Copyright © 2019 Tech Science Press JBD, vol.1, no.1, pp.9-15, 2019

Network Electronic Record Management Based on Linked Data

Jian Qi[1], Yongjun Ren[1, 2, *] and Qirun Wang[3]

Abstract: With the rapid development of e-government, e-commerce and self-media, a large number of electronic network records have been generated on the Internet. How to archive these resources has become a focus of attention in academic circles. Based on the theory of linked data, this paper analyzes the archiving of electronic records, and proposes that the corresponding network electronic records should be collected using the theory of linked data. After the resource description framework and corresponding international standards are utilized to convert the network resources, the electronic network records are stored in common databases to solve the problem of network record archiving management.

Keywords: Electronic record, linked data, storage technology, RDF.

1 Introduction

Electronic records are also called electronic documents, which are generated in digital devices and environments, stored digitally on magnetic tapes, disks, optical disks, etc. Moreover, electronic records rely on digital devices such as computers to read, manipulate and transmit over communications networks. With the rapid development and widespread application of Internet and mobile Internet technologies, a large number of electronic network records have been formed, such as web page records, blog information and so on.

Web pages are multimedia digital information such as texts, images, audio and video, which are directly formed by social organizations or individuals in social practice activities. They are bright and definite original records of social activities in the past and have unique original records. This is the same as the essential nature of the record, determines the record attributes of the web page. The archived network information resources are an essential part of the digital resources of archives, and the resources of the archives together constitute the overall resources of the national archives. The digital archives of a web page can effectively prevent the lack of social memory. As professor Huiling Feng said: The most significant opportunity and challenge for the construction of

[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 (CICAEET), Nanjing University of Information Science & Technology, Nanjing, 210000, China.

[3] School of Engineering and Technology, University of Hertfordshire, Hertford, UK.

[*] Corresponding Author: Yongjun Ren. Email: renyj100@126.com.

JBD. doi:10.32604/jbd.2019.06290 www.techscience.com/jbd

archive resources in today, sera is the rise of digital archive resources, which adds significant new perspectives and new elements to the concept of archive resources. How to identify and recognize in the digital world? How to judge the validity and value of electronic files and files and how to set up the relationship between digital resources and physical resources need to be clarified. Also, the archived web pages carry not only social memories but also many historical evidence are stored, and the voucher values of these web pages reflect visible file attributes. Therefore, how to manage such network electronic files is a frontier issue, and the research work in this paper is carried out.

2 Related work

In November 2012, the National Digital Information Infrastructure Conservation Plan (NDIIPP) of United States released a report "Science in Danger: Building a National Strategy for Preserving Online Scientific Content" [Xu, Zheng and Gao (2017); Wu, Zhang, Xie et al. (2015); Rumsey (2012)], which explicitly promoted the preservation of online scientific content as the United States National strategy. One of the earliest projects in the United States to conduct Web Archive was the Internet Archive, a project led by a non-profit organization that has long-term storage and is publicly available to the public for free, with more types of resources stored, including web pages, music, animation, and other electronic resources. It started saving web resources in 1996 and was open to the public in 2001 when it developed the way back machine. Internet Archive is currently the largest web archive project in the United States and works closely with many governments and libraries under the umbrella of the International Internet Preservation Consortium (IIPC). The acquisition strategy of the IA project in the United States is to collect extensively, that is to use the acquisition software to traverse the URL, but sometimes it is also carried out by combining unique collection or extensive collection with a unique collection for some unexpected problems [Wang and Shi (2013)]. The acquisition software used is Heritrix, which is an exclusive web collection and archiving software. It has a sealed package that can be used directly and can be developed by the source code for two times as required. National libraries of countries such as Britain, Canada, and France, use the software to collect web pages. Internet Archive provides an advanced search function of the URL for the saved web page, and the user can display the result automatically through the way back machine [Wang (2012)]. Internet Archive cooperates with SUN in storage and uses the modular datacenter of SUN. In 1997, the National Library and Archives of Canada started to establish and collect electronic resources. The Canadian National Library and Archives (LAC) started collecting the resources of the Canadian government website in 2005. The strategy of collecting is the government resources within the scope of Canadian domain names. The software for collecting the data is Heritrix, and the indexing software is NUTCHWAX, they are open source. WAYBACK is a software tool, which can be used to organize and reproduce web pages [Wang and Ding (2012)]. The royal archive of the Royal Swedish Library is Kulturarw3 (Cultural Heritage Cubed). The project started in 1996 and adopted a selective acquisition strategy that was mainly collected the web resources in the domain name of se. The collection software is Heritrix, and the web page display is also implemented using WAYBACK software [Guo, Mu and Wang (2015)].

The first task of webpage archiving is webpage intake, and some articles are also called webpage collection or harvesting-page collection based on specific strategies and acquisition software tools. The acquisition strategy is divided into selective acquisition, batch acquisition, and hybrid acquisition. The selective acquisition is the selected collection of objects and the scope of a particular frequency web page collection; the current selection criteria have been adopted, including topics or resource types. The collection strategy is used at PANDORA in Canada, Japan, and Australia libraries. The batch acquisition does not specify a specific topic or resource type, and web pages around the world are collected. Internet Archive (IA) [Brown (2018)] is more famous in the United States. The hybrid acquisition is the simultaneous use of several acquisition strategies. For example, the Library of Congress's MINERVA project includes a method for the selective archiving and collection of global snapshots. The Royal Danish Library takes a multi-pronged approach with three different types of acquisition methods: one year for the domain name “DK” overall global harvesting, high-quality selective harvesting of about 80% of the sites, and thematic harvesting of 23 events each year [Gu, Mu and Wang (2015)]. For the current domestic situation, it is better to adopt a hybrid strategy.

At present, storage technologies are evolving from the original paper, optical disks, and magnetic media and data storage centers to cloud storage. Cloud storage integrates the originally distributed technologies, clustering technologies, gridding technologies, and virtualization technologies into networks. Different heterogeneous storage devices in the application management work together and provide highly scalable mass storage and access [Ren, Shen, Wang et al. (2015)]. Cloud storage [Xu, Jiang, Wang et al. (2018)] is divided into a public cloud, private cloud, and hybrid cloud. ① Public cloud storage. Public cloud storage can be handled by a professional company that provides abundant storage at a low cost. Cloud service providers can allocate separate storage space for each customer. Each customer’s application is private and independent, and public cloud storage can also allocate a portion of storage space as required for private cloud storage. ② Private cloud storage. When private cloud storage is available, the library or archive itself can own or control the infrastructure and can be deployed for different applications. Private cloud storage can be deployed as needed in government departments, libraries or archives data centers. Private cloud storage management, the library or the archives can be responsible for the technical department, but also by a professional cloud management service provider. ③ Hybrid cloud storage. It is a combination of public and private clouds. According to customer needs for access, especially for the need to temporarily configure the larger capacity. In this case, a portion of the capacity can be configured as a private cloud from the public cloud [Ren, Shen, Liu et al. (2016)].

3 Linked data

Linked Data, proposed by Tim Berners-Lee who is the father of the World Wide Web, has now become a research hotspot in many fields such as information management, information systems, computer science, and library science [Ruckhaus, Vidal, Castillo et al. (2013)]. Linked data, as a way to publish data, can be viewed as an implementation of a semantic web. It facilitates the development of the World Wide Web by correlating data from different areas. Linked data is technically not complicated, yet it is making a profound change to the World Wide Web. Currently, Linking Open Data (LOD) project has successfully semi-automatically converted the data (including Wikipedia, geo-datasets, government datasets, etc.) on more than 13 billion traditional web pages into related data and built vast data network. Relevant data make not only many new applications into practice but also provide new opportunities for application in some specific fields [Li (2017)].

Linked Data is a specification recommended by the Internet Society (W3C) to publish and link data, information, and knowledge of all kinds. It aims to build on the existing World Wide Web to create a data mapping all natural, social and spiritual worlds Network that enables the Internet to evolve into a semantic [Wang, Zhang, Zhang et al. (2018)**]**, interoperable sea of knowledge by providing a machine-readable description of everything in the world and its interrelationships so that anyone can access the computing facilities of the entire Internet and Computing power to accurately, efficiently and reliably find, share and utilize these interrelated information and knowledge to a greater extent.

Technically, linked data is a way to publish any “resource” on the World Wide Web. The Semantic Web defines resources as “anything that has a URI identity”, and it is divided into two categories: information resources and non-information resources, which are used to express any information and usually exist in some kind of coded file; non-information resources to refer to all kinds of entities in the world can be all objects of the spiritual world (concept, idea, abstract entity, etc.) created by nature, human society and human consciousness.

4 Network electronic record management based on linked data

RDF (Resource Description Framework) is a series of standard specifications developed by W3C to promote the application of Semantic Web, including RDF abstract model and a set of RDF coding format specifications such as Turtle, N-Triples, JSON, N-Quod, et al. [Elbashir, Aboelhassan, ALI et al. (2018)]. RDF is conducive to data sharing, reuse, and semantic interoperability [Wang, Wang, Guo et al. (2018)], more and more systems and applications began to use RDF to transform the underlying data structure. How to convert the RDB data in the relational database to the RDF data format (RDB2RDF for short) is convenient, fast and accurate, while the data in a large number of legacy systems is stored in the Relational Database (RDB), is often the problem of practical concern. The W3C set up the RDB2RDF working group to develop two standards, DM (Direct Mapping) and R2RML (RDB2RDF Mapping Language), to standardize the implementation of RDB2RDF and became the W3C recommendation in June 2012. For two years, these two standards have been applied in some relational database products and many RDB2RDF tools.

The RDB2RDF working group recommends two RDB2RDF mapping languages: DM and R2RML, which defines various rules for how data in a relational database can be transformed into RDF data, including the generation of URIs, the definition of RDF classes and attributes, the processing of empty nodes, the relationship between data expression and so on. DM is a direct mapping method, which outputs the relational database table structure and data directly into an RDF graph (RDF Graph, which can be seen as a collection of multiple triplets). The RDF graph is a complete reflection of the data structure of a relational database. A relationship table in the database is converted to an RDF class, and a field of table is converted to an RDF property. Moreover, the terms are used to represent classes and properties are aligned with table names and field names in the relational database. DM mapping is usually generated automatically by the program.

Unlike DM, R2RML has a high degree of customization and flexibility. R2RML defines a systematic, logical framework for RDB2RDF mapping. The concept of "Logical Table" is proposed. A table in a relational database, a view, and even a valid SQL query is defined as a "logical table" which breaks through the physical structure of the relational database table. Before the RDF data is generated, the data in the RDB can be calculated, filtered, cleaned, and integrated. It lays the foundation for creating RDF data flexibly on demand without changing the original structure of the database. R2RML mappings typically exist as a human-editable text file encoded in turtle syntax and mapping the data structure of a relational database to an existing ontology vocabulary. Through R2RML mapping, a relational database can be output as an RDF graph. The class names and attribute names used in the RDF graph can be derived from existing ontology vocabularies such as FOAF, DC, SKOS, etc., and the terms in the ontology vocabulary can be flexibly selected according to specific needs. R2RML records can be easily edited by hand to change the mapping rules.

RDB2RDF mapping in the practical application of the system generally has two modes. In some application system architecture design, the RDF dataset derived from the RDB need to be exported to the local, and then into the professional RDF Store (also known as Triple Store). The approach is also called extract-transform-load (ETL), referred to as ETL mode. As the original system data is still updated continuously, the approach often cannot provide the latest data in real time. In another application architecture design, a virtual RDF data view and SPARQL query interface need to provide, and the query result of the RDF data is returned. That is, adding an RDF data encapsulation layer on the original server. While in the data query, the SPARQL query request is usually converted to a SQL query language based on a pre-defined mapping rule. The query result is then converted to RDF data. The process is called the "real-time transformation" mode.

As the storage and management system of data, the database is still used by a large number of application systems for its mature and stable function and performance. Although it is not the most suitable data management system for RDF data, there is still a large amount of data stored in it. For those applications that are still growing and being used, it is best not to change the original system business processes and data structure, and provide RDF data based on semantic RDF data, which requires a timely generation or deriving RDF data from the original relational database. This transformation will involve RDB data format to RDF data format, namely RDB2RDF. For most relational databases, management of RDF data and conversion of RDB2RDF can be done through third-party plug-ins or tools.

Oracle has two options for storing and retrieving RDF data, one based on RDBs and the other based on NoSQL. The representative product based on RDB is Oracle Spatial and Graph, which supports not only geospatial data, RDF data storage management and SPARQL retrieval for the semantic web but also supports related data applications, the creation of native RDF data, importing existing RDF data and RDB2RDF way. That is to say, existing data stored in relational database tables are converted to RDF data in real time according to the pre-defined mapping rules. In the definition of transformation rules, W3C and DM standards are supported. Morph is a suite of semantic technology tools developed and maintained by the Ontology Engineering Working Group, which includes the RDB2RDF component morph-RDB and supports R2RML. In the generation of RDF data, it supports real-time conversion mode and ETL mode. In the real-time conversion mode, it needs to convert each front-end SPARQL query request into the corresponding SQL query request according to the defined R2RML mapping. In ETL mode, it needs to allocate the RDBs according to the defined R2RML mappings. The data is converted to RDF data. Morph-LDP is another component of the morph family, which can be considered as an extension of morph-RDB. Another standard of W3C, the Linked Data Platform (LDP), is an extension of LDP, further clarification and expansion, the provisions of how to get more resources on the URI through the resources of information, but also on the information to be updated, including ways to update and permission control. The purpose of Morph-LDP is to implement LDP based on R2RML so that the data in the RDB can be seamlessly integrated into the LDP environment. The RDF data generated from the RDB according to the R2RML mapping is presented in the LDP standard encapsulation to the foreground, and then it receives and processes the SPARQL query or update request encapsulated by the client regarding the LDP standard.

5 Conclusion

Based on the theory of linked data, this paper analyzes the archiving of electronic records and proposes that the corresponding network electronic records should be collected by using the theory of associated data. After the resource description framework and corresponding international standards are utilized to convert the network resources, the electronic network records are stored in common databases to solve the problem of network record archiving management.

Acknowledgement: This work is supported by the NSFC (61772280), and the PAPD fund from NUIST.

References

Brown, H. L. (2018): Internet Archive (2018): http://archive.org/.

Elbashir, K. M.; Aboelhassan, A. M.; ALI, A. (2018): Mapping relational database to resource description framework using direct mapping method. *Gezira Journal of Engineering and Applied Sciences*, vol. 11, no. 2.

Guo, S.; Mu, J.; Wang, S. (2015): Research on archival resources in cloud storage mode based on Hadoop. *Lantai World*, no. 11, pp. 22-23.

Li, Q. (2017): From geomatics to urban informatics. *Geomatics and Information Science of Wuhan University*, vol. 42, no. 1, pp. 1-6.

Ren, Y.; Shen, J.; Liu, D.; Wang, J.; Kim, J. (2016): Evidential quality preserving of electronic record in cloud storage. *Journal of Internet Technology*, vol. 17, no. 6, pp. 1125-1132.

Ren, Y.; Shen, J.; Wang, J.; Han, J.; Lee, S. (2015): Mutual verifiable provable data auditing in public cloud storage. *Journal of Internet Technology*, vol. 16, no. 2, pp. 317-323.

Ruckhaus, E.; Vidal, M.; Castillo, S.; Burguillos, O.; Baldizan, O. (2013): Analyzing linked data quality with LiQuate. *Lecture Notes in Computer Science*, vol. 8186, pp. 629-638.

Rumsey, A. S. (2012): *Toward a National Strategy for Preserving Online Science [EB/OL]*.

http://www.digitalpreservation.gov/meetings/documents/othermeetings/science-at-risk-NDIIPP-report-nov-2012.pdf.

Wang, F.; Shi, H. (2013): Progress of foreign research and practice in web archive. *Journal of Library Science in China*, vol. 39, no. 204, pp. 36-45.

Wang, M.; Wang, J.; Guo, L.; Harn, L. (2018): Inverted XML access control model based on ontology semantic dependency. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 465-482.

Wang, S. (2012): Development research on internet archive-web archiving project in the United States. *Lantai World*, no. 17, pp. 18-19.

Wang, S.; Ding, Y. (2012): The research on web page filing project in Canadian library. *Archives Science Study*, no. 6, pp. 83-85.

Wang, S.; Zhang, L.; Zhang, Y.; Sun, J.; Pang, C. et al. (2018): Natural language semantic construction based on cloud database. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 603-619.

Wu, Z.; Zhang, Z.; Xie, J.; Hu, J. (2015): Developing web archive system of international institutions based on IIPC open source software. *New Technology of Library and Information Service*, vol. 257, no. 4, pp. 1-9.

Xu, F.; Zheng, Q.; Gao, Y. (2017): Research of web archive scheme based on cloud storage. *Journal of Computer Era*, no. 4, pp. 21-28.

Xu, J.; Jiang, Z.; Wang, A.; Wang, C.; Zhou, F. (2018): Dynamic proofs of retrievability based on partitioning-based square root oblivious RAM. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 589-602.