A Formal Approach to the Modelling of Digital Archives

Rathachai Chawuthai[1], Vilas Wuwongse[2], and Hideaki Takeda[3]

[1] Asian Institute of Technology, Prathumtani, Thailand
rathachai.chawuthai@ait.ac.th
[2] Thammasat University, Prathumtani, Thailand
wvilas@engr.tu.ac.th
[3] National Institute of Informatics, Tokyo, Japan
takeda@nii.ac.jp

Abstract. Understanding digital information for users who have different background knowledge becomes important for archival information systems. Key challenges include contextual knowledge changing over time and across different designated communities, and the lack of the reference knowledge among them. Therefore, in order to help users understand digital archives precisely, an archival system should preserve its resources along with their underlying community knowledge, and support the interpretation of their concepts in the right context together with relevant concepts. This research presents a formal approach to digital archives including ontology for preserving the evolution of contextual knowledge. In addition, a prototype has been developed to demonstrate the feasibility and suitability of the proposed formal approach. It is found that the proposed approach can serve as a framework to enhance the capability of the existing archival information systems with regard to the fulfillment of the need of understanding digital archives.

Keywords: Conceptual model, Digital archive, Formal approach, Logical model, Linked data, Knowledge evolution, Ontology, Semantic web.

1 Introduction

Nowadays, to ensure long-term accessibility of digital information, particularly the born digital information, many organizations that need to preserve digital information for future use have started to develop archival information systems according to some digital preservation approaches, including reference guidelines, and metadata standards such as OAIS[1], and PREMIS[2]. As a result, bit streams of digital resources will be assured to be preserved and originally rendered in the future [1].

Digital preservation techniques ensure that digital content can be stored and rendered correctly. However, the situation can get worse when no one can understand the true meaning that needs to be conveyed from the content of an archived file. A great

[1] http://public.ccsds.org/publications/archive/650x0m2.pdf
[2] http://www.loc.gov/standards/premis

H.-H. Chen and G. Chowdhury (Eds.): ICADL 2012, LNCS 7634, pp. 179–188, 2012.

challenge is to get the reader to understand the original meaning of the concepts, while the reader's background knowledge which depends on specific time and community may be dissimilar to that of the information writer. Fortunately, a theory of digital preservation was introduced for archiving digital objects along with underlying community knowledge (UCK) and providing proper contextual knowledge to a reader for correct understanding [2]. However, the nature of UCK always changes over time. To support the interpretation of a concept of a given point in time, the change of UCK has to be monitored, captured, and stored in the preservation environment [16]. To materialize this theory, this research will introduce a model of knowledge evolution that records the change of concepts and context depending upon time and community.

Not only knowledge evolution but also the association between relevant concepts becomes a feature for understanding digital archives [3]. Linking archival information becomes a key concept for open access to digital information [4]. Many digital libraries, museums and institutional repositories give priority towards exchanging their knowledge and establishing relationship between their resources across repositories. Nevertheless, digital resources across various repositories are difficult to link because different kinds of knowledge schema have to be dealt with. This research also proposes underlying common community knowledge (UCCK) to serve as reference knowledge for all designated communities in order to link relevant concepts together.

As a consequence, the preservation of contextual knowledge and the linkage between proper concepts have become key ideas for understanding archived digital resources correctly. Moreover, the evolution of contextual knowledge and UCCK are considered to support the issues above. In this view, a formal approach to digital archives is introduced to respond to these important requirements in preserving the correct meaning of a digital resource. A formal approach to modeling of digital archives is hence proposed.

Section 2 reviews background, related works, and relevant technology. Next, Section 3 proposes a conceptual model which describes formal definitions of knowledge evolution and UCCK. Section 4 describes a logical model that is a materialization of the conceptual model in RDF. Furthermore, Section 5 shows a prototype and discusses about the usability of the model. Lastly, Section 6 draws conclusions and suggests further enhancements for this formal approach.

2 Literature Review

In order to make digital objects suitable for original understanding by the user in the future, there is a need to capture and represent the content along with context information, which depends much upon the time and space dimensions [16]. Moreover, the correct understanding of digital content requires the correct interpretation of a concept with original context that can be background knowledge of the reader. Therefore, if users have different background knowledge, they may interpret the same concept as dissimilar things. For example, the concept "First floor," which was recorded with a digital archive by an American, may not be interpreted as a ground floor if another person who has different background knowledge, such as a British, reads it. This issue

arises because the writer and the reader are not in the same Designated Community (DC). DC is a particular group of people who share a common particular set of contextual knowledge called UCK [2].

Preserving the correct meaning of digital resources is a difficult task. However, different systems and technologies have been developed in order to fulfill this task. Some of the systems such as CASPAR[3] and SHAMAN[4], which are data archival systems, are hence reviewed. CASPAR Knowledge Manager aims to preserve digital resources together with knowledge as representation information, which maps a data object into more meaningful concept, for example, ASCII definition that describes how a sequence of bits. The model also gives essential information that the user community does not know in order to access the archived object. SHAMAN Context Model preserves the context of a document according to a document process. The context is expressed as metadata stored externally from the document itself. The metadata for managing information of a digital resource can be abstract, session, topic, affiliation, person, contributor, reviewer, process and result. Both CASPAR and SHAMAN can preserve context for digital archives and offer link to relevant objects, while the interpretation of a concept in the different context remains a big challenge. For this purpose, the evolution of contextual knowledge becomes an appropriate way to see the changes that have occurred in the concept through time and in specific communities. This provides a basis for correct interpretation of the concept, which is invaluable to understand the original meaning depicted by the concept [16].

In order to capture, store, and present the knowledge evolution of digital archives, Semantic Web technology, especially Spatio-Temporal RDF and Linked Data, becomes a noble approach to solving this issue. Spatio-temporal RDF takes into account the dynamic character of RDF data in surrounding context and presents things that move, change, appear, and disappear over time [5-6]. Linked Data describes a methodology to connect structured data by linking together relevant concepts with proper semantics. This approach is built upon Semantic Web technology, which results in an ability for computers and machines to read and understand data. Furthermore, it enables data from different sources to be connected and queried [7].

3 Proposed Conceptual Model

As mentioned in the previous sections, knowledge evolution becomes a key player to identify the original contextual knowledge of a concept and to link to relevant concepts. The description knowledge evolution consists of the change of concepts and the transition of relationship between concepts. To present general definitions for knowledge evolution, a conceptual model is proposed based on set and function formalism. The model includes definitions of time, designated community, concept evolution, relationship evolution, and UCCK. In practice, the model can be applied for further development of archival information systems.

---

[3] http://www.casparpreserves.eu

[4] http://www.shaman-ip.eu

3.1 Time and Designated Community

Definition 1. A domain of time intervals ($\mathcal{T}_I$) is a set of pairs of time points, which are elements of the domain of time points ($\mathcal{T}_\mathcal{P}$). The expression of a time interval is the half open interval $[t_s,t_e)$, where t_s is the start time point and t_e is the end time point.

$$\mathcal{T}_I = \{[t_s, t_e) | t_s, t_e \in \mathcal{T}_\mathcal{P} \text{ and } t_s < t_e\}$$

In practice, the contextual knowledge which is recorded by a time interval, for example "[2001, 2006)" indicates the contextual knowledge is valid within the time interval from year 2001 until the time before year 2006. Technically, this research employs the datatype XSD:dateTime to express time points.

Definition 2. To represent the group of concerned communities, a notation $\mathcal{DC}$ is introduced to be a finite set of designated communities each of which is usually represented by a URI.

3.2 Concept Evolution

Concepts keep on changing with time, for example; "Korea" was split into "North Korea" and "South Korea" in 1948 A.D. The change of concepts, known as concept evolution, results in interdependencies between concepts before change and after change, such as, North Korea and South Korea have interdependency with Korea [8]. Particularly, in such scenarios, changes can be classified into Replace, Split, and Merge [9]. To capture a change of a concept, this research introduces a function δ_C which is a mapping from concepts to other concepts along with spatio-temporal data.

Definition 3. Let $\delta_C^{Replace}$ be a function Replace, δ_C^{Merge} be a function Merge, δ_C^{Split} be a function Split, and $\mathcal{C}$ represent the set of domain concepts.

The function **Replace** is an operation which gives the concept a new identification. To signify this operation with spatio-temporal condition, the notation $\left[\delta_C^{Replace}\right]_{\mathcal{T}_I \times \mathcal{DC}}$ is introduced, where the subscripts of the function Replace are the time interval ($\mathcal{T}_I$) and designated community ($\mathcal{DC}$):

$$\left[\delta_C^{Replace}\right]_{\mathcal{T}_I \times \mathcal{DC}} : \mathcal{C} \rightarrow \mathcal{C}$$

The definition shows that the function Replace maps from one concept to another concept. For instance, let dc_1 be an example of a designated community, the renaming of the old concept Burma by the new concept Myanmar in 1989 A.D, can be expressed by:

$$\left[\delta_C^{Replace}\right]_{[1989,UC),dc_1} (Burma) = Myanmar$$

where, *UC* means Until Change, [*1989*, *UC*) is a time interval which starts from year 1989 and ends when the change of the new concept appears again. Next, the function

Merge is an operation which joins two or more concepts with one concept. The definition of this function is similar to the function Replace.

$$\left[\delta_C^{Merge}\right]_{\mathcal{T}_I \times \mathcal{DC}} : 2^{\mathcal{C}} \to \mathcal{C}$$

where, $\mathcal{C}$ is the set of domain concepts and $2^{\mathcal{C}}$ is a power set of domain concepts which indicates that the function Merge maps many concepts into one concept. Lastly, the function **Split** is an operation which divides one concept with two or more concepts. The definition of this function is quite similar to functions Replace and Merge, while this function maps one concept into many concepts.

$$\left[\delta_C^{Split}\right]_{\mathcal{T}_I \times \mathcal{DC}} : \mathcal{C} \to 2^{\mathcal{C}}$$

3.3 Relationship Evolution

A relationship between concepts is represented by a triple that defines a property as a function which maps one resource to another resource. For example, the statement "Pluto is a planet" can be presented as follows:

$$higherClass \mapsto \{(Pluto, Planet)\}$$

where, "higerClass" is a property that maps the concept "Pluto" with the concept "Planet". However, concept relationships might change; for example, Pluto was changed from being a subclass of Planet to being a subclass of Dwarf Planet in the year 2006 A.D. In other words, the statement "Pluto is a Planet" was denied in 2006, and then a relationship $higherClass \mapsto \{(Pluto, DwarfPlanet)\}$ was asserted. The proposed function that maps the transition of the relation is Relationship Evolution.

Definition 4. Let δ_R (Delta of relationship) be a function of relationship evolution, $\mathcal{C}$ be the set of domain concepts, and $\mathcal{P}$ be the set of properties. In addition, to integrate this operation with spatio-temporal condition, a notation $[\delta_R]_{\mathcal{P} \times \mathcal{T}_I \times \mathcal{DC}}$ is introduced, where the subscripts are the Cartesian product of property set ($\mathcal{P}$), time interval set ($\mathcal{T}_I$), and designated community set ($\mathcal{DC}$), and is defined by:

$$[\delta_R]_{\mathcal{P} \times \mathcal{T}_I \times \mathcal{DC}} : \mathcal{C} \times \mathcal{C} \to \mathcal{C} \times \mathcal{C}$$

This definition indicates that a binary relation of concepts is mapped to another binary relation of concepts in order to capture the transition of that relation, such that both domain and range must have either common first entry (subject) or common second entry (object). It means that if an input is (x,y), the output can be (x,z) or (z,y) but cannot be the same (x,y) as the input. For example, the change of being-a-subclass-of Pluto from a planet to a dwarf planet in the year 2006 is expressed by:

$$[\delta_R]_{higherClass,[2006,UC),ucck}\big((Pluto, Planet)\big) = (Pluto, DwarfPlanet)$$

where, *ucck* is a reserved term representing Common Community Knowledge which shares its knowledge with all designated communities.

3.4 Underlying Common Community Knowledge (UCCK)

In practice, a designated community normally archives its UCK using its own schema which may not be compatible with other communities because of the lack of common reference knowledge [10]. The UCCK is introduced to solve this issue by serving as a common storage of facts "publicly known" and providing relationships between conceptual knowledge across variety of UCK [11]. Thus, the relevant concepts across designated communities can be linked together through the UCCK. Practically, the linking of concepts is carried out as a result of concept evolution, for example the splitting of Korea to North Korea and South Korea leads to Korea being linked to North Korea and South Korea. If one repository contains knowledge about Korea and another system has knowledge about North Korea, both repositories can exchange knowledge through the UCCK that records the relationship between these concepts.

4 A Logical Model

To materialize the proposed conceptual model, a logical model is introduced to represent the evolution of knowledge in RDF. The model records the change of concepts and triples together with a certain set of parameters; properties are derived from Event ontology and provenance data. These properties will become reference data and metadata respectively for UCK [12-14]. Based on this idea, an ontology named CKA is expressed with the inclusion of temporal (when), spatial (where), agent (who), casual (why) and other properties, as shown in Fig. 1.

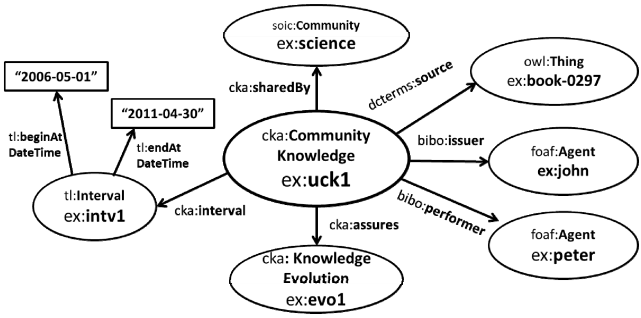

Fig. 1. A logical model for digital archives

This model presents the UCK as an instance of class cka:CommunityKnowledge, which contains a set of parameters: time, community, provenance, and knowledge evolution. First of all, the instance of the time interval (tl:Interval) which is extended from Timeline ontology identifies the beginning (tl:beginAtDateTime) and the end (tl:endAtDateTime) time points (xsd:dateTime). Next, the instance of community (soic:Community) indicates that the knowledge is shared by a designated community. Furthermore, a few instances of provenance data are employed as reference, such as, source of the knowledge (dcterms:source), person who creates the knowledge

(bibo:performer), and person who reports the knowledge (bibo:issuer). Lastly, the instance of knowledge evolution (cka:KnowledgeEvolution) indicates the change of concepts, which consists of two types – concept evolution and relationship evolution.

RDF Model for Concept Evolution: According to the definition of concept evolution (Definition 3), the change of concepts is defined by a relationship between old and new concepts. To denote this fact in RDF, the old and new concepts, which are instances of skos:Concept, are identified to be values of properties cka:oldConcept and cka:newConcept. For example, the splitting of Korea is expressed as follows:

```
ex:sp1    rdf:type        cka:ConceptSpliter ;
          cka:oldConcept  ex:Korea ;
          cka:newConcept  ex:NorthKorea, ex:SouthKorea .
```

RDF Model for Relation Evolution: To present the transition of triple by RDF, this model is developed by applying from the definition of relationship evolution (definition 4). CKA ontology offers properties to specify a subject, an old object and a new object each of which is a skos:Concept, and a relation which is a rdf:Property. This ontology also offers necessary operations, such as, classification, part-whole, and membership [15]. For example, the reclassification of the concept Pluto can be presented by:

```
cka:Reclassification
        rdfs:subClassOf    cka:RelationEvolution ;
        cka:relation       cka:higherClass .
ex:rc1  rdf:type           cka:Reclassification;
        cka:subject        ex:Pluto ;
        cka:oldObject      ex:Planet ;
        cka:newObject      ex:DwarfPlanet .
```

In practice, every custom operation like cka:Reclassification must be a subclass of cka:RelationEvolution, and the predicate of a triple must be a value of cka:relation. Parameters of the custom operation are identified by values of the subject, the object, and the new object, in order to generate the proper RDF statements.

Lastly, in order to preserve digital resources together with correct contextual knowledge, an archivist has to identify a list of concepts along with original designated community and time. The ontology also offers a model for digital archives with UCK called **Application Profile for Archives**, which is modeled as follows:

```
ex:doc1   cka:uckConcept [
    dc:subject      ex:Pluto ;
    tl:atDateTime   "2012-03-01T12:14:00.000+09:00";
    cka:sharedBy    cka:ucck ] .
```

The model uses the widely used term "dc:subject" to indicate a concept, cka:sharedBy to pinpoint a community and tl:atDateTime to identify the original time. Thus, archival information systems, that are enhanced by this model, will have enough information to construct correct contextual knowledge for its digital objects.

5 Implementation and Discussion

The proposed approach originated in an idea to utilize contextual knowledge evolution and UCCK for preserving the interpretation of information for digital archives. In order to verify the feasibility and suitability of this approach, a prototype has been developed, and it has been evaluated with actual requirements.

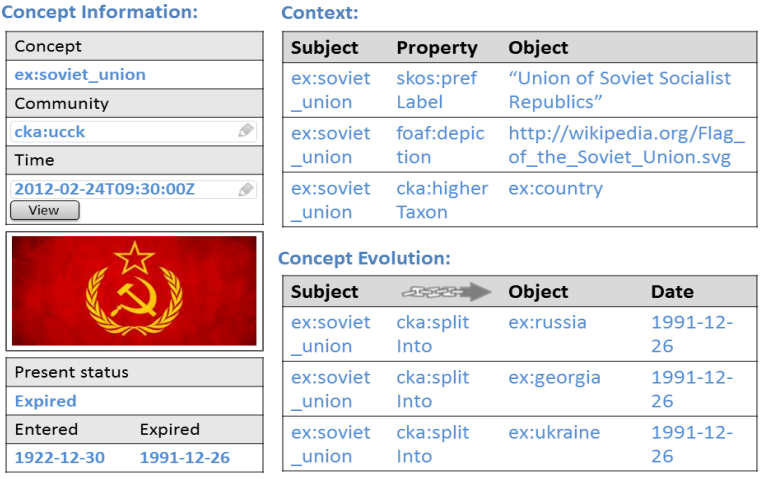

Fig. 2. An example of contextual knowledge of a concept in a particular time and community

In this research, three web application servers are developed and demonstrated. Two of them are representatives of two different UCK; uck1 and uck2. The other one is a representative of UCCK, which contains common knowledge for each UCK, and exchanges knowledge between them. Each server consists of a Contextual Knowledge Manager (CKM) which is used to manage each entry of contextual knowledge. CKM provides a web interface that allows users to record the change of concept or relationship according to the operations of changes from CKA ontology, and then stores them into the triple store. Moreover, CKM offers a service that computes and presents contextual knowledge of a concept and its relevant concepts according to a particular time and community by using CKA ontology and Jena[5] reasoning engine. Fig. 2 shows the precise contextual knowledge of an example concept, Soviet Union, according to the original context and some links to relevant concepts. This service normally is requested by a digital archive application which is a web-based application deployed in each UCK server. The application is enhanced from the well-known document management, Alfresco[6], to preserve digital resources with application profile and illustrate contextual knowledge of each digital object. Its software architecture and its model are improved by following the OAIS guideline and PREMIS metadata; so it can exchange information with other standard archival information systems. In

[5] http://jena.apache.org/
[6] http://www.alfresco.com/

addition, linking relevant concepts across UCK servers is handled by the UCCK server, because the common knowledge from UCCK records the changes of concepts. This prototype works well with a few thousands triples; however, in fact, the number of contextual knowledge in the world may be over billion triples. In this case, it is necessary to redesign the system architecture to be the high-performance platform such as a cloud computing. In summary, the prototype confirms that the approach is feasible in order to service users.

Apart from implementing a prototype, the evaluation of this approach has been carried out by analyzing the proposed model with ability to preserve contextual knowledge and interoperability of digital archives [16]. This research also considers the related models: CASPAR and SHAMAN in order to demonstrate the uniqueness of the proposed model. The result of the evaluation follows:

1. **Ability to preserve understanding of digital information, which is dealt with temporal contextual knowledge:** The proposed model fully supports this requirement by offering the information model that preserves dynamic change of contextual knowledge with time condition. CASPAR model also supports this feature by having static information about technical context of digital resources.
2. **Ability to support information interpretation according to spatial constrains, such as, culture, society, politics, and background knowledge:** CASPAR and the proposed model are better considering this feature because both models are focusing on identifying precise knowledge for particular designated community, whereas, SHAMAN model seems support this requirement by offering common metadata and multi-lingual for use in different countries.
3. **Ability to link digital resources with across digital repositories:** All of these models support this criterion.
4. **Ability to linked concepts and contextual knowledge across digital repositories**: The proposed approach offers links between relevant concepts according to knowledge evolution and UCCK, which is not presented by the other systems.

Moreover, to the best of our knowledge of CAPAR and SHAMAN, we can verify that both models pay attention to only syntactical context and technical context of digital archives. On the contrary, the purposed model places importance to sematic context which emphasizes on keeping understanding particular concept from content of digital archives. Therefore, the proposed model can achieve all requirements of the understanding of digital archives. In conclusion, the result indicates that the proposed model accomplishes the preservation of digital information along with context, and this approach is not presented in any existing models.

6 Conclusion and Future Work

This research proposes the formal approach to the modeling of digital archives. The approach aims to support correct understanding of digital archives. To accomplish this approach, a conceptual model is presented to describe knowledge evolution based on Mathematics formalism. It also introduces UCCK to link relevant concepts in

order to share contextual knowledge across communities. Consequently, the approach can support users who have different background knowledge to understand the original meaning of digital archives. Additionally, the CKA ontology, which is materialized from the conceptual model, is utilized for practical development. The result of the prototype shows that the approach is feasible for real implementation. This approach also presents some unique ideas and techniques, which do not exist in any other works, to satisfy the need of understanding digital archives.

Currently, this approach can only preserve the interpretation of concepts under the change of RDF data. However, in the future, the changes of schema and ontology may affect the original interpretation. Therefore, this approach should be enhanced to support schema and ontology changes by employment of such languages as RDFS and OWL.

References

1. Yuan, L., Banach, M.: Institutional Repositories and Digital Preservation: Assessing Current Practices at Research Libraries (2011)
2. Flouris, G., Meghini, C.: Terminology and Wish List for a Formal Theory of Preservation. In: PV 2007 (2007)
3. Mulwad, V., Finin, T., Syed, Z., Joshi, A.: Using linked data to interpret tables. In: The First International Workshop on Consuming Linked Data, ISWC 2010 (2010)
4. Hodge, G.M.: Digital preservation and permanent access to scientific information: the state of the practice (2004)
5. Buraga, S.C., Ciobanu, G.: A RDF-based Model for Expressing Spatio-Temporal Relations between Web Sites. In: The 3rd International Conference on Web Information Systems Engineering, WISE 2002 (2002)
6. Gutierrez, C., Hurtado, C.A., Vaisman, A.A.: Temporal RDF. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 93–107. Springer, Heidelberg (2005)
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data-The Story So Far. In: IJSWIS, vol. 5(3), pp. 1–22 (2009)
8. Rizzolo, F., Velegrakis, Y., Mylopoulos, J., Bykau, S.: Modeling Concept Evolution: A Historical Perspective. In: Laender, A.H.F., Castano, S., Dayal, U., Casati, F., de Oliveira, J.P.M. (eds.) ER 2009. LNCS, vol. 5829, pp. 331–345. Springer, Heidelberg (2009)
9. Yildiz, B.: Ontology Evolution and Versioning The state of the art (2006)
10. Zeng, M.L., Chan, L.M.: Trends and issues in establishing interoperability among knowledge organization systems. In: JASIST, pp. 377–395 (2004)
11. Halpern, J.Y., Moses, Y.: Knowledge and common knowledge in a distributed environment. In: JACM, vol. 37(3), pp. 549–587 (1990)
12. Bouquet, P., Serafini, S., Stoermer, H.: Introducing Context into RDF Knowledge Bases. In: The 2nd Italian Semantic Web Workshop, SWAP 2005, pp. 14–16 (2005)
13. Raimond, Y., Abdallah, S.: The Event Ontology (2007)
14. Bao, J., Tao, J., McGuinness, D.L., Smart, P.: Context Representation for the Semantic Web. In: Extending the Frontiers of Society On-Line, WebSci 2010 (2010)
15. Stevens, R., Goble, C.A., Bechhofer, S.: Ontology-based Knowledge Representation for Bioinformatics. In: Briefings in Bioinformatics 2000, pp. 398–414 (2000)
16. Chowdhury, G.: From digital libraries to digital preservation research: the importance of users and context. Journal of Documentation 66(2), 207–223 (2010)