European Journal of Technique

journal homepage: https://dergipark.org.tr/en/pub/ejt

Vol.13, No.2, 2023

Research Article

Data Management and Ontology Development for Provenance-Aware Organizations in Linked Data Space

Fatih Soygazi[1*], Tuğkan Tuğlular[2], Oğuz Dikenelli[3]

[1*]Aydın Adnan Menderes University, Computer Engineering Department, Aydin, Turkey. (e-mail: fatih.soygazi@adu.edu.tr).
[2] Izmir Institute of Technology, Computer Engineering Department, Izmir, Turkey. (e-mail: tugkantuglular@iyte.edu.tr).
[3] Ege University, Computer Engineering Department, Izmir, Turkey. (e-mail: oguz.dikenelli@ege.edu.tr).

ARTICLE INFO

Received: Dec., 08. 2023
Revised: Dec.., 26. 2023
Accepted: Dec., 26. 2023

Keywords:
Linked data space
Organization
Ontology
VoID
Metadata

Corresponding author: *Fatih Soygazi*

ISSN: *2536-5010* | e-ISSN: *2536-5134*

DOI: https://doi.org/10.36222/ejt.1402149

ABSTRACT

The importance of tracing the source of shared datasets has become evident. It is also crucial to monitor factors such as trust in the data, especially considering the widespread use of social media. The concept of Linked Data Space needs to be addressed in conjunction with organizations. From this point of view, provenance tracking in organizations, with respect to their origin, needs attention. This study elaborates on the concept of Linked Data Space, introducing the terms Interior Data and Exterior Data to the literature. Additionally, an architecture for Linked Data Space and data management for organizations is defined. Furthermore, the study explains how organizations can access Exterior Data in the Linked Data Space and how provenance metadata and ontologies will be created. These developed methods are illustrated in the News Aggregator Scenario, a main scenario for provenance, demonstrating how it can work in a use case.

1. INTRODUCTION

The Web consisted of HTML pages and the links between these pages until recently. However, the problem here is that the content of the Web pages can be understood only by people and not interpreted by machines. Semantic data models, such as Resource Description Framework (RDF), Resource Description Framework Schema (RDF Schema), and Web Ontology Language (OWL), which enable the machine interpretation of web content, have ensured that information can be processed not only by humans but also by machines. Linked Data (LD) [1, 2] is a set of best practices for publishing and connecting structured data on the Web using the expressed semantic data models. Linked Open Data (LOD) extends the principles of LD by emphasizing the use of open standards and licenses, ensuring that data is openly accessible and usable by anyone without restrictions. The Web of Data (WoD) is an evolution from the traditional Web of documents, embodying LD principles. While LOD is a specific approach to publishing and interlinking data on the Web, focusing on openness and reusability, the WoD is a broader vision of a globally interconnected data space enabled by semantic technologies and standards. The Data Space (DS) is a powerful conceptual model residing on heterogeneous data sources, providing a virtualization layer [3]. Linked Data combines data sources within Linked Dataspaces (LDSs) [4, 5], where applications are executed using the semantic data models and LD standards.

The concept of the DS has already been defined in the field of databases to manage the distributed data utilized by an organization, and the core services in a DS have been determined [5-7]. Some of these services include querying, tracing the data sources, and managing changes to data/metadata [6]. While the concept of the DS has been explained in the literature for LD, the services and integration of these services with data have not been examined in detail.

When the DS is considered in terms of data management at the organizational scale, it is necessary to clarify the relationship between the organization and the services and data it will use. One of the most serious challenges faced by organizations today is the need for an excessive number of interrelated data sources. Relevant services should be available in the DS for organizations to effectively use the data sources they need within a specific domain. The DS should offer solutions that keep users within the organization independent when providing these services, addressing how the data is integrated.

This manuscript proposes a data architecture in terms of organizations. We propose that the data used by organizations is expressed in two different ways: Exterior Data (ED) and Interior Data (ID). ID defines the organization's own data, organizational preferences, and metadata of accessible datasets on the Web needed by the organization. Organizational preference expresses the demand for the dataset to be obtained when accessing ED. For example, an organization may prefer to receive datasets only published in Turkey related to news. ED represents data located in the WoD. It covers whole accessible raw or semantic data on the Web and the metadata of those data.

Provenance is a concept guiding the processes of establishing trustworthiness and ensuring data quality [8]. In this paper, we focus on provenance from LD perspective [9]. Applications that publish and consume LD represent definitions of trust through the use of provenance metadata. Our proposal also concentrates on provenance for ED and ID to conduct a quality assessment of the data architecture for an organization operating within an LDS. We express provenance metadata to extend dataset metadata which belongs to ED. Additionally, provenance definitions related to raw dataset, semantic dataset and dataset metadata are clarified. Furthermore, we built an ontology that considers organizational and provenance concepts to construct ID. The ontology covers terms for querying provenance independent of the domain.

Organizations may have diverse preferences, one of which pertains to provenance. Consequently, another ontology has been developed to represent the provenance preferences of the organizations. Despite the existence of various provenance dimensions and the associated provenance preferences, this ontology addresses sample preference dimensions under different conditions. As a result, the method for establishing organizational preferences is exemplified.

The contributions of this manuscript are:

- the description of the architecture for organizations gathering data from an LDS.
- proposing Interior Data (ID) and Exterior Data (ED) concepts to the LD literature by explaining these concepts in the context of LDS.
- the development of provenance ontologies with respect to ED and ID perspectives.
- the development of a provenance preference ontology.

2. PROVENANCE

A vast amount of data on the Web is created through copying, modifying, or combining. The same data or dataset can be copied or presented at different locations. Datasets can be linked together using RDF links created with different tools. Thus, conflicting copies of the same sets of entities can be linked to each other. It is necessary to consider the quality of the data thus produced. This is important because this type of data can rapidly spread on the web and lead to the WoD being comprised of low-quality data.

When attempting to find an entity, data from various sources linked to many URIs can be returned. At this point, the question that needs to be answered is which links to follow to reach the desired data. The data source to be used should be a source that provides more reliable or up-to-date data. When accessing a data source, it is not sufficient to have only data about the entity. Furthermore, metadata expressed with a dictionary for the discovery of the data source can not also answer questions about how or when the data source was produced. Therefore, data or metadata about data discovery is not sufficient, and additional metadata for provenance tracking is also needed. In this respect, it should be considered how the data and provenance-related metadata should be used in the WoD.

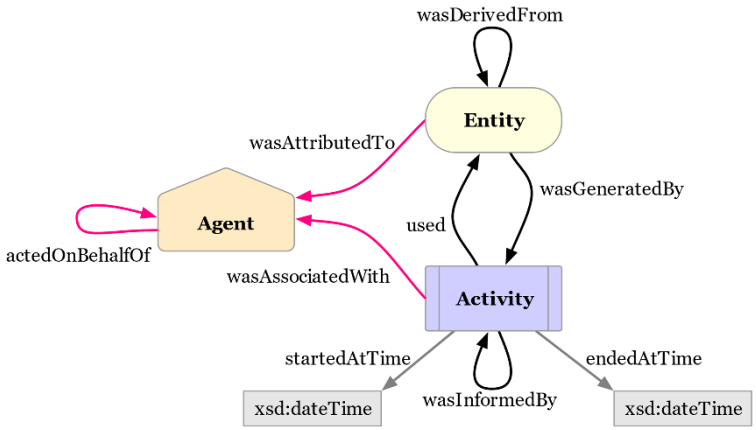

Figure 1. The PROV-O [11] Ontology

There is a need for a provenance data model that expresses how actions such as the creation of provenance data, its publication on the Web, and the access to published data. An ontology is a semantic data model that provides shared representation of knowledge. Ontologies have been developed for the modeling of provenance [10-13]. RDF-based provenance descriptions can be published on the Web and consumed by relevant actors in organizations.

2.1. Provenance ontologies

It is necessary to first clarify the core concepts of the provenance when defining a provenance model. Although core concepts are expressed similarly in all ontologies, the most popular and commony used work is the W3C PROV Ontology (PROV-O) [12, 13]. PROV-O is composed of domain-independent and general-purpose concepts. Its most important reason for being created as a general-purpose ontology is to produce a dictionary that can cover different needs and to encourage studies addressing provenance [14-18]. It represents tracking changes that occur during the creation or updating of sources on the Web. The basic concepts are defined as Entity, Agent, and Activity in PROV-O, as shown in Figure 1. An

entity can be derived from another entity, produced by an activity, or attributed to an agent. For example, let's consider a journalist working at a news agency. When creating a news article, the news entity is attributed to the journalist as the agent. Since the news agency is also expressed as an agent, it can be indicated that the journalist works at the news agency. The news creation activity is associated with the journalist, and this activity is linked to the created news.

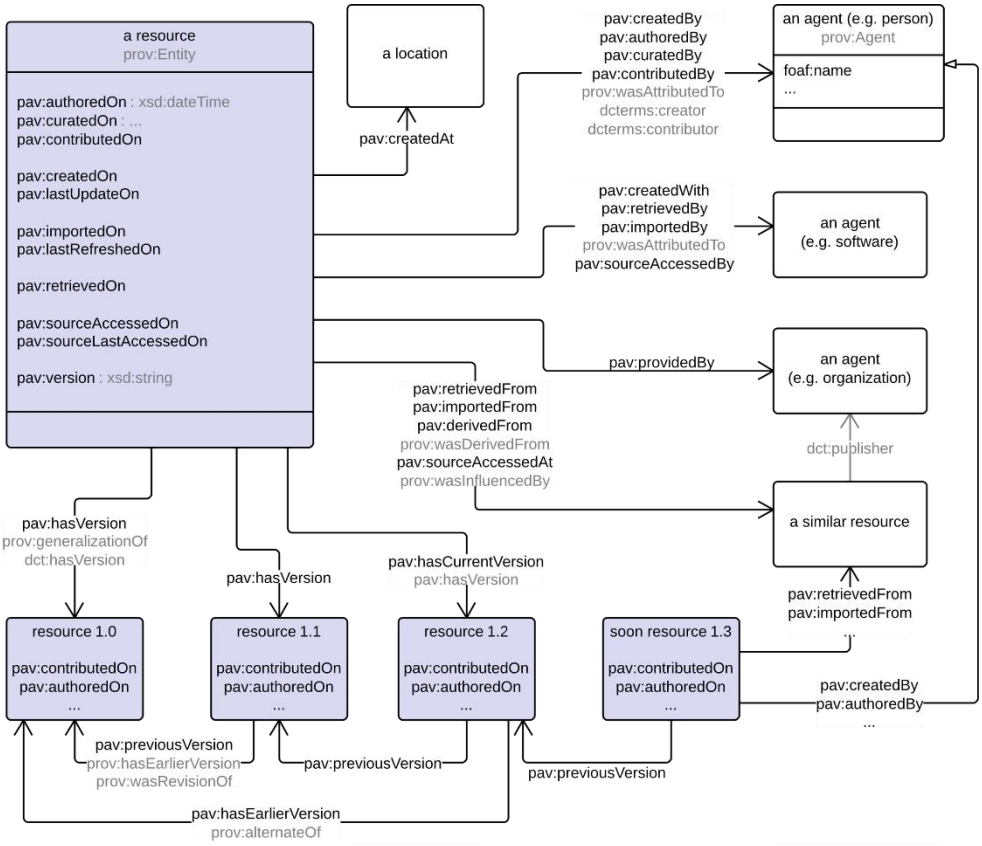

Figure 2. The PAV [19] Ontology

Provenance, Authoring and Versioning (PAV) ontology [19, 20] extends PROV-O to focus on the authoring, editing, and digital creation of data (Figure 2). The PAV ontology has introduced terms to distinguish between the different roles of agents providing content to existing Web-based systems. These roles concern the stages involved in the creation of a document in textual form and its publication on the Web, such as resource authoring, contribution, creation or curation. Hence, the provenance of digital resources can be traced during creation, retrieval or derivation processes. The PAV ontology customizes the general-purpose model of the PROV-O ontology to provide a more comprehensive and interoperable approach. There is an important distinction between authoring and creation of a resource that is described by PAV ontology. Provenance defines the creation of digital data, while Authoring describes the actual creation of data and its related features. For example, although the novel "İnce Memed" was written in 1955 (Authoring), it was first published as an e-book in 2010 (Provenance). Versioning shows the evolution of digital assets over time.

The creation of data, data derivation, data acquisition from another source, versioning, and similar concepts related to origin are found in PROV-O and PAV ontology, so the developers of the PAV ontology have expanded similar terms found in the PROV-O ontology. Especially the *prov:wasAttributedTo* property used for attribution in PROV-O, which has been expanded to include giving existence to the work expressed by the digital resource (*pav:authoredBy*), contribute to the work by the given agent *(pav:contributedBy)*, specify an agent specialist responsible for shaping the expression in an appropriate format (*pav:curatedBy*), create the digital artifact or resource representation (*pav:createdBy*), indicate the software/tool used by the creator when making the digital resource (*pav:createdWith*), describe an entity responsible for importing the data (*pav:importedBy*), and define entity to retrieve the data from a specific source without transformation (*prov:retrievedBy*).

A standard vocabulary is required to define the metadata of the datasets [21] and an extension of this vocabulary with provenance is also necessary. VoID (Vocabulary of Interlinked Datasets), a dictionary enabling the discovery and utilization of linked datasets, defines the metadata of datasets found on the WoD [22, 23]. In addition to a general provenance model created to define the provenance, customized approaches should also be expressed at the dataset level in the context of VoID. The extension of provenance to VoID, VoIDp [24], can be described as an enhanced version of the VoID vocabulary with provenance information. The aim of VoIDp is to assist dataset publishers in providing metadata related to the origin of their datasets so that data-consuming tools (or organizations) can access more reliable and higher-quality data. VoIDp provides a metadata extension within VoID that allows for the storage of information specifically about the origin of the source and modified resulting datasets. However, considering only the source and modified resulting datasets for their origin is not sufficient when filtering among the many datasets available. In this paper, one of our objectives is to define a broader vocabulary that extends VoID, covering more general

and usable situations by considering the provenance metadata of each dataset.

It has been observed that it can be used together with some extensions considering VoID, PROV-O and PAV ontologies to trace provenance in an LDS. In this paper, the ontological adaptation with provenance seems that it is related to Provenance and Versioning expressed in the PAV Ontology. Therefore, the creation of the VoID document and the processes through which the associated dataset was published are developed using the higher-level metadata from PROV-O and PAV ontology. The details of these ontologies are given in Section 4.

3. THE INTEGRATED ARCHITECTURE OF ORGANIZATIONS AND LINKED DATA SPACES FROM PROVENANCE PERSPECTIVE

The proposed conceptual architecture, which considers both an organization and LDS, is illustrated in Figure 3. An LDS encompasses LD services, with organizational LD applications consuming data provided by these services. Organizational LD applications may concurrently use one or more LD services based on their objectives.

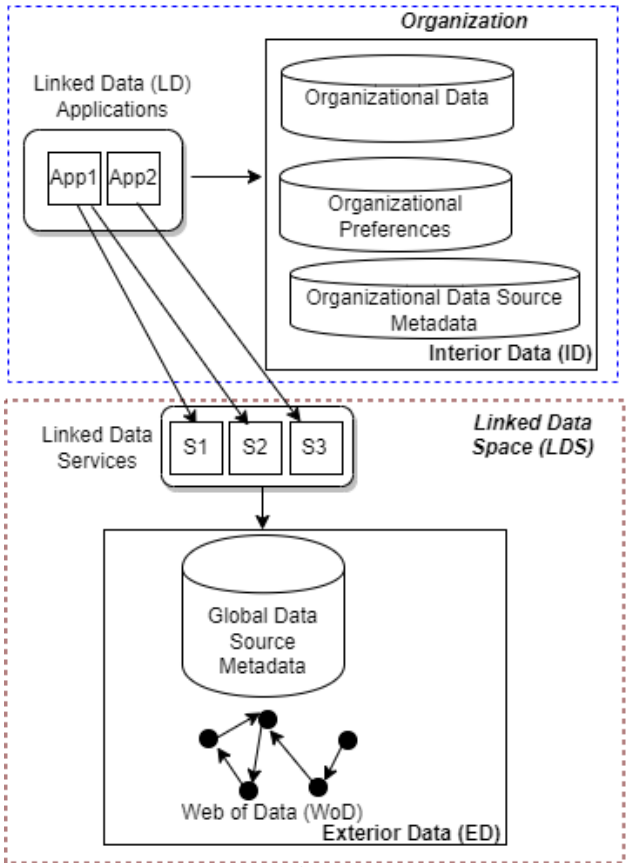

Figure 3. The architecture for integrating an organization and LDS

In order for ED to be discovered by LDS services, it is necessary to define the metadata of the data sources. Organizational Data Source Metadata stores these metadata of the datasets to be accessed by the organization. The data about datasets (metadata) in WoD should be defined by a vocabulary. VoID is used to create examples in Global Data Source Metadata and it is also commonly used by data federation systems [25] to define upper-level standards. Federated query engines [26-28] discover relevant datasets in WoD by querying with the help of VoID. An LDS is responsible for providing the organization with a service related to the selection of these datasets (One of the Linked Data Services from S_1,...,S_n in Figure 3).

When the data consumer (organization in our case) attemps to obtain data on the Web, the requested data may be available from multiple sources. For example, when searching for a photograph of a person for a news article to be published in a newspaper, the most up-to-date accessible photograph is obtained and intended to be published. The data consumer (or organization) similarly demands that the news content and the media related to the news (such as photographs) be current and reliable. The application in the Organizational Domain, which is expected to take into account similar situations, organizational data, organizational preferences, and local VoID metadata, is considered within the ID scope. The metadata of datasets relevant to the organization should be stored locally within the organization, referred to as local VoID metadata. Organizational Data Source Metadata should be able to keep up-to-date by tracking changes in Global Data Source Metadata [29, 30]. This allows the organization to access the data it needs in a way that meets its expectations in a current form. ED encompasses all linked data sources that can be queried via global VoID metadata in Global Data Source Metadata and the data accessed by these global VoID metadata in the WoD.

Provenance is a research direction that can be expanded by adding metadata to ED [31]. In the literature, there are similar extension studies using VoID [32, 33], but these studies are not sufficient to express the provenance in detail. Our manuscript emphasizes semantically expressing organizational preferences in ID with the added provenance metadata in ED. Therefore, a provenance-aware approach has been suggested.

According to Heath and Bizer [5], the data consumption method in LDS involves various LD services, such as access to WoD, data quality assessment and more. Our paper specifically focuses on provenance, and the detailed representation of Figure 3 is provided in Figure 4, considering provenance from an organizational perspective.

In Figure 4, Query Service gathers instances from WoD. The Provenance-Aware Filtering Service filters datasets based on provenance using the Organizational Data, Organizational Preferences, Enriched Organizational Data Source Metadata and stores them as Filtered Organizational Data Source Metadata. The Query Service uses Filtered Organizational Data Source Metadata to query the data in the WoD with respect to organizational expectations. The Monitoring Service is responsible for tracking changes in the Provenance-Enriched Global Datasource Metadata. The objective of the Monitoring Service is to report the modified VoID documents to the Enriched Organizational Data Source Metadata to keep it up-to-date. The actors in the organization are the Organizational User and the Organizational Infrastructure Administrator. The Organizational User accesses WoD via the Query Service and consumes the required data within the organization. The Organizational Infrastructure Administrator executes administrative services with the aim of providing the data to be used in LD applications as organization-specific data. The Administrative Application uses the Administrative Service(s) within each organization, and these services generate

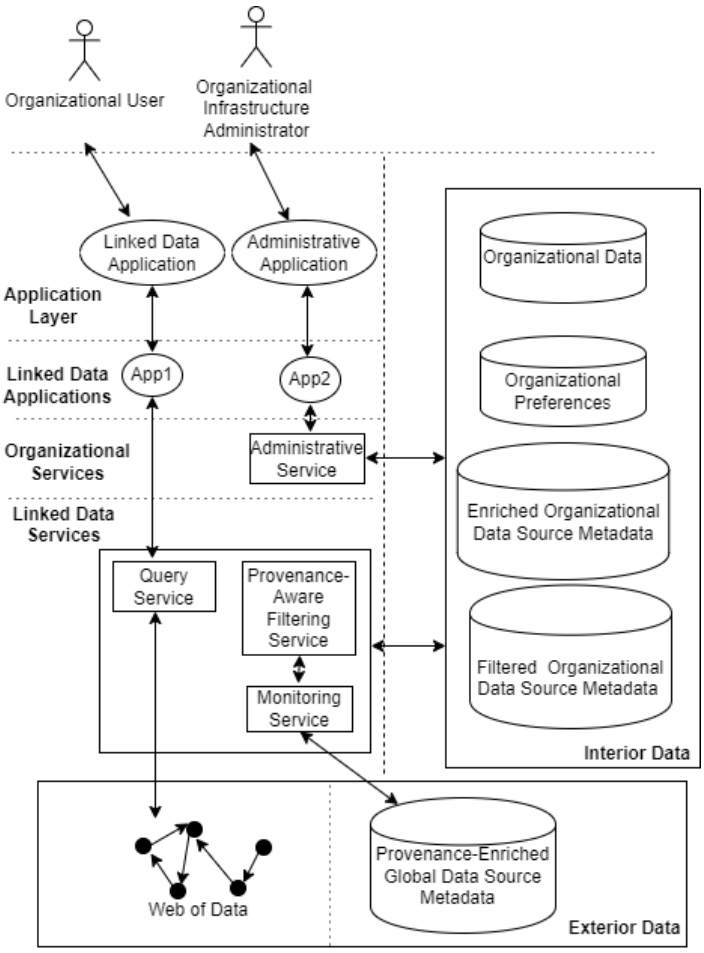

Figure 4. The architecture for integrating an organization and LDS considering provenance

Organizational Data and Organizational Preferences or provide datasets that could be published about that organization.

4. PROPOSED ONTOLOGIES FOR PROVENANCE-AWARE ORGANIZATIONS IN LINKED DATA SPACE

The data on the Web represents all data on the Web independently of semantics. Web of Data (WoD) [34] is a term that emerged to combine data sources containing a large amount of data. WoD aims to create a usable structure by leveraging non-semantic web data and semantic web technologies together.

Raw data is the data expressed in a representation format other than RDF that humans can understand (for example, Wikipedia). Semantic data is data that can also be understood by machines and is represented with RDF (for example, DBpedia). The data on the Web encompasses raw data and semantic data, but there are not many semantic connections between these different types of data. The WoD has emerged as a result of establishing links between these data and making these links explicit and discoverable. For example, documents on the Web are mostly presented with HTML pages. It is observed that raw data is represented on Wikipedia web page in Figure 5. In HTML documents, there are links within the document, and when these links are clicked, access to other documents on the internet is achieved through the HTTP protocol. When considering data on the Web, one should think at the data level, which is a more granular structure than the document level. Therefore, rather than reading an entire document, semantic meaning should be taken into account to elements within sentences, establishing links with relevant entities. In Figure 5, there is a semantic representation of data in triple format for the DBpedia and DBLP datasets. By using the concept of Paul Erdös mentioned in the example, links can be established with relevant entities published in hundreds of datasets across different datasets. When linking all of these relevant elements, it introduces another challange in discovering the required entity in WoD. VoID [22, 23] documents are being created to establish the metadata of datasets for the purpose of improving data discovery. The relationship between the DBpedia dataset and the DBLP dataset, which contains publication information for academics, is established at the dataset level using a VoID document in Figure 5.

It has been observed that in addition to publishing raw data as semantic data and creating VoID descriptions for semantic data, it is also necessary to express the provenance. At this point, one should consider what kind of metadata needs to be provided for each of the three data levels (raw level, semantic level, metadata level).

Dublin Core Metadata refers to a set of standardized metadata elements used to describe digital resources such as documents, images, web pages, and other types of media. The Dublin Core Metadata Initiative (DCMI) [35] developed and maintains these standards to facilitate the discovery, sharing, and management of information resources on the internet. Dublin Core (*dc*) is widely used in various digital libraries, archives, and content management systems. In Figure 5, *dc* is used to illustrate the relationship between entities in DBLP

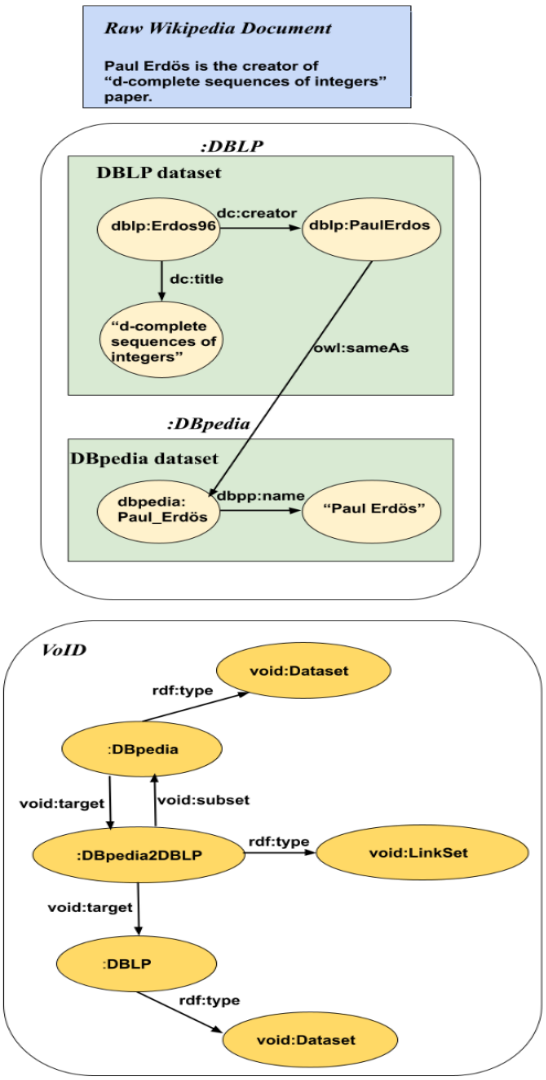

Figure 5. A sample representation of raw data, semantic data and VoID metadata

dataset where the knowledge of researchers are stored semantically. However, there is potential ambiguity when representing the author of the paper, as it is defined as the creator term in *dc*, while in Section 2.1, we defined it as "Authoring" in the PAV ontology. This discrepancy reflects different perspectives among ontology developers. In this paper, we aim to renconcile these viewpoints by considering the data elements and the dataset that includes these data elements.

4.1. Proposed ontologies considering Linked Data Space

Data from different datasets is merged to generate knowledge in the LDS. During this merging process, the links between datasets are taken into account (Figure 5). When dealing with these links, it is necessary to select the dataset that can provide higher quality data from datasets capable of offering similar data. Systems suitable for making such selections should present the user/organization with the most suitable integrated data using data about the provenance, such as where the data came from, when it was obtained, or how the links were created. On the other hand, it is necessary to provide upper-level data (metadata) regarding data creation and versioning, along with establishing links using this data during the dataset description.

The dataset metadata provided by VoID (that is supplied as metadata of datasets in ED) is not sufficient to meet all the requirements, needed on the Web [36]. Dataset publishers should think from a broad perspective to adapt the metadata of datasets along with the data to meet new requirements. Therefore, quality requirements such as provenance should be expressible with the metadata enriched by dataset publishers by using vocabularies such as VoIDp [24], as explained in Section 2.1. However, VoIDp is not sufficient for the requirements in [36] and another proposal is required, which we focus on in this paper.

There is a need for a comprehensive vocabulary that can fit into the working domain of organizations to consider organizational data alongside metadata of datasets in ED. While creating this vocabulary, it is considered that it should be aligned with the fundamental concepts of organizational preferences and organizational data in ID. The same considerations about ID should also be applicable to Enriched Organizational Data Source Metadata. Therefore, it is considered necessary to use ontologies containing established concepts related to provenance and expand them according to the requirements from ED and ID perspectives.

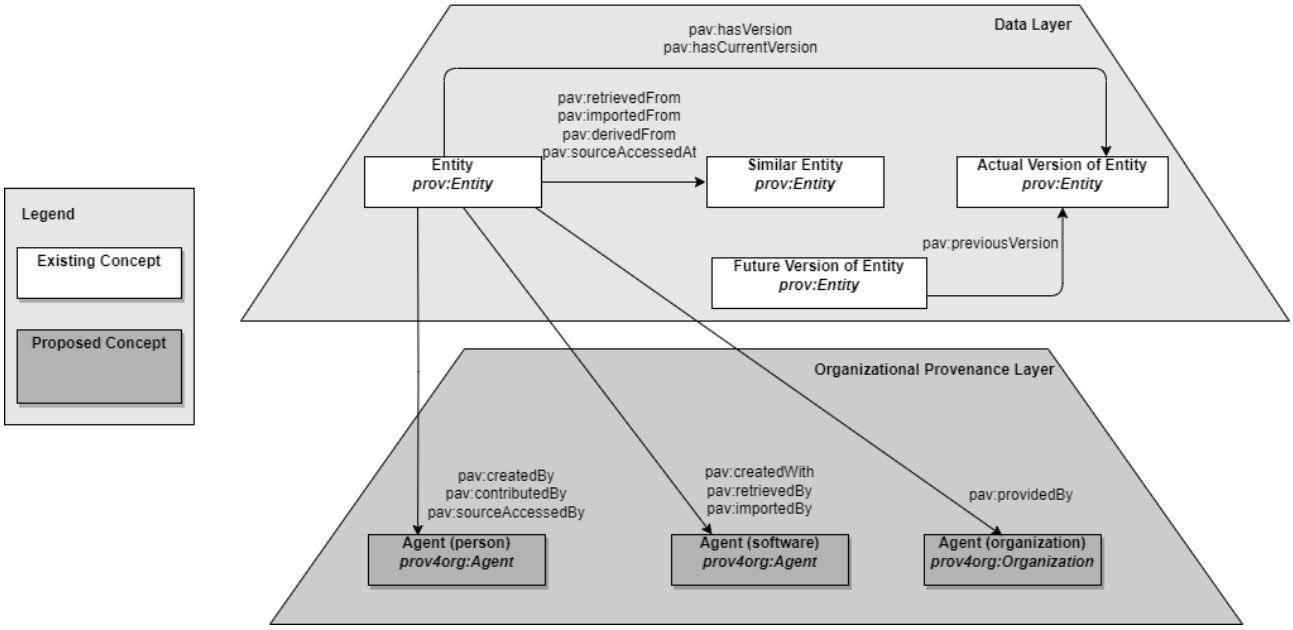

Figure 6. VoID enhancement by using PROV-O and PAV

4.2. Perspective of Exterior Data

Our proposal anticipates that VoID should be extended to define metadata for datasets and provenance expressions for organizational data. With this regard, these enhanced requirements should be considered in a layered manner as the Data Layer and the Organizational Provenance Layer, as shown in Figure 6. The Data Layer consists of the concepts and relationships required to express the provenance of the datasets. The Organizational Provenance Layer expresses the agents (person, software or organization) that are involved in the creation or update of the Entities. While the Organizational Provenance Layer is developed as a contribution, its namespace is given as *prov4org*.

As seen in the Data Layer of Figure 6, all these operations are defined by *prov:Entity*, however the operations on the entities differ. Accordingly, the concept referred to as an entity can be a VoID document, a raw dataset, or a semantically published dataset. During the creation of the provenance information, attention should be paid to three points with respect to these entities. First, attention should be paid to the provenance information of the created VoID metadata, second, if available, the provenance data before the dataset becomes semantic, and third, the expression of the provenance data for the semantically published dataset. The provenance data of the VoID metadata includes who (*pav:createdBy*) or which tool (*pav:createdWith*) created the document, made contributon during creation (*pav:contributedBy*), when it was created (*pav:createdOn*), and when it was last updated (*pav:lastUpdateOn*).

Secondly, the provenance data before the dataset becomes semantic includes the data about the dataset creation process and the agents in this process. The raw dataset mentioned can be obtained from any data source and used without modification. In order to express this situation, the software agent retrieving the data (*pav:retrievedBy*) and the source from which it is retrieved (*pav:retrievedFrom*) are important for knowing who originally created the data. In addition to obtaining and using the raw dataset without making any changes, it may also be possible to have a processed dataset that provides insights and comments about the data. Let's assume that after obtaining data from the E-Government website and processing it, it becomes the data presented by the Turkish Statistical Institute. In this case, the primary source of the data is the E-Government website. To enable provenance tracking, the software accessing this data (*pav:sourceAccessedBy*) and the actual location of this data (*pav:sourceAccessedAt*) should be expressed within VoID. The raw dataset may be modified despite being obtained in the same data format. In this case, metadata regarding from which dataset the dataset is derived (*pav:derivedFrom*) needs to be expressed. The data about the organization providing the source (*pav:providedBy*) should also be expressed as metadata. Defining data about the versions of datasets that are derived with these changes (*pav:hasVersion*, *pav:hasCurrentVersion*, *pav:previousVersion*) helps in understanding whether there are major or minor changes during data modification.

For the third and last consideration, it is necessary to keep data about the transformation of the content of the raw dataset into a semantic format. Specifying the software (*pav:importedBy*) that performs the transformation from the source dataset (*pav:importedFrom*) to the semantic dataset is necessary for provenance tracking. It is also important to specify the creator (*pav:createdBy*) of newly generated semantic datasets, the person who contributed to the creation of the dataset (*pav:contributedBy*), and the software tool (*pav:createdWith*) used in creating the dataset. Just like in the raw dataset, for the semantic dataset, it is important to express the data about the organization providing the primary source (*pav:providedBy*).

4.3. Perspective of Interior Data

The ontology that defines the concepts required for each organization is shown in Figure 7. The details and relationships of our definitions, *prov4org:Agent* and *prov4org:Organization*, using *prov:Agent* in Figure 6 are illustrated in Figure 7. This ontology has been developed by reusing organization ontology (ORG) [37], PROV-O [12, 13], and SWP (Semantic Web Publishing Vocabulary) [38] for provenance adaptation to the organization.

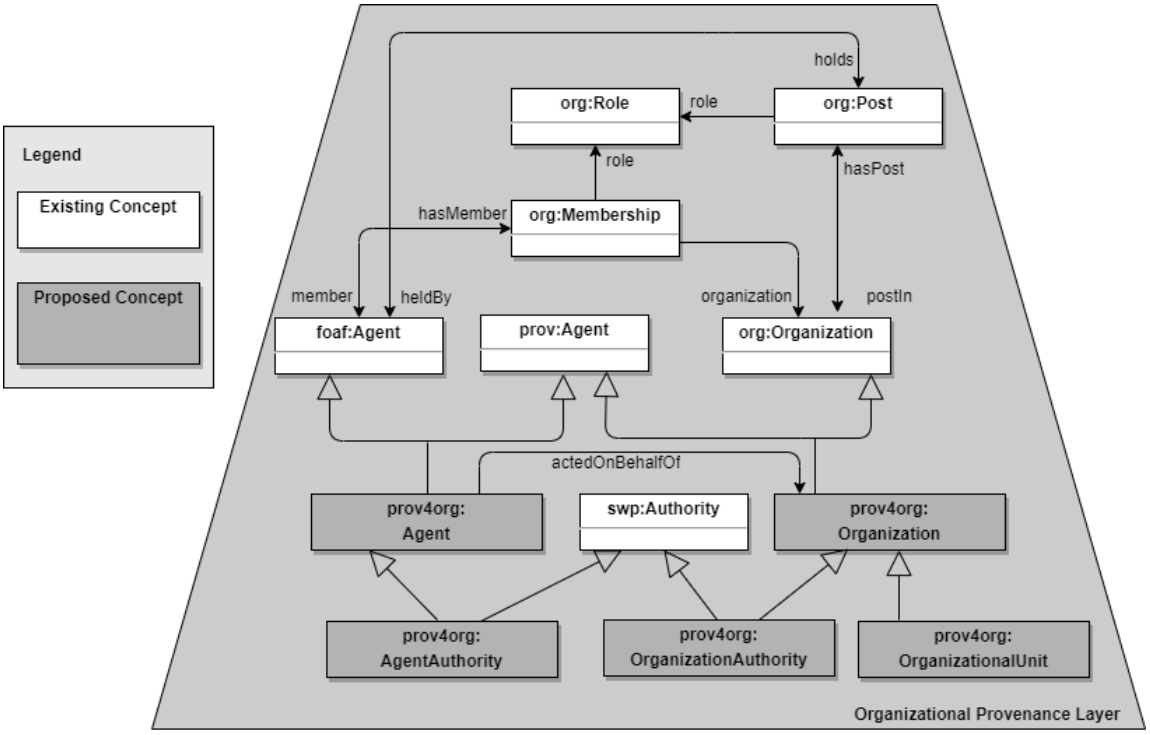

Figure 7. Organizational Provenance Ontology

prov4org:Agent can refer to a person within the organization as well as to a software. When considered as a person (Organizational User in Figure 4), it is necessary to take into account that the agent is a member of the organization (*org:Membership*). The concept of membership represents the hierarchical position within a company. Membership is expressed through a relational structure that includes the individual's role (*org:Role*), the organization he/she belongs to (*org:Organization*), and his/her personal information (*foaf:Agent*). Therefore, every employee in the organization has a membership. It has been observed that the concept of membership needs to considered for tracking the duration of membership and changes in roles (such as employees being promoted over time). The position information (*org:Post*) represents situations in the organizational hierarchy where it is expressed, but it is not necessary for there to be a person assigned to the position. In this sense, a position can exist without being filled by a person. However, membership represents a relationship that connects the organization and the agent and does not exist without them. When creating an organizational profile, usually either a position or membership is preferred for defining it. If an independent structure of individuals working in the organization is desired, positions are defined. If the goal is to keep records of the individuals who constitute the organization and to make queries about their provenance according to their abilities, then the concept of membership needs to be created. Organizational Unit (prov4*org:OrganizationalUnit*) represents smaller units within a large organization, such as departments. Therefore, it is seen that the concept of organizational unit needs to be expressed for the provenance of rules to be created at the organizational unit level for the execution of certain operations organizationally. There are two concepts defined for the expression of trust within the organization, namely, agent authority (*prov4org:AgentAuthority*) and organization authority (*prov4org:OrganizationAuthority*).

Another important point with respect to Interior Data (ID) is taking into account the preferences of the organization itself or other agents within the organization. Therefore, a domain independent ontology is required to express these preferences. Thus, by considering organizational preferences, the most suitable dataset from the datasets in the Exterior Data (ED) can be selected to meet the needs of the organization. When considering preferences, two scenarios are observed: the preferences of the organization (ID) for the Linked Data Space (ED) or the preferences of the Linked Data Space (ED) for the organization (ID). As illustrated in Figure 8, preferences should be adaptable to organizational ontological definitions independently of the domain. The development of a general preference ontology also enables the transfer of individuals' or organizational units' preferences (such as the continuation of an employee's preferences when moving from one organizational unit to another or the use of an organizational unit's preferences in different units).

As illustrated in Figure 8 adapted from Figure 7, *prov4org:Agent* and *prov4org:Organization* demonstrate that preferences can be expressed at the agent or organizational level. Preferences actually represent the constraints of the elements in ID. Therefore, *provpref:Restriction* is used to indicate the constraints on the preference of an agent or organization within an organization, or on another agent, organization, or entity. The filtering condition within the constraint (*provpref:FilterCondition*) indicates the expectation regarding the element on which the preference is made. The filtering condition is associated with different preferences via the blank node. Blank nodes nodes in a graph data structure that does not have an explicit identifier or a value. They are used when a web resource wants to define multiple pieces of data. For example, when expressing a professor, if we assume that the professor worked in different places at different times, this structure can be used to express the institution and department where he/she worked. In our work, we have resorted to the use

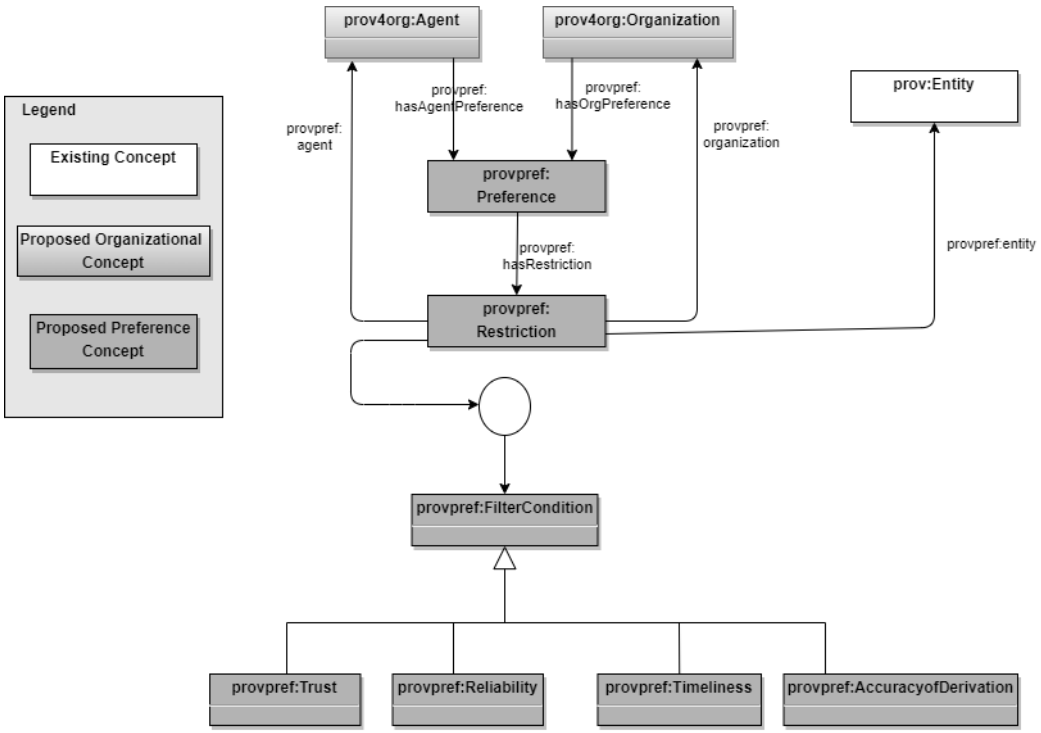

Figure 8. Organizational Preference Ontology

of blank nodes because Agents or Organizations can create preferences with different expectations in various situations. When considering the filter condition in the ontology, the preference types have been inspired by [39], where the terms *provpref:Trust*, *provpref:Reliability*, *provpref:Timeliness*, and *provpref:AccuracyofDerivation* are described. Trust represents the preference of whether the actor involved in generating the data is reliable or not during provenance tracking. Reliability is taken into account at the point of preference for activities during the derivation of the data (such as the creation of raw data, the transformation of raw data into semantic data). The data that is reliable may have been published in an unreliable manner through a different transformation by an activity. In this case, defining the reliability of the activity as a preference by the organization becomes important. Timeliness is related to tracking the times of creation and update of the data and creating preferences for the organization accordingly. Accuracy of Derivation focuses on tracking the changes made to the data since its creation. In data evaluation, the data is preferable if it has been generated with the least number of derivations.

The critical point in this study is to enable provenance tracking for the dataset at the metadata level. Thus, it is possible to create examples of how Agents, Activities, and Entities that affect the metadata of the dataset have made changes. From the perspective of preference, the important point is which features related to Agents, Activities, and Entities that affect the metadata of the dataset should be taken into account by the organization. Figure 9 illustrates how preferences are expressed for defined different filter conditions of the datasets.

Figure 9 illustrates how preferences will be expressed in terms of the datasets. Trust preference is used to express which agents are involved in creating, assisting in the creation, or providing access to the datasets that the organization will use. If the agent is a software agent, trust preferences can also be specified in the stages of data creation, acquisition, and transfer to the organization. If the data is provided by an organization, preferences specific to that organization can also be established. In the Reliability preference, preferences regarding the activity that creates the dataset are specified. The Timeliness preference reveals the preferences regarding the times when the dataset is acquired, created, or accessed by the organization. Additionally, preferences regarding the current version of the dataset can also be defined. The preference for Accuracy of Derivation anticipates obtaining the dataset with the least amount of change by examining the operations that have occurred since the dataset was acquired by the organization.

4.3. Sample use case for provenance

The ontology related to the organizational provenance layer should be created differently for each field of study, as the scope of each organization varies. In order to express the provenance concepts, there are three main scenarios defined in [36]. These scenarios are the News Aggregator Scenario, the Disease Outbreak Scenario, and the Business Contract Scenario. In the scope of this article's case study, the News Aggregator Scenario has been selected. This scenario aims to combine news items obtained from different sources (such as news websites, social media feeds, and news-related images).

The sample data to be generated in the News Aggregator Scenario consists of media such as photographs, videos, or the content of the news produced. For the generated data to be selected based on the origin of the media or news content, it must be licensed or published by a reliable provider. The requirement for selection is for the metadata that defines the provenance to be accessible to all data consumers. Therefore, the provenance metadata needs to be published along with the data. This way, the most suitable data source can be selected from data sources with similar accessible data.

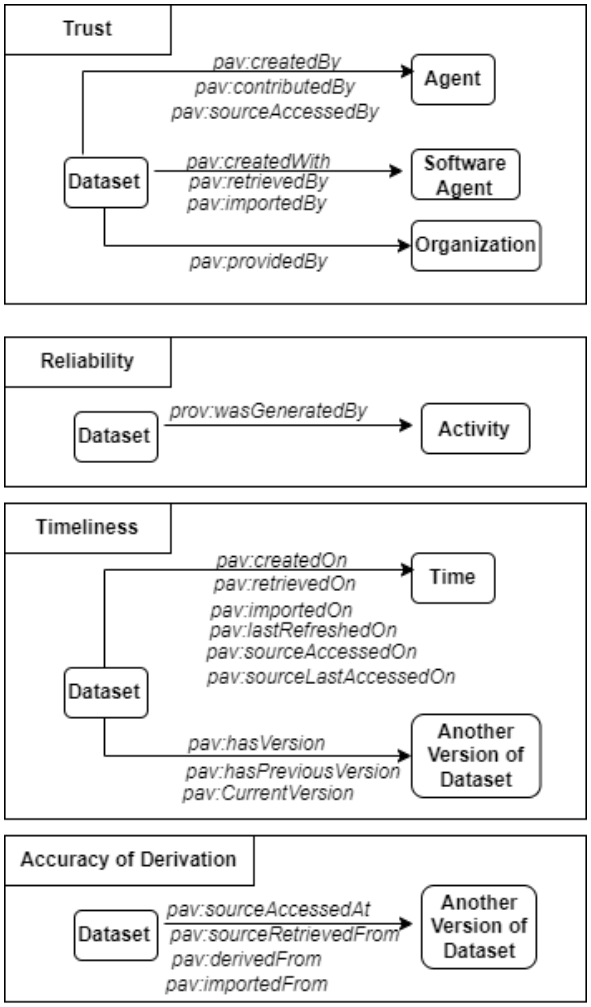

Figure 9. Representation of Various Organizational Filter Condition Preferences

Figure 10 illustrates the provenance metadata to be expressed for the New York Times (NYT) data in News Aggregator Scenario instantiating Figure 6. There can be three types of Agents for creating data and all other operations: user, organization, or software. The important point for raw NYT dataset is the creator of the dataset. However, when looking at the related NYT dataset, the focus should not be on the creation of the data but rather on the user and organization providing the created data or the software converting the raw data into semantic data. Another point is the activity that transforms the data from raw form to semantic data. Entity, Agent and Activity classes with their relationships have all been represented in terms of this use case.

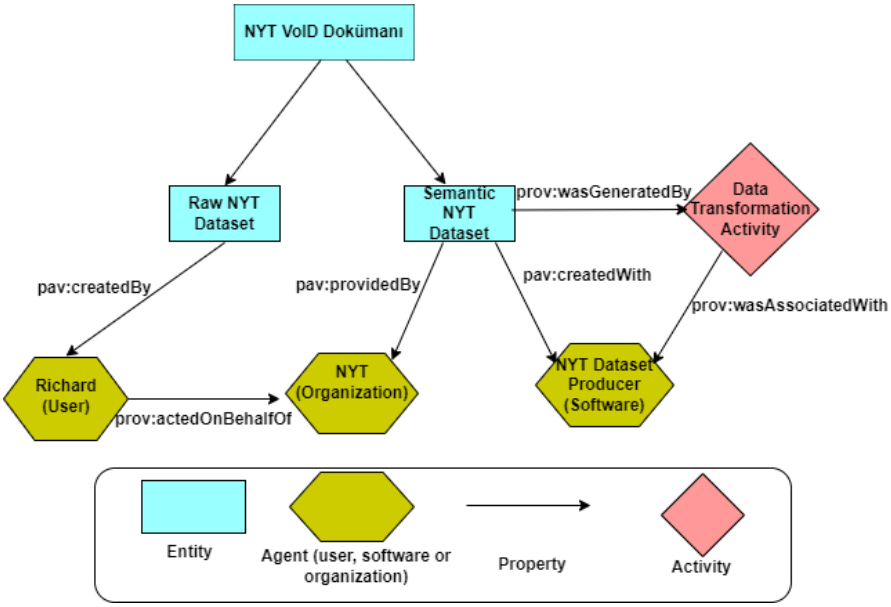

Figure 10. The Scenario Used to Express Provenance on VoID

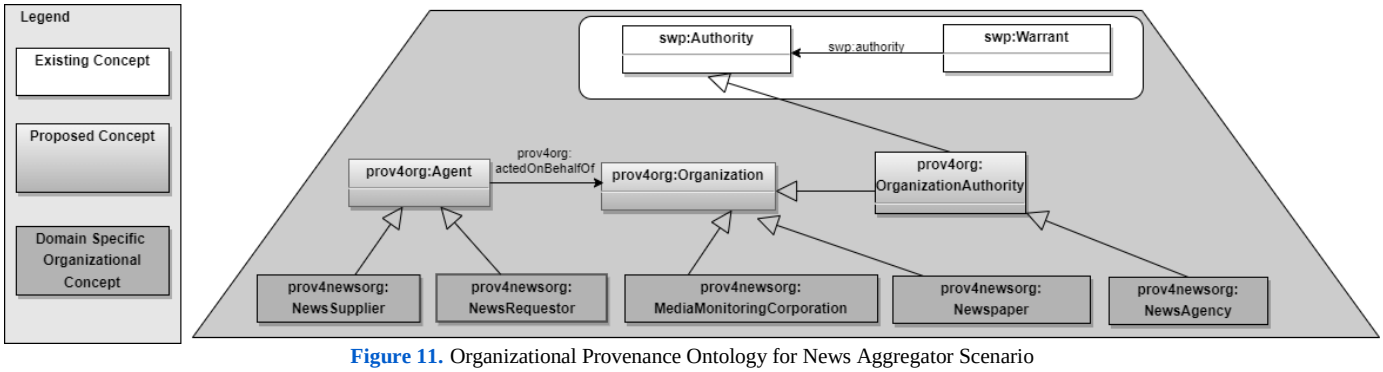

Figure 11. Organizational Provenance Ontology for News Aggregator Scenario

Figure 11 deals with the organizational provenance ontology for the News Aggregator Scenario instantiating the ontology in Figure 7. The agent (*prov4org:Agent*), organization (*prov4org:Organization*), and organizational authority (*prov4org:OrganizationAuthority*) are the fundamental elements that constitute the News Aggregator Scenario in Figure 11. In the scenario, the agent can be expanded as the news supplier (*prov4newsorg:NewsSupplier*) as the person producing the data or the news requestor (*prov4newsorg:NewsRequestor*) as the person consuming the data. The organization can be a newspaper (*prov4newsorg:Newspaper*), a media monitoring corporation (*prov4newsorg:MediaMonitoringCorporation*), or a news agency (*prov4newsorg:NewsAgency*). The organizational authority (*prov4org:OrganizationAuthority*) indicates who is responsible for publishing or modifying the news in the newspaper. In the example, it is anticipated that the news agency will assume the authority role. SWP [38] was used and associated with the organizational provenance ontology to express the concepts related to authority in order to fulfill the trust requirement for the selected scenario. Considering only trust is not sufficient as the sole dimension of provenance where the other dimensions have been mentioned in Figure 9. This ontology has been worked on to be expressed appropriately for the example, so only trust has been taken into account.

5. CONCLUSION

The widespread publication of linked datasets leads to the availability of similar data sources, creating a scenario where the selection of the most suitable and reliable dataset becomes essential. In order to make a selection from similar datasets based on higher quality or personal preferences, data sources need to be considered with certain processes. One of the objectives of this study is to establish processes for selecting datasets using and extending VoID metadata. It is envisioned that VoID should maintain metadata based on provenance-aware quality criteria using additional ontological definitions on VoID. The process of this selection demonstrates how an organization can access the Linked Data Space (LDS) and acquire suitable data or datasets through specific data and processes. In this context, the proposed method is illustrated through a scenario.

Complex rules can be generated for preferences in the organizations within its own dimension or in conjunction with different preferences. The representation of how more preferable datasets can be created will be emphasized by defining the superiority of preferences over each other [39] or by assigning separate weights to provenance preferences [40]. In future work, linked rules and provenance-based preferences will be combined for a more comprehensive provenance tracking.

In future studies, all the proposed methods for provenance will be executed within an LDS, and experiments will be conducted to explore the details of the query process, as well as the applicability and performance of the developed method. Additionally, Trustworthy AI (TAI) will be examined to see how provenance can serve as a means to enhance trustworthiness in LDSs [41]. Furthermore, constructing knowledge graphs and training them in graph neural networks (GNNs) [42], considering provenance might lead to explainability of these graphs by provenance features. Hence, Explainable AI (XAI) also could be examined in our proposed method by projecting the GNNs' decision boundary onto the interpretable feature space [43] from an organizatonal perspective in LDSs.

The actors in the organizations may exhibit certain attitudes that can be learnt by some organizational data. Considering these attitudes [44] may lead to learning and providing data with respect to preference-based biases. Hence, another area for future work could involve learning the attitudes of the actors and adapting the dataset selections in LDSs based on these attitudes.

ACKNOWLEDGEMENT

This study has been conducted as part of the doctoral dissertation titled "Açık Bağlı Veri Sistemlerinde Köken Bazlı Erişim Gerçekleştirimi" [45] in Ege University.

REFERENCES

[1] C. Bizer, T. Heath, T. Berners-Lee, "Linked data: the story so far", Linking the World's Information: Essays on Tim Berners- Lee's Invention of the World Wide Web, Association for Computing Machinery (ACM), New York, NY, United States, pp. 115-143, September 2023.
[2] F. Tekbacak, İ. Korkmaz, "Bağlı veri: veri ağının yapı taşı,"Akademik Bilişim Konferansı, 2015.
[3] E. Curry, S. Scerri, T. Tuikka (ed.), Data Spaces: Design, Deployment and Future Directions. Springer Nature, 2022.
[4] E. Curry, Real-time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems**.** Springer Nature, 2020.
[5] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space. Springer Nature, 2022.
[6] M. Franklin, A. Halevy, D. Maier. "From databases to dataspaces: a new abstraction for information management," ACM Sigmod Record, vol 34, no. 4, pp. 27-33, 2005.
[7] A. Halevy, M. Franklin, D. Maier. "Principles of dataspace systems," Twenty-fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 1-9, 2006.
[8] M. Herschel, R. Diestelkämper, H. Ben Lahmar, "A survey on provenance: what for? what form? what from?," The VLDB Journal, vol. 26, pp. 881-906, 2017.

[9] L.Moreau, P. Groth, Provenance: An Introduction to PROV. Springer Nature, 2022.
[10] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, J. V. den Bussche, "The open provenance model core specification (v1.1)," Future Generation Computer Systems, vol. 27, no. 6, pp. 743-756, 2011.
[11] Internet: O. Hartig, J. Zhao, Provenance Vocabulary Core Ontology Specification, https://trdf.sourceforge.net/provenance/ns.html, 02.12.2023.
[12] Internet: K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao. T.Lebo, S. Sahoo, D. McGuinness (eds.), PROV-O: The PROV Ontology, https://www.w3.org/TR/prov-o/, 02.12.2023.
[13] P. Missier, K. Belhajjame, J. Cheney, "The w3c prov family of specifications for modelling provenance metadata," Proceedings of the 16th International Conference on Extending Database Technology (EDBT'13), Genoa, Italy, pp. 773-776, 18-22 March 2013.
[14] C. Baillie, P. Edwards, E. Pignotti, D. Corsar, "Short paper: assessing the quality of semantic sensor data," Proceedings of the Sixth International Workshop on Semantic Sensor Networks (SSN'13), 1063, pp. 71-76, 22 October 2013.
[15] M. Markovic, P. Edwards, D. Corsar, "Utilising provenance to enhance social computation," 12th International Semantic Web Conference (ISWC 2013), Sydney, NSW, Australia, Springer Berlin Heidelberg, pp. 440-447, October 21-25, 2013.
[16] P. Missier, S. Dey, K. Belhajjame, V. Cuevas-Vicenttín, B. Ludäscher, "{D-prov}: extending the {prov} provenance model with {workflow} structure," 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP 13), pp. 1-7, 2013.
[17] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, j. M. Gomez-Perez, S. Bechhofer, G. Klyne, C. Goble, "Using a suite of ontologies for preserving workflow-centric research objects," Journal of Web Semantics, vol. 32, pp. 16-42, May 2015.
[18] L. McKenna, C. Debruyne, D. O'Sullivan, "Modelling the provenance of linked data interlinks for the library domain", Companion Proceedings of the 2019 World Wide Web Conference (WWW'19**)**, pp. 954-958, May 2019.
[19] Internet: P. Ciccarese, S. Soiland-Reyes, PAV - Provenance, Authoring and Versioning, http://pav-ontology.github.io/pav/, 04.12.2023
[20] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, T. Clark, "PAV ontology: provenance, authoring and versioning," Journal of Biomedical Semantics, vol. 4, pp. 1-22, 2013.
[21] L. Rietveld, W. Beek, R. Hoekstra, S. Schlobach, "Meta-data for a lot of lod," Semantic Web, vol. 8, no. 6, pp. 1067-1080, 2017.
[22] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, "Describing linked datasets-on the design and usage of void, the vocabulary of interlinked datasets," Proceedings of the Linked Data Workshop at WWW09 (LDOW'09), Madrid, Spain, 2009.
[23] Internet: J. Zhao, K. Alexander, M. Hausenblas, R. Cyganiak, Digital Enterprise Research Institute, Vocabulary of Interlinked Datasets (VoID), http://vocab.deri.ie/void, 02.12.2023.
[24] T. Omitola, L. Zuo, C. Gutteridge, I. C. Millard, H. Glaser, N. Gibbins, N. Shadbolt, "Tracing the provenance of linked data using void," Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS'11), pp. 1-7, 2011.
[25] Z. Gu, F. Corcoglioniti, D. Lanti, A. Mosca, G. Xiao, J. Xiong, D. Calvanese, "A systematic overview of data federation systems," Semantic Web, pp. 1-59, 2022.
[26] O. Görlitz, S. Staab, "Splendid: sparql endpoint federation exploiting void descriptions," Second International Workshop on Consuming Linked Data (COLD'11), 782, 2011.
[27] Z. Akar, T. G. Halaç, E. E. Ekinci, O. Dikenelli, "Querying the web of interlinked datasets using void descriptions," Workshop on Linked Data on the Web (LDOW'12), 937, 2012.
[28] L. Heling, M. Acosta, "Federated sparql query processing over heterogeneous linked data fragments," Proceedings of the ACM Web Conference, pp. 1047-1057, Virtual, 25-29 April 2022.
[29] F. Tekbacak, T. Tuglular, O. Dikenelli, "Policies for role based agents in environments with changing ontologies", Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), pp. 1335-1336, Taipei, Taiwan, May 2-6, 2011.
[30] R. C. Erdur, O. Alatli, T. G. Halaç, O. Dikenelli, "Monitoring the dynamism of the linked data space through environment abstraction," 9th International Conference on Semantic Systems (I-SEMANTICS'13), New York, NY, USA, ACM, pp. 81-88, September 2013.
[31] L. F. Sikos, D. Philp, "Provenance-aware knowledge representation: a survey of data models and contextualized knowledge graphs," Data Science and Engineering, vol. 5, pp. 293-316, 2020.
[32] C. Böhm, J. Lorey, F. Naumann, "Creating void descriptions for web-scale data," Journal of Web Semantics, vol. 9, no. 3, pp. 339-345, 2011.
[33] M. Mountantonakis, C. Allocca, P. Fafalios, N. Minadakis, Y. Marketakis, C. Lantzaki, Y. Tzitzikas. "Extending void for expressing connectivity metrics of a semantic warehouse," Proceedings of the PROFILES@ ESWC, Anissaras, Greece, 26 May 2014.
[34] A. Hogan, "Web of data," The Web of Data, Springer International Publishing, pp. 15-57, 2020.
[35] Internet: S. Weibel, J. Kunze, C. Lagoze, M. Wolf, Dublin Core Metadata for Resource Discovery, https://www.rfc-editor.org/rfc/rfc2413, 05.12.2023.
[36] P. Groth, Y. Gil, J. Cheney, S. Miles, "Requirements for provenance on the web," International Journal of Digital Curation, vol. 7, no. 1, pp. 39-56, 2012.
[37] Internet: D. Reynolds (ed.), The Organization Ontology, https://www.w3.org/TR/vocab-org/, 06.12.2023.
[38] C. Bizer, "Semantic web publishing vocabulary (swp) user manual," Freie Universitat Berlin, November 2006.
[39] A. Toniolo, F. Cerutti, N. Oren, T. J. Norman, K. Sycara, "Making informed decisions with provenance and argumentation schemes," Proceedings of the Eleventh International Workshop on Argumentation in Multi-Agent Systems (ArgMAS'14), 2014.
[40] R. Dividino, G. Gröner, S. Scheglmann, M. Thimm, "Ranking rdf with provenance via preference aggregation," 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Galway City, Ireland, Springer Berlin Heidelberg, pp. 154-163, October 8-12, 2012.
[41] A. Kale, T. Nguyen, F. C. Harris Jr, C. Li, J. Zhang, X. Ma, "Provenance documentation to enable explainable and trustworthy ai: a literature review," Data Intelligence, vol. 5, no. 1, pp. 139-162, 2023.
[42] R. Das, M. Soylu, "A key review on graph data science: The power of graphs in scientific studies," Chemometrics and Intelligent Laboratory Systems, vol.240, 104896, 15 September 2023.
[43] K. Mukherjee, J. Wiedemeier, T. Wang, M. Kim, F. Chen, M. Kantarcioglu, K. Jee, "Interpreting gnn-based ids detections using provenance graph structural features," arXiv preprint arXiv:2306.00934, 2023.
[44] M. Soylu, A. Soylu, R. Das, "A new approach to recognizing the use of attitude markers by authors of academic journal articles,", Expert Systems with Applications, vol. 230, 120538, 15 November 2023.
[45] F. Tekbacak, "Açık bağlı veri sistemlerinde köken bazlı erişim gerçekleştirimi," Doctoral dissertation, Computer Engineering Department, Ege University, Izmir, Turkey, 2015.

BIOGRAPHIES

Fatih Soygazi, also known as Fatih Tekbacak, earned a Bachelor of Science and Master of Science degree from the Izmir Institute of Technology, Computer Engineering Department, followed by a Ph.D. from Ege University, Computer Engineering Department. His current research interests are knowledge graphs, linked data, machine learning, deep learning, natural language processing, agent-based software engineering, distributed ledger technologies and knowledge management.

Tuğkan Tuğlular received the B.S., M.S., and Ph.D. degrees in Computer Engineering from Ege University, Turkey, in 1993, 1995, and 1999. He worked as a research associate at Purdue University from 1996 to 1998. He has been with Izmir Institute of Technology since 2000. After becoming an Assistant Professor at Izmir Institute of Technology, he worked as Chief Information Officer in the university from 2003-2007. In addition to his academic duties, he acted as IT advisor to the Rector between 2010-2014. In 2018, he became an Associate Professor in the Department of Computer Engineering of the same university. He has more than 75 publications and an active record of duties with international and national conferences. His current research interests include model-based testing and software quality with machine learning support.

Oğuz Dikenelli currently works as a Professor in Ege University, Computer Engineering Department. His research interests are agent-based software engineering and linked data.