Journal of Cyber Security
DOI:10.32604/jcs.2020.010963
Article

Tech Science Press

Guarantee Mechanism of Data Continuity for Electronic Record Based on Linked Data

Yuyi Huo[1], Shi Zhou[3], Ruiguo Hu[1], Yongjun Ren[1,2,*] and Jinyue Xia[4]

[1]School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, 210000, China
[2]Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing, 210000, China
[3]Changwang School of Honors, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[4]International Business Machines Corporation, New York, 10041 NY 212, USA.
[*]Corresponding Author: Yongjun Ren. Email: renyj100@126.com
Received: 10 April 2020; Accepted: 07 September 2020

Abstract: In the field of electronic record management, especially in the current big data environment, data continuity has become a new topic that is as important as security and needs to be studied. This paper decomposes the data continuity guarantee of electronic record into a set of data protection requirements consisting of data relevance, traceability and comprehensibility, and proposes to use the associated data technology to provide an integrated guarantee mechanism to meet the above three requirements.

Keywords: Electronic record; data continuity; guarantee mechanism

1 Introduction

The amount of electronic record that needs to be stored in a big data environment is growing geometrically. The traditional electronic record management system based on structured small-scale data management technology is difficult to effectively utilize the unstructured large-scale electronic record in the era of big data [1–3]. Therefore, how to ensure the availability of electronic record has become one of the major challenges in the research of electronic record management in the era of big data [4,5].

Traditional electronic record management is often functional or target-driven, that is, according to the business objectives of enterprises or organizations, the electronic record management function requirements are decomposed into many function points, and the electronic record management system is developed by realizing these function points one by one [6,7]. For example, retrieval, statistics, authority control, extract, secret level management, programming interface, offline utilization and other function points are the core function points of electronic record management [8,9]. If the business environment is based on the relative stability of electronic documents, the function-driven design method is indeed the preferred solution [10–13]. However, with the advent of the era of big data, the business activities, service contents and data of organizations are in a constantly changing environment. Data-driven mode is about to become a new mode of electronic record management [14]. In this new development mode, data will become an important driver of electronic record management.

Data continuity refers to a set of data protection measures composed of data relevancy, traceability, comprehensibility, and internal connections. Its purpose is to ensure the availability, credibility and control-ability of data, reduce the risk of data misuse, loss of trust and loss of control, with a focus on data quality assurance-better correlation of data, avoiding fragmentation; enhancing data tracking and identification, providing evidence-based data; having reasonable self-describing information, and maintaining the subject's understanding and control.

Data continuity assurance is the key to solving the big data challenges faced by electronic record management described above. First, the business of record requires that electronic record have machine comprehensibility, otherwise it is difficult to define new services based on record content or optimize existing services. Second, data-driven electronic record management requires that the electronic record itself be in an associated and traceable state. Third, data centralization shifts the focus of electronic record management from computing to data. The association, traceability and semantic understanding of electronic record have become their key activities. Finally, the real-time processing power of electronic record depends not only on the choice of technology (such as stream processing techniques such as Spark), but also on the state of the data. That is, the relevance, traceability, and comprehensibility of electronic record.

At present, although the continuity guarantee of electronic record data is put forward, there is still a lack of technical means to provide guarantee. Aiming at this problem, this paper proposes the continuity guarantee mechanism of electronic record data based on related data.

2 Related Work

2.1 Electronic Record

With the mass production of electronic record, electronic record has gradually replaced paper record as the main form of social records, and electronic record management has become an important part of record management. China's research on electronic record management has gone through more than a decade and is becoming increasingly mature and showing Chinese characteristics.

From a theoretical perspective, compared with paper record, electronic record have many differences in characteristics, among which three aspects deserve special attention: First, the authenticity, integrity and validity of electronic record; The second is the change of the separability of electronic record content and carrier and the way it is managed; The third is the integration of the content, background and structural information of electronic record. Fundamentally, the uniqueness of these three aspects is the essence of electronic record, which is the result of the generation, management, preservation or destruction of record based on the "system", and is directly related to the electronic record management system. The generation and circulation of electronic record are aimed at effectively supporting the business activities between the organization and the users. The management purpose is firstly to obtain business recognition. Considering the various business systems to which electronic record are attached, the key and difficult aspects of authentic and trusted maintenance are business process integration and version control. Important business activities have corresponding rules and regulations for the management of their record, which are used to control the versioning, process control, audit tracking and other work of business documents.

Like security, continuity is an important attribute of data resources and one of the primary tasks of data management. However, the arrival of the era of big data makes the problem of fragmented electronic record, garbage data and data islands increasingly prominent, and the loss of use, credibility and control of electronic record have become a new challenge to the management of electronic record, and the data continuity of electronic record has become an important subject to be studied urgently.

2.2 Linked Data

The associated data uses the RDF (Resource Description Framework) data model, which uses URIs (Uniform Resource Identifiers) to name data entities, publish and deploy instance data and class data, Thus, the data can be revealed and obtained through a hypertext transfer protocol, at the same time, it emphasizes the interrelation of data, the interrelation and the contextual information that is beneficial to people and computers.

Linked data can create links between data from different sources. These data sources may be databases maintained by two geographically located organizations, or they may be different systems within an organization that are not interoperable at the data level. Linked data is machine readable and unambiguous and linked to other external data sets, as well as linked from data from external data sets.

The associated data network is different from the current hypertext network, the basic unit of a hypertext network is a hypertext markup language (HTML) file linked by hyperlinks. Linked data is not simply connecting these files, but using RDF to form a network that links anything in the world, namely the data network, which can be described as a network of online data describing all the entities in the world. The emergence of the associated data network not only expands the current hypertext network, but also discriminates, selects and locates the confusing information resources on the current network. The three modes of data storage are shown in Tab. 1.

Table 1: Three modes of RDF data storage

Storage mode	SG single image mode	MG multi-picture mode	DH node mode
Numbers of nodes	Single SPARQL Endpoint	Single SPARQL Endpoint	Multiple SPARQL Endpoint
Numbers of figures	Single Graph	Multiple Graph	Single/Multiple Graph
Numbers of data sets	Single/Multiple Dataset	Single/Multiple Dataset	Single/Multiple Dataset
Interview method	Remote call	Remote call	Remote call
Data volume distribution	Small amount of data	Large amount of data. Each graph can be used separately	Distributed

3 Problem Description

The difficulty of the main challenges facing the electronic record management system in the era of big data is to achieve the following four transformations: The first is the transition from the document of the business to the commercialization of the record. The second is the transition from target-driven to record-driven. The third is the transition from computation-centric to data-centric. The fourth is the transition from offline processing to real-time processing. There are many problems that need to be addressed to achieve these four transitions, but the most important thing is to focus on the continuity of electronic record management. Data continuity is not only a prerequisite for record business and data driving, but also a core problem that needs to be solved in a data-centric design pattern and real-time processing.

The connotation of electronic record data continuity guarantee is shown in the following table. As can be seen from Tab. 2, the relevance, traceability and comprehensibility are related to each other, from space to time, from structure to semantics, to ensure the integrity and usability of electronic record, so that electronic record can be found with data, and are true and effective. Different from the theory of digital continuity centered on long-term preservation, the theory of data continuity further emphasizes the continuity of the content and semantic level of electronic record.

From the connotation of data continuity guarantee, it can be seen that the following three core issues should be studied in the research of electronic document management for the new challenges of the big data era. The first is the guarantee of the relevance of electronic record. The second is the traceability guarantee of electronic record. The last is the guarantee of comprehensibility of electronic record. In order to provide the above three aspects of protection, this paper uses the associated data technology to provide the corresponding technical support.

Table 2: The connotation of electronic record data continuity guarantee

Meaning/attribute	Relevance	Traceability	Comprehensibility
Basic motivation	Prevent “disuse” of electronic record	Prevent “loss of trust” of electronic record	Prevent “out of control” of electronic record
Main connotation	Continuity between different record objects	Continuity between historical versions of the same record	Continuity between record and their production, management, maintenance subjects (people, computers)

Analysis dimension	Space	Time	Semantics
Main purpose	Open association, cross-domain access	Evidence chain management, credibility assessment, predictive analysis	Self-describing and self-contained information
Key technology	Linked data	Data traceability	Semantic web

4 Guarantee Mechanism of Data Continuity for Electronic Record Based on Linked Data

The theory of connected data provides a theoretical basis for the study of data continuity, especially the relevance of electronic record. The practice of data engineering based on the associated data set has important reference significance for the data continuity of electronic record, especially the implementation method and guarantee mechanism of data relevance.

At the same time, the associated data broadens the theory and technology of data traceability. The data traceability method based on associated data lays a good foundation for data continuity, especially the design of data traceability. At present, there are tracing methods based on annotation, tracing methods based on inverse function, tracing methods based on bit vectors and so on. Among them, the annotation-based tracking method is relatively simple, and it is also relatively easy to implement. At present, there are certain applications.

The Resource Description Framework (RDF) is the cornerstone of the development of the Semantic Web and is a standardized language used to describe metadata for network resources. It is intended to describe the resources and their relationships. The associated data uses the RDF description language, which uses Uniform Resource Identifiers (URIs) to identify things and describe resources with attributes and attribute values. The description of a resource is a statement of the attributes of the resource and the value of the attribute, called a statement. It uses a specific set of terms to express the various parts of the statement. The part of the statement of things used to identify things is called the subject. The part used to distinguish the different attributes of the stated object is called the predicate. The part of the statement that distinguishes the values of the individual attributes is called the object. The object can be either an attribute value or a resource object. The associated data is described as objects whenever possible, which is beneficial for establishing connection of data.

Resource objects in associated data are divided into information resources and non-information resources. Information resources themselves are information, such as pictures, web pages, etc., and generally have representations that can be accessed by HTTP, such as different formats, protocol properties, or natural language. Non-information resources refer to the concept of the real world outside the Web. For non-information resources, the associated data assigns it a Uniform Resource Identifier (URI) that cannot be directly referenced by the HTTP protocol. The URI points to not the non-information resource itself, but the information resource associated with it. The interoperability between resource objects links different resource objects, resource object forms and their information resources with non-information resources, thus forming a wide data network and providing a basis for data sharing. This type of data understanding occurs both within a data set and across data sets.

The main goal of RDF is to provide a framework for enabling different domains to define their own metadata elements, while providing a machine-understandable representation that facilitates data exchange in a big data environment. That is, RDF provides a metadata solution for web data integration. In RDF, a resource can be of any type, a property of a resource is a special kind of resource, a value of a property is also a resource, and even a statement can be a resource, and each resource has a unique URI reference. In order to be able to fuse different metadata sets, RDF is designed to allow anyone to define metadata to describe a particular resource. Since there are more than one attribute of a resource, it is generally a definition of a metadata set, which is the set of words in RDF. It includes various metadata sets such as DC metadata, ontology, classification tables, thesaurus and so on. A vocabulary is also a resource that can be uniquely identified using a URI. Thus, when using RDF to describe resource attributes, you can use a variety of different vocabularies, just by specifying them with a URI.

Since RDF only provides a primary semantic representation, there is no uniform label to support a more specific description of the semantic relationship, therefore, a unified knowledge organization system standard that supports more specific semantic relationships and flexible extensibility needs to be established on the basis of RDF. When the history records the value of the attribute, in order to ensure the consistency of the description and its relevance, the values are specified from a specific vocabulary. This makes it easy to merge and fuse with other RDFS data in the Semantic Web, providing support for interoperability between thesaurus and between the thesaurus and other vocabularies.

5 Conclusion

In the field of electronic record management, especially in the current big data environment, data continuity has become a new topic that is as important as security and needs to be studied. This paper decomposes the data continuity guarantee of electronic record into a set of data protection requirements consisting of data relevance, traceability and comprehensibility, and proposes to use the associated data technology to provide an integrated guarantee mechanism to meet the above three requirements.

Funding Statement: This work is supported by the NSFC (61772280), the national training programs of innovation and entrepreneurship for undergraduates (Nos. 201910300123Y, 202010300200), and the PAPD fund from NUIST. Yongjun Ren is the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. J. Ren, F. J. Zhu, S. P. Kumar, T. Wang, J. Wang *et al.,* "Data query mechanism based on hash computing power of blockchain in Internet of Things," *Sensors*, vol. 20, no. 1, pp. 207, 2020.

[2] Y. P. Liu, Y. J. Ren, C. P. Ge, J. Y. Xia and Q. R. Wang, "A CCA-secure multi-conditional proxy broadcast re-encryption scheme for cloud storage system," *Journal of Information Security and Applications*, vol. 47, pp. 125–131, 2019.

[3] Y. J. Ren, Y. Leng, F. J. Zhu, J. Wang and H. J. Kim, "Data storage mechanism based on blockchain with privacy protection in wireless body area network," *Sensors*, vol. 19, no. 10, pp. 2395, 2019.

[4] C. P. Ge, Z. Liu, J. Xia and L. M. Fang, "Revocable identity-based broadcast proxy re-encryption for data sharing in clouds," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[5] C. P. Ge, S. Willy, Z. Liu, J. Xia, P. Szalachowski *et al.,* "Secure keyword search and data sharing mechanism for cloud computing," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[6] L. M. Fang, Y. Li, X. Y. Yun, Z. Y. Wen, S. L. Ji *et al.,* "A novel authentication scheme to prevent multiple attacks in SDN-based IoT network," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5745–5759, 2019.

[7] Z. G. Qu, Z. Y. Li, G. Xu, S. Y. Wu and X. J. Wang, "Quantum image steganography protocol based on quantum image expansion and grover search algorithm," *IEEE Access*, vol. 7, pp. 50849–50857, 2019.

[8] Z. G. Qu, Z. W. Cheng, W. J. Liu and X. J. Wang, "A novel quantum image steganography algorithm based on exploiting modification direction," *Multimedia Tools and Applications*, vol. 78, pp. 7981–8001, 2019.

[9] S. B. Zhang, Y. Chang, L. L. Yan, Z.W. Sheng, F. Yang *et al.,* "Quantum communication networks and trust management: a survey," *CMC*, vol. 61, no. 3, pp. 1145–1174, 2019.

[10] Y. Chang, S. B. Zhang, G. G. Wan, L. L. Yan, Y. Zhang *et al.,* "Practical two-way QKD-based quantum private query with better performance in user privacy," *International Journal of Theoretical Physics*, vol. 58, no. 7, pp. 2069–2080, 2019.

[11] Y. Chang, S. Zhang, L. Yani, G. Han, H. Song *et al.,* "A quantum authorization management protocol based on EPR pairs," *Computer, Material & Continua*, vol. 59, no. 3, pp. 1005–1014, 2019.

[12] Y. C. Mao, J. H. Zhang, H. Qi and L. B. Wang, "DNN-MVL: DNN-Multi-View-Learning-based recover block missing data in a dam safety monitoring system," *Sensors*, vol. 19, no. 13, pp. 2895, 2019.

[13] W. Zhao, J. Liu, H. Guo and T. Hara, "Edge-node-assisted transmitting for the cloud-centric internet of things," *IEEE Network*, vol. 32, no. 3, pp. 101–107, 2018.

[14] Z. Sun, Y. R. Bi, S. L. Chen, B. Hu, F. Xiang *et al.,* "Designing and optimization of fuzzy sliding mode controller for nonlinear systems," *Computer, Material & Continua*, vol. 61, no. 1, pp. 119–128, 2019.