Constructing an Event-Centric Provenance Management Framework for Long-Term Preservation Systems

Authors: Wu Zhenxin, Wenyan Li, Jiang Shiyin

Date: 2016-04-20T00:00:00+00:00

Abstract

Purpose/Significance: This research establishes a provenance management framework for long-term preservation systems to ensure the authenticity, reliability, and usability of archived data through effective provenance information management. Methods/Process: Based on the digital object preservation lifecycle, provenance events are defined; following the OAIS preservation workflow, the provenance management framework is designed; and with events as the core, both the provenance management functional model and provenance information model are developed. Results/Conclusion: The design of an event-based provenance management framework for preservation systems has been preliminarily completed. It complies with relevant standards in the preservation domain while accommodating practical requirements, demonstrating good universality and feasibility for long-term preservation systems; however, its effectiveness and practicality remain to be further validated.

Full Text

Preamble

Volume 60, Issue 6, March 2016
Constructing an Event-centric Provenance Management Framework for Long-term Preservation Systems
Wu Zhenxin[1], Li Wenyan[2], Jiang Shiyin[1]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] King Channels Digital Technology (Beijing) Co., Ltd., Beijing 100096

Abstract: [Purpose/significance] This study establishes a provenance management framework for long-term preservation systems to ensure the authenticity, reliability, and usability of archived data through effective provenance information management. [Method/process] Based on the digital object preservation lifecycle, we define provenance events; following the OAIS preservation workflow, we design the provenance management framework; and with events as the core, we design both the provenance management functional model and the provenance information model. [Result/conclusion] We have preliminarily completed the design of an event-based provenance management framework for preservation systems that not only complies with relevant standards in the preservation field but also accommodates practical requirements, demonstrating good universality and feasibility for long-term preservation systems, though its effectiveness and practicality require further verification.

Keywords: provenance information, provenance event, long-term preservation, lifecycle management, OAIS, PREMIS, PROV semantic model
Classification number: G250.76
DOI: 10.13266/j.issn.0252-3116.2016.06.014

Digital object provenance records the change history of digital objects. Through provenance information, people can comprehensively understand the changes that occur after a digital object' s creation, including the reasons, timing, location, and personnel involved—7W information (what, where, who, when, which, why, how).

Long-term preservation systems for digital resources, as a special class of data management systems, ensure that digital objects remain usable by target user communities after extended periods through ingestion, preservation, management, and other administrative actions. They face greater responsibilities and challenges regarding data authenticity, reliability, and usability. In long-term preservation systems, provenance information can serve multiple functions. On one hand, preservation systems must manage and maintain digital object usability over considerable time; on the other, they must resist the impact of technological changes on both digital objects and the preservation system itself. Format migration, media migration, and technology updates are common preservation strategies. Therefore, whether changes occur to the objects themselves or their environments, provenance can meticulously document these alterations and maintain associations between digital objects before and after such changes. This approach enables preservation systems to effectively manage versions and

derivatives, provide evidence for digital object authenticity and system trustworthiness certification, and support rights management and responsibility attribution. Thus, provenance holds even greater significance for preservation systems.

The authors previously comprehensively summarized and analyzed provenance research in long-term preservation in the article "Research and Application of Provenance Technology in Long-term Preservation" [1], initially proposing a provenance management framework. This paper builds upon that work, outlining how to further improve the design of the long-term preservation system provenance management framework and related functions.

1 Analysis of Provenance Research at Home and Abroad

Foreign research on data provenance technology began relatively early, emerging in the 1990s and spanning multiple fields and disciplines including computer science, geographic systems, biology, and finance. It can be divided into foundational theory and application aspects.

Foundational provenance research includes definitions, organizational models, descriptive vocabularies, and serialization formats. Models represent one of the hottest topics in provenance research. Current general models include W3C PROV-DM, OPM, Provenir, and CRMdig. Additionally, some metadata schemes and ontological vocabularies provide terms for expressing provenance, such as DC metadata, VoID vocabulary, and Provenance Vocabulary. The most representative is PROV-DM, which was released as a provenance standard by W3C in April 2013. This model can be compatible with other provenance models and has gained unanimous recognition across industries, making the exchange and transmission of provenance information between systems—especially large-scale promotion and use in WEB environments—possible, and driving the standardization process of provenance.

Application research on provenance includes technologies, tools, systems, and frameworks for provenance capture, storage, query, and visualization. Numerous tools and systems support provenance management, such as Taverna, VisTrails, REDUX, and VDS, all of which can capture, store, and browse provenance in workflow environments. Y. L. Simmhan et al. [2-3] provided a relatively comprehensive review of these. Provenance frameworks represent another research focus. R. Bose et al. [4] proposed a conceptual framework for scientific workflow provenance management that can automatically capture provenance information from workflow systems, record it as metadata, and store it in corresponding databases for query. B. R. Barkstrom et al. [5] designed a computational framework for earth science data provenance tracking to trace three types of activities in earth science data: creation, custodial history, and intellectual property history, presenting them as directed acyclic graphs. DERI (Digital Enterprise Research Institute) released the general-purpose provenance management framework and system Prov4J [6], which uses semantic web standards

and tools to manage provenance and helps users develop provenance-aware applications [7].

In recent years, domestic research on provenance has also begun, primarily involving provenance technology reviews [8-9], provenance expression, and provenance model analysis. Model research is particularly prominent, with domestic scholars attempting to modify existing models or propose new ones [10], such as security improvements to OPM (Open Provenance Model) [12] and a data provenance model based on DNA double helix structure [13]. However, relatively few domestic studies have developed provenance management practices, particularly process-level provenance management, as well as provenance research and application in specific domains.

In the long-term preservation field, OAIS and PREMIS standards provide conceptual definitions and explanations of provenance but lack practical content such as what to record, relevant technologies, and provenance management strategies. Projects such as DAITSS, CASPAR, and APASEN [1] have begun related explorations, and domestic scholars [14-16] have also started research on provenance capture. Overall, the preservation field still lacks comprehensive, systematic analysis and research on provenance. This paper provides a comprehensive analysis of provenance content, capture, storage, and encapsulation from the perspective of the preservation lifecycle, proposing a comprehensive and complete provenance management framework that we hope will provide useful reference for long-term preservation.

2 Basic Design Approach for the Long-term Preservation System Provenance Management Framework

From the application of provenance in existing long-term preservation systems, we find that different systems record provenance in different ways, with varying content and emphasis, including important operations, responsible persons, timing, and equipment. Meanwhile, OAIS and PREMIS have not provided detailed specifications for provenance description. However, the PREMIS framework and long-term preservation practice have gradually formed an event-centric approach to recording provenance. Therefore, this paper chooses to manage provenance with events as the core, defining provenance as the history of content information that reveals relevant changes occurring after creation [17].

Provenance information management runs through the entire lifecycle of digital objects in OAIS systems. Considering various factors, this paper follows these design principles in the provenance management framework design:

(1) **Based on OAIS.** OAIS is the universal standard for long-term preservation, providing basic processes and events, making it the fundamental starting point for this research.

(2) **Based on the digital object preservation lifecycle.** This paper takes the submission of digital objects to the preservation system as the start-

ing point, implementing provenance collection and management for all changes throughout the entire preservation lifecycle. Provenance information before submission can be submitted to the long-term preservation system in a standardized manner by content producers based on consensus with the preservation party, which is also consistent with OAIS requirements.

(3) **Recording provenance information with events as the core.** During long-term preservation, digital objects generate various events through management activities, and these events typically accompany provenance information. It can be said that events are the primary drivers of object changes, connecting various types of state changes. As events accumulate, the event chain occurring on objects can dynamically present the state changes of preserved objects.

(4) **Interoperability.** The information model is important content in system software design. To enhance interoperability of provenance information between different systems, provenance management should use information models to organize and manage data.

(5) **Universality.** This framework aims to provide general functional processes, models, and other content references for organizing and managing provenance information in long-term preservation, independent of specific technical implementations.

To clearly explain the design principles and key content of the entire management framework, this paper first proposes a provenance event list based on preservation lifecycle management, identifying which events generate provenance information and at which processes management should be implemented. It then uses the OAIS preservation system functional model diagram to further clarify the embedding relationship between the provenance management module and the preservation system functional framework. In the functional model section of provenance information management, it clarifies how provenance management functions can be integrated into OAIS management processes for effective management. Finally, through the provenance information model, it defines the basic components and structure of provenance information for organization and preservation. These four parts basically explain the what, when, where, and how questions of provenance management in preservation systems.

3 Provenance Event List for Preservation Lifecycle Management

Events are the primary drivers of object changes, connecting various types of state changes through events. As events accumulate, the event chain occurring on objects can dynamically present state changes of preserved objects. Therefore, this paper defines provenance events as identifiable actions in preservation systems that involve or affect at least one object or agent. These differ from common computer events such as clicking, double-clicking, or form loading; they

are operations or operation sets defined by preservation systems for processing objects, such as compressing files, ingesting information packages, or creating objects. Identifying which events should be recorded as provenance events is the core issue of the provenance management framework constructed in this paper.

PREMIS [18], as the preservation metadata standard in the preservation field, defines five basic entities, with event being one of them. Therefore, using events to record provenance facilitates description using preservation metadata. PREMIS defines 15 preservation events: creation, deaccession, decompression, decryption, deletion, digital signature validation, dissemination, fixity check, ingestion, message digest calculation, migration, normalization, replication, validation, and virus check. However, these events are not specifically designed as provenance events, and there are overlaps between different events, such as creation and normalization. Some events are ambiguous and may cause 歧义 in application, such as validation. Therefore, in practical application, it is necessary to clearly define which events in long-term preservation systems should be recorded as provenance events.

OAIS states that provenance is the history of content information, showing the origin of content information, changes that have occurred since creation, and changes in custodial responsibility since creation. This definition implies two important criteria for selecting provenance events: "time" and "change." "Time" refers to the preservation lifecycle of digital objects. In OAIS, the entire preservation lifecycle includes six preservation processes: ingestion, archival storage, data management, administration, preservation planning, and access. "Change" is the basis for determining whether an event is a provenance event.

In summary, the following aspects should be considered when selecting provenance events: Events that lead to the initial creation of content objects—a process from nothing to something. Events that cause changes to content objects themselves or can capture naturally occurring changes during long-term preservation, involving content, structure, quantity, format, location, metadata, and custodial responsibility. Events that lead to the creation of new version objects. Although the digital object content itself may not have changed (which is fundamental to long-term preservation), creating new objects closely related to the original, such as copies or different format versions, is beneficial for object reuse. Events that cause changes in digital object rights and management authority. Events that lead to the disappearance of digital objects.

Based on these principles, we have selected provenance events from the processes included in OAIS as shown in Table 1 .

4 Provenance Management Framework Embedded in OAIS Preservation Processes

As shown in Figure 1 [Figure 1: see original paper], provenance events involve all processes and functional modules of OAIS. Provenance information management needs to be embedded throughout the complete preservation system workflow.

In Figure 1, the central part is the provenance management module, which must be embedded into various OAIS processes to dynamically monitor events in each process. According to the pre-configured provenance event list in provenance management, it captures provenance events of digital objects. It then organizes events captured by the preservation system and external events provided by producers into standardized provenance according to the corresponding provenance model, stores them in appropriate formats (encapsulation and storage), and manages them long-term through the preservation management module to ensure provenance integrity, understandability, and long-term accessibility. The application module provides provenance usage for users or other long-term preservation modules (such as authenticity management) according to preservation system requirements.

5 Event-centric Provenance Management Functional Model

Figure 2 [Figure 2: see original paper] clearly shows the relationships and data flows among various functional modules of provenance management. This model contains four basic sub-functional modules: capture, organization, preservation management, and application, with capture, organization, and preservation management being the focus of the functional model.

5.1 Capture Module

The event configuration function is responsible for pre-defining and configuring the types of events that provenance management needs to capture. This function is completed before capture, with preservation system administrators summarizing operations that need to be recorded as provenance based on the functions included in the preservation system, defining them in detail, and configuring the provenance event list into computer-readable formats such as database tables or XML files.

The event monitoring function is responsible for dynamically monitoring all events occurring in the preservation system. When an event matches a predefined event in the provenance event list, it triggers the organization module, passing event information such as event content, event time, operation objects, and equipment used to the organization module.

5.2 Organization Module

Upon receiving event messages, the organization module adds them to the provenance record task queue for extraction use. This asynchronous recording approach both minimizes impact on original system progress and reduces server burden. The provenance record task queue includes two types of events: automatically captured internal preservation system events and external events provided by content producers.

The extraction function reads event information from the task queue in sequence

and generates standardized provenance by organizing event information according to the system-defined information model (such as XML schema).

5.3 Preservation Management Module

The storage management function receives provenance from the organization module and stores it in relevant ways while maintaining associations between provenance digital objects. The version management function manages various versions of provenance, primarily creating copies or supporting format migration according to preservation plans and policies. The provenance audit function regularly performs fixity checks, format checks, and replica checks for each version of provenance. This function is triggered by two types of tasks: regular inspection tasks and tasks triggered by new provenance addition, provenance backup, and provenance version changes.

5.4 Application Module

The application module provides standard interface calls for other system modules (such as audit trails) to use provenance information. The query processing function directly receives user provenance requests, then calls relevant provenance interfaces to return provenance in specific formats to users. The user interaction function provides users with visualization interfaces, such as web pages. When users initiate provenance query or download request messages through the interaction interface, they are passed to the query processing module, which returns the requested specific-format provenance to users through the interaction interface after underlying processing.

6 Event-centric Provenance Information Model

Using an information model not only enables effective organization of managed data but also facilitates long-term preservation, management, and reuse. In OAIS, digital objects include both content information and representation information. Therefore, provenance must record changes to both content information and representation information. Provenance information should include the following relevant content:

(1) **Event.** Events drive changes to digital objects. Detailed event description is the key content of provenance information, including not only event identifiers, detailed descriptions, time, event types, processing equipment, processing results, locations, and reasons, but also basic information about responsible persons and objects involved in the event.

(2) **Digital Object.** Digital objects involved in events need to be completely recorded. The two are associated by referencing the digital object' s identifier in the event, but descriptive metadata of the digital object is not included. If an event simultaneously associates with two or more digital objects, it means all objects possess this provenance information and

should include identifiers of all digital objects.

(3) **Agent Content.** In the narrow sense, agent refers to the operator of an event. Here, agent has a broader meaning, including four subtypes: Person (individual), Organization (institution), Software (software), and Device (physical equipment).

(4) **Relationships Between Digital Objects.** Event operations on digital objects may lead to the creation of new version objects, such as copies or different format versions. Although relationships between digital objects may not be directly recorded in provenance information, version changes of digital objects can be indirectly derived by analyzing the nature of relevant events and the input/output objects involved.

The W7 semantic model [19] provides a 思路 for recording provenance information from seven dimensions, comprehensively explaining the content of provenance information and serving as an important reference for building provenance models. Based on the W7 semantic model and the event provenance content summarized above, this paper designs the event provenance information model shown in Figure 3 [Figure 3: see original paper]. This model describes the content elements that provenance information should include from an event perspective and designs the content concepts included in each provenance event from seven dimensions.

In this information model, events as the core connect various types of information recording digital object changes, covering basic elements of provenance events: Object, Agent, EventID, Date, Reason, Task, Detail, EventOutcome, Category, EventType, and Location. Each element corresponds to one dimension of the W7 model, as shown in Table 2 .

Although Agent and Object are part of the event, the descriptive metadata schemes for both are not within the scope of the provenance management framework model. This model only describes the concepts that provenance events should include.

During system implementation, the authors reused PREMIS OWL [20] and W3C PROV-O [21] to implement the provenance organization model and used RDF for provenance encapsulation. This paper does not cover provenance storage and encapsulation strategies, which will be detailed in another paper by the authors.

Currently, the framework system prototype has just been completed, and its effectiveness and practicality still require further verification. Overall, the design of the event-based provenance management framework for preservation systems not only follows relevant standards in the preservation field but also references many international project practices [1], considering both theoretical and practical aspects, demonstrating good universality and feasibility for long-term preservation information systems. Based on the digital object preservation lifecycle, this framework proposes an approach embedded in OAIS processes and inte-

grated with relevant preservation events. Preservation domain researchers have fully recognized the important role of provenance in long-term preservation and continue exploring how to effectively manage and use it. We hope this research and effort can provide useful reference for relevant personnel in theoretical research and practice of provenance management in long-term preservation.

References:

[1] Wu Zhenxin, Li Wenyan. Research and Application of Provenance Technology in Long-term Preservation[J]. Library and Information Service, 2015, 59(8): 118-125.
[2] Simmhan YL, Plale B, Gannon D. A survey of data provenance in e-science[J]. ACM SIGMOD record, 2005, 34(3): 31-36.
[3] Simmhan YL, Plale B, Gannon D. A survey of data provenance techniques[EB/OL]. [2016-02-16]. http://www.cs.indiana.edu/pub/techreports/TR618.pdf.
[4] Bose R. A conceptual framework for composing and managing scientific data lineage[EB/OL]. [2016-02-16]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1029701.
[5] Barkstrom BR. A mathematical framework for earth science data provenance tracing[J]. Earth science informatics, 2010, 3(3): 167-196.
[6] Prov4J (Provenance for the Web)[EB/OL]. [2015-05-13]. http://prov4j.org/.
[7] Freitas A, Legandre A, O' Riain S, et al. Prov4J: a semantic Web framework for generic provenance management[EB/OL]. [2016-02-16]. http://people.csail.mit.edu/pcm/temp/ISWC/workshops/SWPM2010/paper_8.pdf.
[8] Dai Chaofan, Wang Tao, Zhang Pengcheng. Review of data provenance technology development[J]. Computer Application Research, 2010(9): 3215-3221.
[9] Shen Zhihong, Zhang Xiaolin. Review of semantic web data provenance expression models[J]. New Technology of Library and Information Service, 2011(4): 1-8.
[10] Li Wenyan, Wu Zhenxin. Research and analysis of provenance information models and the PROV standard[J]. Information Studies: Theory & Application, 2015, 38(4): 23-29.
[11] Chen Ying. A data provenance model based on DNA double helix structure[J]. New Technology of Library and Information Service, 2008(10): 11-15.
[12] Liu Tong. Security provenance research based on OPM[D]. Zibo: Shandong University of Technology, 2013.
[13] Chen Ying. A data provenance model based on DNA double helix structure[J]. New Technology of Library and Information Service, 2008(10): 11-15.
[14] Zhu Yi. Research on provenance-aware technology in long-term digital preservation[D]. Wuhan: Huazhong University of Science and Technology, 2013.
[15] Yu Hui. Design and implementation of provenance-aware system for digital archives preservation[D]. Wuhan: Huazhong University of Science and Technology, 2012.

[16] Qin Leihua, Wu Xucheng, Xiao Bo. Research on provenance-aware technology for long-term digital resource preservation[J]. Modern Educational Technology, 2013(12): 102-106.
[17] Consultative Committee for Space Data Systems. Reference model for an open archival information system (OAIS): CCSDS 650.0-M-2[EB/OL]. [2016-02-16]. http://public.ccsds.org/publications/archive/650x0m2.pdf.
[18] PREMIS data dictionary for preservation metadata, version 2.0[S/OL]. [2015-05-13]. http://www.loc.gov/standards/premis/v2/index.html.
[19] Ram S, Liu J. A semantic foundation for provenance management[J]. Journal on data semantics, 2012, 1(1): 11-17.
[20] PREMIS Editorial Committee. PREMIS OWL ontology 2.2 now available[EB/OL]. [2015-03-31]. http://www.loc.gov/standards/premis/ontology-announcement.html.
[21] PROV-O: The PROV ontology[EB/OL]. [2015-05-17]. https://www.w3.org/TR/2013/REC-prov-o-20130430/.

Author Contributions:
Wu Zhenxin: Proposed research ideas, designed paper framework, revised paper, finalized manuscript;
Li Wenyan: Queried and collected relevant papers and materials, wrote initial draft;
Jiang Shiyin: Queried and collected relevant papers and materials.

Note: Figure translations are in progress. See original paper for figures.