International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 5 376 – 379

Data Representation Model for Management and Distribution of Scientific Data

Focusing on Management and Distisribution of Visible Korean

Pyung Kim
Dept. of Computer Education
Jeonju National University of Education
Jeonju, Korea
pyung@jnue.kr

***Abstract*—**Scientific tools and computer simulations enable rapid creation of various types of data and a number of studies have been conducted on data provenance and web-based data representation models to enhance the distribution, reproduction and reusability of scientific data. Ontology is a knowledge representation model, which is also used as data and workflow technology for data provenance. In this study, as part of managing and distributing for scientific data studies, metadata and data representation model were defined for the management and distribution of Visible Korean online. In addition, additional metadata required for re-distributing the user data created through the Visible Korean study is defined using an ontology-based data representation model, and an RDFa-based web page generation method is proposed to search and extract data from existing web pages. This study enables to manage and distribute the Visible Korean online, which has been managed and distributed offline, and a virtuous recycling of distributing research results as wells.

Keywords*–*Data Representation Model; Scientific Data; Data Provenance

*****

I. INTRODUCTION

Scientific tools and computer simulations enable the rapid creation of various types of data and as types of scientific data become more diverse and a vast amount of scientific data is rapidly generated, various techniques are required to analyze and manage the data[1]. Various studies have been actively carried out to store, manage and distribute the scientific data rapidly increasing in its types and quantity. In order for scientific data to be searched and distributed online, data representation models are required and for data reuse, utilization and distribution, a metadata describing the scientific data and workflows must be defined. Researches on data provenance that supports the whole process of generation, distribution, and utilization of scientific data have been actively conducted [2,3].

Ontology is a knowledge representation model, which is also used as data and workflow technology for data provenance [4,5]. It can express not only the attributes of data but also the data generation and conversion process through workflow. Also, as it is used as a standard model for data disclosure through the web, we also use the ontology to represent metadata and workflow of scientific data.

Visible Korean is a project to support the construction of anatomical image of human body and its utilization in research. Though the anatomical images constructed as a result of the project have been used for various researches, they are currently distributed through offline, which hinders efficient management of the research results [6]. In this study, we have modeled ontology to describe metadata and workflow around the data of Visible Korean for the management and distribution of scientific data.

Chapter 2 introduces scientific data, data provenance to support generation to management of the scientific data, Visible Korean project, and data representation model. Chapter 3 describes metadata and DB schema for management and distribution of Visible Korean. Chapter 4 describes how to improve the management and distribution process of Visible Korean and an ontology modeling for metadata and workflow representation. Chapter 5 discusses results of this study and future studies.

II. RELATED STUDIES

This chapter describes scientific data and data provenance, Visible Korean, and studies related to data representation for metadata distribution on the web.

A. *Scientific Data and Data Provenance*

Scientific data includes all data produced in the course of conducting a study, inclusive of statistical values, formulas, images, charts, documents, and so on. The amount of scientific data is rapidly increasing in various fields, and researches are being conducted to prevent problems in searching and using scientific data and to improve the reusability of scientific data[7]. Data provenance technology enables data reuse, data trust, and data representation through data and workflow representations. Recently, studies about data and workflows have been conducted in various environments such as spatial data [8], distributed relational tables [9], and data provision in the cloud environment [10].

B. *Visible Korean*

KISTI (Korea Institute of Science and Technology Information), having been working with Ajou University since 2000 under the support of the Korean government, has produced a continuous sectional image of a Korean by cutting the entire body of a Korean male body with a continuous cutter and taking images of sections. The sectional images are provided as CT, MRI, sectioned images and 3D images to domestic and foreign users. In 2007, continuous sectional images of a male head and a female pelvis were produced and in 2009, sectional images of a female whole body were produced. In 2012, segmented images of human anatomy for male and female whole bodies and 3-D images of continuous sectional images were produced [6].

Figure 1. DataSet of Visible Korean [6]

At Visible Korean website, researches can access sectional images of the whole body of a male (Visible Male), head of a male (Visible Head), pelvis of a female (Visible Pelvis), whole body of a female (Visible Female), whole bodies of a male and a female (3D visualization and development of viewer/software).Anatomy-related studies have been conducted continuously using the data from Visible Korean[11].

C. Data Representation Model

Ontology is one of knowledge representation model, which is actively used as a standard model to disclose and distribute information on the web. To represent metadata and workflow for management and distribution of scientific data, ontology-based data provenance studies have also been conducted [5].

One of the methods to distribute structured data like scientific data through Internet is to use Microdata that incorporates metadata in the existing web pages. Search engines, web collectors, and web browsers can extract and process microdata from web pages.Microdata is an attempt to provide a simpler way of annotating HTML elements with machine-readable tags than the similar approaches of using RDFa(Resource Description Framework in Attributes) and microformats [12]. RDFa is a W3C Recommendation that adds a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding rich metadata within Web documents [13].

III. Management and Distributionof Visible Korean

Visible Korean provides data offline, upon receiving online data requests from domestic and foreign researches through its website. Therefore, it takes a lot of time for researches to apply for data and to receive it, which hinders efficient statistic management of data forwarding and usage.In this study, we design a database required for online processing of a series of processes from searching Visible Korean data and forwarding the required data.

In order for users to search images faster, it should be possible to filter various search conditions, download the search results immediately, and manage various statistical data.

A. Metadata of Visible Korean

To management and distribute the images of Visible Korean, information about users, cadavers, images, organs and download history should be managed. For this, we designed twelve (12) tables as shown in Table 1.

TABLE I. Table List of Visible Korean

Table	Description
User	User Information
UserStatus	Data Usage Agreement of Visible Korean
UserFile	Data Usage File
Cadavers	Cadavers Information
ImageSet	Description of Image Set
Image	Description of Image
Organ	Organ and Part of Body Information
OrganImage	Mapping of Organ and Image
OrganImageSet	Mapping of Organ and Image Set
DownloadImage	Download Status Information of Image
DownloadImageSet	Download Status Information of Image Set
RelatedImage	Mapping of Image and Image

B. DB Scehma of Visible Korean

ER diagram of Visible Korean is designed as shown in Figure 2. User download history and agreement to use, organs and cadavers, images and image sets, and download data are all interconnected, it is possible to manage user status, and data search and download history.

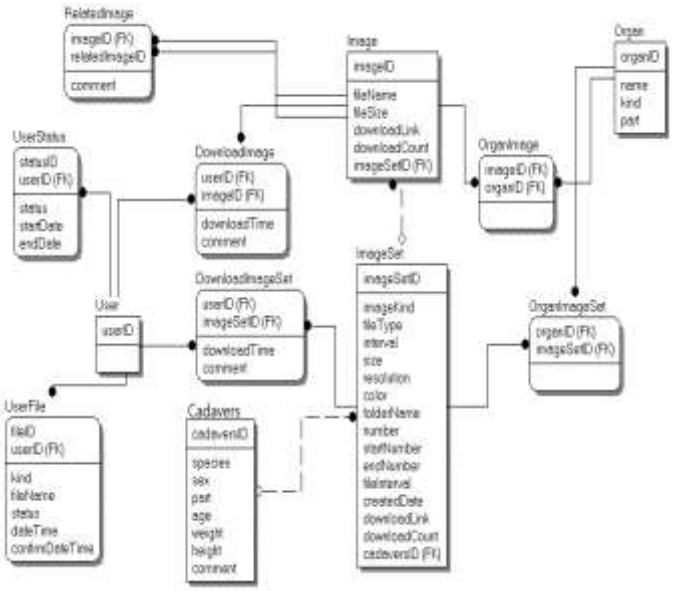

Figure 2. ER diagram of Visible Korean

IV. Data Represnetation Model for Visible Korean

In order to apply to access data in Visible Korean online and to provide it online, the existing offline processing process needs to be changed. Also, in order to distribute the research results created by researches through the website, additional information about the created data, metadata and workflows needs to be managed.

A. Overall Process for Distribution of Visible Korean

In order to use data in Visible Korean, a user should submit a data usage application to KISTI and, once the application for data usage is approved valid, KISTI provides the required data

on human body to the user offline. The user should submit a data usage report once a year. When the data usage period expires, the user should delete the corresponding data and submit the evidence to KISTI. As this series of processes are done offline, it is difficult to manage data provision and utilization efficiently. Also, as the existing service method does not provide a function for users to distribute his or her research results, it is difficult for users to share their research results.

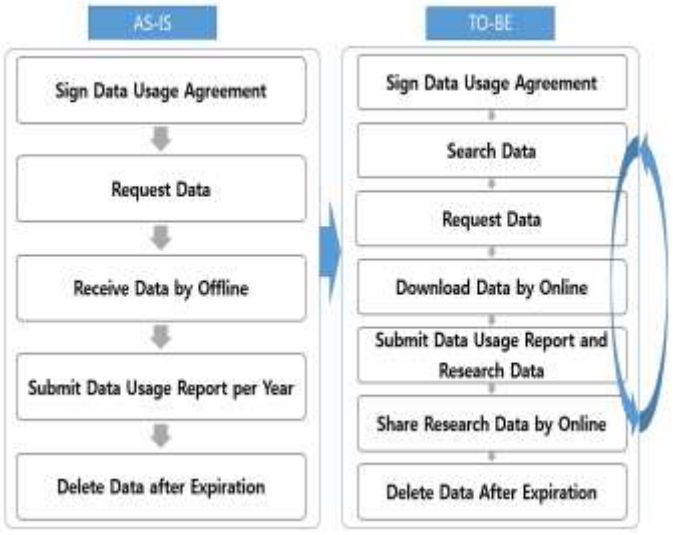

Figure 3. As-Is vs. To-Be Process for Distribution of Visible Korean

In this study, we improves the existing process as shown in Figure 3, to enable online data request, distribution, and research data sharing.

- Sign Data Usage Agreement: The user submits an agreement for using data in Visible Korean. In this step, the user also decides whether to share his or her research results.
- Search Data: It is possible to search organs, systems, cadavers, users, and metadata about related software, and research results shared by the users. Metadata is provided in RDFa format, and the users can extract and process the metadata from the machine-readable website.
- Request Data: The user requests necessary data, depending on the data usage agreement for the corresponding data and whether the user agrees to share his/her research results.
- Download Data by Online: A Download link is provided for the user to download the requested data in a zip file. Time available for downloading the file can be limited.
- Submit Data Usage Report and Research Data: The user should submit a usage report to Visible Korean, and may submit his/her research results and metadata about the results as well.To ensure representation and reliability of the research results, information about the software and workflow used for generating the research results may be incorporated in the metadata.
- Share Research Data by Online: Metadata about the research results shared by users is converted into RDFa format and disclosed on the web, and the research results are saved on the server to be accessed through search.
- Delete Data After Expiration:Once the usage agreement term expires, the user should delete the corresponding data and submit the evidence. The term for agreement to share his/her research results expires, the corresponding data should be deleted so that it will not be searched or downloaded.

Users can continue searching data and sharing research results during the data usage agreement term.

B. Ontology Modeling for Visible Korean

To manage and distribute data in Visible Korean, and to share research results created by users, the DB schema presented in Chapter 3 should be expanded. For this, we propose a metadata expansion model using ontology, as shown in Figure 4.

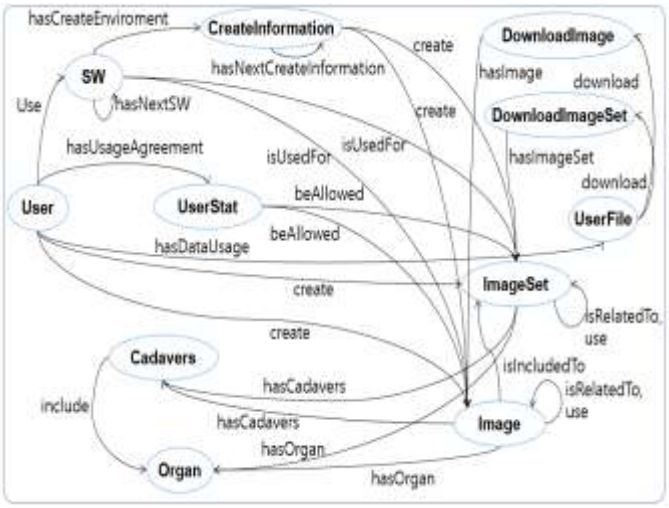

Figure 4. Ontology Modeling for Data and Workflow of Visible Korean

To store the research results and the metadata and workflow for the results, the following relations are additionally defined in the ontology.

- User-use-SW: Stores information about the software used to create the research results, and represents the relation between the user and the software used.
- SW-hasCreateEnvironment-CreateInformation: Represents the software used to create research results and the software use environment. To represent the workflow used for creating the research result, the SW-CreateInformation defines a recursive relationship(hasNextSW, hasNextCreateInformation). Data usage reports entered by the users and the information about research results are classified into metadata about research results and workflows to store.
- CreateInformation-create-ImageSet, CreateInformation-create-Image: Represents the image information created by users. As Visible Korean deals with images and image data only, the research results are also limited to images and image data.
- User-create-ImageSet, User-create-Image: Defines users, image sets and images, to represent the relationship between users and research results created by them.
- User-isRelatedTo-User, ImageSet-isRelatedTo-ImageSet, Image-isRelatedTo-Image : Defines relationships between users, image sets and images.

User can easily find different images for the same organand user can also find related research results.

- User-use-User, ImageSet-use-ImageSet, Image-use-Image : These relationships represent the data used to generate the research result.

C. Data Representation for Visible Korean

To represent the meaning of data in the existing web environment and to extract and process information required by machine-readable program, RDFa can be used. RDFa uses HTML tags but it can express necessary meanings within the tags.

In this study, we provided information onusers, cadavers, organs, and images, which are represented using ontology, and metadata and workflow information about the research results created by the users, through web pages in RDFa format. The users can search necessary contents and share and reproduce research results created by other users as well.

V. CONCLUSION

As various simulations and R&Ds through scientific tools and computers have been activated, various research results are being created rapidly and various studies are being actively conducted to validate and reuse the research results. Scientific data as well as workflows, including all data created in the course of the research, should be described in terms of data provenance.

In this study, we designed a DB schema and improved distribution process in order to convert offline management and distribution of the information about cadaver anatomy, which is one kind of scientific data, into online management and distribution. To do this, we defined metadata, designed the DB schema, and modeled ontology to manage research results and the workflows created by the researchers. Also, as a way to represent data on the web, we disclosed scientific data in machine-readable data format using RDFa, which makes it possible to search and extract the data in the web environment.

In future studies, a management and distribution system for Visible Korean shall be developed by applying the data schema and process proposed in this study, and methods to reflect diverse needs arising in the course of using the Visible Korean system in the scientific data should be developed as well.

ACKNOWLEDGMENT

This work was supported by Jeonju National University of Education Research Grant.

REFERENCES

[1] JimGray, David T. Liu. "Scientific data management in the coming decade." ACM SIGMOD Record 34(4), 2005, pp. 34-41, 2005.

[2] Yogesh L. Simmhan, Beth Plale, Dennis Gannon. "A survey of data provenance techniques." Computer Science Department, Indiana University, Bloomington IN 47405, 2005.

[3] Ang Chen, et al. "Data provenance at internet scale: architecture, experiences, and the road ahead." The biennial conference on innovative data systems research (CIDR'17), 2017.

[4] Satya S. Sahoo, Amit Sheth, Cory Henson. "Semantic provenance for escience: Managing the deluge of scientific data." IEEE Internet Computing 12(4), 2008.

[5] Costa, Gabriella Castro Barbosa, Cláudia ML Werner, Regina Braga. "Software Process Performance Improvement Using Data Provenance and Ontology." International Conference on Business Process Management. Springer International Publishing, pp. 55-71, 2016.

[6] Visible Korean Homepage, http://vkh3.kisti.re.kr/

[7] Suntae Kim, Sunhwa Hahn, Taeyoung Lee, Yong Kim. "A Study on a Model for Using and Preserving Scientific Data." JOURNAL OF THE KOREAN BIBLIA SOCIETY FOR LIBRARY AND INFORMATION SCIENCE 21(4), pp. 81-93, 2010.

[8] Liping Di, et al. "Geoscience data provenance: An overview." IEEE Transactions on Geoscience and Remote Sensing 51(11), pp. 5065-5072, 2013.

[9] Chad Vicknair, et al. "A comparison of a graph database and a relational database: a data provenance perspective." Proceedings of the 48th annual Southeast regional conference. ACM, 2010.

[10] Boris Glavic. "Big data provenance: Challenges and implications for benchmarking." Specifying big data benchmarks. Springer Berlin Heidelberg, pp.72-80, 2014.

[11] Chun Hui Suen, et al. "S2logger: End-to-end data tracking mechanism for cloud data provenance." Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on. IEEE, 2013.

[12] Beom Sun Chung, et al. "Virtual dissection table including the Visible Korean images, complemented by free software of the same data." Int. j. morphol 33(2), pp. 440-445, 2015.

[13] Microdata, Wikipedia, https://en.wikipedia.org/wiki/Microdata_(HTML)

[14] RDFa, Wikipedia, https://en.wikipedia.org/wiki/RDFa