Implementing Data Provenance in Health Data Analytics Software

Shen Xu[1], Elliot Fairweather[1], Toby Rogers[2], Vasa Curcin[1]

[1] King's College London, London, United Kingdom
[2] Imosphere Ltd, Nottingham, United Kingdom
shen.xu@kcl.ac.uk

Abstract. Data provenance is a technique that describes the history of digital objects. In health applications, it can be used to deliver auditability and transparency, leading to increased trust in software. When implementing provenance in end-user scenarios, on top of standard provenance requirements, it is important to properly contextualize the provenance features within the domain and ensure their usability. We have developed a novel user interface, embedded into Imolytics data analysis tool and based on our Provenance Template technology, to help the end-user consume provenance information. In this demonstration, we shall demonstrate how the interface can be used to examine the audit trail of analysis results to spot when the two analytical methods start producing different results. In addition to the novel provenance UI, this is the first implementation of standard-based data provenance in a commercial data analytics software tool.

Keywords: data provenance; system demo.

1 Introduction

Atmolytics is a data analytics tool aimed at health and social care that focuses on simplifying the data insight process by providing users with a set of generic apps that can be customized into powerful interactive reports. The reports operate on cohorts – patient data sets that are created from various databases integrated into Atmolytics internal data warehouse. This model has proven effective for a range of use cases, from genetical studies and cancer centres to primary care repositories for epidemiological studies. In the Atmolytics architecture, an enterprise service bus is used to receive data tasks before they are distributed over several farms for processing. Programmatic calls within Atmolytics are invoking a RESTful API upon the provernance server. The provenance services correspond to standard actions in the system and are implemented using abstract *provenance templates* which get instantiated during API service calls with concrete data and persisted into the provenance data store. Provenance capturing is triggered by a controller in Atmolytics – a Targeted Activity. After a new graph segment, denoting a specific Atmolytics component, is created, it is linked into the overall provenance graph by grafting the new nodes onto the existing structure.

The architecture of the Provenance Template Server is shown in Figure 1. At the core of the system is the model component. Provenance documents are represented as graphs in which vertices and edges are typed and annotated with key-value pairs. The graph itself may also have key-value properties. Serialisation and deserialisation to PROV data formats is accomplished using the parsers provided by ProvToolbox library. Substitutions also form part of the model and parsers to both a proposed PROV-N format and JSON are given in the implementation. The template instantiation algorithm by which new fragment documents are generated from templates and substitutions is also defined within the model component.

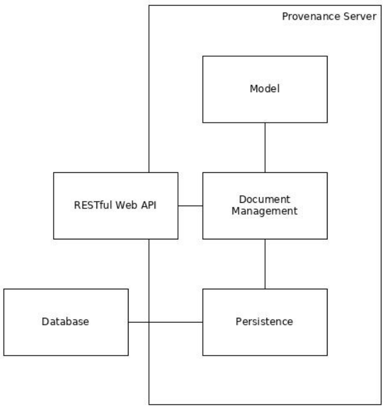


Figure 1: Provenance template server architecture

Storage of data in the system is abstracted by a persistence component to enable the use of different database technologies. Here, a Neo4j graph database is used but a relational database, SPARQL-enabled or alternative graph database could be used either instead or concurrently. The metadata of documents is stored and updated separately to the graph data itself to facilitate indexing and other administrative operations upon the documents stored.

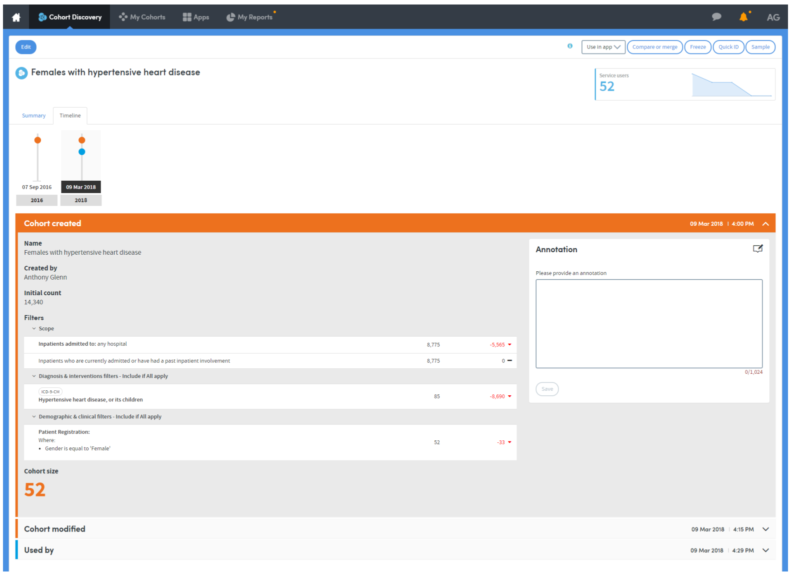
The system is accessed via the document management component. This controls and executes operations such as the creation of new target documents, namespace management, the registering of templates, and the generation and merging of new fragment documents. Fragment generation is achieved through interaction with the model component. Operations requiring the import, export or update of document data and metadata are supported via the persistence component.

Access to the document management interface is provided via a RESTful web service. Documents and substitutions are passed to and from the server encoded as JSON and analysis is conducted by running Cypher queries over the underlying Neo4j database. The specifics of a higher-level query interface for the system, agnostic to a particular storage solution, is an area of ongoing research.

2 Use case

Atmolytics system is based on the use of patient cohorts for data analysis, which are updated over time to reflect new additions to the data set. These cohorts can be created either as dynamic queries or as static patient lists generated through set operations. A common problem users are facing is to confuse the two and then later noticing that the latter does not get automatically updated when new patients are added to the cohort. In our demonstration, we shall demonstrate how provenance data captured in Atmolytics can be used to trace the origin of this problem.

Taking as an example an investigation of female hypertensive heart disease (Figure 2), a cohort could be created in two ways: 1) by creating a query for a group of female patients with added criteria for presence of hypertensive heart disease, 2) by creating a group of female patients and another group of hypertensive heart disease patients, and then applying the subgroup function to create a static list which is the intersection of the two groups. The results of the two approaches will initially be identical, however over time, the cohort sizes might change.



Females with hypertensive heart disease

Service users
52

Cohort created 09 Mar 2018 | 4:00 PM

Name
Females with hypertensive heart disease

Created by
Anthony Glenn

Initial count
14,340

Filters

Scope		
Inpatients admitted to: any hospital	8,775	-5,565
Inpatients who are currently admitted or have had a past inpatient involvement	8,775	0
Diagnosis & interventions filters - Include if All apply		
ICD-9-CM Hypertensive heart disease, or its children	85	-8,690
Demographic & clinical filters - Include if All apply		
Patient Registration: Where: • Gender is equal to 'Female'	52	-33

Cohort size
52

Annotation

Please provide an annotation

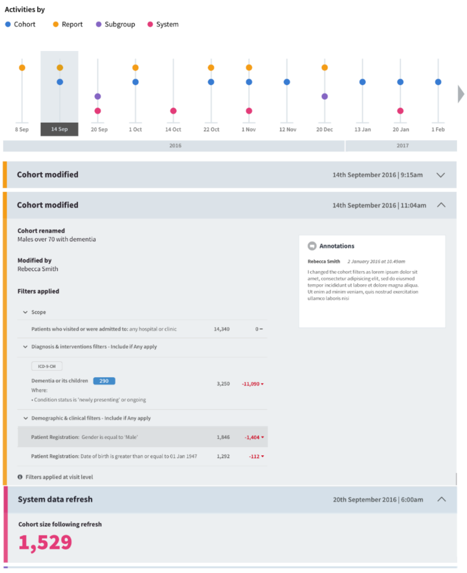
Cohort modified 09 Mar 2018 | 4:15 PM

Used by 09 Mar 2018 | 4:29 PM

Figure 2: Cohort of females with hypertensive heart disease

To that goal, the demonstration will show:

a. The Atmolytics use case of diverging cohorts
b. How provenance is captured through the Provenance Template Server
c. How provenance data is visualized in Atmolytics
d. How this visualization is used to address the use case.



Cohort modified 14th September 2016 | 9:15am

Cohort modified 14th September 2016 | 11:04am

Cohort renamed
Males over 70 with dementia

Modified by
Rebecca Smith

Filters applied

Scope

Patients who visited or were admitted to: any hospital or clinic	14,340	0 –

Diagnosis & interventions filters - Include if Any apply

ICD-9-CM

Dementia or its children 290 Where: • Condition status is 'newly presenting' or ongoing	3,250	-11,090 ▾

Demographic & clinical filters - Include if Any apply

Patient Registration: Gender is equal to 'Male'	1,846	-1,404 ▾
Patient Registration: Date of birth is greater than or equal to 01 Jan 1947	1,292	-112 ▾

Filters applied at visit level

Annotations

Rebecca Smith *2 January 2016 at 10.49am*

I changed the cohort filters as lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi

System data refresh 20th September 2016 | 6:00am

Cohort size following refresh

1,529

Figure 3: Timeline-based Activity Chain visualization of provenance data

Figure 3 shows the temporal provenance view, visualizing relevant events on a timeline, but also allowing free text to be shown alongside the provenance information. This facilitates the justification of activities while reviewing the origin of results or patient cohorts. Following user feedback, the interface highlights the changes between activities, e.g. cohort updates, change of base cohort size etc. The events history provenance reporting will also be demonstrated on a separate data example.

3 Summary

Atmolytics adopted a data provenance approach to implementing the auditing capabilities required by their users and the evolving legislative landscape. The new features help improve end-users' trust in their results and data exploration performed.The users' response to initial provenance reporting functionality has been positive in initial evaluations, and usability studies are continuing.

References

1. Curcin, V. et al.: Templates as a method for implementing data provenance in decision support systems. J. Biomed. Inform. 65, 1–21 (2017).
2. Xu S. et al.: Application of Data Provenance in Healthcare Analytics Software: Information Visualisation of User Activities. In: Proceedings in AIMA Informatics Summits 2018.
3. Xu, S. et al.: Capturing Provenance of Visual Analytics in Social Care Needs. In: Informatics for Health 2017. p. 2 , Manchester (2016).