Application of Data Provenance in Healthcare Analytics Software: Information Visualisation of User Activities

Shen Xu, PhD¹, Toby Rogers, BSc², Elliot Fairweather, PhD¹, Anthony Glenn, BA², James Curran², Vasa Curcin, PhD¹,
¹King’s College London, London, United Kingdom; ²Imosphere Ltd, Nottingham, United Kingdom

Abstract

Data provenance is a technique that describes the history of digital objects. In health data settings, it can be used to deliver auditability and transparency, and to achieve trust in a software system. However, implementing data provenance in analytics software at an enterprise level presents a different set of challenges from the research environments where data provenance was originally devised. In this paper, the challenges of reporting provenance information to the user is presented. Provenance captured from analytics software can be large and complex and visualizing a series of tasks over a long period can be overwhelming even for a domain expert, requiring visual aggregation mechanisms that fit with complex human cognitive activities involved in the process. This research studied how provenance-based reporting can be integrated into a health data analytics software, using the example of Atmolytics visual reporting tool.

Introduction

Because of the increasing complexity of analytical and data tasks, the aim of analytics software is to devise and construct visual abstractions in addition to multifaceted information so to provide usable options. New data protection laws and regulations from the EU have demanded and stipulated that computer-assisted choices must be explainable¹. Furthermore, this requirement tackles the “black box” created around the reasoning behind important actions, increasing the risk of both human error and systemic errors, and engendering mistrust in analytics².

Data provenance is a technique that describes the history of digital objects, where they came from, how they came to be in their present state, who or what acted upon them. Provenance maintains the integrity of the digital objects, e.g. the results of a data analysis engender greater trust if their provenance shows how they were obtained. In health data settings, it can be used to deliver auditability and transparency, and to achieve trust in a software system. Specifically, in Learning Health Systems, it is applicable across a range of applications³. Provenance data is commonly represented as directed acyclical graphs, with vertices being data entities, processes operating on those data and agents controlling those processes, and edges being causal relationships between the different types of vertices (e.g. data produced by a task, task utilising some data, etc).

Despite the interest in the field, there is still much work that needs to be done to make provenance practical. One important step is to obtain a meaningful answer to the query, “how was this object produced?” which may be hindered by large volumes of collected provenance data. On an enterprise level of system development, the query needs to be answered by responding to a rational justification as well as regarding a semantic viewpoint⁴. Rational justifications in this sense are not able to be automatically derived from software processes, indeed they demand input from the user; however, the user’s rationale can be assisted or enabled through provenance data capturing. The exemplar software used in this paper, Atmolytics, is a healthcare focused analytics system, interacting with patient datasets to provide answers to specific user questions. When loaded with social care data and care cost information, Atmolytics can be used by a care services director who would like to know if there are service users whose needs can be met in a more cost-effective way.

It is clear that the research community is aware of the necessity of supporting provenance; according to the taxonomy of provenance technology⁵, many researchers have developed tools and systems that help provenance dissemination, analysing records both workflows⁶–⁸ and reasoning processes⁹–¹⁴. Tools designed to support provenance across domains, particularly using an automated capture method, typically capture event-based provenance (e.g. clicks, drags, and key-presses) at a very granular level, consequently increasing computational cost and potentially overwhelming end-users.
Our research focuses on insight provenance, which shares the view of Gotz and Zhou\textsuperscript{13}, which described how the HAVEST system captures the history of actions during financial analysis activities by identifying a set of semantic units of user activity. This research further provides links between semantic actions and data entities, then represents them in an organized timeline to facilitate consumption of provenance data by end-users. Such difficulties and challenges are covered in the second section of this paper, with the focus on existing designs and challenges within the user interface (UI) design. The Atmolytics analytics software use case shall be outlined in the third section. After the user test results have been shown and assessed, a conclusion shall be drawn.

**Challenges of visualising provenance**

Provenance of digital scientific objects is metadata that can be used to determine attribution, to establish causal relationships between objects, to find common tasks and parameters that produced similar results, as well as for establishing a comprehensive audit trail to assist a researcher wanting to reuse a particular data set\textsuperscript{15}.

However, considerable effort is required to ensure the usability of provenance. To illustrate this, consider a typical provenance question: “What is the means by which the object in question was created?”. Answering this question may potentially be hindered by large amounts of provenance data collected. According to Macko et al. (2013), introducing local clustering into provenance graphs enables the identification of a significant semantic task through aggregation. This may be realised through the utilisation of online and offline metrics and a detection algorithm that may generate significant or informative results. As an example, when commencing with a seed node, S, only those nodes from S’s ancestry corresponding to semantically significant actions are present in the cluster – those that are produced by S which may also be built to assess and evaluate among and between numerous thresholds. Data provenance generation was introduced by Peng Chen et al. (2014) through the use of logical time, though this was achieved without provenance volume control generated at the source\textsuperscript{16}.

Finely grained provenance capture of digital scientific data could yield a graph that the user may find overwhelming in detail. This huge amount of data can be processed in a number of ways, e.g. by caching certain content or by progressing and furthering provenance perspectives\textsuperscript{17,18}. Provenance capture may also be throttled so that the number of provenance data generated at the source can be controlled\textsuperscript{19}. Regardless of these techniques, visualisation techniques are crucial in interpreting provenance data\textsuperscript{20}.

An absence or insufficient system engineering perspective must be accounted for, and this is the assumption of practicable concerns and considerations regarding provenance application. Though it was devised with the intention of an engineering approach, the temporal representation herein used the same logical time. The main difficulties and challenges of implementing provenance in visual data analytics are detailed below.

- **Design patterns (multiple)**. Before a particular design pattern can be followed, a system communication backbone must first be utilised (an example of this is an enterprise service bus). Web applications and web services may also possess certain design patterns (for instance, the MVC/PAC). Indeed, a particular design pattern may also apply to database access (such as an entity framework). Several different patterns for design demand varied and differentiated granularity levels. A multistate approach is most popular for web services and web applications, and, in addition, the MVC (model-view-control) and presentation-attraction-control (referred to hereafter as PAC) are generally implemented in addition to the multi-layered architecture\textsuperscript{21}. Service-oriented architecture (SOA) provenance data was studied by Tsai et al. (2007), taking the data from a service framework viewpoint; one must consider web service/application data provenance on top of SOA. Indeed, enterprise software can process numerous data at the source, at the destination of the data and in intermediary nodes. Certain information does not need to be represented to the software user, for example the data flow between model view controller.

- **Flexible combined service.** So that several different aims and intentions can be realised or met, services may be used in parallel or combined in some other manner.

- **Security.** To ensure data consistency among the collated and recorded set, sensitive data can be hidden or obfuscated.

- **Classification of dynamic data.** Not all data need to be traced within analytics software. Therefore, only those significant or important data need to be tracked. As a result, the classification of data by their criticality regarding provenance is implied as a result.
- **Data literacy.** The capacity to comprehend the meaning and importance of data is referred to as data literacy. This incorporates the means by which charts and graphs are read and understood, the right extrapolations from the available data and the acknowledgement of data use regarding inappropriate or incorrect manners. Software users, even those using advanced analytics packages, have a wide range of data literacy.

So that such provenance may be differentiated from that of a scientific data provenance research study, comparison of traditional scientific and analytic software data provenance is shown in Table 1 below.

**Table 1. Comparison of analytics software data provenance and scientific data provenance.**

| Features            | Scientific data provenance | Analytics software data provenance |
|---------------------|---------------------------|-----------------------------------|
| Design patterns     | N/A                       | Many design patterns can be involved, for instance, SOA and/or MVC which is usual for web applications driven by data. |
| Services            | N/A                       | Combination of methods by the user for adherence to different aims and goals. |
| Security            | Traditionally utilised security mechanisms such as covert channels, digital signatures, encryption, kernel authentication for security and authorisation. | Same problems tackled, though within an SOA system context, particularly regarding data routing with content-based data. |
| Data classification | Generally static          | Potentially dynamic. |
| Data literacy       | High exposure level       | Assorted. |

**User Activity Driven Solutions for Analytics Software**

Utilisation of provenance in regard to a particular domain is dependent on granularity level. Coarse- and fine-grained as concepts are relative to data being observed. For example, when applying data provenance to a relational database, tuple-level is referred to as fine-grained while relation-level is referred to as coarse-grained data. Meanwhile, two granularities can be determined regarding SOA provenance:

- **Fine-grained provenance:** This refers to intra-system data movement tracing, such as where the data originated and where it is headed, the time at which the data was created, the rationale concerning the dataset and the data's termination.
- **Coarse-grained provenance:** This alludes to workflow processing and the data generated by it.

Within the context of Web application which is on top of SOA, whether the user is more concerned with the system processes’ ancestry or whether or not their colleague has arrived at their conclusion is the main issue in question. According to Roberts et al. (2014), provenance may be categorised into the following strata: *data level*, which concerns inter database movement; *the analysis level*, which concerns intra-systematic interactions; and finally, the *reasoning level*, which concerns the decision-making process, the reasoning and the thinking that comprise the function of the analytic stages. The provenance question, per these stages and this approach, may be translated into two different user stories. Further, it may be considered according to these three aforementioned levels. This can be seen in Table 2.
Table 2. Provenance Question (interpreted at different levels).

| Provenance query                                           | User scenario                                                                 | Levels                                                                 | Level of granularity |
|------------------------------------------------------------|-------------------------------------------------------------------------------|------------------------------------------------------------------------|----------------------|
| “By what means was the object created?”                    | As a budget controller. I wished to see how my colleagues generated a list of candidate service users so that my confidence regarding configuration cost adjustment can be increased. | What were the reasoning steps as well as these assumptions made by experts? [Level of Reasoning] | Coarse-grain         |
|                                                            | I wish to determine, as a data manager, why there are two population sizes for the identical cohort and I want to determine whether there are incorrect figures or processes therein. | What were the data extraction processes? [Level of Analysis] | Fine-grain           |

From Table 2, we can see that the same provenance question can be answered from different perspectives with different foci. Analytic provenance (for example, select and click) was included in our initial design of provenance template. Nevertheless, such interactions could be aggregated into abstract process through the relation of intangible systematic element, thus a bridge of system activities to reasoning steps can be ‘constructed’. Additionally, other studies have presented a machine-learning method that incorporates inference of reasoning provenance taken from log data of user interactions. The authors of this study believe that, as it is such a challenge to capture reasoning provenance data since it is fit within “the mind” of the analyst in question and more tacit than overt, though conversely the inter-system analyst interactions (analytic provenance) may actually be captured with relative ease, thereby allowing for action-expressed intentions. The endeavours of this research study concerned low-level classification, like scrolling, to reasoning stages regarding instance seeking. The significance of reasoning processes in analytics software is thus addressed.

According to the requirements of the SOA, this research extends the analytics software requirements with suggested mechanisms. This, it is hoped, shall bring about a more practicable use of and regarding data provenance. The requirements and mechanisms of this are summarised in Table 3. The mechanisms proposed, as per a system engineering perspective, adhere to a provenance template methodology as devised by another study, and it brings about a system architecture for provenance captured and displaying for web applications. Herein the representation of data provenance for software users constitutes the primary research focus.

Table 3. Suggested mechanisms.

| Item | Feature | Type | Analytics software requirements | Proposed mechanisms |
|------|---------|------|---------------------------------|---------------------|
| A    | Design patterns (multiple) | MVC/Entity Framework/SOA, and others | Should permit flexible capturing means | Provenance Abstracted Activity/template (W3C standard complied) + Selected feature |
| B    | Services (Flexible combination) | Potential recombining of methods to meet different ends | Flexible capturing mechanism required | Provenance grafting/template method |
| C    | Security | Disguising data | Should allow encryption on data sent across boundaries. | Pre-processing Encrypt and Decrypt/data |
| D    | Authentication mechanism needed to guarantee data is taken from correct sources | Data/template accessibility capturing provenance |
| E    | Different access needs to be provided to different users | Provenance |
so to guarantee data security

Capturing/template agent node security

Data volume and assorted data literacy

Confidence/trust judgement concerning reasoning level

Representation allowed by adhering to the logical steps taken by the user

Logical temporal stamp (1); Input justification (2)

Judgement of confidence/trust above that of analysis level

Support should be provided for logical steps

Feature list (3) + Highlighted difference (4)

Judgement of confidence/trust above that of data level

Support should be provided for data sources

Summary of Linked System Activity

Partitioning of Provenance Data Graphs

There are two reasons why an annotated provenance graph is not the best visual solution for representing provenance data. Firstly, with potentially thousands of attributes and nodes in a graph, scoping the relevant items becomes difficult. Second, placement of non-structural and structural information within a single status display yields unwieldy results. As a result of this, we have introduced a graph partitioning mechanism based on temporal representation and user activities, with the aim of producing confidence in a given data product within an analytics solution. Further relevant information concerning the system-based activities and related automated processes is also included.

A provenance graph contains three different types of nodes: agent, activity and entity nodes. Our partitioning is total and creates a series of non-overlapping node subsets according to the logical temporal stamp and the targeted identifier. A partition of a provenance graph is therefore, more exactly, seen to be $G=(V, E)$. Here, the set of agent nodes is denoted by $V^a$, while the set of entity nodes are denoted using $V^e$, while the activity node set is denoted by $V^a; V^a \cup V^e \cup V^w = V$ is used to denote the total set of nodes wherein $E$ denotes all edges, which can be seen to be denoted per: $G=(V, E)$ provenance graph, $V$ is then partitioned into several subsets, per $V_1, V_2$ etc. Thus:

1. $V_1, V_2, \ldots, V_k \in V$ and $\bigcup_{i=1}^{k} V_i = U$
2. $\forall i \neq j \text{ and } 1 \leq i, j \leq k, V_i \cap V_j = \emptyset$
3. $\forall a, b \in V^{ac}\subseteq V_i$, we must have $\text{LTS}(a) = \text{LTS}(b)$

LTS is used to denote the logical temporal stamp. Here a function assumes an activity node to be an output and thus generated an output stamp. Additionally, each node is annotated with the stamp according to its adherence to certain relationships. One-day-long provenance graphs are made from the present testing stamp (which is one day long).

Figure 1. Activity driven representation of data provenance.
Every one of the subsets is arranged into an ordered list as such: $\{V_1, V_2, \ldots, V_k\}$. An “appears prior to” relationship was suggested by Chen et al.\textsuperscript{16}. The approach used herein is to arrange such subsets into a timeline arrangement per (logical stamp) real-time ordering. Hence throughout the capturing process, the subsets will be ordered sequentially. This implicit order is then undertaken as the user interacts with the web app. This is in contrast to some other studies in the provenance field, many researchers that ignore the temporal ordering and instead use significant computational activities to re-engineer it from the bottom-up. The activity design is shown in Figure 1, to be read from left to right, showing our design of activity-based temporal representation.
User tests and analytics software case study

In the Atmolytics architecture, an enterprise service bus is used to receive data tasks before they are distributed over several farms for processing, enabling multiple shared databases to be used to complete a single task. The means by which a provenance infrastructure can be embedded into the Atmolytics system architecture is depicted in Figure 2, based on programmatic calls within Atmolytics invoking a RESTful API upon the provenance server. The provenance services correspond to standard actions in the system and are implemented using abstract provenance templates\textsuperscript{11, 12} which get instantiated during API service calls with concrete data and persisted into the provenance data store\textsuperscript{27} (B in Figure 2). Provenance capturing (A in Figure 2) is triggered by a controller in Atmolytics – a Targeted Activity – after which the object continues to the method used for data processing. Content may be encrypted prior to its being sent to data storage regarding every individual data slot. This enables our proposed mechanism for security requirements in table 3, Item C. The use of provenance templates enables our proposed mechanism for design patterns and security in table 3, Items A, D, E. After a new graph is created (C in Figure 2), it is typically linked into the existing graph by means of grafting the new nodes onto the existing structure. This supports the loosely-coupled services feature in table 3, Item B.

![Figure 2. Simplified representation of web application processes and provenance capturing (adapted from Xu et al 2016\textsuperscript{30})](image)

A: Provenance capture triggered from a process in Atmolytics; B: Abstract provenance template gets instantiated with the details provided by the process\textsuperscript{28, 29}, C: New provenance data gets grafted onto existing data in the provenance data storage\textsuperscript{11, 12}.
The initial round (comprised of five users) of the user advisory group is completed, according to present progress. This intends to ascertain a collated sentiment or opinion as well as design feedback above that of the implementation of provenance and thus assist the system regarding its trust/confidence in data products as well as increase transparency. Every one of the tests is approximately two hours in length and the analytics software demonstrative exemplar is covered in each of the sessions. Feedback regarding the designs is provided as and when the user interacts with the mocked-up interface.

Figure 3. User experience review mock-up of user interface (https://invis.io/NQDCUN6DS shows the 2nd round user interface). F1 is the logical temporal stamp; F2 is the activities’ justification; G3 is the prechosen feature list; G4 underlines the changes; and H is the linked system activity summary, refer to table 3.

Looking at Figure 3, F1 shows the revised design of temporal representation, because during our review, some users found the aggregated nodes showing number of activities confusing. Compared to Figure 1, all nodes are of equal size
and allocate them distinctive, representing the requirement F in Table 3. Annotation/Justification (F2) allows user free text to be shown alongside the provenance information. Data exploration is a form of unstructured problem solving, so activities should be justified in case of reviewing. Together with F1, the reasoning steps and justification provide a higher-level auditing of data exploration.

G3 denotes the feature space of each activity, describing how the data is affected by the user-selected features. According to the user feedback, we have highlighted the changes between activities (G4). H represents system activities involving system processes linked to the changes of cohort, for example updates, changed base cohort size etc.

The users were also asked about the sentiments and opinions concerning the implementation of our provenance infrastructure. Their opinions are summarised in Table 4.

Table 4. Summary of user feedback. (5 Strongly agree, 4 Agree, 3 Neutral, 2 Disagree, 1 Strongly disagree).

| Questions                                                                 | A | B | C | D | E |
|---------------------------------------------------------------------------|---|---|---|---|---|
| 1 The provenance information displayed helps you understand how the result is produced. | 4 | 4 | 5 | 4 | 4 |
| 2 The provenance information displayed provides transparency to the processes of producing outputs. | 4 | 4 | 5 | 3 | 4 |
| 3 The provenance information displayed improves the trust/confidence of outputs. | 4 | 2 | 4 | 3 | 4 |
| 4 The provenance feature captures the decisions involved in producing an output. | 4 | 5 | 3 | 4 | 4 |
| 5 The provenance feature is useful to your work. | 4 | 4 | 5 | 4 | 4 |

The initial feedback from these users indicates that our approach aligns well with users’ own cognitive model. For example, all users during the test session recognize and understand well of displayed activities. One user reported that “If accompanying annotation/comments it can make clear the rationale”. In addition, all users welcomed a future review of the design to see the development of provenance feature.

Notably, user B disagreed with the confidence/trust statement. “At a basic level yes but the problem as discussed is the lack of transparency back to the original distributed data sources themselves. The outputs are dependent on that data so ideally you need to have visibility at least of the import/mapping process to those data sources to have trust in the report outputs”. This relates to the stage in the data lifecycle at which we start capturing the provenance. In our current implementation, this begins once the data sets are in the Atmolytics system. However, we are currently exploring the implementation of provenance during the Extract-Transform-Load stage during which original data is being transformed and loaded into Atmolytics, and manually annotating the loaded data with relevant ethics and governance information.

Conclusion

Provenance is a critical requirement for analytic applications, but the methodologies of many existing implementations, typically manual browsing of recorded provenance or pre-defined queries of provenance data, have fundamental limitations. This study aims to establish a flexible method of data provenance capture and visualization within analytics software, using the Atmolytics health data analytics software as an exemplar. As part of the study, the objective of this paper is to develop a new approach based on semantic actions that combines the benefits of both approaches while avoiding their deficiencies.

The research concentrates on the various analytic processes’ reasoning steps, with the consequential design being more appropriate to a broader provenance community, and potentially driving wider appreciation and adoption of provenance. A positive response to the activity-driven temporal representation of data provenance is confirmed in our initial evaluation, and user studies are continuing.

Several challenges remain to be addressed in future work. In particular, our approach represents provenance artefacts generated by system-level processes in the same way as user-generated events, but the abstraction required to do this may cause a loss of fidelity for some such processes; further development could produce an activity map that has capacity to increase the granularity of the activity to address this. Finally, we must perform more comprehensive system performance evaluation to fully evaluate our system development in addition to user studies.
References

1. Association for Computing Machinery US Public Policy Council (USACM). Statement on algorithmic transparency and accountability. USACM Press releases. 2017;(January 12):1–2.

2. Broadhurst K, White S, Fish S, Munro E, Fletcher K, Lincoln H. Ten pitfalls and how to avoid them: What research tells us. 2010;(September):1–42. Available from: papers2://publication/uuid/3ABE0587-8B27-4BEB-9A22-F3B4B34D4C85

3. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learn Heal Syst [Internet]. 2017 Apr 1;1(2):e10019–n/a. Available from: http://dx.doi.org/10.1002/hr2.10019

4. Shen Xu, Kecheng Liu, Llewellyn C.M. Tang, Weizi Li. A Framework for Integrating Syntax, Semantics and Pragmatics for Computer-aided Professional Practice: With Application of Costing in Construction Industry. Comput Ind. 2016;83c:28–45.

5. Simmhan YL, Plale B, Gannon D. A Survey of Data Provenance Techniques Technical Report IUB-CS-TR618. Science (80-) [Internet]. 2005;47405(3):1–25. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70.6294&rep=rep1&type=pdf

6. Parker SG, Johnson CR. SCIRun: A Scientific Programming Environment for Computational Steering. In: Proceedings of the 1995 ACM/IEEE conference on Supercomputing. 1995. p. 52.

7. Davidson SB, Freire J. Provenance and scientific workflows. Proc 2008 ACM SIGMOD Int Conf Manag data - SIGMOD ’08 [Internet]. 2008;1345. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-57149126952&partnerID=Z0tx3y1

8. Bavoil L, Callahan SP, Crossno PJ, Freire J, Scheidegger CE, Silva T, et al. VisTrails: Enabling Interactive Multiple-View Visualizations. In: IEEE Conference on Visualization [Internet]. Minneapolis, MN: IEEE; 2005. p. 135–42. Available from: http://ieeexplore.ieee.org/document/1532788/citations

9. Kreuseler M. A History Mechanism for Visual Data Mining. In: Proceeding of IEEE Symposium on Information Visualization. 2004. p. 49–56.

10. Jankun-Kelly TJ, Ma K-L, Gertz M. A Model for the Visualization Exploration Process. In: Proceeding of IEEE Visualization. Boston, MA, USA: IEEE; 2002. p. 323–30.

11. I2 I. Analyst’s Notebook [Internet]. IBM; Available from: https://www.ibm.com/uk-en/marketplace/analysts-notebook

12. Eccles R, Kapler T, Harper R, Wright W. Stories in GeoTime. In: Proceeding of IEEE VAST. 2007. p. 19–26.

13. Gotz D, Zhou MX. Characterizing Users’ Visual Analytic Activity for Insight Provenance. In: IEEE Symposium on Visual Analytics Science and Technology. Columbus, Ohio, USA: IEEE; 2008. p. 123–30.

14. Dow W, Jeong DH, Stukes F, Ribarsky W, Lipford HR. Recovering reasoning process from User Interactions. In: IEEE Computer Graphics & Applications. 2009.

15. Bechhofer S, Goble C, Buchan I. Research Objects: Towards Exchange and Reuse of Digital Knowledge Research Objects: Towards Exchange and Reuse of Digital Knowledge. 2010;(August 2017).

16. Chen P, Plale B, Aktas MS. Temporal representation for mining scientific data provenance. Futur Gener Comput Syst [Internet]. 2014;36:363–78. Available from: http://dx.doi.org/10.1016/j.future.2013.09.032

17. Zhao J, Wroe C, Goble C, Stevens R, Quan D, Greenwood M. Using Semantic Web Technologies for Representing E-science Provenance. ISWC 2004, Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. 2004;3298:92–106. Available from: http://twiki.ipaw.info/bin/view/Challenge/MINDSWAP

18. Gehani A, Park M, Lafayette W. Efficient Querying of Distributed Provenance Stores.

19. Braun U, Garfinkel S, Holland DA, Muniswamy-Reddy K-K, Seltzer MI. Issues in Automatic Provenance Collection. In: Moreau L, Foster I, editors. Provenance and Annotation of Data: International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 171–83. Available from: https://doi.org/10.1007/11890850_18

20. Silva CT, Freire J, Callahan SP. Provenance for Visualizations: Reproducibility and Beyond. Vol. 9, Computing in Science & Engineering. 2007. p. 82–9.

21. Patrick C. Literature Review Service Frameworks and Architectural Design Patterns in Web Development. 2014;(May).

22. Carlson JR, Fosmire M, Nelson MRS. Determining Data Information Literacy Needs: A Study of Students and Research Faculty. 2011;
23. Ridsdale C, Rothwell J, Smit M, Ali-Hassan H, Bliemel M, Irvine D, et al. Strategies and Best Practices for Data Literacy Education. 2016;
24. Cheney J, Chiticariu L, Tan W-C. Provenance in Databases: Why, How, and Where. Found Trends Databases. 2007;1(4):379–474.
25. Roberts JCC, Keim D, Hanratty T, Rowlingson RRR, Walker R, Hall M, et al. From Ill-Defined Problems to Informed Decisions. EuroVis Work Vis Anal [Internet]. 2014;7–11. Available from: http://hdl.handle.net/10.2312/eurova.20141138.007-011
26. Kodagoda N, Pontis S, Simmie D, Attfield S, Wong BLW, Blandford A, et al. Using Machine Learning to Infer Reasoning Provenance From User Interaction Log Data: Based on the Data/Frame Theory of Sensemaking. J Cogn Eng Decis Mak [Internet]. 2017;11(1):19. Available from: http://edm.sagepub.com/lookup/doi/10.1177/1555343416672782
27. Curcin V, Fairweather E, Danger R, Corrigan D. Templates as a method for implementing data provenance in decision support systems. J Biomed Inform [Internet]. 2017;65:1–21. Available from: http://dx.doi.org/10.1016/j.jbi.2016.10.022
28. Lebo T, Sahoo S, McGuinness D. PROV-O: The PROV Ontology [Internet]. World Wide Web consortium (W3C). 2013 [cited 2016 Jun 7]. Available from: https://www.w3.org/TR/2013/REC-prov-o-20130430/
29. Moreau L, Missier P, Belhajjame K, B’Far R, Cheney J, Coppens S, et al. PROV-DM: The PROV Data Model [Internet]. W3C. 2013 [cited 2016 Jan 1]. Available from: https://www.w3.org/TR/prov-dm/
30. Xu S, Rogers T, Curcin V. Capturing Provenance of Visual Analytics in Social Care Needs. In: Informatics for Health 2017. Manchester; 2016. p. 2.