Studying the impact of provenance-driven visualization

The notion of a data-driven approach to provenance and provenance-driven visualization illustrated above is a departure from the dominant approach of visualizing the *content* of historical records, to visualizing the context in which these have been collected and modified. This approach also departs from traditional ways of representing provenance (text or diagrams) in that it is informed by data we gathered systematically about each individual record's transformations.

In order to explore the potential of our data-driven approach to provenance disclosure, we conducted a qualitative study. We were particularly interested in the following questions: What types of insights and interpretations do participants gather from our provenance-driven visualization?, Does provenance-driven visualization promote transparency?, and Can it raise awareness of the labor and different layers of interpretation that are inherent in historical document collections?

5.1 Study approach

Evaluating how visualizations inform user insights and interpretations is a complex challenge, since analysis processes and outputs are difficult to capture (Lam *et al.*, 2012), especially in a study context that is typically constrained by time. In order to start addressing our research questions above, we designed a qualitative study that exposed participants to two independent visualizations that sit at opposite ends of a content—provenance continuum. Our study should not be understood as a comparative appraisal where one condition is tested against the other. Instead, inspired by previous work in the context of personal visualization (Thudt *et al.*, 2016), we consider the two visualizations as probes to trigger situated reflections on two different aspects of historical data that can be made visible and, in consequence, on the impacts, these may have on potential explorations and interpretations of the records.

The provenance-driven visualization we showed participants is the one described in the previous section (see Fig. 3). The, more traditional, content-driven visualization focused on the geo-temporal aspects of the historical student records (see Fig. 5).

Geo-temporal visualizations are commonly used to provide an overview of historical document collections (Brizzi, 2013; Jenkins *et al.*, 2013; Edelstein *et al.*, 2017; Schwinges, 2018; Conroy, 2021). The map view shows the geographical distribution of the students' birth locations (see Fig. 5.1) while the timeline view depicts the number of graduates by year (see Fig. 5.2). Individual student records are shown in the 'Record View' (see Fig. 5.3). All views are linked; hovering over the circles on the map acts as a filter on the timeline

and the 'Record View'. Similarly, hovering over the timeline updates the map and the 'Record View'.

The geo-temporal visualization focuses purely on the content of the student records, while in the provenance-driven visualization, all views represent the provenance aspects of the records. The exploration of the records' content is supported, but only through the lens of provenance. In both visualizations we provided provenance information in textual form on-demand: An 'Info' button on the very left brings up a textual overview of the transformations the records have undergone. This text was identical for both visualizations.

5.2 Participants

Our study goal was to gain insights from users with different disciplinary perspectives in order to explore the merit of provenance-driven visualization and future research directions. We therefore recruited 24 participants from a range of backgrounds that can be divided into three groups. Eight participants had a background in computer science, specializing in human–computer interaction and visualization. Eight participants' background was in history and/or archiving practices with expertise in interpreting historical document collections. Eight participants had no background in either computer science or history/archiving, therefore coming to the historical records and the visualizations from an educated general interest perspective.

5.3 Study procedure

Each study session took approximately 1 h and consisted of four phases.

5.3.1 Pre-questionnaire

Participants were first asked to fill out a questionnaire about their professional background and their experience with digitized historical materials.

5.3.2 Visualization Exploration I and interview

Participants were then introduced to either the traditional 'content' visualization (see Fig. 5) or the provenance-driven visualization (see Fig. 1). We counterbalanced the number of participants who first interacted with the traditional versus the provenance-based visualization to avoid study confounds due to learning effects. We provided a brief introduction to the visualization at hand, including where to find additional information about the records (i.e. the provenance information provided in textual form). We then let participants explore the visualization freely, based on their own curiosity. Participants interacted for 15–20 min with the visualization which was followed by an interview that focused on their understanding of the visualization, insights about the records they had gathered, and on questions the visualization had raised. We also asked questions about the discoverability of the information sought.

5.3.3 Visualization Exploration II and interview

The participants were then exposed to the other of the two visualizations in a repeat of the brief introduction and interview processes.

5.3.4 Final interview

After exposing the participants to both visualizations, we interviewed them about their overall experience with the two visualizations.

5.4 Data collection and analysis

Participants had filled in and submitted the online form in advance of trialing the two visualizations. We recorded all participant interactions via screen capture, and all interviews were audio recorded.

Our data analysis mostly focused on the audio recordings which were fully transcribed and analyzed using a thematic analysis approach (Boyatzis, 1998; Guest *et al.*, 2012). The thematic analysis and development of coding for the emerging (and repeating) participant reactions in the interviews were conducted by two researchers on the team. Codes were informed directly by the interview questions, but also emerged from the interview data by capturing the participants' experience with and reactions to the two visualizations. This coding was discussed with and then refined by all four members of the research team. The findings deriving from the thematic analysis were enhanced and contextualized with the participant observations that were gathered from their interactions with both visualizations in the final interview.

6 Results

Our study findings shed light on the participants' first impressions of the provenance-driven visualization and its potential value. In particular, they raise interesting questions regarding the role of text and context in visualizing provenance, and how the mindset of the participants influenced their experience and expectations toward provenance-driven visualization. Below, we outline our findings in detail. We indicate the participants' background as follows: 'CS' stands for a background in computer science, 'H' for history, and 'G' for no expertise in related fields.

6.1 Approach and first impressions of the visualizations

Our observations of participants interacting with the two visualizations and their interview statements shed light on their approach and experience of the traditional and provenance-driven visualizations.

6.1.1 Traditional visualization

Not surprisingly, when engaging with the traditional visualization, participants gravitated toward the geographical aspects of the records highlighted on the map. For example, participants were interested in students' countries of origin. Accordingly, when asked what caught their attention, participants frequently mentioned discoveries of geographic nature: 'I was surprised that there is someone from there (Australia). I thought it was strange, but then you explained about the British Empire.' [G1]; 'I found that there were no Catalans or Italians.' [CS4]. Participants also paid attention to visual patterns on the timeline visualization: 'What happened in 1862? That's such a huge jump.' [CS1]. The exploration of the traditional visualization was often described as a 'stroll through' the records [G4], or 'jumping around different places, looking at names and dates.' [CS2]. No participants expressed difficulties understanding how to navigate the traditional visualization or how to interpret the visual encodings in play. Participants described the visualization as 'clear', 'self-explanatory', and 'intuitive'. Eight participants explicitly attributed this to the fact that the traditional visualization felt familiar: 'It's a familiar paradigm; it invites you to go in and start clicking at things and zooming around and exploring things, and you can immediately see things which are interesting from the geographic perspective at least.' [CS2]; 'I've seen that kind of visualization before, that kind of map-based, timeline-based visualization, so it seems more familiar.' The traditional visualization was described as 'focused' in terms of the type of data attributes that were represented. Participants stated that this, paired with the familiar visual paradigm (map) would make this visualization suitable not only for experts familiar with such records, but also for general audiences.

6.1.2 Provenance-driven visualization

With the provenance-driven visualization, we observed a larger variety of exploration approaches. Participants focused on individual transformation layers that looked familiar (e.g. the timeline views at the bottom, see Fig. 3) as well as those that triggered their curiosity because of their unfamiliar look (e.g. the digitization and the database layer, see Fig. 3.4 and 3.5). Exploring the interconnectedness of the visualization layers and the individual records in their different forms was another common point of focus. Participants described their exploration of the provenance-driven visualization as 'browsing' or 'rummaging around'. Participants found the visualization to be 'initially confusing'; some even experienced it as '... scary. I was like: 'what is all this'? That's because I am not used to seeing information like that.' [G1]. Participants emphasized that the supporting text was necessary for the overall understanding of the functionality and the intention of the visualization. They often indicated that their initial confusion was caused by the complexity of the visualization and the (unexpected) amount of information shown. It was primarily due to these factors that participants suggested the target audience for the provenance-driven visualization as expert historians, archivists, or computer/data scientists. However, after spending some time with the visualization, including through some additional explanations, participants typically explored and were able to interpret the presented information without problems.

6.2 Values of visualizing provenance

The 23/24 participants explicitly stated that they saw value in the idea of visually representing provenance. Below, we describe positive and negative aspects of the provenance-driven visualization.

6.2.1 Triggering critical questions

The provenance-driven visualization triggered questions about the historical records and the context in which the corresponding data was extracted: 'Why does it [the data] exist in different arrangements over time? ... Who are these different people that are being named here? [people involved in the transformations]' Questions such as these were common, which indicates that the provenance-driven visualization made participants pause and think not only about the data represented, but also about the people and the processes involved behind it. One participant mentioned that the complexity of the visualization might be key here, as it mirrors the complexity of the transformation processes: 'I think bringing in the complexity of the records themselves, I know, that makes for a complex visualization and needs for complicated answers somehow. But it's being transparent about where certain data comes from, and the messiness—and I know not everyone wants to see that—but I would hope some people want to see it, so I think it gets that across.' [H6].

6.2.2 Making transformations visible

Visualization can make (higher-level) patterns in data visible (Card, 1999), and participants highlighted the usefulness of providing visual overviews of provenance-related information: 'To understand what was recorded at each period or how it was recorded. Even if you physically cannot check the books [the original, transcription, alphabetical index etc.], you can still play with what they contained.' [CS4]; '... It shows the process of digitization, and how data is treated, and how you can work with data and interact. And I guess it's interesting how archiving developed

over time, and I think it's a very interesting aspect.' [G4]. Two participants also emphasized that even if they were not particularly interested in the records, representing information visually 'fuels more your curiosity, and I think that makes the data more interesting.' [CS6].

6.2.3 Ethical dimensions

Three participants saw value in representing provenance for ethical reasons, be it through visualization or otherwise. For example, CS2, a participant involved in processing historical records, highlighted the ethical value of provenance disclosure as the laborious work is often overlooked: 'It is useful to know about these things [provenance]. I've been involved in enough things when people transcribe stuff and digitize it, to know how much goes into it, and it's not necessarily always appreciated what goes into that.' [C2]. Participants further highlighted that provenance is also about making the people responsible for the processes, their labor, and curatorial decisions visible: 'To see the process that people have, through time, done with the information that they have—I can't even imagine [the visualization] without that. To be able to see how that process has happened historically is massive. It takes you to a completely different journey, and it's not just about the students. It's also about the people that decided to keep these records, and what happens to the records as time went on.' [G1]. Statements such as this highlight the importance of disclosing provenance from a collection-overview perspective, including the transformations that precede data acquisition as well as decisions and the data behind these transformations.

6.2.4 Transparency and validity

When asked about the value of provenance-driven visualization, six participants emphasized the importance of transparency in terms of the data and its sources. For example, comparing the two visualizations, one participant stated: '... the second visualization [provenance-driven visualization] is more transparent about the data, where it came from, and how it interrelates. In the first [traditional] visualization, the raw data was a bit obscure.' Another participant commented: 'I think that it is valuable to see how the data from historical documents has changed. Being able to see the process is like seeing the 'chain of custody' of criminal evidence—you'll be able to see, if the data has been corrupted or enriched.' [CS8]. One participant emphasized that making the transformations transparent—through a visualization or otherwise—is crucial for ensuring the data's validity: '[.] as this [digitization] becomes a more common process, it is very important to keep a record of how we got from point A to point B. For the validity of it, provenance of it. You just want to make sure that every time, data is handled or presented in a different way; that something isn't omitted or added. Or, if it is, that there's a note on what was omitted or added and who did it.' [G2]. Another participant highlighted provenance disclosure, through visualization or otherwise, as important to acknowledging and thus ensuring data accuracy: '... if you don't know the sources you can't tell how accurate the data is.' [CS5].

However, some statements also indicate a misinterpretation of the provenance-driven visualization, overrating its ability to support transparency: 'I think it [the provenance-driven visualization] is valuable in that it's the whole story. ... It shows the whole model. ... It prompts further investigation, so from a historian point of view, to have the whole story there is valuable.' Statements such as this demonstrate the power of visualization which can provide the impression of objectivity; of showing the 'full picture', when it can ever only be as complete as the data behind it.

6.2.5 Critical voices

Two out of our 24 participants remained skeptical about the value of visualizing provenance. One participant—a historian—found the representation of provenance-related information in the form of a visual overview 'unnecessary: I don't see an advantage of the visualization over just using the current [textual] version of the biographical register, because I am pretty much searching for individual sources that I need to get more information from.' *[H7]*. Another participant acknowledged the potential value of visualizing provenance, but only as an added bonus: 'It might be enough to have it [provenance] in terms of context, but [the visualization] would be maybe for some people an added bonus.' [G7].

6.3 The role of text

Although we provided the textual information on-demand in both visualizations, participants experienced the role of text in the two visualizations differently.

6.3.1 Provenance represented through text

In the traditional visualization where provenance was represented in textual form on-demand, this information was largely ignored by participants. While all participants referred to this text at least once, most of them only briefly skimmed it or did not read it at all. Questions regarding provenance while viewing the traditional visualization did not come up, apart from a single inquiry about the information that was captured about students: 'This was in the records that people told the university?' [G4]. It is unlikely that this lack of provenance-related questions reflects on the participants' level of interest in

provenance. Participants' statements rather seem to indicate that the familiar look-and-feel of the traditional visualization simply did not trigger such questions. As outlined in Section 6.1, participants frequently commented on the 'familiar paradigm' of the traditional visualization, and one participant suggested that familiarity and intuitiveness were the contributing factors for not reading the supporting textual description of provenance: 'I didn't look too much into the textual description so much, because I am familiar with it. ... If I were new to it [to the records], I'd probably wouldn't have looked at it [the textual description], I think. I'd just assume what was going on—which may not be accurate, but it [the traditional visualization] is more intuitive.'

6.3.2 Text as a support mechanism in a provenance-driven visualization

In contrast to the traditional visualization where the textual description did not play a big role, the same text was experienced by all participants as an important support mechanism in the provenance-driven visualization: '... without that [the textual provenance information]. I would be quite lost. I thought it was really important to have. But with this one [the traditional visualization], ... that would kind of just be added information.' [G1]; 'I think it [the textual description] helps a lot for this [provenance-driven] visualization, so that you know what exactly you are looking at.' As mentioned earlier, the provenance-driven visualization was initially experienced as unfamiliar and complex, even confusing, and participants turned to the textual description for more information about the representation of the different transformation layers.

6.3.3 Textual versus visual representation of provenance

Finally, when asked about the preferred way of being exposed to provenance (text or visualization), participants acknowledged the appeal of visualization over text: 'The visualization gives you more insights and understanding because you can play with it.' [CS4]; 'I think, the people are quite lazy, and probably they won't read any of that [text], because they just want to see how it [the visualization] works, and they want to play with it.' Some statements indicate that providing text as an optional addition to the visualization of provenance might be useful: 'It's useful to have the context of the sources as the background. But then it's really interesting to have the visualization.' ; 'I think the visualization [provenance visualization] shows you more clearly the results of the work, but you still need to read the information to know where it [the underlying data] came from. I imagine a lot of people don't care where it comes from, and they are just interested in the end result. So they don't have to read it [the textual information], if they don't want to. But if you are interested, it's there for you to see.' [H5]

6.4 Historical context versus provenance

While there were few comments about provenance with the traditional visualization, both visualizations equally triggered questions regarding historical context: 'I'd like to see historical events happening.' [C3], or 'Why would someone come from Australia to Scotland?' [H8]. Participants were also interested in social aspects such as students' gender, the background of female students, or students' general social circumstances: 'I really enjoyed the small bits about the people's lives ... Probably, I would like to find out more about them, especially the female students' [G1].

This interest in the historical context of the records stands in contrast to the seeming lack of interest in provenance in the traditional visualization. We define *historical context* here as the cultural, political, or social circumstances that may have affected students listed in the University Records at the time when they were alive. It might include wars, colonialism, and parliament acts related to education or questions of race and gender. In contrast, provenance provides a context about the treatment of records over time, rather than about the people represented by the records themselves. Provenance, to a large extent, is detached from historical context, although certain historical events may be relevant in that they may have had an influence on archiving, transformation, and curation processes.

Given the historical nature of the records, the participants' curiosity in the historical circumstances of former students in the Records may not be surprising, but our findings raise the question of how to better promote provenance as another crucial source of context, enhancing how, as well as what we know about them.

6.5 Content- versus provenance-based mindset

We asked participants how they would describe each of the two visualizations to someone who has not seen them, in order to gain insights into the aspects of the visualizations that stood out to participants, and to reveal their expectations and ultimately their mindset when exploring visualizations of historical records.

6.5.1 Descriptions of the traditional visualization

Describing the traditional visualization, 15/24 participants focused on the actual information shown, how this data is encoded, and how one can interact with the visualization: 'It's a map with points on it representing students that had matriculated from the university of St Andrews within the dates it covers. And the size of each point represents the number of students that

matriculated from that point on the map.' [CS6]. 'You got a map of the world, and there are circles of different sizes. And in the circle it tells you how many people graduated from St Andrews University between 1747 and 1897. And you can zoom in or zoom out.' [G8]. Some participants highlighted the types of insights that can be extracted from the visualization: 'It's really geographical, you can see where everyone has come from, and you can see what years they were here; you can see how many came each year.'

Provenance was never mentioned in the descriptions of the traditional visualization and, as stated earlier, it was rarely mentioned in corresponding interviews overall. Instead, participants focused on the content of the records that one can extract from the visualization.

6.5.2. Descriptions of the provenance-based visualization

Provenance-related aspects came up more frequently when participants described the provenance-driven visualizations. About 10/24 participants mentioned provenance in their descriptions as follows: 'It's a visualization that attempts to capture the transformations of information from the physical entity to digital entity and all the stages involved.'. 'It represents the different stages of the dataset as it's gone from original handwritten details to a fully digitized version.' [CS6]. 'It depicts changes in the records of students in St Andrews from a certain period, and it shows how it [the data] evolved over time, or how it was developed over time, with basic information about the names of graduates that were first standardized and extended over time and digitized.' [G4].

These descriptions emphasize processes of record-keeping and how these changed over time, which indicates that participants understood, and to some extent valued the provenance perspectives provided by the visualization: 'The feature that I really liked was that you can follow the record through time. That's very nice. ... I specifically like the representation of the changes on the right side, so you can follow what happened to the records.' [G4].

6.5.3 A content-centric mindset

Participants' statements across all groups also indicate that even though they were aware of the purpose of the provenance-driven visualization, their expectations about the insights it can provide were still driven by a *content-centric mindset*. By this, we mean that participants expected to see information related to the students behind the records (e.g. education, age and birth place, parentage, or careers after university), rather than information about the modification of the records over time. This is visible in the participants' judgment of features in the provenance-driven visualization. Participants focused on how well these features facilitate content retrieval, rather than how provenance information is communicated. For example, some participants thought of the different layers in the provenance-driven visualization as a content search tool: 'I would describe it as a search tool that allows you to look for alumni of the university using a range of different specifications.' [H7]; 'I think the alphabetical order [of Smart's layer; see Fig. 1.3] is good, because you can find students.' [G8]; 'The histogram showing what had been done by Robert Smart ... showed how he reorganized it alphabetically. ... Obviously you can use it to filter by letter.' [CS2]. While participants acknowledged the purpose of the provenance-based visualization, they expected to be able to search and explore the data for its content-related interests.

Further indicating a content-centered mindset, 7/24 participants found the first two layers of the visualization (see Fig. 3.1 and 3.2) to be redundant, because these do not provide additional insights: 'In terms of content I can't see any difference.' [CS2]; '... what you might need is this one [original layer] or that one [transcription layer] ... these two are kind of superfluous.' Interestingly though, the latter participant also valued how the two layers visually highlight the transcription effort of Maitland-Anderson: 'I think what you did with the original matriculation roll and the transcription was very clever, because there is no big change between the two. But it really makes the point.' [H1].

Some participants bluntly stated to be more interested in learning about the students behind the records than about provenance: 'I see this [provenance-driven] visualization more about capturing the technical innovations of taking a handwritten source from a previous period and turning it into a digital surrogate. And that's something I find sort of interesting, but I am more interested in the historical context.' Similarly, another participant plainly stated: 'I find content more interesting than the process.' [CS2].

6.5.4 Promoting a provenance-centered mindset

Despite the dominance of a content-centered mindset, statements by 9/24 participants indicate that the provenance-driven visualization raised interest in questions regarding process: 'The third layer, was that Robert Smart? Gosh, that's a lot of work. How it's gone from a name to far more comprehensive, we've got occupation, death. That has been collected from all sorts of different sources.' ; 'It [provenance-driven visualization] is very informative to me, because it provides context. ... To see the process that people have through time done with the information that they have. ... To be able to see how that process has happened historically is massive. Because it takes you to a

completely different journey, and it's not just about the students. It's also about the people that decided to keep these records and what happened to the records as time went on.' [G1].

7 Discussion

In the light of our study results, we now discuss the merit of our proposed approach alongside its limitations and open questions, and how these may be addressed by future research on provenance-driven visualization.

7.1 The potential of provenance-driven visualization

Visualizations are known for their 'rhetorical power' (D'Ignazio and Klein, 2020) and can hide the interpretative and constructive layers upon which the data is built (Drucker, 2011). We argue that a data-driven approach to provenance disclosure and, with it, the visualization of provenance, can expose these layers and the labor that went into creating them.

Our study shows that a provenance-driven visualization can promote critical perspectives on the historical record collection by raising awareness about provenance as an essential part of contextual information. Our findings indicate that provenance is easily overlooked if not explicitly put at the center of attention—which can be achieved through visualization. Visualizing provenance can therefore add value to many projects in the Digital Humanities and beyond by promoting critical perspectives and interpretation of document collections, be it historical records or other sources of information. After all, knowing the context which surrounds items in a collection and its visualization can have far-reaching implications regarding the ways in which items and the collection as a whole—also through the visualization—are interrogated, explored, and interpreted.

However, our findings also identify potential points of friction that provenance-driven visualization may introduce. For example, participants suggest that while provenance is considered important, the interest in historical context and in the lives of students represented by the records seems stronger. Moreover, our study indicates that people are likely to approach a provenance-driven visualization with a 'content-centric' mindset, looking for information within the records, rather than the processes the records have gone through. This raises the question of whether there are other ways to visualize provenance that further promote interest in provenance-related questions. For example, our prototype shows the transformation steps records have gone through to promote transparency. While our findings suggest that this alone promoted critical questions among some participants, other visualization elements could be introduced that not only highlight but, rather, critique some of the transformations that were made. Storytelling elements, for example, could emphasize potential issues introduced by certain transformation steps (or even the lack of certain steps and measures).

Also, there are interesting parallels between our findings and research in the area of uncertainty visualization. Hullmann *et al.* report that the authors of visualization acknowledge the value of visualizing uncertainty, yet they rarely incorporate this into their visualizations (Hullman, 2019). Additional research is required to further investigate this paradox in the context of provenance visualization and beyond.

Nevertheless, visualizing provenance opens up possibilities for new ways of thinking about what in fact can, and should, be visualized. A focus shift toward visualizing processes may move forward research not only in DH, but also in the field of visualization, leading to novel visualization techniques that challenge common points of focus when it comes to visualizing historical records and the mindsets with which people approach such visualizations.

7.2 Designing visualizations with provenance in mind

Our prototype is only one illustration of what a provenance-driven visualization could look like and how it could function. Comments from participants raised a number of considerations that can inform future design explorations in this area.

7.2.1 The issue of complexity

Visualizing provenance can result in complex visualizations. Our prototype presents key transformation steps in individual visualization layers, and the participants' reactions indicate that they needed time to understand what was shown. Future work needs to explore additional visualization approaches to promote easy entry points for exploration; our visual representation choices are, admittedly, highly abstract. Closely connected to this is the question of how to integrate provenance into content-driven visualizations that make use of familiar visualization techniques (e.g. the map view we provide in our content-driven visualization). The integration of provenance into content-driven visualizations may lead to more visual complexity.

However, as also indicated by some participants, the goal should not necessarily be the simplification of provenance-driven visualization above all else. In fact, the portrayal of provenance may require intricate and complex visual representations in order to do justice to the layers of transformations and links between them. Visualization can only provide glimpses into the often

vastly complex transformations historical documents undergo. Depicting these 'glimpses' too simplistically can further obscure these processes and the interpretation of the corresponding records. In other words, we need to find a balance between promoting exploration and interpretation by providing digestible entry points into provenance-driven visualization while avoiding an oversimplification of issues that are complex.

7.2.2 Textual versus visual approaches

While common approaches disclose provenance information through separate text-only format, our prototype almost exclusively uses visual elements to represent provenance more integrally. Our findings indicate that both approaches have limitations. Provenance presented 'merely' as text runs the risk of being overlooked. In contrast, we show that visualization can put provenance-related perspectives into focus, even though the interpretation of such potentially complex visual representations (see above) can be difficult without textual explanation. In fact, our participants often consulted the textual description of provenance to fully understand the provenance-driven visualization. The combination of textual and visual approaches to provenance disclosure could be a pathway for future exploration. Visualization can be used to provide interactive overviews of transformation steps; interactive labels along with textual descriptions embedded in the visualization can provide the necessary context and explanations to facilitate interpretation. Furthermore, guided tours through the provenance-driven visualization might facilitate and promote the exploration from multiple perspectives.

7.2.3 Provenance as overview versus individual records

Providing both overview and details of the data is a common paradigm in data visualization (Shneiderman, 2003), and our prototype combines both perspectives. However, some participants pointed out that an overview, which shows provenance through aggregated views, can be problematic because it requires a high-level abstraction of transformation steps which can increase complexity. Also, decisions about what to aggregate have to be made. We aggregated records based on their organizational changes only, although other options exist, for example, highlighting the time spent on transformations from one record form into another. On the other hand, when highlighting provenance on an individual record basis, high-level patterns across records (e.g. of the consequences of structural changes) remain invisible. Further design explorations are required to investigate how best to combine overview and detailed views in provenance-driven visualization in ways that leverage the unique advantages of both perspectives.

7.2.4 Curated vignettes of provenance or the big picture?

Related to the aspects above, the design of provenance-driven visualization raises the question of how much provenance information needs to be disclosed. Do all transformations that the given historical records have gone through deserve the same amount of attention/space on the screen? Some transformations might be of higher impact than others. For example, Robert Smart's content expansion and alphabetization of the records had more significant impact than the transcription undertaken by Maitland-Anderson. Instead of giving each transformation the same screen space, we could have focused more on higher impact transformations, for example, by giving greater prominence to more significant record changes.

This idea of curated vignettes of course highlights curatorial decision-making, and therefore related ethical questions that need to be explored in the future. If our approach to provenance disclosure aims to promote ethical research, should not all the transformations deserve the same amount of attention? Every transformation creates an artifact with a merit of its own. Therefore, any measure that orders the transformations by perceived importance will introduce bias. However, provenance-driven visualization itself inevitably adds interpretation and subjectivity. Therefore, rather than avoiding subjectivity, the more productive question is how (better) to highlight the curatorial decisions that were made during the design process.

7.3 Limitations

We also acknowledge that a data-driven approach to provenance disclosure comes with limitations.

7.3.1 Resources and feasibility

Researching and documenting past forms of historical records can be time-consuming and costly. For example, the archival work, digitization, and programming required to expand existing records with provenance-related information is labor- and resource-intensive and may require additional expertise (e.g. in paleography). In some cases, depending on the type and age of historical record collections, it may not be feasible to uncover past transformations at all, or only partially.

7.3.2 Visualizing provenance

Making provenance visible in a data-driven approach, including inherent uncertainties in the records and their transformations is a challenge in itself. Structural transformation steps can be visually represented by experimenting with different spatial layouts—as visible in our prototype—while other aspects such as changes in material form are more challenging to represent at a record-level.

Further studies are required to explore when and how a data-driven approach to provenance disclosure is feasible. We argue for taking this approach into consideration as a minimum for its far-reaching implications for a more critical and ethical approach to the representation and exploration of historical record collections.

8 Conclusion

In this paper, we introduce the data-driven approach to provenance disclosure and provenance-driven visualization of historical record collections as an alternative to abstract diagrams or separate text (box). By focusing on gathering qualitative data about past forms of historical records to unearth layers of transformations and interpretations reveals the many people behind them. In comparison to diagrams or text used traditionally for this purpose (Hinrichs *et al.*, 2015; Edelstein *et al.*, 2017; Hyvönen *et al.*, 2017), a data-driven approach allows provenance disclosure at different levels of granularity. These include aggregated views on transformations of the records, as well as detail views on changes in individual records. A data-driven approach to provenance disclosure promotes user agency in encouraging the interactive navigation and exploration of the different (visual) provenance perspectives.

Through a study conducted on a historical record collection we have illustrated one way of applying the data-driven approach to provenance disclosure and provenance-driven visualization in practice. We characterize the transformation steps these records have gone through, and based on this data, develop a working prototype of a provenance-driven visualization which we have studied in-use with participants across a spectrum of backgrounds. Our study findings suggest that this approach can (1) add transparency to the visualization of historical records, and thus give rise to accuracy and validity of the research around this collection, (2) highlight ethical dimensions by paying tribute to the people, curatorial processes, and decisions that were made during the transformation of the records, and (3) introduce a shift toward the question of what can be visualized: rather than focusing on content, provenance-driven visualization aims at visualizing transformation processes.

Besides highlighting the potential of the data-driven approach to provenance disclosure and provenance-driven visualization, our case study also underlines challenges and limitations of this approach that ought to be investigated in future research. We need to further evaluate the merit of this potentially resource-intensive approach. Furthermore, the approach raises design questions related to integrating data content and provenance into a single visualization, combining text, and visual elements to more strongly support onboarding and exploration, and merging visual overviews with provenance information at an individual record level. We hope this paper will initiate further explorations in this area, leveraging approaches on provenance disclosure from both the sciences and the (digital) humanities.

Funding

My PhD research was funded by the Alfred Dunhill Links Foundation Postgraduate Scholarship.

Notes

1. https://www.ifar.org/provenance.php.
2. https://museum.cornell.edu/provenance-guidelines-resources.
3. https://en.unesco.org/fighttrafficking/1970.
4. https://arts.st-andrews.ac.uk/biographical-register/.
5. https://github.com/sebastian-meier/d3.sketchy/blob/master/README.md.

References

American Alliance of Museums. (2022). *Unlawful Appropriation of Objects during the Nazi Era*. Washington, DC: American Alliance of Museums.

Berg-Fulton, T., Newbury, D., and Snyder, T. (2015). Art Tracks: Visualizing the stories and lifespan of an artwork. *MW2015: Museums and the Web*, 8–11.

Anderson, E. W., Ahrens, J. P., Heitmann, K., Habib, S., and Silva, C. T. (2008). Provenance in comparative analysis: a study in cosmology. *Computing in Science & Engineering*, **10**(3): 30–7.

Anderson, E. W., Preston, G. A., and Silva, C. T. (2007). Towards development of a circuit based treatment for impaired memory: a multidisciplinary approach. In *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, Kohala Coast, HI, May 2007, pp. 302–5. doi: 10.1109/CNE.2007.369671

Barker, A. and Van Hemert, J. (2007). Scientific workflow: a survey and research directions. *International Conference on Parallel Processing and Applied Mathematics*, Berlin, Germany: Springer, pp. 746–53.

Bartrum, G. (2000). Research into wartime provenance at the British Museum. *British Museum Magazine*, **37**: 13–15.

Bavoil, L., Callahan, S., Crossno, P., Freire, J., Scheidegger, C., Silva, C., and Vo, H. (2005). VisTrails: enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization*. pp. 135–42. doi: 10.1109/VISUAL.2005.1532788

Betti, A., Gerrits, D., Speckmann, B., and Van Den Berg, H. (2014). GlamMap: Visualising Library Metadata. Proceedings of VALA.

Boer, V. D., Oomen, J., Inel, O., Aroyo, L., Staveren, E. V., Helmich, W., and Beurs, D. D. (2015). DIVE into the event-based browsing of linked historical media. *Journal of Web Semantics*, **35**: 152–8.

Boyatzis, R. E. (1998). *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks, CA: Sage Publications.

Brizzi, G. P. U. D. B. (2013). *ASFE: Amore Scientiae Facti Sunt Exules*. Bologna: Alma Mater Studiorum, Università di Bologna. CRR-MM. doi: 10.6092/UNIBO/ASFE

Buneman, P., Khanna, S., and Wang-Chiew T. (2001). Why and where: a characterization of data provenance. In *International Conference on Database Theory*, Berlin, Germany: Springer, pp. 316–30.

Capodieci, A., D'Aprile, D., Elia, G., Grippa, F., and Mainetti, L. (2015). Visualizing cultural digital resources using social network analysis. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, vol. **01**, pp. 186–94. IEEE.

Card, M. (1999). *Readings in Information Visualization: Using Vision to Think*. Burlington, MA: Morgan Kaufmann.

Conroy, M. (2021). Networks, maps, and time: visualizing historical networks using Palladio. *Digital Humanities Quarterly*, **15**(1).

Correll, M. (2018). Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

D'Ignazio, C. and Klein, L. (2020). *Data Feminism. Strong Ideas*. Cambridge, MA: MIT Press.

D'Ignazio, C. and Klein, L. F. (2016). Feminist data visualization. *Workshop on Visualization for the Digital Humanities (VIS4DH)*. Piscataway, NJ: IEEE.

Dörk, M., Feng, P., Collins, C., and Carpendale, S. (2013). Critical infovis: exploring the politics of visualization. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 2189–98.

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, **5**(1): 1–21.

Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., and Coleman, N. (2017). Historical research in a digital age: reflections from the mapping the republic of letters project. *American Historical Review*, **122**(2): 400–24.

Feigenbaum, G., Reist, I., and Reist, I. J. (2012). *Provenance: An Alternate History of Art*. Los Angeles, CA: Getty Publications.

Fekete, J. D. and Freire, J. (2020). Exploring reproducibility in visualization. *IEEE Computer Graphics and Applications*, **40**(5): 108–19.

Freedman, M. H., Gukelberger, J., Hastings, M. B., Trebst, S., Troyer, M., and Wang, Z. (2012). Galois Conjugates of Topological Phases. *Physical Review B*, **85**(4), 045414.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: a survey. *Computing in Science Engineering*, **10**(3): 11–21.

Freire, J. and Silva, C. T. (2012). Making computations and publications reproducible with vistrails. *Computing in Science Engineering*, **14**(4): 18–25.

GenWiki (2023). *Universitätsmatrikel – GenWiki*. https://wiki.genealogy.net/index.php?title=Universit%C3%A4tsmatrikel&oldid=2352481 (accessed 12 April 2023).

Guest, G., MacQueen, K., and Namey, E. (2012). *Introduction to Applied Thematic Analysis*. Thousand Oaks, CA: Sage Publications.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., and Coates, C. M. (2015). Trading consequences: a case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities*, **30**(1): i50–75.

Hullman, J. (2019). Why authors don't visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, **26**(1): 130–9.

Hullman, J. and Diakopoulos, N. (2011). Visualization rhetoric: framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, **17**(12): 2231–40.

Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., and Sirola L. (2017). Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web. In *International Conference on Language, Data and Knowledge*, pp. 113–9.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2017). Visual text analysis in digital humanities: visual text analysis in digital humanities. *Computer Graphics Forum*, **36**(6):226–50.

Jenkins, N., Meeks, E., and Murray, S. (2013). *Kindred Britain*.

Kouroupas, M. P. (1995). US efforts to protect cultural property: implementation of the 1970 UNESCO convention. *African Arts*, **28**(4): 32–41.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2012). Empirical studies in information visualization: seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, **18**(9): 1520–36.

Lamqaddam, H., Vande Moere, A., Vanden Abeele, V., Brosens, K., and Verbert, K. (2021). 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): 1084–94.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J., and Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, **18**(10): 1039–65.

Maitland Anderson, J. (1905). *The Matriculation Roll of the University of St Andrews, 1747–1897*. Edinburgh: Blackwood.

Maitland-Anderson, J. (1926). *Early Records of the University of St Andrews: The Graduation Roll, 1413 to 1579, and the Matriculation Roll, 1473 to 1579*. Edinburgh: T. and A. Constable for the Scottish History Society.

Merriam-Webster (2022). *Provenance*. Springfield, MA: Merriam-Webster.

Missier, P., Belhajjame, K., and Cheney, J. (2013). The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, *Association for Computing Machinery*, New York, NY, p. 773–6.

Mäkelä, E., Hyvönen, E., and Ruotsalo, T. (2012). How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. *Semantic Web*, **3**(1): 85–109.

Morisette, J. T., Jarnevich, C. S., Holcombe, T. R., Talbert, C. B., Ignizio, D., Talbert, M. K., Silva, C., Koop, D., Swanson, A., and Young, N. E. (2013). VisTrails SAHM: visualization

and workflow management for species habitat modeling. *Ecography*, **36**(2): 129–35. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0587.2012.07815.x.

Schwinges, R. C. (2018). The repertorium academicum germanicum (rag) and the geography of German universities and academics. *Geographies of the University*, **12**: 23.

Shneiderman, B. (2003). The eyes have it: a task by data type taxonomy for information visualizations. *The Craft of Information Visualization*. Amsterdam, Netherlands: Elsevier, pp. 364–71.

Smart, R. N. (2004). *Biographical Register of the University of St Andrews, 1747–1897*. North Street, St Andrews: St Andrews University Library.

Smart R. N. (2012). *Alphabetical register of the students, graduates and officials of the University of St Andrews 1579–1747*. North Street, St Andrews: University of St Andrews Library.

Stitz, H., Luger, S., Streit, M., and Gehlenborg, N. (2016). Avocado: visualization of workflow-derived data provenance for reproducible biomedical research. *Computer Graphics Forum (EuroVis '16)*, **35**(3):481–90.

The Metropolitan Museum of Art. (2022). *Provenance Research Project*. New York, NY: The Metropolitan Museum of Art.

Thudt, A., Baur, D., Huron, S., and Carpendale, S. (2016). Visual mementos: reflecting memories with personal data. *IEEE Transactions on Visualization and Computer Graphics*, **22**(1):369–78.

Tufte, E. R. (2006). *Beautiful Evidence*. Cheshire, CT: Graphics Press.

Vancisin, T., Crawford, A., Orr, M., and Hinrichs, U. (2018). From people to pixels: visualizing historical university records. In *Proceedings of the 5th Biennial Transdisciplinary Imaging Conference 2018 (Transimage 2018)*, pp. 41–57.

Vancisin, T., Orr, M., and Hinrichs, U. (2020a). Externalizing transformations of historical documents: Opportunities for provenance-driven visualization. In *Proceedings of the 5th Workshop on Visualization for the Digital Humanities (VIS4DH2020)*.

Vancisin, T., Orr, M., and Hinrichs, U. (2020b). Illuminating past labor: making transformation processes of historical documents visible. In *Proceedings of the ADHO Conference on Digital Humanities (DH2020) (Long Presentation)*.

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., and Mayr, E. (2019). Visualization of cultural heritage collection data: state of the art and future challenges. *IEEE Transactions on Visualization and Computer Graphics*, **25**(6): 2311–30.