Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

SKO Types: An entity-based scientific knowledge objects metadata schema

Xu Hao, Giunchiglia Fausto

February 2015

Technical Report # DISI-18-005

Published in: Journal of Knowledge Management

SKO Types: an entity-based scientific knowledge objects metadata schema

Hao Xu and Fausto Giunchiglia


Hao Xu is based at College of Computer Science and Technology, Jilin University, Changchun, China. Fausto Giunchiglia is based at Department of Information Engineering and Computer Science, University of Trento, Trento, Italy.

Received 3 November 2014
Revised 3 November 2014
Accepted 3 November 2014

This work is supported by the National Natural Science Foundation of China (No. 61300147), China Postdoctoral Science Foundation (No. 2014M551185), European Project "Liquid Publication" and "Bridging the Gap" Erasmus Mundus European Programme.

Abstract

Purpose – *This paper aims to propose an entity-based scientific metadata schema, i.e. Scientific Knowledge Object (SKO) Types. During the past 50 years, many metadata schemas have been developed in a variety of disciplines. However, current scientific metadata schemas focus on describing data, but not entities. They are descriptive, but few of them are structural and administrative.*

Design/methodology/approach – *To describe entities in scientific knowledge, the theory of SKO Types is proposed. SKO Types is an entity-based theory for representing and linking SKOs. It defines entities, relationships between entities and attributes of each entity in the scientific domain.*

Findings – *In scientific knowledge management, SKO Types serves as the basis for relating entities, entity components, aggregated entities, relationships and attributes to various tasks, e.g. linked entity, rhetorical structuring, strategic reading, semantic annotating, etc., that users may perform when consulting ubiquitous SKOs.*

Originality/value – *SKO Types can be widely applied in various digital libraries and scientific knowledge management systems, while for the existing legacy of scientific publications and their associated metadata schemas.*

Keywords *Semantic annotation, Entity oriented, Metadata schema, Scientific knowledge object*

Paper type *Research paper*

1. Introduction

Metadata are generally defined as "data about data" or "information about data", which is used to facilitate resource discovery, e-resources organization, interoperability, digital identification, archiving and preservation. There are three main types of metadata, i.e. descriptive metadata, structural metadata and administrative metadata (National Information Standards Organization, 2004).

During the past 50 years, many metadata schemas and concept models have been developed in a variety of disciplines. Standards for metadata in digital libraries include Dublin Core, Encoded Archival Description (EAD) (Pitti, 2005), Machine-Readable Catalogue (MARC) bibliographic records (Delsey, 2002), Metadata Encoding and Transmission Standard (METS) (Cantara, 2005), PREservation Metadata: Implementation Strategies (PREMIS) schema (Guenther, 2004), Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) (Lagoze *et al.*, 2002), the CIDOC conceptual reference module (CIDOC-CRM) (Doerr, 2003), Functional Requirements for Bibliographic Records (FRBR) (ONeill, 2002), Common European Research Information Format (CERIF), etc. Moreover, Friend of a Friend (FOAF) defines an open, decentralized technology and metadata schema for connecting social websites and the people they describe. Learning Object Metadata (LOM) (Learning Object Metadata Working Group, 2000) focuses on learning objects, digital or non-digital, and their management, location and evaluation. VIVO, an open-source, semantic web tool for research discovery, supports finding people and the research they do based on open linked data and VIVO-ISF ontology. In addition to this, major search engines, such as Google, Yahoo and Bing, also

provide their own metadata schemas for archiving and searching. Those aforementioned standards constitute the metadata foundation for scientific publication management.

However, current scientific metadata schemas focus on describing data, but not entities. They are descriptive, but few of them are structural and administrative. They provide a rare mechanism for linking entities and describing relationships between them.

In this paper, the authors propose an entity-based scientific metadata schema, i.e. Scientific Knowledge Object (SKO) Types, that specifies sets of bibliographically related entities, relationships, attributes and services, intended to describe ubiquitous scientific knowledge objects semantically, and to facilitate their dissemination, collaboration, evolution and reuse.

2. SKO Types definition

SKO Types is an entity-oriented theory for representing and linking SKOs by defining entities, relationships between entities and attributes of each entity in the scientific domain. In SKO management, SKO Types serves as the basis for relating entities, entity components, aggregated entities, relationships and attributes to various tasks, e.g. linked entity, rhetorical structuring, strategic reading, semantic annotating, etc., that users may perform when consulting ubiquitous SKOs.

2.1 SKO

An SKO, an abbreviation for Scientific Knowledge Object, is a type of entity of intellectual and artistic endeavour, which is defined as:

$$\text{SKO} = < \text{T}, \{\text{A}\}, \{\text{R}\}, \{\text{S}\} >$$

where

- T Is one of the entity types in an SKO hierarchy.
- {A} Is a non-empty set of attributes A, while there are several mandatory attributes, e.g. URI.
- {R} Is a set of relationships R.
- {S} Is a set of services S.

Figure 1 illustrates the entity types in an SKO hierarchy. SKO, as an entity type, has been divided into two subtypes, i.e. MonoSKO and MultiSKO. MonoSKO comprises paper and monograph, while MultiSKO consists of journal issue, proceedings and article collections. Furthermore, paper contains subtypes of article, tech report, comment and review. Monograph includes book, booklet and thesis.

In this hierarchy tree, the father entities are more generic than the children entities. In addition, the lattice makes the children nodes inherit all the attributes, relationships and services that their ancestors have.

2.2 SKO set

The SKO Types model permits us to represent aggregated SKOs as a whole, i.e. SKO set, and the component SKO as an integral unit, i.e. SKO nodes, in the same way as SKOs.

From a logical perspective, SKO sets and SKO nodes share the same characteristics as SKOs. For example, they express scientific knowledge, and they also have subject, author/editor, publisher, etc.

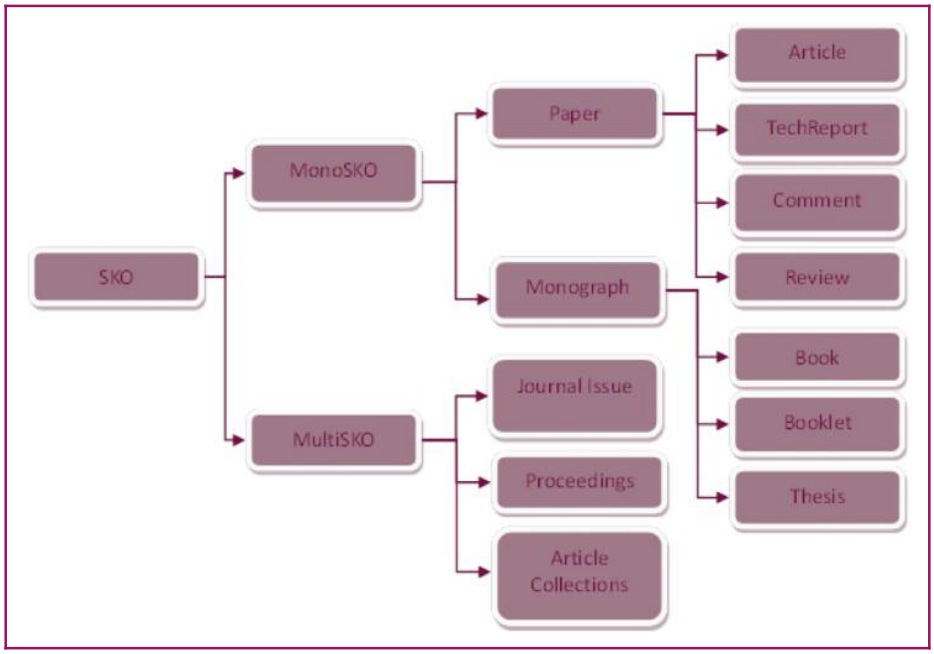
An SKO set is a set of SKOs whose attributes answer a query, and it is defined as:

$$\text{SKO Set} = < \text{N}, \{\text{T}\}, \text{Q}, \{\text{R}\}, \{\text{S}\} >$$

where

- N Is the name of the SKO set.
- {T} Is a set of entity types that the elements in this SKO set must belong to.
- Q Is the query Q = < {A} > where {A} is a set of attributes.

Figure 1 Entity types in an SKO hierarchy



{R} Is a set of relationships R.
{S} Is a set of services S.

As shown in Figure 2, the authors define three types of SKO sets at the first level, i.e. Liquid Journal, Conference Call for Papers and Simple Query, where Simple Query can be done using topics or categories.

2.3 SKO node

An SKO node is a component entity encapsulated in SKOs that semantically represent scientific knowledge as an integral unit.

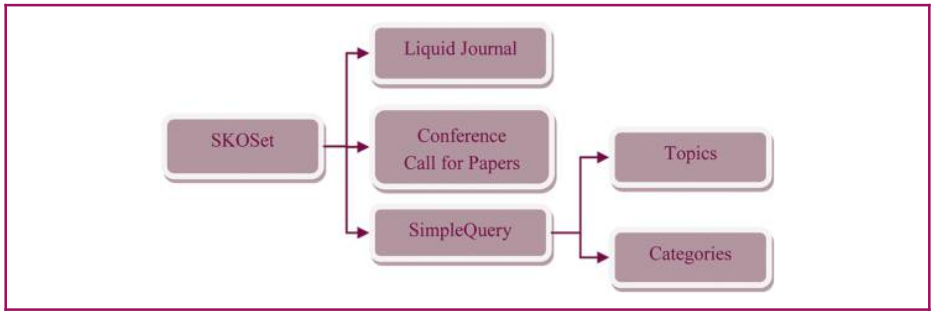
An SKO node is defined as:

$$\text{SKO node} = < N,\ T, \{A\},\ \{R\},\ \{S\} >$$

where

N Is the name of the SKO node.
T Is the type of SKO that the SKO node belongs to.

Figure 2 SKO set types and subtypes



{A} Is a set of attributes.
{R} Is a set of relationships R.
{S} Is a set of services S.

Figure 3 describes the types of SKO nodes. The first level includes text chunk, video, audio and data. Text chunk can be further divided into two groups, namely, syntactic partition and rhetorical partition. Syntactic partition comprises chapter, section, paragraph, sentence, figure, formula and table. Rhetorical partition comprises state of the art, problem statement, solution, discussion, methods, material, results and evaluation.

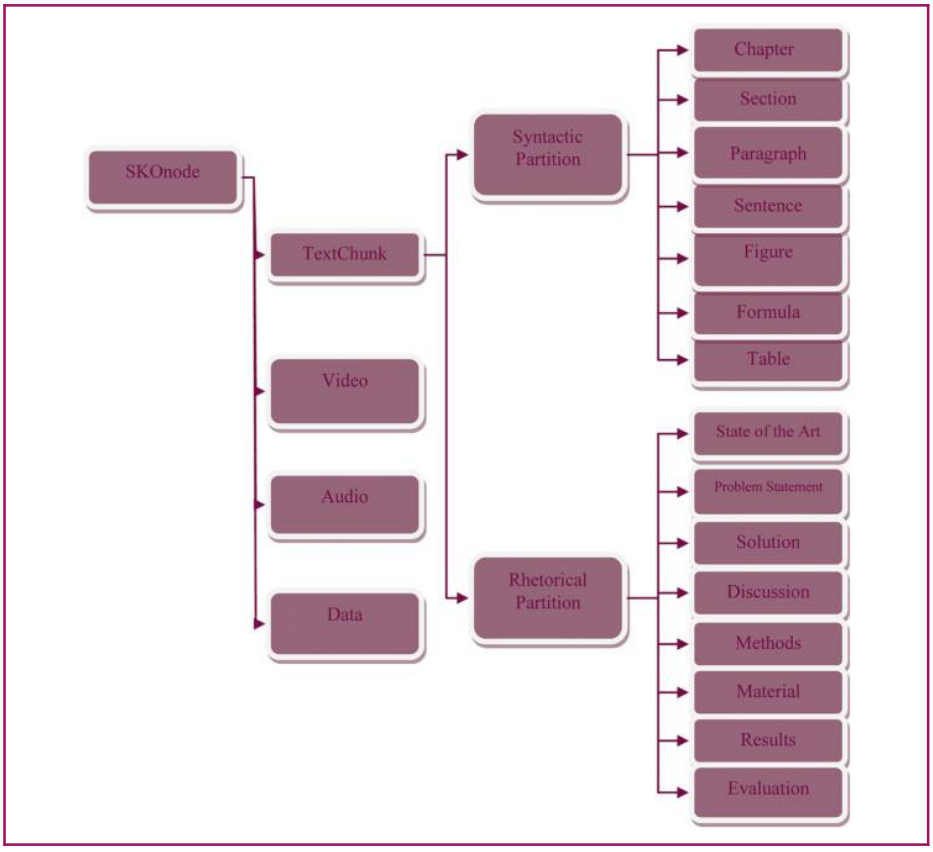
An SKO node is the smallest object in SKO Types that:

- has a unique identifier;
- was created independently;
- can be cited independently;
- can be reused autonomously;
- can be published or distributed separately; and
- has separable copyright.

2.4 SKO-related entities

In the scientific universe, there are several other entities which are tightly related to SKOs, SKO sets or SKO nodes that are responsible for the production, dissemination or

Figure 3 SKO node types and subtypes



custodianship of knowledge, such as researcher, conference, institution and project. Generally speaking, an entity can be defined as:

$$\text{Entity} = < T, \{A\} >$$

where

T Is one of the entity types.
{A} Is a set of attributes A.

Actually, researcher is a role of person, conference and project are subtypes of event and institution is a subtype of organization.

3. Relationships

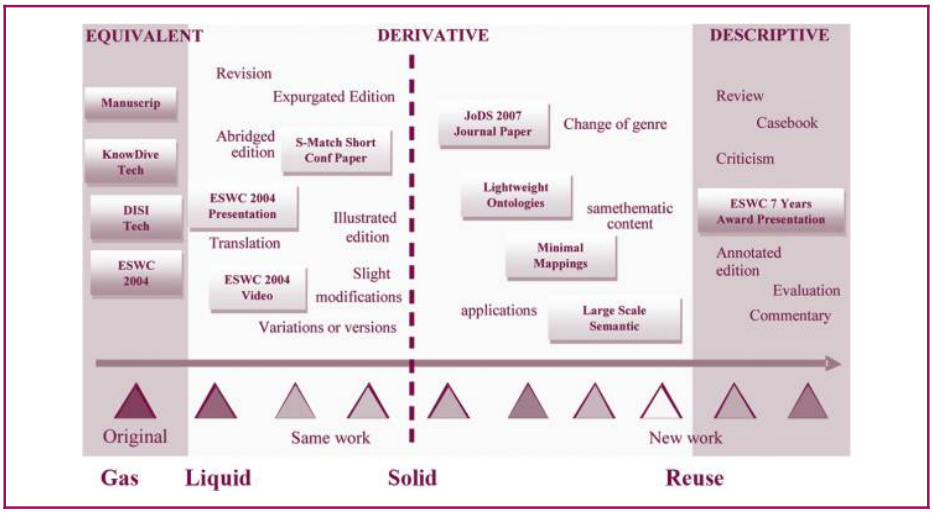
Relationships abound in the scientific world. These may be educational, economic, social, legal, etc. The relationships addressed herein are restricted to those involved in the representation and management of SKOs, including:

- *Syntactic relationships*: Text structure, hyperlink.
- *Content relationships*: Equivalent, derivative, descriptive, sequential, accompanying, shared characteristic.
- *Whole/part relationships*: Whole-whole, whole-part, part-whole, part-part.
- *Rhetorical relationships*: State of the art, problem statement, solution, discussion, material, methods, results, evaluation.
- *Entity relationships*: Relationships between SKO and SKO-related entities.

Note that these five categories are not necessarily mutually exclusive, and the authors have endeavoured to attain and keep alignment with other relevant terminology systems such as FRBR, Semantic Publishing and Referencing Ontologies (SPAR), etc. In SKO Types, a relationship is viewed as a particular kind of attribute, i.e. a relational attribute.

One of the distinctive features of SKO theory is that it keeps evolving during its entire life cycle, namely, gas, liquid and solid. Figure 4 gives a concrete story of the work "S-Match". When the ideas and manuscripts of S-Match are discussed and distributed internally in the KnowDive group, it exists in the gas stage. The milestone of its liquefaction is when it is published openly to communities with modalities of a DISI tech report and an European Semantic Web Conference (ESWC) conference paper. Then, more SKOs are derived from

Figure 4 Family of SKOs



the original work of "S-Match", such as an abridged edition, a conference presentation or some slight modifications, while all of these are based on the same work (semantic) and become more stable. Along with its solidification, "S-Match" keeps evolving and being reused in terms of new work or topics, e.g. lightweight ontologies, minimal mapping, large-scale semantic matching, etc. In addition, more descriptive SKOs appear, including review, evaluation, annotations, commentary, etc.

4. Attributes

Each of the entities defined in SKO types has associated with it a set of attributes. An attribute A is defined as:

$$A = < N, V >$$

where

N Is is an attribute name.
{A} V is an attribute value.

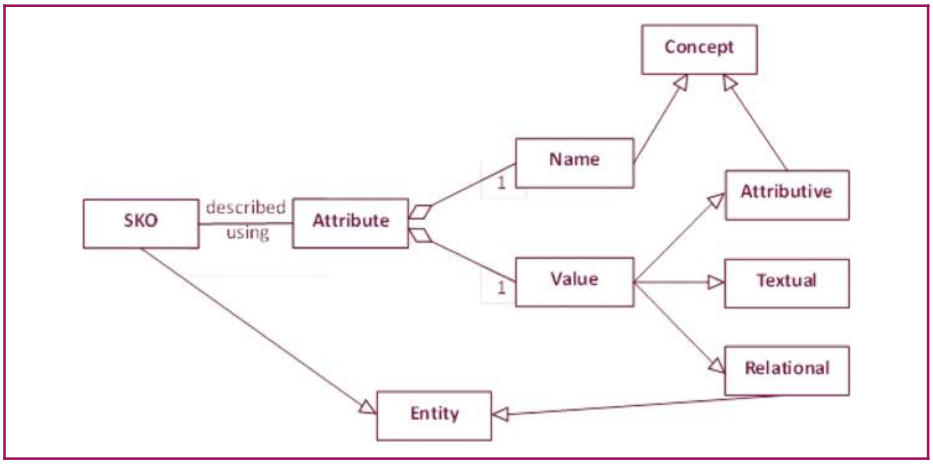
In SKO Types, an attribute name is a concept, which means that there cannot exist two attributes with the same name. The attribute value domain consists of Boolean, integer, float, date, duration, semantic-less string (SLS), semantic string (SS), entity and URL. Note that an attribute definition allows multiple values and polymorphism, in which the data type domain can be a single data type, an array or a list of different data types. For example, the attribute value of "author" is "researcher [] or organization []".

Figure 5 specifies an abstract model for SKO Types. It defines the nature of the elements used and illustrates how those elements are combined to create structured knowledge representation. The model is presented here using a Unified Modeling Language (UML) class diagram.

The attributes defined for SKO Types were derived from a comparative analysis of state-of-the-art metadata schemas such as Dublin Core (DC), FOAF, LOM, etc. The scope of attributes included in SKO theory is intended to be comprehensive but not exhaustive.

For the focus of this research, the attributes for the other entities conference, project, researcher and institution include only those that are conventionally displayed as part of the scientific knowledge per se. Additional logical attributes are not included in this paper.

Figure 5 The abstract model for SKO types



Related attributes are grouped into six categories as follows:

1. The general category groups the general information that describes the SKO as a whole.
2. The life cycle category groups the features related to the history and current state of this SKO, and those who have affected this SKO during its evolution.
3. The relational category groups features that define the relationship between the SKO and other entities.
4. The technical category groups the technical requirements and technical characteristics of the SKO.
5. The rights category groups the intellectual property rights, authorship, copyrights and conditions of use for the SKO.
6. The meta-metadata category groups information about the metadata instance itself, rather than the SKO that the metadata instance describes.

Each attribute is specified by the following properties:

- *ID*: The unique identifier of an attribute.
- *Name*: The name of an attribute in natural language (NL).
- *Data type domain*: Boolean, integer, float, date, duration, SLS, SS, entity and URL.

Table I SKO Types specification: general category

Name	*Data type*	*Whole/Part*	*Reference*	*Description*	*Example*
Identifier	URL	W and P	DC	An unambiguous reference to the resource within a given context	www.liquidpub.org/doc/SKOTypesV1.9
Description	SS	W and P	DC	An account of the resource	This work is a branch of EType Theory
Language	SS	W and P	DC	A language of the resource	English
Keywords	SS []	W and P	DC: subject	The topic of the resource	Taxonomy mapping, semantic matching, mapping evaluation
Coverage	SS	W and P	DC	The spatial or temporal topic of the resource, the spatial applicability of the resource or the jurisdiction under which the resource is relevant	Sixteenth-nineteenth century, Italy
Creator	Person[] or organization[]	W and P	DC	An entity primarily responsible for making the resource	Hao Xu
Source	URL	W and P	DC	A related resource from which the described resource is derived	www.sweb.com/0001.pdf
Title	SS	W and P	DC	A name given to the resource	SKO Types Version 2.0
Alternative	SS	W and P	DC	An alternative name for the resource	SKO Types Version 2.0
Pattern	SS	W	DC: conform to	An established standard to which the described resource conforms	SKO pattern 001
Author	Person[] or organization[]	W and P	DC: contributor	A set of authors of this SKO	Fausto Giunchiglia, Ronald Chenu
Editor	Person[] or organization[]	W and P	DC: contributor	A set of editors of this SKO. Note: sometimes there is no author for an SKO like an article collection, but editors	Hao Xu
References	SKO[] or SKO node[]	P	DC	A related resource that is referenced, cited, or otherwise pointed to by the described resource. Note: internal reference is form part-to-part, while external one is from part-to-whole	SKO definition V3.0
Serialization	URL	W		An SKOs serialization	Skotypes.serial.xml

Source: Xu (2011)

Table II Comparison between SKO Types and Dublin Core

Dublin core element	*SKO type attribute*	*Whole/Part*	*Date type*	*Category*	*Note*
Contributor	Author editor	W and P	Person[] or organization[]	General	DC Basic Element
Coverage	Coverage	W and P	Formula	General	DC Basic Element
Creator	Creator	W and P	Person or organization	General	DC Basic Element
Date	Date of solidification, date of publication	W and P	Date	Life cycle	DC Basic Element
Description	Description	W and P	Formula	General	DC Basic Element
Format	Format	W and P	Formula	Technical	DC Basic Element
Identifier	Identifier	W and P	SURL	General	DC Basic Element
Language	Language	W and P	Formula	General	DC Basic Element
Publisher	Publisher	W	Person or organization	Life cycle	DC Basic Element
Relation					DC Basic Element (all the relational attributes in SKO Types)
Rights	Copyrights	W and P	Formula	Intellectual property	DC Basic Element
Source	Source	W and P	SURI	General	DC Basic Element
Subject	Keywords	W and P	Formula []	General	DC Basic Element
Title	Title	W and P	Formula	General	DC Basic Element
Type	Kind	W and P	Enumeration <formula>	General	DC Basic Element W: see BibTex P: see LaTex
Abstract					
Access rights	Access rights	W and P	Person[]	Technical	
Accrual method	Conditions	W	Formula	Life cycle	For SKO sets
Accrual periodicity					
Accrual policy					
Alternative	Alternative	W and P	Formula	General	
Audience					
Available					Service(T)
Bibliographic citation					Service(G)
Conforms to	Pattern	W	Formula	General	
Created	Created	W and P	Date	Life cycle	
Date accepted	Date accepted	W	Date	Life cycle	SKOs in SKO sets
Date copyrighted	Date copyrighted	W and P	Date	Life cycle	
Date submitted	Date submitted	W	Date[]	Life cycle	SKOs in SKO sets
Education level					
Extent					
Has format					Service(T)
Has part					Service(G)
Has version	Has version	W and P	SKO	Life cycle	

(*continued*)

Table II

Dublin core element	*SKO type attribute*	*Whole/Part*	*Date type*	*Category*	*Note*
Instructional method					
Is format of					Service(T)
Is part of					Service(G)
Is referenced by					Service(G)
Is replaced by					
Is required By					
Issued					
Is version of					Service(L)
License	License	W and P	Formula	Intellectual property	
Mediator					
Medium					
Modified	Modified	W and P	Date	Life cycle	
Provenance					
References	References	W and P	SKO[]	General	
Replaces					
Requires					
Rights holder					Service(R)
Spatial					
Table of contents					SKO node type
Temporal					
Valid					Service(L)
	Serialization	W		General	
	State	W and P	Enumeration <formula>	Life cycle	
	Submitted to	W and P	SURL[]	Life cycle	

- *Kind*: Strictly mandatory, mandatory, suggested, permanent, temporal, computed, transitive, symmetric.
- *Overrides*: Specifies a more general attribute name that this attribute "over sides".
- *Reference*: For example, Dublin Core, SALT, FOAF, etc.
- *Description*: A brief account of an attribute in NL.
- *Concept ID*: The name of an attribute in FL.
- *Whole/part*: Indicates an attribute may apply in SKOs, SKO sets or SKO nodes.
- *Example*: Indicates when and how to use an attribute.

Table I gives an example of the excerpt version of SKO Types specification, which is being encoded and used in the SWeb system and AISN platform (Giunchiglia *et al.*, 2010a, 2010b).

5. Discussion

Interoperability is one of the most important factors that should be considered during the practical development and implementation processes, as the SKO Types, along with the SKO Patterns and SKO TeX, will be mainly applied in various digital libraries, while for the existing legacy of scientific publications and their associated metadata schemas, it is required to build up a compatible mechanism (Xu, 2010a, 2010b, 2010c, 2010d). This will be one in which the original metadata can be imported into our system on the one hand, generated according to the SKO Types metadata schema, while on the other hand, to promote the proposed standard, the authors hope to provide more convenient updating methods for harmonizing with different kinds of libraries.

So far, the authors have already compared and matched SKO Types with the current metadata standards such as Dublin Core, LaTeX and BiBTeXin several mainstreams. Table II shows the comparison between SKO Types and Dublin Core.

6. Conclusion

In this paper, the authors proposed an entity-oriented theory, i.e. SKO Types, for representing and linking SKO. The authors defined SKO-related entities, relationships between these entities and attributes of each entity in the scientific domain. Such schema will serve as the basis for relating entities, entity components, aggregated entities, relationships and attributes to various tasks, e.g. linked entity, rhetorical structuring, strategic reading, semantic annotating, etc., that users may perform when consulting ubiquitous SKOs.

References

Cantara, L. (2005), "Mets: the metadata encoding and transmission standard", *Metadata: A Cataloger's Primer*, Routledge, London, p. 237.

Delsey, T. (2002), "Functional analysis of the MARC 21 bibliographic and holdings formats", Library of Congress, Network Development, and MARC StandardsOffice, available at: www.loc.gov/marc/marc-functional-analysis/original_source/analysis.pdf

Doerr, M. (2003), "The cidoc conceptual reference module: an ontological approach tosemantic interoperability of metadata", *AI magazine*, Vol. 24 No. 3, p. 75.

Giunchiglia, F., Xu, H., Birukou, A. and Chenu, R. (2010a), "Scientific knowledge object patterns", *Proceedings of 15th European Conference on Pattern Languages of Program*, EuroPLoP 2010, Irsee Monastery, Bavaria.

Giunchiglia, F., Chenu-Abente, R., Xu, H. and Kharkevich, U. (2010b), "A metadata-enabled scientific discourse platform", E-Print, University of Trento, Trento, Technical Report # DISI-10-069, available at: http://eprints.biblio.unitn.it/archive/00001939/01/069.pdf (accessed 10 October 2014).

Guenther, R. (2004), "Premis-preservation metadata implementation strategies update 2:Core elements for metadata to support digital preservation", *RLG DigiNews*, Vol. 8 No. 6.

Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. (2002), "Open archives initiative protocol for metadata harvesting-v. 2.0", available at: www.openarchives.org/OAI/openarchivesprotocol.html

IEEE Learning Technology Standards Committee (2002), 1484.12.1-2002 - IEEE Standard for Learning Object Metadata, available at: http://standards.ieee.org/findstds/standard/1484.12.1-2002.html

National Information Standards Organization (2004), *Understanding Metadata*, NISO Press, Bethesda.

ONeill, E.T. (2002), "Frbr: functional requirements for bibliographic records", *Library Resources & Technical Services*, Vol. 46 No. 4, pp. 150-159.

Pitti, D.V. (2005), "Encoded archival description: an introduction and overview", *ESARBICA Journal: Journal of the Eastern and Southern Africa Regional Branch of the International Council on Archives*, Vol. 20 No. 1.

Xu, H. (2010a), "Managing ubiquitous scientific knowledge on semantic web", *AST/UCMA/ISA/ACN*, in Miyazaki, Lecture Notes in Computer Science, Vol. 6059, Springer, Kanto, pp. 421-430.

Xu, H. (2010b), "A semantic pattern for scientific discourse representation", *Journal of Computational Information Systems*, Vol. 6 No. 13, pp. 4223-4228.

Xu, H. (2010c), "A semantic pattern approach to managing scientific publications", *Automation of Software Test*, in Miyazaki, Lecture Notes in Computer Science, Vol. 6059, Springer, Kanto, pp. 431-434.

Xu, H. (2010d), "New format and framework for managing scientific knowledge", *Future Generation Information Technology*, Lecture Notes in Computer Science, Vol. 6485, Springer, Jeju Island, pp. 290-293.

Xu, H. (2011), "Managing ubiquitous scientific knowledge objects", PhD Dissertation, University of Trento, Trento.

About the authors

Hao Xu is an Associate Professor at College of Computer Science and Technology, Jilin University, China, and a Courtesy Professor of KnowDive group, University of Trento, Italy. He received his PhD in information and communication technologies (ICT) in November 2011 at the Department of Computer Science and Information Engineering, University of Trento, Italy. He has published more than 20 indexed international journal and conference papers. His research interests include semantic annotation, data analysis and gamification. Hao Xu is the corresponding author and can be contacted at: xuhao@jlu.edu.cn

Fausto Giunchiglia is Full Professor of computer science at the University of Trento, Italy. His recent areas of interest are the use of semantics for managing knowledge diversity in the large and social computations, i.e. how to study and exploit the impact of ICT on organizations, people and society, towards the construction of a better society. He has published around 50 journal papers and more than 200 publications overall; has been a speaker at more than 30 invited talks; and has chaired around 10 international events. He has actively participated in many EU-funded projects and acted as coordinator for KnowledgeWeb, OpenKnowledge, Insemtives, LivingKnowledge and SmartSociety.