Using Provenance in Data Analytics for Seismology: Challenges and Directions

Umberto Souza da Costa, Javier Alfonso Espinosa-Oviedo, Martin Musicante, Genoveva Vargas-Solar, José-Luis Zechinelli-Martini

Using Provenance in Data Analytics for Seismology: Challenges and Directions*

Umberto S. da Costa[4], Javier A. Espinosa-Oviedo[3], Martin A. Musicante[4], Genoveva Vargas-Solar[1], and José-Luis Zechinelli-Martini[2]

[1] CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205,69622 Villeurbanne, France
genoveva.vargas-solar@cnrs.fr
[2] Fundación Universidad de las Américas Puebla, 72820 San Andrés Cholula, Mexico
joseluis.zechinelli@udlap.mx
[3] CPE Lyon, LIRIS, Université de Lyon, Villeurbanne, France
javier.espinosa-oviedo@cpe.fr
[4] Universidade Federal Rio Grande do Norte, DIMAp, Brazil
{umberto,mam}@dimap.ufrn.br

Abstract. We analyze data and meta-data modeling challenges to provide curated collections that can be easy to explore. Data exploration can use provenance tools to give insight into the conditions in which data are collected. We are concerned with data curation and exploration in seismic geophysics. We believe that the tasks involved in graph exploration depend highly on the knowledge domain. The discussion about possible solutions is driven by the hypothesis that graphs can be well-adapted data models to represent, explore and analyze seismic data. Given that data curation is done by human agents, it is essential to provide automatic tools to add provenance to seismic data.

Keywords: data curation · metadata extraction · graph database design · data exploration · graph analytics and querying

1 Introduction

With the unprecedented volumes of heterogeneous data automatically collected and available, data collections must be heavily transformed before consumers can use them. Data curation approaches [1] have defined strategies and protocols to maintain data collections and contribute to preparing and integrating datasets to perform analytics tasks. Curation tasks include extracting meta-data and integrating semantic information. Data-driven analytics and experimentation have quality requirements concerning the validity of the data and results. For example, Geophysics experiments and analyses to study the seismic behavior of specific zones rely on data collections produced by observation stations (seismographs) and labeled manually by experts. Observation stations are located

* This work is partially funded by the project ADAGEO funded by the CNRS EDI program. Authors are listed in alphabetical order.

in different environments and submitted to conditions that can perturb sensed data. Their location determines the signals detected, for instance, near a mine, the sea, or a crowded urban space. Studying seismic behavior using the data collected from these sensors must consider this provenance meta-data to better drive conclusions. For instance, there was an extreme seismic event in Puebla city, located near the seismic zone along a geological fault line of the *Sierra Madre Occidental*, and it is a mine zone close to a risk zone of the Popocatepetl volcano. Geophysicists must compare data from different observation stations to classify an observation as an earthquake or a human-generated seismic event.

Exploratory search allows us to discover and understand relevant data to answer the informational needs of users. This data exploration task is critical for driving conclusions. Therefore, data and provenance meta-data should be associated to guide the construction of datasets. Next-generation data management engines should aid the user in understanding the data collections' content and guide to exploring data.

This position paper analyses data and meta-data modeling challenge to provide curated collections that can be easily explored. Data exploration can be done considering data provenance to give insight into the conditions in which data are collected. For instance, the device used to collect data, the human/synthetic agents that analyse data, etc. We focus on geophysics data curation and exploration because we believe these tasks depend highly on the knowledge domain. We postulate that:

- Graph data models can be well adapted for representing and storing the data and (provenance) meta-data mesh.
- Curation processes can consider enriching/completing graphs (data and their provenance) using link discovery techniques.
- Exploration can be done (semi) declaratively using domain-specific languages that can traverse and analyze graphs to extract the portions of data pertinent to given analytics tasks.

These postulates call for addressing graph database design, processing, and querying challenges discussed in this paper and put in perspective with graph modeling, processing, and querying existing work. Accordingly, the remainder of this paper is organized as follows. Section 2 introduces related work about data provenance and quality and data exploration techniques. Section 3 describes a motivation example that shows the need for a provenance-guided exploration for selecting data for performing geophysics analytics processes. Section 4 introduces our vision for modeling geophysics data collections and provenance using graphs. It proposes the general lines for defining a domain-specific graph exploration language for geophysics data collections. Finally, Section 5 concludes the paper and discusses research directions and opportunities.

2 Related work

Provenance refers to sources of information involved in producing or delivering a product. The provenance role is crucial in deciding whether the information

is reliable, integrating it with other diverse sources of information, and giving credit to its originators when using it. Provenance, therefore, provides a critical basis for assessing authenticity, enabling booth trust and reproducibility.

Despite its importance for the usability of information, tracking provenance, sharing data, and integrating them from multiple sources according to their provenance remains an open issue [13]. Several workflow management systems and Semantic Web systems have been developed and are actively used by communities of scientists. Many of these systems implement some form of provenance tracking internally and have begun standardising some common representations for provenance data to allow for the exchange and integration [5].

Provenance models for databases A database can be broadly understood as a data repository that enables queries. Often, information about the data itself is also stored, indicating, in addition to its value, its type, and other restriction rules [2]. Data collected about data routinely may fall into the category of provenance information, *e.g.* creation date, creator, instrument or software used, data processing methods, etc. In this way, good data management practices are the basis for accurately recording provenance. Provenance is recorded as metadata that may include items used to compile provenance information: plain text files, spreadsheets, file names, databases, etc. Provenance data can be represented using different data models (relational, graph, documents) and are usually associated with the data items to which they refer.

The PROV-DM model [5] is an extension of the PROV model [8] released by the Provenance Interchange Working Group and recommended by the W3C. PROV-DM promotes the interoperable exchange of provenance information in heterogeneous environments. PROV-DM is defined using an abstract relational model and an OWL ontology, with various serializations including RDF and XML. PROV is generic and domain-independent. It provides extension points through which such systems and applications can extend PROV for their purposes [4].

Provenance-based querying on graphs Graph databases are a specific database type that falls under the NoSQL (Not Only SQL) category. Graph databases are composed of data items (the nodes of the graph) and links between them (the edges of the graph). Properties may be associated with the nodes. Popular graph database systems include RDF/SPARQL [12] and Neo4J [6]. The aspect of providing provenance explanations for query results has been mostly neglected. Based on query rewriting, the method, SPARQLprov [3] computes "how-provenance polynomials" for SPARQL queries over knowledge graphs. SPARQLprov has been evaluated on "real" and synthetic data. Results show that it incurs good runtime overhead w.r.t. the original query, competing with state-of-the-art solutions for how-provenance computation. The work in [9] establishes a translation between a formalism for dynamic programming over hypergraphs and the computation of semiring-based provenance for Datalog programs. The approach proposes a new method for computing provenance for a specific class of semirings.

Provenance in scientific workflows Scientific workflows implement online data-driven experiments of specific experimental sciences (biology, geosciences, physics, etc.). Scientific workflow provenance, both for the data they derive and their specification, is essential to allow for reproducibility, sharing, and knowledge reuse in the scientific community. Harvesting provenance for streaming workflows introduces challenges related to the characteristics of scientific workflows. Executing these workflows produces a high rate of updates and promotes a large distribution of the tasks spread across several institutional infrastructures. Since activities are often externalized can be an obstacle to enabling provenance metadata extraction procedures. According to the target experimental science, maintaining and managing provenance in scientific workflows must be often specialized. For example, the work proposed by A. Spinuso et al. [11] is an example of an approach dealing with provenance in seismological processing workflows. The work by S. Bowers et al. [1] captures the dependencies of data and collection creation events on preexisting data and collections and embeds these provenance records within the data stream. A provenance query engine operates on self-contained workflow traces representing serializations of the output data stream for particular workflow runs. The bottom line is analyzing the large amounts of provenance data generated by workflow executions and extracting valuable knowledge of this data [7].

Discussion According to the taxonomy proposed in [10] provenance can be used for describing processes and data under a fine or coarse-grained approach. It can be represented by syntactic and semantic information by annotating entities. Provenance data can be disseminated through visual graphs, queries, and services (providers). The work argues about provenance modeling for tagging data and processes using syntactic and numerical meta-data. We believe that provenance models must be close to the knowledge domain of data and processes. Therefore, the perimeter of our proposal is defined by seismologic data and provenance meta-data.

This work addresses the questions and open issues associated with the provenance of data-driven geophysics and seismology scientific workflows —our work models provenance regarding the data, activities, and agents participating in the workflow. We focus on modeling the meta-data used to deal with these components' provenance to curate seismic data and guide their exploration.

3 Motivation use case

Let us now present a motivational example. We focus on the data produced by a set of seismographs distributed across a large geographic region. Our data originated from seismographs in the Brazilian northeast in our case. The data produced by the seismographs are used to generate a weekly bulletin identifying seismic activity during a period of interest and create a database of seismic data to be used by researchers and students. The database should be tagged with provenance metadata.

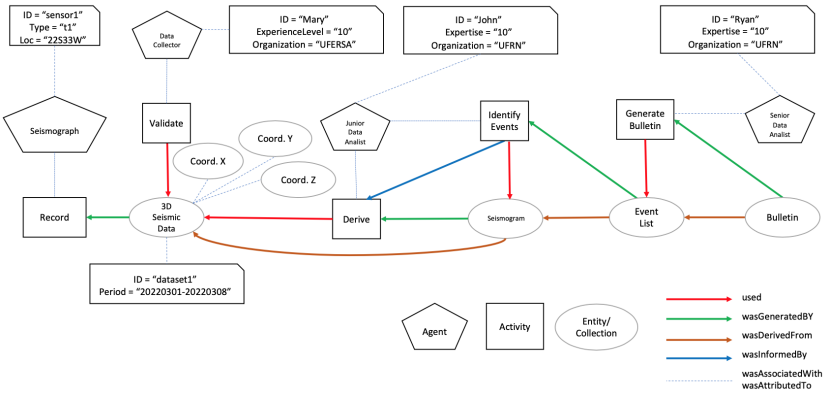

Fig. 1. Provenance Graph for Seismic Bulletins.

Our provenance graph describes activities and the entities they produce. Several agents may affect this process. The provenance model introduces types and their relationships. Figure 1 shows a provenance graph for our scenario. The agents in charge of activities in this example are:

Seismographs: Devices that produce raw data.
Data Collector: Human agent, responsible for retrieving and validating the data from seismographs.
Junior Data Analyst: Human agent that identifies seismic events from the collected data. She produces a list of the seismic events in the bulletin.
Senior Data Analyst: Human-agent, responsible for checking the list of events produced by the Junior Analyst and generating the bulletin to be published.

Entities in our context correspond to data items and data collections. They are depicted as ovals in Figure 1. In our example, we have the following:

3D Seismic Data: This collection comprises files that record seismic movements for a given period and location. One file for each dimension (North-South, East-West, and Up-Down). Each file contains a list of pairs formed by a timestamp and an integer value. Typical sampling rates range between 50 and 4000 records per second.
Seismogram: Corresponds to graphical representations of ground movements. These graphics depict the data acquired by seismographs of a given station, being analysed by a Junior Data Analyst to identify geological events, like earthquake and mining explosions, and their magnitudes (see Figure 2).

Figure 2 shows two vertical lines marked as P wave (in red) and S wave (in green). The analyst creates these lines to indicate, respectively, the arrival to the sensor of the *primary* and *secondary* waves of a seismic event. Both waves are created at the same time by the seism. The primary wave travels through any media and arrives faster at the sensor. The secondary wave travels only through

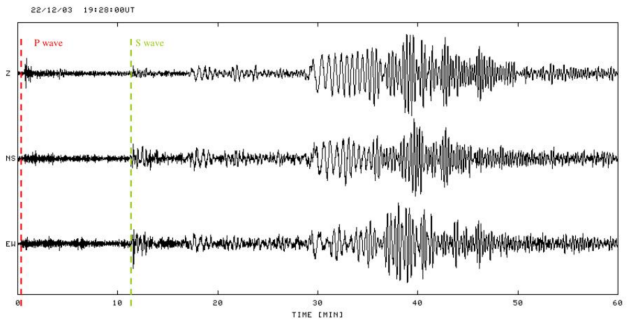

Fig. 2. Example of Seismogram. Source: `https://image3.slideserve.com/6627149/seismogram-example-l.jpg`.

solid media and arrives later. The seismic analyst is in charge of identifying the seismic event by identifying the moments of arrival o these waves and the end of the event.

Event List: This is the result of the analysis of the Junior Data Analyst, reporting the events identified on Seismograms. The event list also includes the location of the seism epicenter, calculated by triangulation from the data registered by several seismographs.

Bulletin: The bulletin is the final document reporting the detected events and their properties, including locations and magnitudes, along a given period. This document needs to be certified by a Senior Data Analysts before its release to the general public.

So far, the data has been used to generate a bulletin from seismic activity. However, other problems can be solved using the collected data by seismographs. We describe an example in the following lines.

Using one sensor data to locate the epicenter of a seism. This process uses the data collected by just one seismograph. The direction from where the seism originated is calculated by considering the initial movement detected by the hardware on each dimension (north-sud, east-west, and up-down). The distance from the epicenter to the sensor is given by the difference (in time) among the waves P and S, and the soil class around the seismograph. Notice that the P and S waves originated simultaneously at the epicenter. The P wave is faster than the S wave.

4 Exploring geophysics data guided by provenance

Geophysics data exploration is an analytics process that seeks to process data collection content (produced by metrology devices) for answering factual and analytics queries. Since the exploration is intended to support the manual or

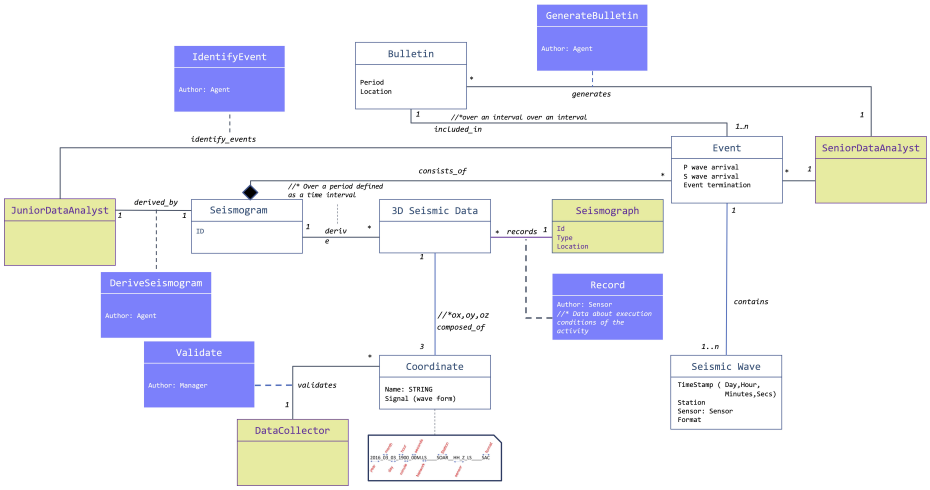

Fig. 3. Provenance based seismic events data

semi-automatic identification of geophysics events, it is critical to produce results that answer queries, but that report the provenance of identified events and the conditions in which data and events are observed, detected, validated, and disseminated.

The first challenge is identifying and modeling the seismic and provenance concepts. We adopt a database-oriented approach for building a database of geophysical data and metadata with the following steps: (1) Designing the data and meta-data according to a real case described in the previous section (UML class diagram); (2) Transform the design into seismic and provenance property graphs; (3) Identifying exploration query types that can be asked on top of the seismic and provenance graphs.

Seismic and provenance data design Figure 3 illustrates the UML class diagram that combines the concepts of the seismic data (in white) and the associated provenance concepts (in violet and green). The diagram models the concepts representing seismic events detection, including the type of agents and actions that produce, validate, and process the data. These entities represent the meta-data used for tagging the seismic data with provenance.

Regarding data, in the central concept is an 3D Seismic Wave, that is built by processing 3 Coordinates (x,y,z). A Seismogram is derived with a set of 3D Seismic Waves. It contains the information an analyst uses to identify seismic events and build an Events list. An Event refers to a Seism or other earth movement. Each event has a duration and contains timestamps for its P and S waves and Termination. Events produced in an interval are reported in a Bulletin.

Provenance meta-data are of three types agents Data Collector, Junior Data Scientist and Senior Data Scientist (green), and Actions (violet). In UML, verbs (i.e., actions) are modelled by relations between classes, in consequence, as shown in

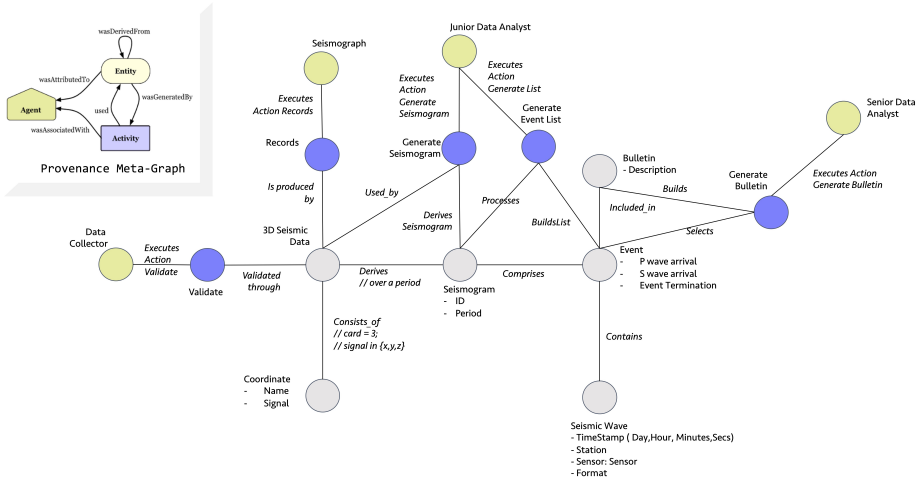

Fig. 4. Provenance and seismic graphs.

the diagram, Action's are associated to relations and Agents's are associated with (perform) actions. For example, a Seismogram is derived from a 3D Seismic Data by an agent of type Junior Scientist.

Seismic and provenance property graphs Once we have modeled the concepts representing the data and meta-data of seismic events detection and dissemination, we assume that graphs are the most adapted data model for storing these geophysics (meta)-data. Graphs can be then explored with factual and analytics queries that can produce results and associated explanations with provenance. We adopt a properties graph model to model seismic data and provenance meta-data for a first approach. Then it is possible to identify the type of queries that can be asked and evaluated on top of this type of graph.

Figure 4 shows a sample of the property graph schema of two interconnected graphs representing seismic data and provenance meta-data. The principle of our design is to separate data and meta-data to open the possibility of associating different meta-data "spaces" with the same data.

Notice that an alternative design strategy would be to add properties to the relations and nodes of the data graph. This second strategy is closer to relational provenance approaches, where attributes are added to the relational schema to tag tuples with provenance meta-data. The seismic data graph (nodes in grey in Figure 4) is defined as follows:

```
consists_of(3DSeismicData, Coordinate)
derives(Seismogram, 3DSeismicData)
comprises(Seismogram, Event)
included_in(Event, Bulletin)
contains(SeismicWave, Event)
```

The provenance meta-data graph represents five action types performed on data during the event detection process (done manually by geophysics techni-

cians or scientists). The provenance meta-data graph (nodes in green and violet in the figure) is connected with the seismic graph, and it is defined as follows:

```
Builds(GenerateBulletin, Bulletin)
Selects(GenerateBulletin, Event)
ExecutesActionGenerateBulletin(SeniorDataScientist, GenerateBulletin)

BuildsList(GenerateEventList, Event)
Processes(GenerateEventList, Sesimogram)
ExecutesActionGenerateList(JuniorDataAnalyst, GenerateEventList)

ExecutesActionGenerateSeismogram(JuniorDataAnalyst, GenerateSeismogram)
ExecutesActionRecords(Seismogram,Records)
ExecutesActionValidate(DataCollector,Validate)

DerivesSeismogram(GenerateSeismogram, Seismogram)
UsedBy(GenerateSeismogram, 3DSeismicDiagram)
ValidatedThrough(Validate, 3DSeismicDiagram)
```

It is through actions that the graph is connected with the seismic graph.

Expressing provenance based queries for exploring seismic graphs The event detection and analysis process within the datasets collected by seismic stations is exploratory in geophysics. Collected data are periodically downloaded from stations and archived so geophysicists can explore signals and detect events produced in specific intervals and locations. Analytics results are plots, event histories, and bulletins. The manual and meticulous process performed on independent data batches makes it challenging to correlate the observations on different data collections and prevents discovering hidden knowledge. Since the detection is manual, it is essential to know who processes data, detects events, and produces bulletins. In general, examples of exploration queries are:

- Graph traversal, for example, *Which seismograms were used to detect the events produced in Natal reported in the bulletin in January? How many data collectors are located in Natal, and which junior analysts processed their collected coordinates to detect events?*
- Graph analytics can be used to answer the following types of queries:

  Community detection: *Which locations have the highest number of detected events? Who are the junior scientists participating in that detection? Which data collectors produced the noisiest readings?*
  Centrality: *Which are the seismograms with 3D seismic data where junior scientists detected the most number of events?*
  Similarity: *Which events have similar intensity and are located in the same region in subsequent intervals produced by the same data collector?*
  Heuristic link prediction: *Are events reported in the march bulletin in a particular region related to those detected in a close area by other data collectors?*
  Pathfinding and Search: *Who are the junior data scientists that have detected events from waves sensed by data collectors in the same region?*

Towards a curation and exploration environment for seismic graphs with provenance Figure 5 illustrates the functional architecture of a possible environment that can provide tools to curate and explore seismic data. The environment must provide services to curate data and processes performed to process it, like extracting content, generating plots, observing and detecting events at intervals, and producing bulletins. From a curation perspective, the environment must archive, process data to extract content, allow agents to explore them, and keeps track of their actions. For example, who performs which action on which data? The system must coordinate a cyclic process that collects, archives, processes, and produces new data and provenance meta-data. Since some curation tasks cannot be completely automatic, junior scientists execute seismogram production and event detection; the environment must provide simple, well-adapted tools with well-adapted human-in-the-loop strategies. The environment is a li-

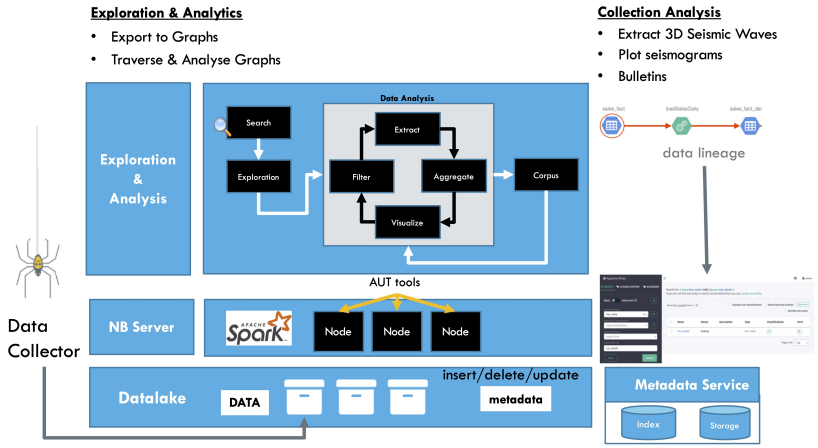

Fig. 5. Curation and exploration environment for seismic graphs with provenance

brary with seismic, experiments data, and meta-data that can be continuously curated (with new metadata) and where exploration and analytics pipelines can be performed. Thanks to the choice of graphs, exploration can lead to discovering patterns that can improve the knowledge about seismicity in different regions and its implications. We believe that the environment must rely on a domain-specific exploration queries engine. Thereby, queries can be expressed by geophysicists and then evaluated to produce results explained with provenance meta-data.

Challenges and Open Issues Associating provenance meta-data to seismic data corresponds to a **curation process**. It focuses on activities and agents that act on data during the event detection process. The curation process keeps track of the conditions in which seismic events are detected and disseminated in bulletins. **First challenge:** The design of the graphs is the first challenge to address. Therefore, we consider two types of open issues. First, formally expressing the

graphs to have a solid and sound representation. Then profile and study their properties to estimate storage, querying, and processing implications.

Second challenge: Junior (senior) data scientists do the curation process manually, and some aspects like the **quality of data and meta-data** deserve to be modeled and assessed.

Third challenge: Storing data and meta-data in a graph database enables the **exploration** of raw and processed data reported in plots and bulletins. In seismology, this feature is essential because it can establish connections across the data reported within bulletins. Storing and maintaining bulletins can build a history and perform analytics to **understand and discover seismic patterns**. Identifying and characterizing the exploration process and queries ad hoc for seismology is an open issue. Geophysicists have not formally expressed how they operate on data to solve questions and study different phenomena and implications. As questions are characterized, it will be possible to define exploration and processing operators and how provenance can contribute to implementing specialized exploration processes.

Fourth challenge: The expression of exploration queries can combine path traversal, aggregation, and analytics operations. Query languages like Cypher enable the expression path traversal queries for property graphs. Extensions with data science cartridges like the one proposed by Neo4J allow to "declaratively" define pipelines that can apply machine learning models on graphs. However, we believe that for the seismic and, in general, the geophysics sciences with particular exploration requirements, it can be interesting to specify domain-specific query languages that integrate models and operations used in the discipline.

5 Conclusion and Future Work

This paper introduced problems, challenges, and open issues regarding the provenance-guided curation and exploration of geophysics data collections. We motivated the problem through a use case regarding seismic data collected by observation stations in the northeast region of Brazil. We exhibited the requirements regarding curation and exploration and highlighted the importance of provenance to ensure seismic events detection, validation, archival, and dissemination. Graphs can provide an intuitive, rich and mathematical way of curating and exploring data. The curation and exploration processes must be specialized for seismic data and analysis. We showed the general architecture of an environment that can implement our approach. As future work, we have to improve the provenance graph in order to broaden the scope, as well as include more details. Also, we count on proposing and implementing a Domain Specific Language for the curation and exploration of geophysical data sets. Other open issues like dealing with data volume, velocity, and variety will be experimented with in the context of the project ADAGEO (`https://adageo.github.io/`).

References

1. Bowers, S., McPhillips, T.M., Ludäscher, B.: Provenance in collection-oriented scientific workflows. Concurrency and Computation: Practice and Experience **20**(5), 519–529 (2008)
2. Cheney, J., Chiticariu, L., Tan, W.c.: Provenance in databases: Why, how, and where. Foundations and Trends in Databases **1**, 379–474 (01 2009). https://doi.org/10.1561/1900000006
3. Hernández, D., Galárraga, L., Hose, K.: Computing how-provenance for sparql queries via query rewriting. Proceedings of the VLDB Endowment **14**(13), 3389–3401 (2021)
4. Missier, P., Dey, S., Belhajjame, K., Cuevas-Vicenttin, V., Ludäscher, B.: D-PROV: Extending the PROV provenance model with Workflow structure. In: 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP 13). USENIX Association, Lombard, IL (Apr 2013), `https://www.usenix.org/conference/tapp13/technical-sessions/presentation/missier`
5. Moreau, L., Missier, P., Belhajjame, K., Far, R.B., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.: Prov-dm: The prov data model (2013), `http://www.w3.org/TR/prov-dm/`, world Wide Web Consortium (W3C)
6. Neo4j: Neo4j - The World's Leading Graph Database (2012), `http://neo4j.org/`
7. Oliveira, W., Oliveira, D.D., Braganholo, V.: Provenance analytics for workflow-based computational experiments: A survey. ACM Computing Surveys (CSUR) **51**(3), 1–25 (2018)
8. Paul Groth, L.M.: Prov-overview (2013), `https://www.w3.org/TR/prov-overview/`, world Wide Web Consortium (W3C)
9. Ramusat, Y., Maniu, S., Senellart, P.: A practical dynamic programming approach to datalog provenance computation. arXiv preprint arXiv:2112.01132 (2021)
10. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. ACM Sigmod Record **34**(3), 31–36 (2005)
11. Spinuso, A., Cheney, J., Atkinson, M.: Provenance for seismological processing pipelines in a distributed streaming workflow. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops. pp. 307–312 (2013)
12. W3C: SPARQL 1.1 query language (2012), `https://www.w3.org/TR/2012/PR-sparql11-query-20121108/`
13. Wang, J., Crawl, D., Purawat, S., Nguyen, M., Altintas, I.: Big data provenance: Challenges, state of the art and opportunities. In: 2015 IEEE International Conference on Big Data (Big Data). pp. 2509–2516 (2015). https://doi.org/10.1109/BigData.2015.7364047