Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute
InformingScience.org

IJIKM.org

Volume 17, 2022

Automatic Generation of Temporal Data Provenance From Biodiversity Information Systems

Zaenal Akbar*	National Research and Innovation Agency (BRIN), Bogor, Indonesia	zaen005@brin.go.id
Dadan R. Saleh	National Research and Innovation Agency (BRIN), Bogor, Indonesia	dada023@brin.go.id
Yulia Aris Kartika	National Research and Innovation Agency (BRIN), Bogor, Indonesia	yuli027@brin.go.id
Widya Fatriasari	National Research and Innovation Agency (BRIN), Bogor, Indonesia	widy003@brin.go.id
Adilla A. Krisnadhi	Universitas Indonesia, Depok, Indonesia	adila@cs.ui.ac.id
Deded Sarip Nawawi	IPB University, Bogor, Indonesia	dsnawawi@apps.ipb.ac.id

* Corresponding author

Abstract

Aim/Purpose: Although the significance of data provenance has been recognized in a variety of sectors, there is currently no standardized technique or approach for gathering data provenance. The present automated technique mostly employs workflow-based strategies. Unfortunately, the majority of current information systems do not embrace the strategy, particularly biodiversity information systems in which data is acquired by a variety of persons using a wide range of equipment, tools, and protocols.

Background: This article presents an automated technique for producing temporal data provenance that is independent of biodiversity information systems. The approach is dependent on the changes in contextual information of data items. By mapping the modifications to a schema, a standardized representation of data provenance may be created. Consequently, temporal information may be automatically inferred.

Accepting Editor Geoffrey Z. Liu │ Received: April 3, 2022│ Revised: June 18, July 17, 2022 │
Accepted: July 20, 2022.

Cite as: Akbar, Z., Saleh, D. R., Kartika, Y. A., Fatriasari, W., Krisnadhi, A. A., & Nawawi, D. S. (2022). Automatic generation of temporal data provenance from biodiversity information systems. *Interdisciplinary Journal of Information, Knowledge, and Management, 17,* 361-385. https://doi.org/10.28945/5003

(CC BY-NC 4.0) This article is licensed to you under a Creative Commons Attribution-NonCommercial 4.0 International License. When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

Methodology
The research methodology consists of three main activities: database event detection, event-schema mapping, and temporal information inference. First, a list of events will be detected from databases. After that, the detected events will be mapped to an ontology, so a common representation of data provenance will be obtained. Based on the derived data provenance, rule-based reasoning will be automatically used to infer temporal information. Consequently, a temporal provenance will be produced.

Contribution
This paper provides a new method for generating data provenance automatically without interfering with the existing biodiversity information system. In addition to this, it does not mandate that any information system adheres to any particular form. Ontology and the rule-based system as the core components of the solution have been confirmed to be highly valuable in biodiversity science.

Findings
Detaching the solution from any biodiversity information system provides scalability in the implementation. Based on the evaluation of a typical biodiversity information system for species traits of plants, a high number of temporal information can be generated to the highest degree possible. Using rules to encode different types of knowledge provides high flexibility to generate temporal information, enabling different temporal-based analyses and reasoning.

Recommendations for Practitioners
The strategy is based on the contextual information of data items, yet most information systems simply save the most recent ones. As a result, in order for the solution to function properly, database snapshots must be stored on a frequent basis. Furthermore, a more practical technique for recording changes in contextual information would be preferable.

Recommendations for Researchers
The capability to uniformly represent events using a schema has paved the way for automatic inference of temporal information. Therefore, a richer representation of temporal information should be investigated further. Also, this work demonstrates that rule-based inference provides flexibility to encode different types of knowledge from experts. Consequently, a variety of temporal-based data analyses and reasoning can be performed. Therefore, it will be better to investigate multiple domain-oriented knowledge using the solution.

Impact on Society
Using a typical information system to store and manage biodiversity data has not prohibited us from generating data provenance. Since there is no restriction on the type of information system, our solution has a high potential to be widely adopted.

Future Research
The data analysis of this work was limited to species traits data. However, there are other types of biodiversity data, including genetic composition, species population, and community composition. In the future, this work will be expanded to cover all those types of biodiversity data. The ultimate goal is to have a standard methodology or strategy for collecting provenance from any biodiversity data regardless of how the data was stored or managed.

Keywords
temporal data provenance, biodiversity, ontology, rule-based reasoning

INTRODUCTION

In many Big Data applications, data quality has become a major challenge. Such applications are challenged by data conflict, incomplete, imprecise, subjective, redundant, biased, or noisy, which could lead to confusion or misinformation that will reduce the quality of produced insights (Wilkinson et al., 2016). As the volume of data is expanding with a promoting diversity of data types, the need for a way to preserve data quality is accelerating (Pérez et al., 2018). Data reusability, on the other hand, is crucial because it allows data to be identified and reused in subsequent investigations (Wilkinson et al., 2016). Applying new analytic methods to archived data, alone or in combination with newly collected data, is another leading practice for knowledge discovery and innovation.

Collecting and evaluating data provenance is an effective way to preserve data quality and enable data reusability. Data provenance can be seen as information describing digital data's production process (Herschel et al., 2017). This data includes meta-data on the entities, processes, and people who worked on the project. In general, data provenance allows the identification of a particular data source, including analyzing the executed transformation of the data (Pérez et al., 2018; Stefanowski et al., 2017). By analyzing this information, multiple useful knowledge can be extracted that can be used for understandability (e.g., to identify how the data was obtained), reproducibility (e.g., to check if a prior result can be confirmed), and quality (e.g., to reveal some quality issues in the data) (Herschel et al., 2017). More than that, provenance also can be applied for security/privacy (e.g., to protect data against unauthorized access and to ensure data integrity), verification (e.g., to verify the trustiness of the data production process), and repeatability (e.g., to enable to repeat the study) (Pérez et al., 2018).

In biodiversity science, scientists have discovered and documented the world's biodiversity, typically in the form of digital collections or specimens. To provide an integrative analysis, the collections that are available from different sources need to be harmonized and coordinated regarding structure, format, and annotation (Sansone et al., 2012). To appreciate the diversity of life and the conditions in which it exists on the planet in a reliable and comprehensive manner, one must make appropriate use of a mix of such collections (Lannom et al., 2020). This circumstance poses certain issues for research infrastructures and data services in terms of making data discoverable, searchable, interoperable, and reusable, as well as doing so in a manner that makes use of the help of machines for greater productivity (Weigel et al., 2020). In an effort to standardize and unify biodiversity data, the Essential Biodiversity Variables (EBV) were established to capture a minimal number of criterion variables (Kissling, Ahumada, et al., 2018). Furthermore, the EBV manufacturing process should be traceable from the result back to the raw data, with the ability to repeat the process and document the data's origin and what has been done with it (Hardisty et al., 2019).

The concept of time as an extra dimension of data provenance is reflected in the concept of temporal provenance, which is one sort of provenance. Temporal data provenance depicts the evolution of data provenance over time. It refers to the encoded temporal information in a data provenance model, representing time, intervals, or versioning (Beheshti et al., 2012). This temporal representation is crucial for representing the development of a piece of data with several states through time. More than that, temporal representation of data can be used further for data mining tasks (P. Chen et al., 2014), for example, pattern generation, finding variants, or discovering more descriptive knowledge of provenance clusters. In this case, logical time can be used to reduce the feature space of the provenance such that data mining tasks can be performed effectively. This kind of time-related data is crucial for improving search results and the overall information retrieval experience for the end user (Alonso et al., 2007). Furthermore, temporal information provides the capability for temporal reasoning, for example, for activity recognition (Zhang et al., 2020), learning non-linearly evolving entity representations over time (Trivedi et al., 2017), and visual question answering generation (Jang et al., 2017).

CHALLENGES AND MOTIVATION

The source of biodiversity data is typically divided into three categories: genetic diversity, species diversity, and ecosystem diversity. As a basis of biodiversity monitoring programs worldwide, the EBV introduces six commonalities classes: genetic composition, species population, species traits, community composition, ecosystem structure, and ecosystem function (Pereira et al., 2013). Due to its broad scope, this work is limited to species traits data. Generally, trait data collection will start with raw traits' measurements through specimen collections, in situ monitoring, or remote sensing (Kissling, Walls et al., 2018). After that, the data will be validated through data cleaning and quality control before being published as trait datasets. The aggregation of data from numerous sources is fraught with difficulties, such as broad heterogeneity in collection and sampling methods, the lack of individual or population level measurements, systematic and temporally contiguous in situ collections, and so forth.

Based on the issues in biodiversity data management stated above, the mechanism for automatically collecting such provenances is difficult. It will be hard to physically gather them for several reasons, including:

1) There is a variety of methods and tools for data collection. Each scientist would use the best method and tool available to him/her.
2) Data collection can be performed by multiple scientists from different groups or institutions. Collected data can be stored in separate databases in various formats. Forcing scientists to use similar structures of the database is not practical.
3) In most cases, temporal information about data collection or processing is not attached to the data itself. Instead, this information resides in the data storage or management system, such as a Biodiversity Information System (BIS).

With these constraints in mind, data provenance may be gathered in a decentralized way. A credible reference to digital biodiversity data enables decentralized archiving and data sharing, allowing for long-term data accessibility (Elliott et al., 2020). In this regard, when integrating data from multiple BISs, the overall quality of the integrated data will be determined by data provenance collected from every BIS.

Research Scope

This work introduces a solution to automatically collect temporal data provenance from distributed BISs. The solution can be integrated into databases maintained by a generic BIS. It combines three methods: event detection from databases, a uniform representation of events through an ontology, and automatic temporal information inference through a rule-based inference engine. Since collecting provenance remains lacking a standard model or a roadmap (Sarikhani & Wendelborn, 2018), the scope of this work is limited to the generation of temporal data provenance from generic databases maintained by BISs. The main research question is to what extent the database events and their uniform representation contribute to the automatic generation of temporal provenance.

The rest of the paper is organized as follows: a few related works will be listed and discussed in the Literature Review section, followed by a detailed explanation of the proposed solution in the Computational Approach section and how it was implemented in the Implementation section. After describing the performed experiment in the section, the Results and Findings will be discussed. Finally, a few conclusions and limitations, as well as future works of this study, will be clarified in the Conclusions section.

Literature Review

This section discusses the two most relevant automatic data provenance capturing mechanisms: workflow-based and templated-based. After listing and discussing related works from each mechanism, the research gaps will be outlined and discussed at the end of this section.

Workflow-Based Provenance Capture

In general, there are three sorts of provenance capture mechanisms: workflow-based, process-based, and operating system-based (De Meester et al., 2017). The workflow-based capture mechanisms are widely used because they provide a simple programming model that allows a sequence of tasks to be composed by connecting one task's outputs to the inputs. In this case, the provenance can be generated automatically based on functions or processes performed on a fraction of the data.

The workflow-based capture mechanisms are highly achievable in specific architecture and controlled environments, where automatic capturing can be performed in the background of a workflow system (Weigel et al., 2020). Capturing provenance from such systems requires a model of provenance (Sarikhani & Wendelborn, 2018). The model should cover several characteristics of provenance capture techniques such as workflow orientation, collecting stages, degree of abstraction, retrospective and prospective, granularity, accessibility, architectural layers, coupling strategy, and time and kind of instrumentations. As an example, data transformation workflows that are based on MapReduce (a distributed programming model for processing large data sets by applying maps and reducing procedures in parallel) tasks can be used as a platform for provenance capture in this specific architecture (Stefanowski et al., 2017; Wang et al., 2015).

Computational experiment workflows are another setting that may be utilized as a platform for provenance capture. With the advent of scientific workflow management systems, provenance can be automatically captured and stored during data production and consumption within a given scientific experiment (Oliveira et al., 2018). For example, different components for efficient data processing tied together in a computational bioinformatics workflow can be used to track multiple transformations performed on the data (Kanwal et al., 2017).

Furthermore, workflow-based provenance capture mechanisms have also been developed based on a specific computing platform; for example, Apache Spark (Guedes et al., 2020; Rajmohan et al., 2019), interactive notebooks (Carvalho et al., 2017), and a specific workflow management system for particle-based simulations (Horsch et al., 2020).

Template-Based Provenance Capture

A model-driven service interface, so-called provenance templates, can be utilized to automatically capture data provenance. Provenance templates may be seen as abstractions with domain meaning that can be readily translated to the activities of client software tools (Curcin et al., 2017). Practically, a template will be represented as a graph that consists of specific entities (as a representation of states of the data), specific activities that produce and consume such entities, and particular agents associated in some capacity with entities or activities. Every time new data is created, the relevant template will be consulted to ensure its validity. Implementing a decision support system in the health domain has shown potential for integrating trust into computerized systems, enabling transparency and auditability (Curcin et al., 2017). In software engineering, a template-based provenance capture mechanism can also be implemented by mapping the structural diagrams of designed applications into provenance templates (Sáenz-Adán et al., 2018). Furthermore, a service for uploading and disseminating provenance templates has been established in the field of environmental and earth sciences, which can be used to build uniform provenance traces from input data in accordance with standards (Magagna et al., 2020).

Research Gap

Provenance meta-data, information system provenance, workflow provenance, and data provenance are a few examples of provenances that may be collected (Herschel et al., 2017). Data provenance is the most precise sort of provenance, and it pertains to the particular data pieces and the actions they go through. Most of the existing provenance capture mechanisms discussed in the previous sub-sections rely mainly on the functions or processes performed on a fraction of data. However, functions and processes can vary from one case to another; therefore, having one generic architecture to facilitate an automatic generation of provenance will be challenging. Furthermore, most of those functions and processes are performed on a collection of data, not at the individual level of item data. As a consequence, workflow-based and template-based mechanisms produce coarse-grained data provenance granularity.

COMPUTATIONAL APPROACH

In this section, the computational strategy for automatically producing the temporal provenance of biodiversity data is presented. First, the procedural architecture of this study is outlined, followed by the three primary actions that contributed to the proposed solution.

PROCEDURAL ARCHITECTURE

Figure 1 shows the procedural architecture of this work to transform biodiversity databases as input to temporal provenance as output. It consists of three main activities: database event detection, event-schema mapping, and temporal information inference. First, a list of events will be detected from the databases, typically maintained by biodiversity information systems. Second, the detected events will be mapped to an ontology, so a common representation of data provenance will be obtained. Third, based on the obtained data provenance, rule-based reasoning will automatically be applied to infer temporal information. As a result, a temporal provenance will be generated.

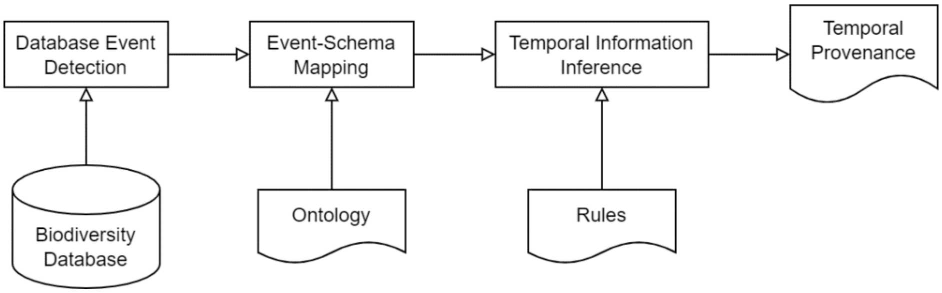


Figure 1. The procedural architecture

As a combination of three methods, the solution takes advantage of each method from event detection methods that have been successfully applied in various applications. An event is a noteworthy, out-of-the-ordinary occurrence of action compared to typical patterns of conduct (Kerman et al., 2009). In this situation, examining the system's statuses may help with event detection. Statistical, probabilistic, artificial intelligence, machine learning, and hybrid approaches are some of the event detection methods. The value of specified parameters that exceed a threshold value, which may be established based on past parameter values, is monitored using a statistical approach. The likelihood of an event occurring and other relevant factors are computed using a probabilistic technique. On the other hand, AI and machine learning methods work by modeling the system based on data training. Furthermore, it is also possible to combine those techniques.

Further, a uniform representation of data through an ontology – shared, explicit and formal conceptualizations of a domain (Gruber, 1995) – has been widely used to guarantee consistency among multiple systems. Multiple systems are able to interact with one another regarding a domain of discourse when they use a common ontology rather than having to necessarily operate on a universally shared theory. Furthermore, data inference through rules would automatically generate temporal information. As a kind of knowledge representation, rules may take several forms, including reactive rules, which include the invocation of actions in response to events and actionable circumstances (Paschke & Kozlenkov, 2009). Production rules, for example, typically represented in the IF Condition THEN Action format can derive new data as action whenever a particular condition is fulfilled.

The proposed solution combines the advantages of each method as follows:

1) Multiple sources of events can be uniformly represented across multiple systems by using a common ontology.

2) Multiple mapping can be performed without changing the ontology by using a mapping mechanism. In this case, the solution should be scalable very well. This benefits biodiversity data collection where methods, tools, and databases vary.
3) A piece of new information can be inferred automatically based on the current situation using reactive rules. Therefore, generating temporal provenance can be performed automatically.
4) A combination of three methods would automatically enable provenance collection mechanisms.

In the next subsections, the three primary phases of the solution will be discussed in depth.

DATABASE EVENT DETECTION

There are two essential components of provenance (Freire et al., 2008):

1) The process description (or series of procedures) that, in conjunction with input data and parameters, led to the generation of the data product.
2) User-defined documentation of information. It is not automatically collected, but key choices and notes are recorded. This information is often presented as annotations.

Typically, the sequences of lines of the data generation process are described according to where the data were copied from (e.g., what characteristics of which tuples), why the data was created (based on what source data), and how the data were changed (to produce the result) (Herschel et al., 2017). The proposed event detection method relies on data items available in transactional databases.

Definition 3.1 (Data Items). A data item is the smallest representation of data, represented as triple (k, v, l), where k is key, v for value, and l for the label.

Definition 3.2 (Contextual Information). Contextual information is the identity of things that gives context to data. For example, an identity about the person who performs certain actions to the data or an identity about when the data was created or modified.

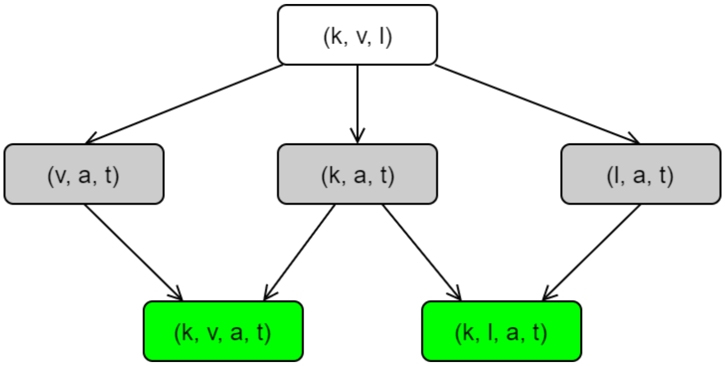


Figure 2. Combine data and contextual information

The suggested method focused on biodiversity information systems, in which various linked data points for producing provenance were removed from the data. For instance, contextual information on the individual who modified the data and the time of the update was maintained in another database. In order for this technique of generic data storage to function, the data and contextual information are integrated, with the mechanism for event detection relying on this combination. Figure 2 depicts how data items are combined with contextual information for authorship (a) and for temporal (t). Every member of a data item triple (the white box) can be combined with this two contextual information, such that key, value, and label are contextually represented as (k, a,t), (v, a,t), (l, a,t) respectively (gray boxes). Further, this three contextual information can be combined further to generate another two contextual information (green boxes) that are represented as (k,v, a,t), (k,l, a,t).

To illustrate the combination process, let's take, for example, an item data as follow: *isLeaf(x)* ∧ *hasColor(x, green)*. Adding an authorship context to the data would produce something like *isLeaf(x)* ∧ *hasColor(x, green)* ∧ *hasAuthor(A)* to indicate that authorA declared that the color of the leaf was green. When another author amends the fact, we will have another fact to indicate what parts of the data were changed and who was responsible for the change. For instance, *isLeaf(x)* ∧ *hasColor(x, yellow)* ∧ *hasAuthor(B)* to indicate that author B has declared that the color of the leaf is yellow. To this point, we can distinguish the data modification and the person responsible for the modification. However, we still miss which one is declared first, second, and so on. Therefore, temporal contextual information would enable us to do that. In this case, adding the temporal as follow: *isLeaf(x)* ∧ *hasColor(x, yellow)* ∧ *hasAuthor(B)* ∧ *hasTime(C)* to indicate that the fact was declared by author B on time C.

As stated in the Introduction section, the target of this study is generic biodiversity information systems. Minimal contextual information is present in such systems. In most instances, the system just maintains the most recent authorship and temporal context information. Consequently, after integrating the data and its contextual information, as previously described, the event detection approach will determine whether any contextual information has changed. If a change is detected, it will further detect whether the authorship, temporal, or both contexts were changed. Based on this detection, a set of relevant attributes (key-value pairs) will be generated to be mapped further with the common ontology. The key-value pairs can be stored in any format (for example, XML, CSV, JSON), which the mapping engine consumes. These key-value pairs bridge the databases and the mapping where only necessary events will be detected. It is not required to map the database directly to a schema because the solution needs to identify different kinds of activities. Instead, the identification will be performed by the event detection process.

Event-Schema Mapping

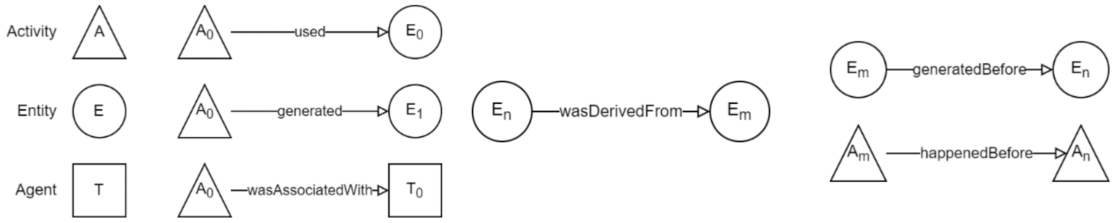
A schema mapping process aims to align an item from one schema to a relevant item in another schema. In general, the objective of mapping is to align data from numerous schemas to a common schema in order to produce a coherent representation of the data. The alignment is typically defined through mapping rules that map item data and required data transformation.

There are several widely used common representations for data provenance, including the Open Provenance Model (OPM) (Moreau et al., 2011) and PROV-O (https://www.w3.org/TR/prov-o/). The OPM characterizes what caused "things" through nodes of a directed graph. Three kinds of nodes are Artifact, Process, and Agent. An artifact represents an immutable piece of the state of a digital representation. A process represents an action or series of actions performed on or caused by artifacts. An agent represents the entity that performs a process. Since 2013, the W3C has proposed PROV-O as another standard format for data provenance. It presents three fundamental classes: Entity, Activity, and Agent. A physical, digital, or other item is an entity. An activity is anything that takes place throughout time and has an effect on entities. An agent is anything that bears responsibility for the occurrence of an action.

In this study, PROV-O is employed as the standard data provenance form for biodiversity data. In this situation, an action may include consuming, processing, converting, altering, transporting, using, or producing biological specimen entities. As noted in the Introduction section, the scattered nature of data collecting posed challenges. Multiple individuals are producing various measurements from the same specimen. An individual measures and collects different characteristics from the specimens using different measurements or types of equipment or protocols. It is the responsibility of every individual to enter obtained data into a database. Representing these activities using PROV-O cannot be performed straightforwardly for several reasons. First, multiple individuals are working on the same specimen; therefore, it is necessary to ensure the entity of those multiple activities refers to the same object. Next, when entering data into the database, the order of data is unknown since each individual could enter data at any time. The solution overcomes these situations with a flexible schema mapping approach through mapping rules in the form of RML (Dimou et al., 2014).

Temporal Information Inference

In this study, the automatic generation of temporal information is implemented using rule-based reasoning. Rules are a kind of knowledge representation that may express deductive information, such as logical connections, and so facilitate inference (Chowdhary, 2020). Due to its simplicity in codifying the knowledge of human experts, rule-based reasoning systems have been widely used in various knowledge-intensive expert systems. For example, a rule-based system has been used for legal reasoning (Liu et al., 2021), safety assessment (Tang et al., 2020), emergency management (Jain et al., 2021), and online communication (Akbar et al., 2014). Specifically, in the biodiversity research area, rule-based systems are also widely used, for example, for predicting the impact of land-use changes on biodiversity (Scolozzi & Geneletti, 2011), molecular biodiversity database management (Pannarale et al., 2012), or for generating linked biodiversity data (Akbar et al., 2020).



(a) Asserted relationships

(b) Inferred causality relationship

(c) Inferred temporal relationships

Figure 3. Asserted and inferred relationships

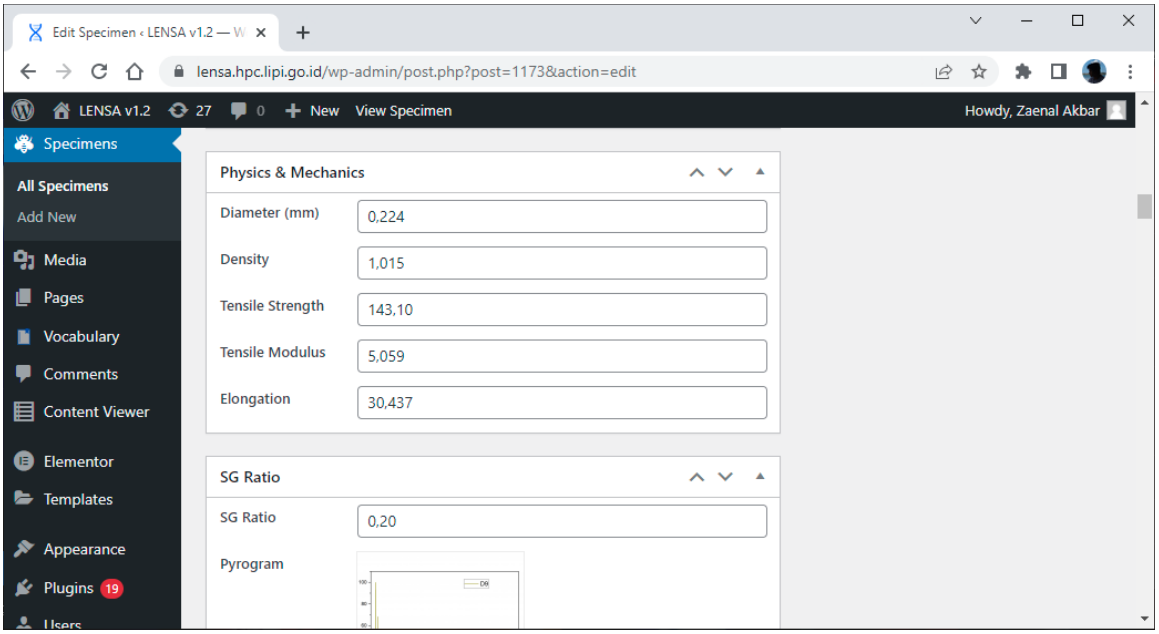
Figure 3 shows the capability of the schema to infer causality and temporal relationships. Triangles, circles, and squares represent activities, entities, and agents. By using facts that are explicitly defined (as shown in Figure 3a), it will infer more information based on causality (Figure 3b) and time (Figure 3c) relationships. For causality relationships, it will be able to define if an entity was derived from another entity and if it was defined by using a specific activity. Further, for time relationships, it will be possible to determine if an activity happened before another activity and if an entity was generated before another entity.

This work adopts the Semantic Web Rule Language (SWRL) (Horrocks et al., 2004) to codify the rules into an inference engine. The language supports interoperability on multiple systems on the web (O'Connor et al., 2005). Furthermore, it has been successfully integrated with domain ontology in various use cases; for example, for anti-diabetic drug selection (R.-C. Chen et al., 2012), underwater robots (Zhai et al., 2018), as well as integrated product design (Abadi et al., 2018).

Implementation

This section presents the suggested solution's implementation prototype. How the intended biodiversity data were gathered and maintained in a database is described first, followed by how event detection, mapping, and temporal inference were implemented.

The data was collected from Lensa, a website portal for disseminating the characteristics of natural fiber extracted from various plants in Indonesia. The portal collects various natural fiber characteristics that are measured or extracted from plants. Multiple groups of scientists collected the characteristics using different kinds of scientific instruments. Specimens from different parts of sampled plants will be distributed to different laboratories, where a unique identification number will be assigned to each sample. Following that, each laboratory examines the received sample further and enters the findings into the database. It is impossible to predict which analysis will be entered first. The portal was built using a generic content management system where data is stored in a MariaDB database server.



Physics & Mechanics

Diameter (mm)	0,224
Density	1,015
Tensile Strength	143,10
Tensile Modulus	5,059
Elongation	30,437

SG Ratio

SG Ratio	0,20
Pyrogram	

Figure 4. The integrated user interface for data collection

Figure 4 depicts the integrated user interface for data collection, which was used to gather several kinds of measurable attributes. Fields and specimens are the two forms of data that are required. Field data contains the value for a single feature, but specimen data has a collection of fields that define the essential qualities of a given specimen. The portal also stored the time when a specimen or field was created and updated. The updated time refers to the last update performed on the specimen or field. Further, it stored the author who created or updated the relevant specimen or field. For analysis, the time and the author information were used as the authorship and temporal contexts, respectively.

To detect multiple events, multiple queries were performed on the database using Structured Query Language (SQL) to capture the current situation of contextual information for every field and specimen, as shown in Table 1. In this example, the queries will collect the identification number (*id*), author (*post_author*), date when the data item was created (*post_date*), and when it was modified (*post_modified*) for field and specimen data, respectively. Such queries are performed regularly to capture modifications to contextual information over time. A contextual changed event is detected whenever one of the following conditions is fulfilled:

1. The value of the "post_modified" attribute is different from the value of the "post_date" attribute from the current query
2. The value of the "post_modified" attribute is different from the value of the "post_modified" attribute from the last query

Table 1. Example of queries to a database

No.	Query
1	SELECT ID, post_author, post_date, post_modified FROM \`wpw0_posts\` WHERE post_type = 'field' AND post_status = 'publish'
2	SELECT ID, post_author, post_date, post_modified FROM \`wpw0_posts\` WHERE post_type = 'specimen' AND post_status = 'publish'

As a result, all collected information for every detected event will be exported as a Comma-Separated Values (CSV) file to be processed further.

A uniform schema is required to represent collected contextual information from the event detection method explained above. As mentioned in the Computational Approach section, PROV-O was used as the base for the provenance schema, combined with Schema.org (https://schema.org/) vocabulary and Temporal Provenance Model (TPM) (Beheshti et al., 2012). It is also required to define rules to infer the temporal data provenance automatically. The schema and rules will be defined using the Protégé Editor (Musen, 2015) with pellet reasoner (Sirin et al., 2007), supporting SWRL as shown in Figure 5.

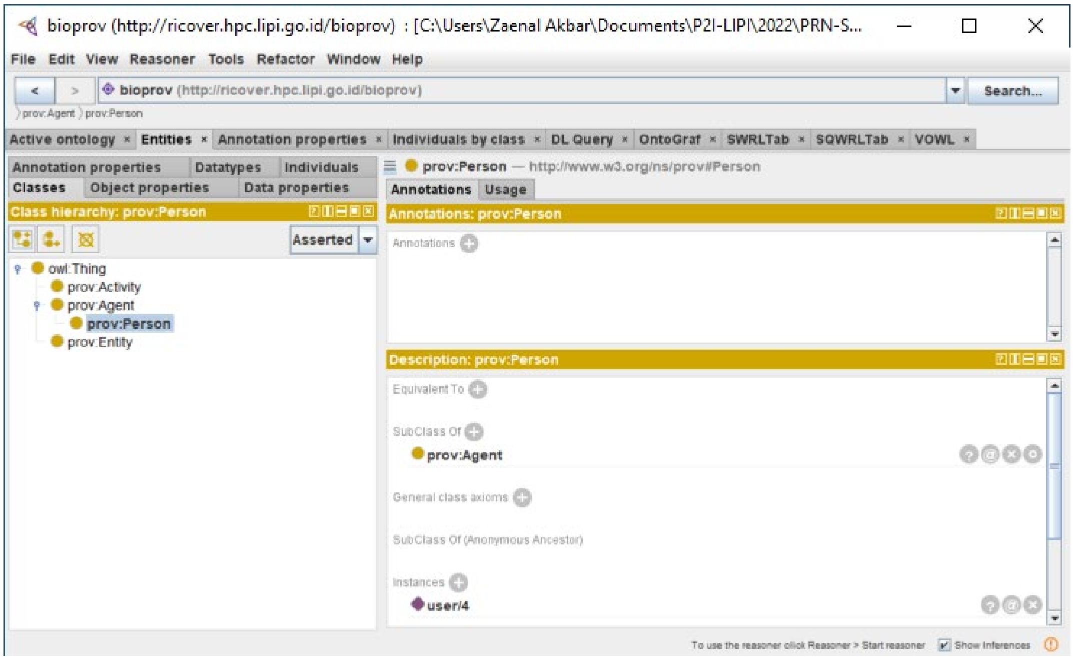


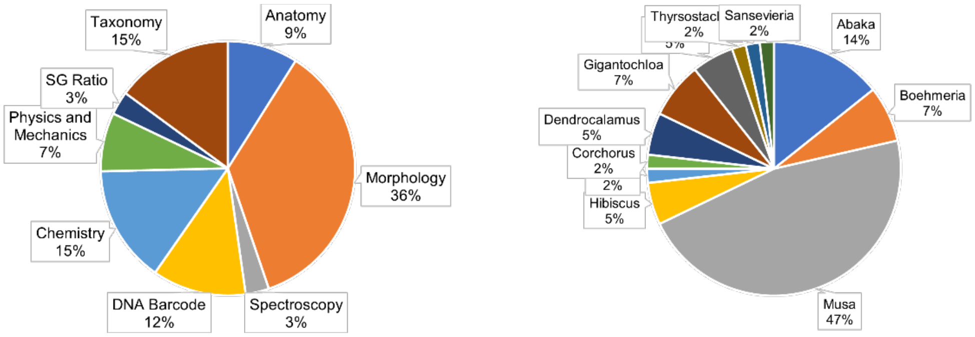
Figure 5. Schema definition using the Protégé Editor

Following that, mapping rules that match the information of detected events to the established schema will be generated to provide associated data provenance for all event participants. The mapping rules were defined using a declarative rules language called RML (https://rml.io/), where a rule will be used to align data fields/columns from the generated CSV files to the schema. Furthermore, RMLMapper (https://github.com/RMLio/rmlmapper-java) was utilized to execute the produced RML rules. As a result, relevant triples (subject-predicate-object) of data provenance in the form of Resource Description Framework (RDF) (https://www.w3.org/TR/rdf11-concepts/) were generated.

Finally, all generated triples of individual data were loaded into the Protégé Editor with the defined schema and specified inference rules. The selected reasoner is then executed such that a few new data items were successfully inferred as the temporal data provenance. These steps need to be performed regularly to capture the history of temporal data provenance over time.

Methods

This section details the performed experiment for automatically producing temporal data provenance using the suggested solution. First, data from a standard biodiversity information system was collected. After that, a data provenance schema was built, and mapping rules were established. Finally, a rule engine automatically performed the temporal provenance generation using several defined inference rules.



(a) The distribution of data fields (based on plants' characteristics)

(b) The distribution of data specimens (based on plants' genus)

Figure 6. The distribution of data fields and specimens in the dataset

DATASET

The database of Lensa consists of multiple types of natural fiber characteristics, including Plant Anatomy, Morphology, Spectroscopy, DNA Barcode, Physics and Mechanics, Chemistry, and Plant Taxonomy. It was chosen as the source of the dataset for this study due to a few reasons. First, the database was constructed using a generic information system which is the main target of this study. Second, most data records in the database have received multiple updates, representing the data changes over time to be extracted as the temporal data for this study.

Figure 6a shows the distribution of fields in the dataset. There are 67 fields in total, distributed in 8 groups of fiber characteristics. The number of fields ranges from 2 to 20, where most of the fields belong to the group of Morphology (36%), followed by Plant Taxonomy (15%), Chemistry (15%), DNA Barcode (12%), and so on. Furthermore, the distribution of specimens in the dataset is shown in Figure 6b. Most of the specimens (47%) fall into the plant genus of Musa, followed by Abaka (14%), Boehmeria (7%), Gigantochloa (7%), and so on.

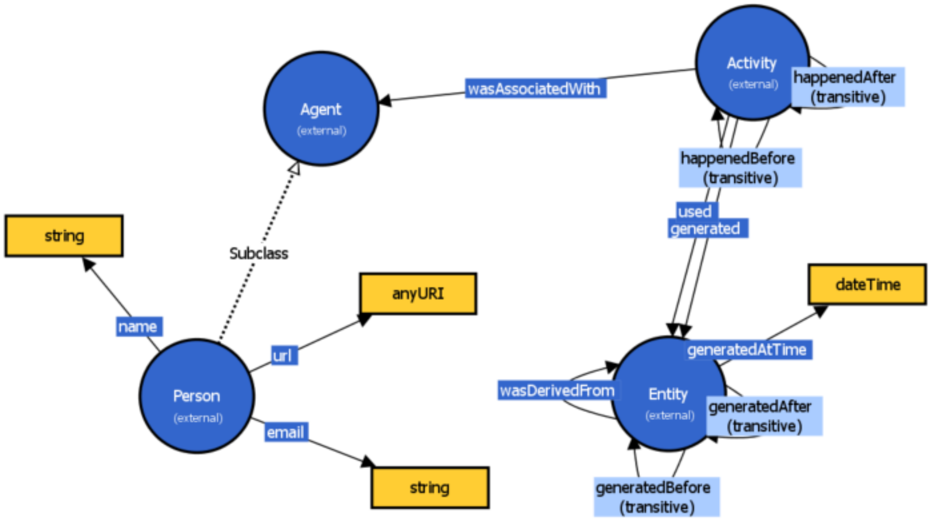


Figure 7. Data provenance schema for biodiversity data

DATA PROVENANCE SCHEMA

Figure 7 shows the constructed data provenance schema. It consists of four classes of PROV-O, namely "Activity," "Entity," and "Person," which is a sub-class of "Agent." As indicated in Table 2, a variety of objects and data characteristics have been utilized.

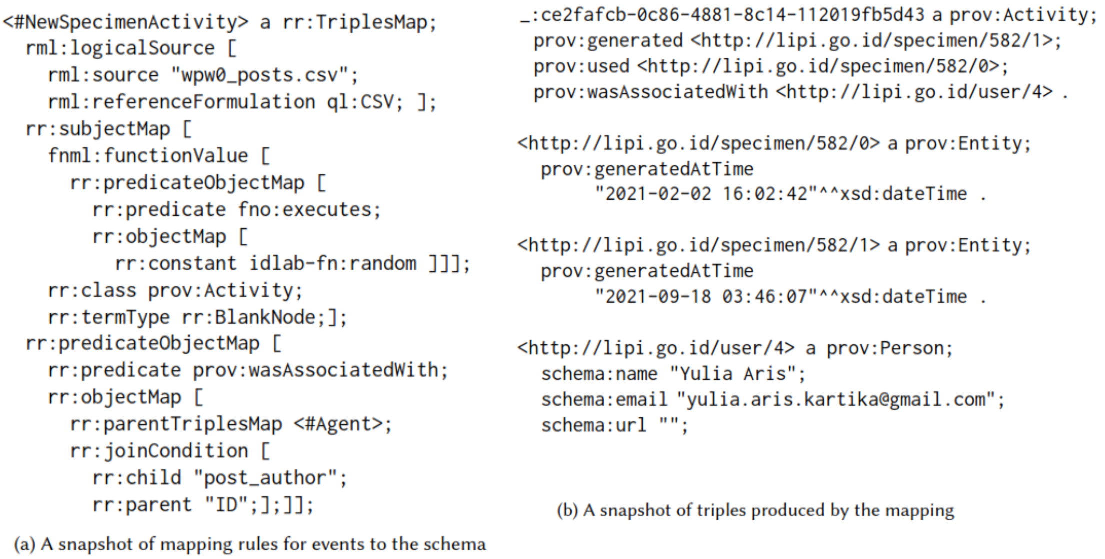
Table 2. The properties of the data provenance schema

No.	Property	Vocabulary	Notes
1	used	PROV-O	
2	generated	PROV-O	
3	wasAssociatedWith	PROV-O	
4	wasDerivedFrom	PROV-O	
5	name	Schema.org	
6	email	Schema.org	
7	url	Schema.org	
8	happenedBefore	TPM	
9	happenedAfter	-	Inverse of happenedBefore
10	generatedBefore	-	
11	generatedAfter	-	Inverse of generatedBefore

The schema describes how data items of specimens and fields were produced or altered over time. An activity may use a field or specimen and generate a field or specimen that is related. Moreover, the time at which a field or specimen was generated or modified must be indicated. Additionally, each action is connected with the individual who conducted the activity. Even though the schema is heavily focused on specific use cases, it has the capability to capture the most important aspect of data provenance, namely the data production history. In addition, four object attributes are supplied to indicate temporal information, such as how actions were executed or entities were created.

Event-Schema Mapping

A mapping rule aligns each detected event with the specified schema and generates provenances for all persons involved in the event. Figure 8a displays an illustration of a mapping rule. In this example, the triple subject is a blank node belonging to the "Activity" class. The triple predicate will be a property "wasAssociatedWith" which relates the activity with an "Agent," in our case, a person. A second relationship "generated" will relate the activity with an "Entity," in our case, a field or specimen. Figure 8b shows a snapshot of the generated RDF. It depicts an activity that was performed by a person (identified by <*http://lipi.go.id/user/4*>) that used an entity (identified by <*http://lipi.go.id/specimen/582/0*>) and generated another entity (identified by <*http://lipi.go.id/specimen/582/1*>).



```
<#NewSpecimenActivity> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "wpw0_posts.csv";
    rml:referenceFormulation ql:CSV; ];
  rr:subjectMap [
    fnml:functionValue [
      rr:predicateObjectMap [
        rr:predicate fno:executes;
        rr:objectMap [
          rr:constant idlab-fn:random ]]];
    rr:class prov:Activity;
    rr:termType rr:BlankNode;];
  rr:predicateObjectMap [
    rr:predicate prov:wasAssociatedWith;
    rr:objectMap [
      rr:parentTriplesMap <#Agent>;
      rr:joinCondition [
        rr:child "post_author";
        rr:parent "ID";];]];
```

(a) A snapshot of mapping rules for events to the schema

```
_:ce2fafcb-0c86-4881-8c14-112019fb5d43 a prov:Activity;
  prov:generated <http://lipi.go.id/specimen/582/1>;
  prov:used <http://lipi.go.id/specimen/582/0>;
  prov:wasAssociatedWith <http://lipi.go.id/user/4> .

<http://lipi.go.id/specimen/582/0> a prov:Entity;
  prov:generatedAtTime
       "2021-02-02 16:02:42"^^xsd:dateTime .

<http://lipi.go.id/specimen/582/1> a prov:Entity;
  prov:generatedAtTime
       "2021-09-18 03:46:07"^^xsd:dateTime .

<http://lipi.go.id/user/4> a prov:Person;
  schema:name "Yulia Aris";
  schema:email "yulia.aris.kartika@gmail.com";
  schema:url "";
```

(b) A snapshot of triples produced by the mapping

Figure 8. An example of a mapping rule and its results

Rules-Based Inference

As mentioned in the Computational Approach section, temporal information is automatically inferred using a rule-based reasoning approach, where rules were codified using SWRL. Table 3 shows rules to infer causality and temporal relationships. The first rule depicts that if an activity has used an entity e_1 and generated an entity e_2, it is possible to infer that e_2 was derived from the entity e_1. The second rule depicts the same condition as the first rule but with a different consequence: e_1 was generated before e_2. The third rule will use the new information generated by the second rule (determine which entity was generated first) and infer which activity happened first.

Table 3. Rules to infer causality and temporal relationships

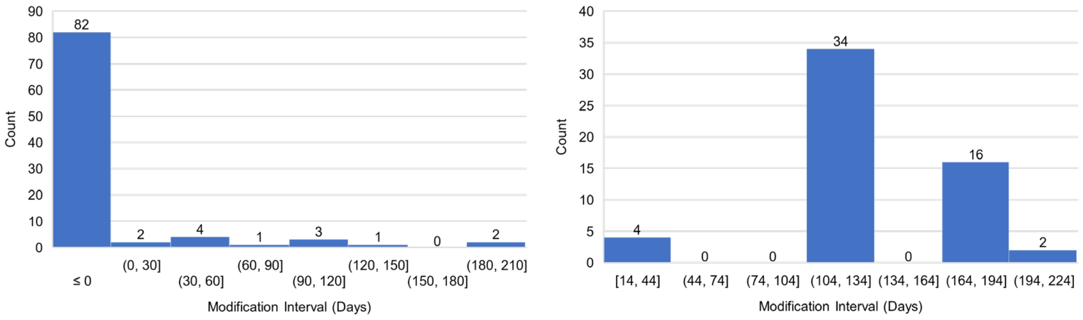
No.	Rules
1	prov:Activity(?a) ∧ prov:used(?a, ?e1) ∧ prov:generated(?a, ?e2) → prov:wasDerivedFrom(?e2, ?e1)
2	prov:Activity(?a) ∧ prov:used(?a, ?e1) ∧ prov:generated(?a, ?e2) → bioprov:generatedBefore(?e1, ?e2)
3	bioprov:generatedBefore(?e1, ?e2) ∧ prov:generated(?a1, ?e1) ∧ prov:used(?a2, ?e1) → bioprov:happenedBefore(?a1, ?a2)

Results and Findings

The outcomes of the conducted experiments and discoveries are presented in this section. First, how the event detection and schema mapping solution works are described. After that, how to produce the temporal data is explained before moving on to the discussion of data analysis and results.

Event Detection and Mapping

Two performance analyses were performed on the event detection solution: how the temporal context was updated until a point in time and within a period. In the first analysis, the time interval between the last update of a field or a specimen to its creation was computed for every detected modification. The interval represents how long a field or a specimen received updates. The analysis was performed on a snapshot of the database captured on 26/8/2021, consisting of 95 fields and 56 specimens. Figure 9a and Figure 9b show the findings of the data modification analysis for fields and specimens, respectively. We found that 86% of fields were never updated once they were created. Further, 13 data modification events were detected where only 2% of fields have received at least one update and were performed within the first month since its creation. On the other hand, all specimens received at least one update over time. 61% of specimens were updated within four months after their creation, and only 7% were updated within 2-5 weeks after their creation.

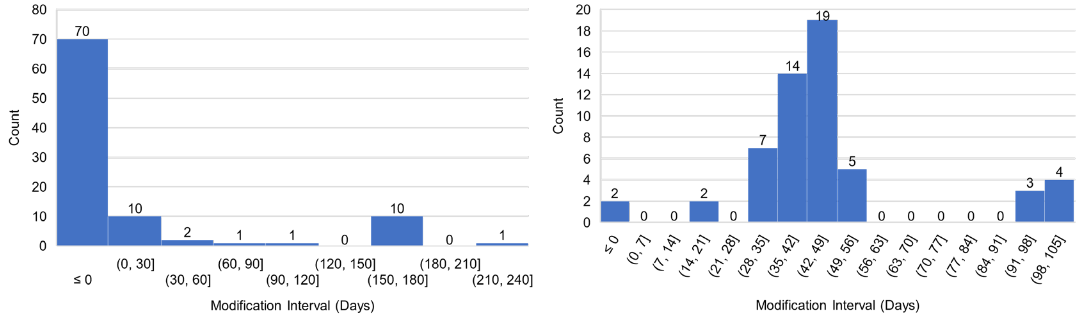


(a) Detected events from data of fields

(b) Detected events from data of specimens

Figure 9. Detected events from data fields and specimens (until August 2021)

For the second analysis, two database snapshots were captured at two different times, on 26/8/2021, and 25/11/2021. Then, the captured snapshots were compared where any modification on the last update will be detected for every field or specimen. Once again, computed the time interval between the last update on every field or specimen from both snapshots. The results are shown in Figure 10a for fields and Figure 10b for specimens. In total, 25 cases of data alteration of fields were identified, and almost every specimen was altered. While 40% of field data modifications occurred in the first month, the majority of specimen updates did not occur until the second month. Seven specimens, for instance, were updated three months later, while a field was updated nine months later. Notably, any newly added fields or specimens over this period were also detected. Four fields were discovered to have been added, while one was removed. There was no addition or deletion observed for the specimens.



(a) Detected events from data of fields

(b) Detected events from data of specimens

Figure 10. Detected events from data fields and specimens (August-November 2021)

Both cases demonstrated how multiple data modification events could be detected using contextual information. However, apart from the cases, events are scattered over some time. This fact indicates that it is crucial to separate event detection from schema mapping to provide scalability in the implementation. Furthermore, each event has different situations concerning data provenance and should be standardized. The data provenance can be uniformly represented as a standardized data model RDF by mapping all related information to a schema. Table 4 shows the number of triples of RDF produced from the database snapshot on 25/11/2021. In total, more than 300 activities were extracted successfully, which were performed by 16 persons.

Table 4. Total number of triples of produced data provenance (status on 25/11/2021)

No.	URI	#Triples
1	http://www.w3.org/ns/prov#Activity	**308**
2	http://www.w3.org/ns/prov#Entity	**308**
3	http://www.w3.org/ns/prov#Person	**16**
4	http://www.w3.org/ns/prov#generatedAtTime	**308**
5	http://www.w3.org/ns/prov#generated	**308**
6	http://www.w3.org/ns/prov#wasAssociatedWith	**308**
7	http://www.w3.org/ns/prov#used	**154**
8	http://schema.org/email	**16**
9	http://schema.org/name	**16**
10	http://schema.org/url	**1**

Inferred Temporal Data Provenance

Ultimately, the performance of the proposed solution can be determined by comparing the number of inferred data provenance triples produced to the original (asserted) ones. Given the number of asserted triples as $N_{asserted}$ and the total number of triples (i.e., asserted plus inferred) as $N_{asserted+inferred}$, then the performance was computed as the percentage change of the number of triples using the following formula: $\frac{N_{asserted+inferred}-N_{asserted}}{N_{asserted}} \times 100\ \%$

To illustrate the performance computation process, a data record from the dataset is selected randomly as an example. Then, the number of triples produced before and after the reasoner is activated can be compared using the Protégé Editor. The comparison result is shown in Figure 11. This figure referred to information defined explicitly (indicated as asserted) and implicit information (indicated as inferred). From three entities (entity/582-0, entity/582-1, entity/582-2) and three activities (activity/0, activity/1, activity/2), the number of triples produced has increased by 66% to 300% (172% on average) when a reasoner is activated (asserted + inferred).

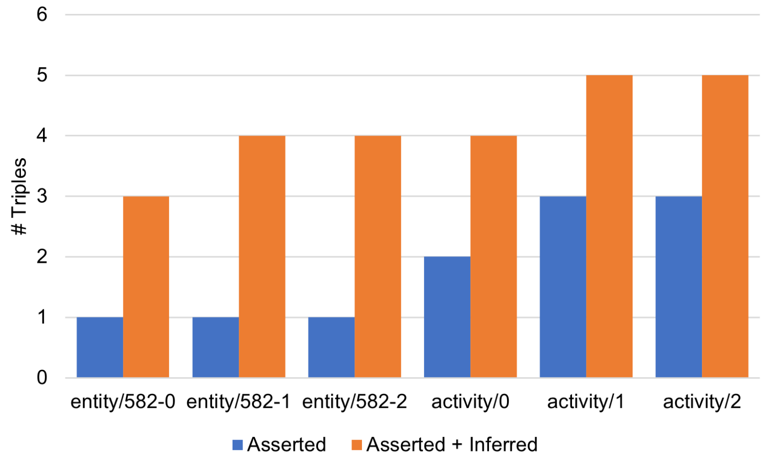


Figure 11. An example of comparing the number of triples produced without and with an inference engine

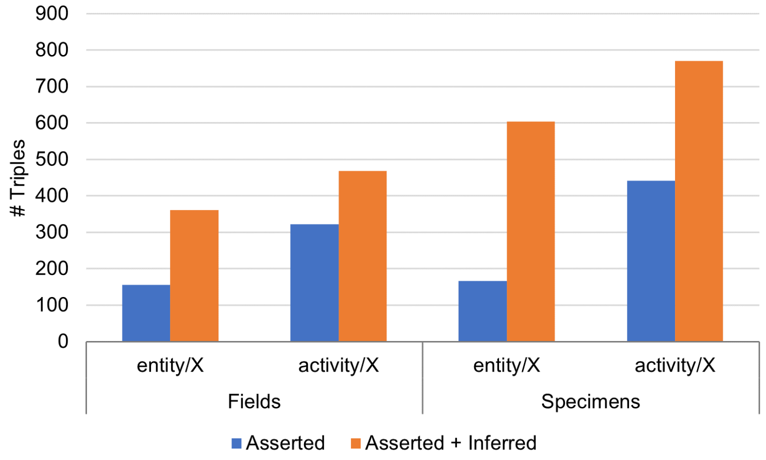


Figure 12. The comparison of the number of triples produced on the dataset, without and with an inference engine

Finally, the solution's performance on the dataset was computed using two database snapshots, captured on 26/8/2021, and 25/11/2021. The results are shown in Figure 12 for both data fields and specimens. The solution increased the number of triples produced by 133% and 45% for entities and activities, respectively, from data fields. Furthermore, it produced an increase in the number of triples by 264% and 74% for entities and activities, respectively, from data specimens. Most of the automatically generated (inferred) triples are about temporal information. Mixing the inference capability of both ontology and rules has enabled us to automatically generate richer information, especially temporal information.

DISCUSSION

The separation between database event detection, event-schema mapping, and temporal information inference has provided flexibility and scalability for implementation. It can be implemented on different schema and multiple data modification events with minimum effort. Moreover, mapping rules, as well as inference rules, can be defined effectively by using standardized languages.

From the implementation of data of natural fiber characteristics collected by Lensa, a few issues can be discussed as follow:

1) The proposed solution has enabled an automatic extraction of temporal data provenance from a biodiversity database maintained by a typical information system. Since most of the existing biodiversity databases are maintained using generic content management systems, the solution is expected to receive wide adoption. Moreover, multiple personnel may undertake the data gathering process; capturing the provenance at the individual data level is vital, particularly when the same specimen is processed with various equipment, tools, or protocols.
2) In most databases, only the last updated contextual information will be stored. Therefore, it is crucial to ensure that this extraction approach will be executed regularly, for example, daily. When the rate of data modifications increases, the approach should be executed directly after new data is inserted or after existing data is updated. A real-time provenance extraction captures fine-grained data production processes.
3) Once all events are collected, mapping rules definition can be declared. However, finding a mechanism to identify the various versions of an entity remains challenging. For example, when an entity has been modified several times, it will generate several entities with similar identification. Therefore, when an entity is referred, it must be identified which version was referred. One straightforward solution is to use different identification for every version of the entity, but a better solution should be achievable.
4) Even though the event detection and schema mapping were decoupled, there is no standardized format for the output from the event detection yet. In the current implementation, the person who creates the mapping rules should also be familiar with the event detection output. Therefore, a subsequent work is to standardize the event detection output, for example, through a common representation such as ontology. This way, schema mapping will be more convenient through an ontology mapping or alignment approach.
5) By using rules, various types of temporal information can be inferred. It provides flexibility and modularity to encode different types of knowledge from experts, where different temporal-based data analyses and reasoning can be performed.

This study has established the capability of automating the generation of temporal data provenances. Even though it was built in the biodiversity domain, it has the capability to be deployed in other domains with comparable use cases. In addition to its potential, this study reveals various future enhancement alternatives.

The comparison of the obtained results to existing relevant studies can be described as follows:

1) As demonstrated in the previous sub-section, the proposed solution produced a fine-grained provenance granularity up to the individual level of data items with several limitations. By using multiple snapshots of the database, richer provenance can be obtained. Compared with workflow-based and template-based solutions, the proposed solution fills the gap by providing more detailed provenance information. The workflow-based solutions rely on specific architectures and controlled environments to capture data provenance in the background (Weigel et al., 2020). For example, such solutions can be applied to a specific architecture (Stefanowski et al., 2017; Wang et al., 2015) or within a given scientific experiment (Oliveira et al., 2018), or on a specific computing platform (Carvalho et al., 2017; Guedes et al., 2020; Rajmohan et al., 2019) or a specific workflow management system (Horsch et al., 2020). Similar to workflow-based solutions, template-based solutions, which rely on the definition of provenance templates, are highly domain-specific. Therefore, such solutions perfectly fit in well-defined domains such as health (Curcin et al., 2017), software engineering (Sáenz-Adán et al., 2018), and environmental and earth sciences (Magagna et al., 2020).
2) The separation of ontology and event detection provides scalability. This aspect is essential in Big Data because multiple users collect and manage data with various methods, tools, and protocols.
3) The integration between ontology-based and rule-based reasoning systems maximizes the generation of multiple types of causality and temporal information. The solution takes advantage of the reasoning capabilities of both systems.
4) To answer the research question, based on the provenance hierarchy (Herschel et al., 2017), the solution which relies on the separation of event detection and provenance generation has reached the highest degree possible, where provenance can be generated at the most elementary level of data. In contrast, the existing solutions, mainly based on workflow and template, would only generate data provenance up to the workflow provenance level.

CONCLUSION

Data provenance provides data accountability by providing details about the data's production process. It includes the data source and any transformations conducted on the data, as well as the contributor (organization, person, or software) to the process. Especially when reusing data from multiple sources, provenance fulfilled the needs for data auditing, authenticity checking, and quality measures in data. Temporal data provenance encodes temporal information in data provenance such that more advanced temporal reasoning or analysis can be performed. In the field of biodiversity, it remains challenging to provide provenance during data collection, especially on the level of individual traits, where data provenance is rarely documented.

This study presents an automatic solution for generating temporal data provenance from biodiversity databases. The solution aims to eliminate the arduous effort required to collect provenance manually. Three processes comprise the solution: database event detection, unification of provenance representation through schema mapping, and temporal information inference. The solution separates the operations of event detection and provenance creation. The separation allows for the flexible alignment of observed events into a shared representation of provenance using a defined mapping strategy. The separation enables the solution to be integrated into most of the generic biodiversity information systems. When integrating with a new biodiversity information system, just the database event detection procedure must be modified; all other procedures may stay unchanged. Multiple databases with varying formats, semantics, and values have been established, therefore, the proposed solution is well suited to the present circumstance.

In the current implementation, a data provenance schema was constructed by adopting entities and properties from several schema or vocabularies, namely PROV-O, Schema.org, and Temporal Provenance Model. After that, a rule-based reasoning system was utilized to infer causality and temporal relationships. As a result, the solution mixes the reasoning capability of ontology and a rule-based system. The solution was evaluated with data collection of characteristics of natural fiber of plants in Indonesia. In this case, specimens from several plants will be analyzed using several scientific equipment and tools. The measurement result will then be entered into a database through a biodiversity information system. Since multiple individuals work on the same specimen, reflected as the same record in the database, it is necessary to ensure who and when data modification was performed. The assessment findings demonstrated that numerous data provenance sources might be captured by taking several database snapshots. Each source then drives the event detection process to generate related objects to be mapped to a schema to represent multiple types of provenance entities uniformly. Based on the collected data provenance, temporal information will be generated automatically using rule-based reasoning. As a consequence, the created system can capture temporal data provenance to a great degree automatically throughout the data gathering process. Based on the provenance hierarchy (Herschel et al., 2017), the solution is able to model the highest degree possible, namely data provenance at the individual level.

The solution was intended for generic biodiversity information systems, which possess a few limitations, including:

1. While a piece of biodiversity data may have the needed information for provenance, an information system could also store several other relevant data. For instance, information about the individual who created or edited a data record may be stored and maintained in a separate data table. Most of the provenance-relevant information exists in multiple data tables maintained by the information system.
2. Most information systems maintain the most recent changes of temporal information, such as the date on which a piece of data was modified. Whenever numerous changes are conducted, the information system will only record the most recent ones.

In the future, we would like to increase the extraction rate of the solution or even perform the extraction in real-time, directly after a change in the database is detected. One promising solution is deploying reactive rules (Paschke & Kozlenkov, 2009), which would trigger actions whenever the database receives any updates. Furthermore, extending the existing schema would enable multiple temporal-based data analyses on different types of biodiversity data. Therefore, experimentation with multiple biodiversity information systems would extend the solution as an integral part of a Big Data biodiversity management system.

Acknowledgments
This study was funded by Program Research of National Research Priorities (PRN) flagship for University F.Y. 2021 with contract number 7457/IT3.L1/PT.01.03/M/B/2021. It was also supported by Productive Innovative Research Activities (RISPRO) Mandatory Themed Education Fund Management Institution (LPDP), National Research Priorities (PRN) F.Y. 2020/2021, Ministry of Finance in collaboration with the Ministry of Research and Technology/National Research and Innovation Agency (Kemenristek-BRIN), Republic of Indonesia, under grant No. 273/E1/PRN/2020 and B-4528/IPH/KS.02.04/VII/2020. It was also supported by the Degree by Research Programme - Indonesian Institute of Sciences. The authors appreciate the facilities and scientific and technical assistance provided by the Advanced Characterization Laboratories Cibinong-Integrated Laboratory of Bioproducts, Indonesian Institute of Sciences via E-Layanan Sains, Lembaga Ilmu Pengetahuan Indonesia.

REFERENCES

Abadi, A., Ben-Azza, H., & Sekkat, S. (2018). Improving integrated product design using SWRL rules expression and ontology-based reasoning. *Procedia Computer Science*, *127*, 416-425. https://doi.org/10.1016/j.procs.2018.01.139

Akbar, Z., García, J. M., Toma, I., & Fensel, D. (2014). On using semantically-aware rules for efficient online communication. In A. Bikakis, P. Fodor, & D. Roman (Eds.), *Rules on the web. From theory to applications* (Vol. 8620, pp. 37-51). Springer International Publishing. https://doi.org/10.1007/978-3-319-09870-8_3

Akbar, Z., Kartika, Y. A., Ridwan Saleh, D., Mustika, H. F., & Parningotan Manik, L. (2020, November). On using declarative generation rules to deliver linked biodiversity data. *Proceedings of the 2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), Tangerang, Indonesia,* 267–272. https://doi.org/10.1109/ICRAMET51080.2020.9298573

Alonso, O., Gertz, M., & Baeza-Yates, R. (2007). On the value of temporal information in information retrieval. *ACM SIGIR Forum*, *41*(2), 35-41. https://doi.org/10.1145/1328964.1328968

Amanqui, F. K., De Nies, T., Dimou, A., Verborgh, R., Mannens, E., Van De Walle, R., & Moreira, D. (2016, June). A model of provenance applied to biodiversity datasets. *Proceedings of the 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, France,* 235-240. https://doi.org/10.1109/WETICE.2016.59

Beheshti, S.-M.-R., Motahari-Nezhad, H. R., & Benatallah, B. (2012). Temporal provenance model (TPM): model and query language. *ArXiv Preprint ArXiv:1211.5009*. https://doi.org/10.48550/arXiv.1211.5009

Carvalho, L. A. M. C., Wang, R., Gil, Y., & Garijo, D. (2017, December). NiW: Converting notebooks into workflows to capture dataflow and provenance. *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017), Austin, Texas*, 12-16. http://ceur-ws.org/Vol-2065/paper04.pdf

Chen, P., Plale, B., & Aktas, M. S. (2014). Temporal representation for mining scientific data provenance. *Future Generation Computer Systems*, *36*, 363-378. https://doi.org/10.1016/j.future.2013.09.032

Chen, R.-C., Huang, Y.-H., Bau, C.-T., & Chen, S.-M. (2012). A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. *Expert Systems with Applications*, *39*(4), 3995-4006. https://doi.org/10.1016/j.eswa.2011.09.061

Chowdhary, K. R. (2020). Rule based reasoning. In *Fundamentals of Artificial Intelligence* (pp. 89-109). Springer. https://doi.org/10.1007/978-81-322-3972-7_4

Curcin, V., Fairweather, E., Danger, R., & Corrigan, D. (2017). Templates as a method for implementing data provenance in decision support systems. *Journal of Biomedical Informatics*, *65*, 1-21. https://doi.org/10.1016/j.jbi.2016.10.022

De Meester, B., Dimou, A., Verborgh, R., & Mannens, E. (2017). Detailed provenance capture of data processing. *Proceedings of the 1st Workshop on Enabling Open Semantic Science Co-Located with 16th International Semantic Web Conference*, *1931,* 1-8. http://ceur-ws.org/Vol-1931/paper-05.pdf

Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. (2014). RML: A generic language for integrated RDF mappings of heterogeneous data. In C. Bizer, T. Heath, S. Auer, & T. Berners-Lee (Eds.), *Proceedings of the 7th Workshop on Linked Data on the Web*, *1184*, 5. http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf

Elliott, M. J., Poelen, J. H., & Fortes, J. A. B. (2020). Toward reliable biodiversity dataset references. *Ecological Informatics*, *59*, 101132. https://doi.org/10.1016/j.ecoinf.2020.101132

Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science Engineering*, *10*(3), 11-21. https://doi.org/10.1109/MCSE.2008.79

Groth, P. (2013). Transparency and reliability in the data supply chain. *IEEE Internet Computing*, *17*(2), 69-71. https://doi.org/10.1109/MIC.2013.41

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5), 907-928. https://doi.org/10.1006/ijhc.1995.1081

Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). Big Data: Promises and problems. *Computer*, *48*(3), 20-23. https://doi.org/10.1109/MC.2015.62

Guedes, T., Martins, L. B., Falci, M. L. F., Silva, V., Ocaña, K. A. C. S., Mattoso, M., Bedo, M., & de Oliveira, D. (2020). Capturing and analyzing provenance from Spark-based scientific workflows with SAMbA-RaP. *Future Generation Computer Systems*, *112*, 658-669. https://doi.org/10.1016/j.future.2020.05.031

Güntsch, A., Groom, Q., Hyam, R., Chagnoux, S., Röpert, D., Berendsohn, W. G., Casino, A., Droege, G., Gerritsen, W., Holetschek, J., Marhold, K., Mergen, P., Rainer, H., Smith, V. S., & Triebel, D. (2018). Standardised globally unique specimen identifiers. *Biodiversity Information Science and Standards*, *2*, e26658. https://doi.org/10.3897/biss.2.26658

Hardisty, A. R., Michener, W. K., Agosti, D., Alonso García, E., Bastin, L., Belbin, L., Bowser, A., Buttigieg, P. L., Canhos, D. A. L., Egloff, W., De Giovanni, R., Figueira, R., Groom, Q., Guralnick, R. P., Hobern, D., Hugo, W., Koureas, D., Ji, L., Los, W., … Kissling, W. D. (2019). The Bari Manifesto: An interoperability framework for essential biodiversity variables. *Ecological Informatics*, *49*, 22-31. https://doi.org/10.1016/j.ecoinf.2018.11.003

Haug, A., Zachariassen, F., & Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)*, *4*(2), 168-193. https://doi.org/10.3926/jiem.2011.v4n2.p168-193

Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from? *The VLDB Journal*, *26*(6), 881-906. https://doi.org/10.1007/s00778-017-0486-1

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., & others. (2004). SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member Submission*, *21*(79), 1-31. https://www.w3.org/Submission/SWRL/

Horsch, M. T., Niethammer, C., Boccardo, G., Carbone, P., Chiacchiera, S., Chiricotto, M., Elliott, J. D., Lobaskin, V., Neumann, P., Schiffels, P., Seaton, M. A., Todorov, I. T., Vrabec, J., & Cavalcanti, W. L. (2020). Semantic interoperability and characterization of data provenance in computational molecular engineering. *Journal of Chemical & Engineering Data*, *65*(3), 1313-1329. https://doi.org/10.1021/acs.jced.9b00739

Jain, S., Mehla, S., & Wagner, J. (2021). Ontology-supported rule-based reasoning for emergency management. In S. Jain, V. Jain, & V. E. Balas (Eds.), *Web semantics* (pp. 117-128). Academic Press. https://doi.org/10.1016/B978-0-12-822468-7.00017-1

Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017, July). TGIF-QA: Toward spatio-temporal reasoning in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii*. https://doi.org/10.1109/CVPR.2017.149

Kanwal, S., Khan, F. Z., Lonie, A., & Sinnott, R. O. (2017). Investigating reproducibility and tracking provenance - A genomic workflow case study. *BMC Bioinformatics*, *18*(1), 337. https://doi.org/10.1186/s12859-017-1747-0

Kerman, M. C., Jiang, W., Blumberg, A. F., & Buttrey, S. E. (2009, June). Event detection challenges, methods, and applications in natural and artificial systems. *Proceedings of the 14th International Command and Control Research and Technology Symposium (ICCRTS), Washington, D.C.*

Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernández, N., García, E. A., Guralnick, R. P., Isaac, N. J. B., Kelling, S., Los, W., McRae, L., Mihoub, J.-B., Obst, M., Santamaria, M., Skidmore, A. K., Williams, K. J., Agosti, D., Amariles, D., Arvanitidis, C., … Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, *93*(1), 600-625. https://doi.org/10.1111/brv.12359

Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., Amengual, J., Basset, A., van Bodegom, P. M., Cornelissen, J. H. C., Denny, E. G., Deudero, S., Egloff, W., Elmendorf, S. C., Alonso García, E., Jones, K. D., Jones, O. R., Lavorel, S., Lear, D., … Guralnick, R. P. (2018). Towards global data products of Essential Biodiversity Variables on species traits. *Nature Ecology & Evolution*, *2*(10), 1531–1540. https://doi.org/10.1038/s41559-018-0667-3

Lannom, L., Koureas, D., & Hardisty, A. R. (2020). FAIR data and services in biodiversity science and geoscience. *Data Intelligence*, *2*(1-2), 122-130. https://doi.org/10.1162/dint_a_00034

Liu, Q., Islam, B., & Governatori, G. (2021). Towards an efficient rule-based framework for legal reasoning. *Knowledge-Based Systems*, *224*, 107082. https://doi.org/10.1016/j.knosys.2021.107082

Magagna, B., Goldfarb, D., Martin, P., Atkinson, M., Koulouzis, S., & Zhao, Z. (2020). Data provenance. In Z. Zhao, & M. Hellström (Eds.), *Towards interoperable research infrastructures for environmental and earth sciences* (pp. 208-225). Springer. https://doi.org/10.1007/978-3-030-52829-4_12

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., & Van den Bussche, J. (2011). The open provenance model core specification (v1.1). *Future Generation Computer Systems*, *27*(6), 743-756. https://doi.org/10.1016/j.future.2010.07.005

Musen, M. A. (2015). The Protégé Project: A look back and a look forward. *AI Matters*, *1*(4), 4-12. https://doi.org/10.1145/2757001.2757003

O'Connor, M., Knublauch, H., Tu, S., Grosof, B., Dean, M., Grosso, W., & Musen, M. (2005). Supporting rule system interoperability on the semantic web with SWRL. In Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen (Eds.), *The Semantic Web – ISWC 2005* (pp. 974-986). Springer. https://doi.org/10.1007/11574620_69

Oliveira, W., De Oliveira, D., & Braganholo, V. (2018). Provenance analytics for workflow-based computational experiments: A survey. *ACM Computing Surveys*, *51*(3), 1-25. https://doi.org/10.1145/3184900

Pannarale, P., Catalano, D., De Caro, G., Grillo, G., Leo, P., Pappadà, G., Rubino, F., Scioscia, G., & Licciulli, F. (2012). GIDL: A rule based expert system for GenBank intelligent data loading into the molecular biodiversity database. *BMC Bioinformatics*, *13*(4), S4. https://doi.org/10.1186/1471-2105-13-S4-S4

Paschke, A., & Kozlenkov, A. (2009). Rule-based event processing and reaction rules. In G. Governatori, J. Hall, & A. Paschke (Eds.), *Rule interchange and applications* (pp. 53-66). Springer. https://doi.org/10.1007/978-3-642-04985-9_8

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*(6117), 277-278. https://doi.org/10.1126/science.1229931

Pérez, B., Rubio, J., & Sáenz-Adán, C. (2018). A systematic review of provenance systems. *Knowledge and Information Systems*, *57*(3), 495-543. https://doi.org/10.1007/s10115-018-1164-3

Rajmohan, C., Lohia, P., Gupta, H., Brahma, S., Hernandez, M., & Mehta, S. (2019, July). On efficiently processing workflow provenance queries in Spark. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, Texas,* 1443-1452. https://doi.org/10.1109/ICDCS.2019.00144

Sáenz-Adán, C., Moreau, L., Pérez, B., Miles, S., García-Izquierdo, F. J. (2018). Automating provenance capture in software engineering with UML2PROV. In K. Belhajjame, A. Gehani, & P. Alper (Eds.), *Provenance and annotation of data and processes*. Springer. https://doi.org/10.1007/978-3-319-98379-0_5

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., … Hide, W. (2012). Toward interoperable bioscience data. *Nature Genetics*, *44*(2), 121-126. https://doi.org/10.1038/ng.1054

Sarikhani, M., & Wendelborn, A. (2018). Mechanisms for provenance collection in scientific workflow systems. *Computing*, *100*(5), 439-472. https://doi.org/10.1007/s00607-017-0578-1

Scolozzi, R., & Geneletti, D. (2011). Spatial rule-based assessment of habitat potential to predict impact of land use changes on biodiversity at municipal scale. *Environmental Management*, *47*(3), 368-383. https://doi.org/10.1007/s00267-011-9613-8

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, *5*(2), 51-53. https://doi.org/10.1016/j.websem.2007.03.004

Stefanowski, J., Krawiec, K., & Wrembel, R. (2017). Exploring complex and big data. *International Journal of Applied Mathematics and Computer Science*, *27*(4), 669-679. https://doi.org/10.1515/amcs-2017-0046

Tang, S.-W., Zhou, Z.-J., Hu, C.-H., Zhao, F.-J., & Cao, Y. (2020). A new evidential reasoning rule-based safety assessment method with sensor reliability for complex systems. *IEEE Transactions on Cybernetics*, 1-12. https://doi.org/10.1109/TCYB.2020.3015664

Trivedi, R., Dai, H., Wang, Y., & Song, L. (2017). Know-Evolve: Deep temporal reasoning for dynamic knowledge graphs. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 3462-3471). https://proceedings.mlr.press/v70/trivedi17a.html

Wang, J., Crawl, D., Purawat, S., Nguyen, M., & Altintas, I. (2015, October-November). Big data provenance: Challenges, state of the art and opportunities. Proceedings of the *2015 IEEE International Conference on Big Data (Big Data)*, *Santa Clara, CA,* 2509-2516. https://doi.org/10.1109/BigData.2015.7364047

Weigel, T., Schwardmann, U., Klump, J., Bendoukha, S., & Quick, R. (2020). Making data and workflows findable for machines. *Data Intelligence*, *2*(1-2), 40-46. https://doi.org/10.1162/dint_a_00026

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zhai, Z., Martínez Ortega, J.-F., Lucas Martínez, N., & Castillejo, P. (2018). A rule-based reasoner for underwater robots using OWL and SWRL. *Sensors*, *18*(10). https://doi.org/10.3390/s18103481

Zhang, J., Shen, F., Xu, X., & Shen, H. T. (2020). Temporal reasoning graph for activity recognition. *IEEE Transactions on Image Processing*, *29*, 5491-5506. https://doi.org/10.1109/TIP.2020.2985219

AUTHORS

Zaenal Akbar received a B.S. in Informatics Engineering from STMIK Dipanegara, Indonesia in 2001, an M.S. in Informatics Engineering from the Institut Teknologi Sepuluh Nopember, Indonesia in 2004, and a Ph.D. in Computer Science from the University of Innsbruck, Austria, in 2018. He joined the National Research and Innovation Agency (BRIN) Indonesia in 2005 and currently serves as a senior researcher at the Research Center for Computing. His research interests include knowledge discovery from large-scale data, knowledge modeling, and knowledge engineering in various knowledge-rich domains, such as biodiversity.

Dadan R. Saleh received a B.S. degree in electrical engineering from Jendral Ahmad Yani University, Cimahi, Indonesia, in 2001, and an M.S. degree in informatics from Bandung Institute of Technology, Bandung, Indonesia, in 2012. He is currently pursuing a Ph.D. degree in computer science with the Universitas Indonesia, Jakarta, Indonesia. He also works as a Researcher and member of the Knowledge Computing Research Group at the Research Center for Computing, National Research and Innovation Agency (BRIN). His research interests include information systems and knowledge engineering.

Yulia Aris Kartika received a B.S. degree in Computer Science from Pakuan University, Bogor, Indonesia, in 2006, and an M.S. degree in Computer Science from Universitas Indonesia, Depok, Indonesia, in 2011. She works as a Researcher and member of the Knowledge Computing Research Group at the Research Center for Computing, National Research and Innovation Agency (BRIN). Her research interests include biodiversity informatics, data mining, information systems, and knowledge engineering.

Dr. Widya Fatriasari is an associate professor in the Research Center for Biomass and Bioproducts, National Research and Innovation Agency (BRIN)-Indonesia. She has been the research group head of Polyphenol Based Bioproduct since 2019. She has obtained a Doctor of Forest Product Technology from Bogor Agriculture Institute and pursued her Bachelor's, Master's, and Doctor's degrees from IPB University-Indonesia. Her scientific area interests include natural fiber and composite technology, biomass conversion, biorefinery, and biopolymer. More than 140 national and international publications, 17 patents, and one copyright have been produced during her research career (2006 – present). She holds a Certificate of Competence from LSP Quantum HRM International (SNI ISO/IEC 17024:2012) as an Indonesian National Research Reviewer, as a trainer in recycling technology of pulp and paper and supervises undergraduate and master graduate student. She has also become a reviewer in a national journal, international proceedings, and some reputable international journals. Some national and international research funding has been successfully obtained, and she accomplished the achievement as a productive researcher in R.C. for Biomaterials LIPI in 2013 and was selected to represent its R.C. for selection as the best researcher of LIPI in 2016. She became the best performance achievement researcher in 2019, while in 2019-2020, she achieved the best head of research groups in R.C. Biomaterials LIPI. In 2020 and 2021, Widya still presented her best research performance in second place. In 2021, she performed outstandingly in LIPI (Indonesian Institute for Sciences) as the highest patent producer. HIMPENINDO award with criteria outstanding young researcher in the technology field was also achieved in 2021 by Dr. Widya. In the same year, the head of R.C. for Biomaterials gave an award to Dr. Widya as a researcher with the highest Intellectual Property Rights Achievement and Licenses 2021. Her Scopus id and orchid ID are 56690604600, and https://orcid.org/0000-0002-5166-9498 with the Google Scholar link is https://scholar.google.co.id/citations?user=YoeHDZ8AAAAJ&hl=

Adila A. Krisnadhi received a B.S. in Computer Science from Universitas Indonesia in 2002, an M.Sc. in Computational Logic from Technische Universität Dresden, Germany in 2007, and a Ph.D. in Computer Science from Wright State University, USA in 2015. He joined as a permanent faculty member at the Faculty of Computer Science, Universitas Indonesia from 2007. Since 2019, he also holds a position of Co-Director of Tokopedia-UI AI Center of Excellence. He has authored/co-authored more than 100 publications in the area of semantic technologies, knowledge representation, and machine learning. His research interests include knowledge graph construction, reasoning and question answering over knowledge graphs and ontologies, as well as knowledge modeling and engineering.

Deded Sarip Nawawi, M.Sc., Ph.D. is a lecturer and researcher at IPB University, Bogor, Indonesia. He is an associate professor in the field of Biomaterial Sciences and Technology. He graduated with a bachelor's degree from IPB University, a master's degree from George August University, Gottingen, Germany, and a doctorate from the Department of Biomaterials Sciences at the University of Tokyo, Japan. His scientific interests were wood chemistry, biomass conversion, and wood chemical component utilization. As of 2021, he has 25 Scopus indexed publications with H-index 8 (Scopus ID: 6505726267, ORCID: https://orcid.org/0000-0001-8367-0349), and he is involved in 2 national patents. He is ahead of the Wood Chemistry Laboratory and the head of the Forest Products Department, Faculty of Forestry and Environment, IPB University. He is also the editor of the *Journal of Tropical Wood Science and Technology* in the period of 2018 to 2021 and a reviewer of other journals.