CDL (Concept Description Language): A Common Language for Semantic Computing

Toshio YOKOI
Tokyo University of Technology
1404-1 Katakura,Hachioji,
Tokyo,192-0982
+81-426-37-2435
yokoi@media.teu.ac.jp

Hiroshi UCHIDA
UNDL Foundation
53-70, Jingumae 5-chome
Shibuya-ku, Tokyo 150-8925
+81-3-3499-2811
uchida@undl.org

Koiti HASIDA
ITRI, AIST and CREST, JST
AIST Tokyo Waterfront
2-41-6, Aomi, Koto-ku,
Tokyo 135-0064
+81-3-3599-8211
hasida.k@aist.go.jp

Hiroshi YASUHARA
Oki Electric Industry Co., Ltd.& ISeC
550-5 Higashi Asakawa
Hachioji, Tokyo 193-8550
+81-426-62-6786
yasuhara223@oki.com

Meiying ZHU
UNDL Foun. & ISeC
53-70, Jingumae 5-chome
Shibuya-ku, Tokyo 150-8925
+81-3-3499-2811
zhu@undl.org

ABSTRACT

CDL(Concept Description Language) is a computer language to describe concept structure of content. CDL consists of a variety of family languages which are based on nested network data model and concept definition dictionary. In this paper, CDL.core which is a core of CDL family languages and CDL.nl which is a basic part of conceptualization of a natural language semantics are presented. CDL is a common language for Semantic Computing in Japan and we will propose to be used as a next generation language for intelligent WWW.

Categories and Subject Descriptors: I.2.4 [**Knowledge Representation Formalisms and Methods**] *Representation languages, Semantic networks*

General Terms: Languages

Keywords: Concept description, Content, Semantic Computing, Semantic Web

1. Semantic Computing Initiative in Japan

The Semantic Computing (SeC) Initiative in Japan is based on the consideration that the following three information and communication technologies(ICT) are very important to establish and maintain intelligent information infrastructures in the 21st century and is being carried out for the purpose of doing research and development of the basic technology (1) while linking with (2) and (3):

(1) Semantic Computing

Technology to realize an intelligent and creative infrastructure

(2) Ubiquitous Computing

Technology to realize an anytime, anywhere, and anything infrastructure

(3) Secure Computing

Technology to realize a secure, stable, and reliable infrastructure

Speaking from the historical view points, the SeC is the realization of spirits in the leading book[1] edited by Marvin L. Minsky. To be concrete, although linking with the R&D activities on Semantic Web which is currently proceeding mainly in the U.S. and Europe, the SeC Initiative replaces "Web" with "Computing" for the following reasons:

– to cover all content formats including Web content

– to cover content as a whole including content body as well as metadata

– to regard content as a basic representation form for all information and knowledge

– to regard computation and communication mechanism for content as the most general one for information and knowledge

With rapidly expanding genres of Web content through the spread of broadband taken into consideration, the SeC Initiative is also positioned as an enabler for the next-generation Semantic Web.

The SeC Initiative is carried out with its themes assigned to various organizations and what was established in January 2004 as an organization to coordinate the whole is the nonprofit organization, the ISeC (Institute of Semantic Computing, http://www.instsec.org). In the ISeC, CDL (Concept Description Language)[2] is being developed to be a common language for the promotion of the SeC Initiative. The first version of the basic specification is currently being developed and the discussion for concrete implementations has started.

WWW 2005, May 10--14, 2005, Chiba, Japan.

2. SeC and CDL

SeC is a new technology to link information processing in human/society with that in computer seamlessly by sharing semantics or having a common semantic world between computer and human/society. The technology to link them seamlessly is a one which regards the form of information in human/society as document (generally content) and which enables a common understanding of document between computer and human.

CDL describes a semantic world or conceptual world shared between human/society and computer. Though being a computer language, CDL is substantially different from existing ones. Existing computer languages have been designed from the viewpoint of computer mechanisms (computation mechanism, data structure, program structure, etc.). CDL, however, is designed (for the first time) from the viewpoint of human/society and of multimedia/content which human/society uses. Still being a computer language, however, CDL will succeed the useful results of existing computer languages.

Furthermore, CDL is designed to play the following roles of an intermediary:

(1) Intermediate language among media: to intermediate among multiple languages, between natural language and formal language (mathematical language, programming language, etc.), and between image/visual media and language media, etc.

(2) Intermediate language from shallow semantic processing on the conceptual level to deeper semantic processing/knowledge processing

(3) Intermediate language between syntactic document structure processing (XML) and semantic document structure processing

What represents well organized information (including knowledge and emotion) by using multimedia is document (generally content). Content means a form of externalizing information in human/society. On the other hand, the forms of information in computer are software and data. With computer sharing a semantic world with human/society, content also becomes software for computer. For this purpose, CDL is designed to describe a range from concept structure of each medium to one of content continuously.

3. CDL and CDD (Concept Definition Dictionary)

The whole CDL (Concept Description Language) and its positioning are shown in Fig.1. The whole CDL is named CDLs. What play a central role within CDLs are CDL.core, which constitutes a core part, and CDL.nl, which constitutes a common part to natural languages. On the CDL.core, CDL describing the conceptualization of media's semantics, CDL describing the conceptualization of objective worlds and content, etc. are developed. This is quite parallel to the way in which various markup languages (markup tag-sets) for media and content are developed on XML .

While each tag-set on XML marks up syntactic structure of content, CDL performs description (including markup function) of semantic structure of content. While XML defines a tag-set for each purpose through DTD (Document Type Definition) or XMLSchema, CDL defines a concept-set for each purpose through CDD (Concept Definition Dictionary). What describes definitions in CDD is CDL.core.

While XML makes document structure human-visible through XSL (XML Stylesheet Language) and browser, CDL makes concept structure human-readable through translation dictionaries and sentence generation program. For XML, various types of software including editor, browser, and checker have been developed. For CDL, on the other hand, a software suite called CDL system including translation dictionaries and sentence generation program will be developed.

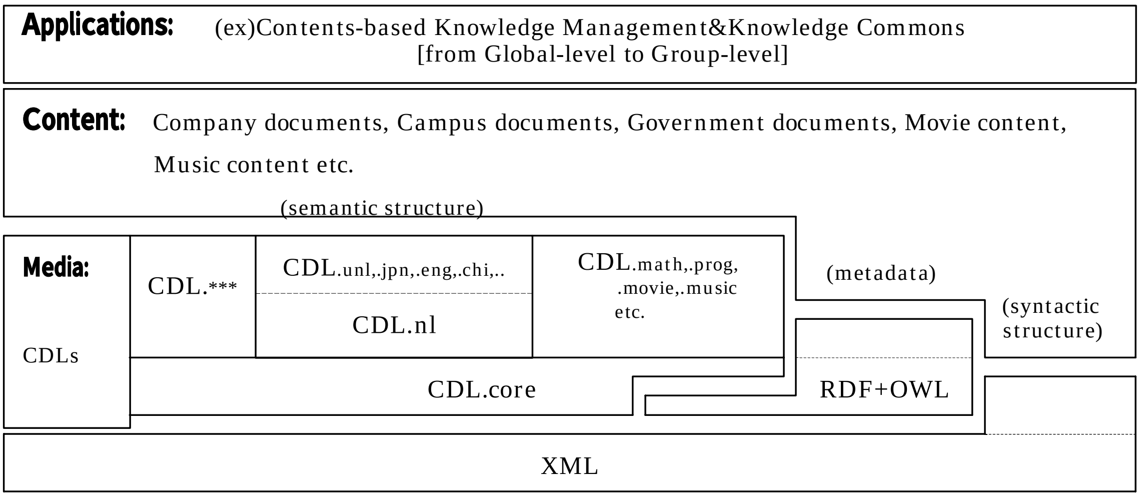
Applications:
(ex)Contents-based Knowledge Management&Knowledge Commons
[from Global-level to Group-level]
Content:
Company documents, Campus documents, Government documents, Movie content, Music content etc.
(semantic structure)
Media:
CDLs
CDL.***
CDL.unl,.jpn,.eng,.chi,..
CDL.nl
CDL.math,.prog, .movie,.music etc.
(metadata)
(syntactic structure)
CDL.core
RDF+OWL
XML

Figure 1. CDLs Layer

CDL.core:

Language part to be a basis for all. It is a meta-language to describe individual CDD's concept definitions. In CDD.core, vocabularies and sentence patterns for this purpose are defined. These definitions are a priori.

CDL.nl:

In CDD.nl, concepts common to all natural languages are defined. Concepts which describe the conceptualization of word, phrase, sentence, and text semantics are defined.

CDL.unl:

The CDL version of UNL (Universal Networking Language)[3], the international common language which is being promoted and developed by United Nations University and the UNDL Foundation (http://www.undl.org/). In CDD.unl, global common concepts to describe information and knowledge are defined.

CDL.jpn, CDL.eng, CDL.chi, etc.:

CDL for natural languages such as Japanese, English, and Chinese. For each language, CDD to define concepts of thing and event is provided. Furthermore, for each natural language, through the development of CDD for daily use language, CDD for each terminology or each community, and CDD for a controlled language, CDL for various language usages can be defined.

CDL.math, CDL.prog, CDL.movie, CDL.music, etc.:

CDL for mathematical formula, programming language, video and image, musical sound, etc. Concepts peculiar to each medium and concepts incorporated from natural language constitute its CDD.

CDL.***:

For different object worlds, different types of CDD.*** for their conceptualizations are defined. Linking with CDD for media, CDL for each object world and each document format is defined.

4. Comparison between CDL and other languages

CDLs will be positioned through their comparison with related technologies and similar trends.

[RDF+OWL]

First, the comparison with the Semantic Web. Though CDL.core roughly corresponds to the RDF+OWL[5] [6]part in the Semantic Web, the design principles are in contrast with each other. The design principle of RDF+OWL is to describe property of resource. On the other hand, that of CDL is to describe conceptual structure of objective content and world. Content and world are also regarded as resource. Therefore, the function of RDF+OWL is to describe property of content and world, that is, to make metadata of content and world machine-understandable. On the other hand, CDL has a purpose of making concept structure of content and world machine-understandable. However, structure description and property description are utterly different but complementary as well. Therefore, CDL and RFD+OWL are utterly different and complementary. Accordingly, in developing the specification of CDL, similar parts in the specification of RDF+OWL is utilized as much as possible and the implementation method on RDF+OWL is also proposed.

[XML]

As for content structure, XML provides markup function of syntactic structure. Syntactic structure of content is approximated by nested tree structure. On the other hand, concept structure of content is approximated by nested network structure. Therefore, CDL provides functions in an integrated way to be able to handle network structure in a comprehensive and focused manner which XML can handle only in a specific and supplementary manner.

Since syntactic structure guides semantic structure, CDL requires a function to add semantic structure to XML element tag structure, that is, a markup function. With semantics of content parts approximated by concepts in CDD (Concept Definition Dictionary), description data to be processed as concept structure separate from content is obtained. In terms of markup function, CDL finds the following new positioning among existing markup languages:

HTML: markup of visual structure of content

XML: markup of syntactic structure of content

CDL: description and markup of concept structure of content

[MPEG7]

As for standardization of markup of structures of content body, there is a precedent result called MPEG-7[7]. A part particularly related to CDL is MDS (Multimedia Description Scheme). Through the incorporation of Linguistic DS into MDS with a focus on markup of syntactic information in natural language, an international standard was compiled on July 1, 2004. However, the specification still has the following defects:

(1) Due to the interests of the existing MPEG, a focus is placed on movie, music, and sound contents and the positioning of language content is weak. For this reason, the function to handle linkages among media which normally should be a central function of MDS is poor.

(2) Being unable to discard the approach to include all markup information in annotation, trying to write all by using only XML has resulted in the unnecessarily huge specification with poor perspective.

What overcomes the defects of MPEG-7 is CDL. CDL changes the viewpoint and focuses on the description of concept structure. For this purpose, an approach from natural language which manages conceptual linkages among media is regarded as important. And not annotation but extended ontology which is description, markup based on description, or CDD (Concept Definition Dictionary) underlying them is placed in the center.

5. CDL.core and CDL.nl

5.1 CDLs common

This is the specification of common language features to all CDLs families. It includes the specification of model of concept and concept structure, basic data model underlying them (basic syntax of language), dictionary structure common to all CDD, etc. CDLs common specification roughly corresponds to RDF Model&Syntax in RDF+OWL.

[Model of concept and concept structure]

The two most basic description elements to describe concepts are Entity and Relation. A structure called directed graph in which Entity is regarded as a node and Relation is regarded as an arc, that is, network structure represents a concept structure. Entity and Relation can have an arbitrary number of Attribute-Value pairs added to describe their properties in detail.

Entity can be classified into an elemental thing, that is, Elemental Entity, and a composite thing, that is, Composite Entity. Composite Entity is a hyper node which contains network structure of Entity and Relation within it. Unlike a hyper node in graph theory, however, nodes inside and outside Composite Entity may be linked with each other by a direct arc.

Relation can be also divided into Elemental Relation and Composite Relation. Composite Relation is a hyper arc or macro arc which is defined by the network structure of Entity and Relation.

Attribute follows Relation, while Value follows Elemental Entity. However, Value is never linked with any other Entity by Relation or Attribute. That is, Attribute-Value pair is local description of its Entity and Relation.

On the level of language specification, arbitrary network structure and arbitrary hyper nested structure are assumed as concept structures. However, these structures may presuppose appropriate controls. Syntactic structure to be described is a tree structure. Generally, syntactic structure guides semantic structure, that is, concept structure. Therefore, the network structure which is a concept structure reflects tree structure in some way or other. With the reflections taken into consideration, efficient network processing program can be implemented. And the most general ones of the reflections are incorporated explicitly into language specification of CDL.

[Basic data model]

As for symbolic encoding methods of concept and description scheme of elemental concept, the following four types are set up:

(1) Concept realization(instantiation) from concept definitions

Concept realization from concept definitions in CDD (Concept Definition Dictionary). Concept identifier (*ConceptID*) in concept definition with realization suffixes (*#integer*, etc) affixed.

(2) Textual concept which text represents (*TextualConcept*)

Concept which text written in languages represents. It is what a standard reader recalls when he reads the text in an ordinary situation. Languages include natural languages in countries and formal languages.

(3) Quotation concept which quoted sentence, phrase, and word represent(*QuatationConcept*)

Concept which sentence, phrase, word quoted from text represent. It is concept which is read from these sentence, phrase, and word in a general context.

(4) Literal concept

What represents unique concept wherever it is, such as numbers including integer, real number, etc. and identifiers including URI, etc.; its notation itself is regarded as concept identifier.

A common formal language to describe all CDL expressions is the basic data model. It is a graphical or textual notation form which represents a model of concept structure. As notation forms, the following five types are set up. (1) and (2) are original forms, while (3), (4), and (5) are forms for implementation:

(1) Graph notation

Graphic notation to describe model of concept structure directly

(2) Text notation

Textual notation of concept structure by using simple syntactic sugar

This notation is converted into XML notation, RDF notation, and UNL notation.

(3) XML notation

XML text notation of concept structure

This leads to CDL implementation on XML.

(4) RDF notation

Textual notation through RDF basic data model

This leads to the implementation on RDF.

(5) UNL representation

Textual notation through UNL triples

This leads to the implementation on UNL.

[Description scheme of concept]

Top layers of all CDDs (Concept Definition Dictionaries) have the same structure. This means that CDD.xxx, the Concept Definition Dictionary of CDL.xxx, has the following structure with the concept definition called Concept.xxx on its top:

Concept.xxx

・Entity

・・ElementalEntity

・・CompositeEntity

・Relation

・・ElementalRelation

・・CompositeRelation

・Attribute

Basic data model is provided for the concepts on the top layers. For example, the basic data model for CompositeEntity is as follows:

[Graph notation]

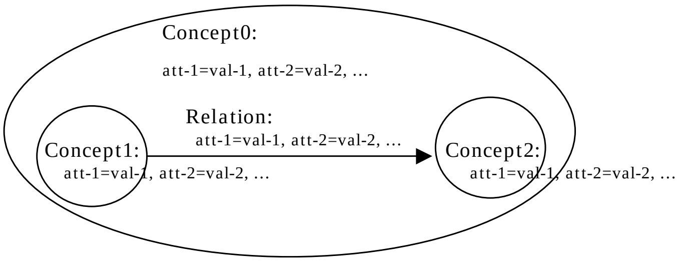
Concept0:
att-1=val-1, att-2=val-2, …
Relation:
att-1=val-1, att-2=val-2, …
Concept1:
att-1=val-1, att-2=val-2, …
Concept2:
att-1=val-1, att-2=val-2, …

Figure2. Graphical Notation

[Text notation]

｛*Concept0*　*attribute*=*value* ⋯ :

　　｛*Concept1*　*attribute*=*value* ⋯ : ・・・｝

　　｛*Concept2*　*attribute*=*value* ⋯ : ・・・｝

　　｛*Relation*　*attribute*=*value* ⋯ : ｝

　　・・・

　　・・・

　　Concept1－*Relation*→*Concept2*

　　・・・

　　・・・｝

5.2 CDL.core

CDL.core is a meta language to define CDDs for all CDLs. Meta vocabulary and meta sentence patterns for the meta language are defined in the Concept Definition Dictionary for CDL.core. These definitions are a priori, however. The relation between CDL.core and each Concept Definition Dictionary of CDL.xxx to be defined, that is, the relation between CDD.core and CDD.xxx is as follows:

Concept.core

・Entity

・・Concept.xxx

・・・Entity

・・・Relation

・・・Attribute

・Relation

・Attribute

In CDL.core, concepts (Entity, Relation, Attribute) of CDL.xxx are all regarded as entity concepts, that is, objectified. Through the utilization of language for CDL.core, that is, entity concept, relation concept, and attribute concept of CDL.core, those of CDL.xxx are objectified, that is, the definitions of them are asserted as Entities. Sentence format for the assertion is Concept Schema. The Concept Schema includes the following vocabularies:

ConcpetDefinition: assertion to put a group of concept definitions together

NewName: declare all ConceptID names in a lump to be used in definition

ConceptSchema: assert semantic relation of concept to be defined to the known concepts

EntitySchema: assert what is to be defined in entity concept

RelationSchema: assert what is to be defined relation concept

AttributionSchema: assert what is to be defined in attribution concept

Constraint: assert a constraint among elements in definition

5.3 CDL.nl

[Concept model of natural language semantics]

The concept model(conceptualization) of semantics common to every natural language is defined as follows. As for semantics expressed in natural languages, there are many theories and opinions regarding their levels and viewpoints. The semantics to be covered here is the surface level semantics as the most general one. Through the conceptualization of surface level semantics, the deeper level one will be approached effectively.

Conceptualizations of sentence and text are individually described as Composite Entities. A Composite Entity of text contains several Composit Entities of a sentence as nodes, which are linked by two types of relations. One is the conjunctive relation which describes conjunctive structure through conjunctions, etc., that is the relation which links sentence Entities each other. The other is the reference relation which describes reference structure through referring expression, substitution expression, etc., that is the relation which links the node within one sentence with another sentence or the node within that.

A sentence is divided into propositional part and modal part, which is then conceptualized into Composite Entity of a sentence. Propositional part is described by Entities and Relations, while modal part (aspectuality, temporality, polarity, sentense-modality, discource-modality) is described by Attribute-Value pairs. Propositional part is the conceptual structure in which Entities, that is, predicate, predicate modifier, and several case components are linked by relations called case relations. If being composed of composite words, phrase, and clauses, each of those components is conceptualized as Composite Entity. Within Sentence Composite Entity, the relation which represents reference structure inside as well as case relation is described.

[Concept Definition Dictionary: CDD.nl]

In the Concept Definition Dictionary of CDL.nl (CDD.nl), concepts based on the concept model are defined. The relationship between CDD.nl and CDD.core is as follows:

Concept.core

・Entity

・・Concept.nl

・・・ Entity

・・・Relation

・・・Attribute

・Relation

・Attibute

CDD.jpn, CDD.eng, CDD.chi, etc. corresponding to CDL.jpn, CDL.eng, CDL.chi, etc. for natural languages are composed in a way to refine the lower layer concepts by regarding CDD.nl as a common framework. For each national language, controlled languages can be designed according to its fields and uses. For each controlled language, CDDs are developed, all of which are composed by regarding CDD.nl as a common framework.

6. The mission of CDL

Using CDL as a common computer language, various R&D activities of the Semantic Computing will start. The activities are expected mainly in three directions: creation and accumulation of contents which are described and marked up by CDL, research on the new method of semantic representation and knowledge representation, and research&development toward the Semantic ICT.

The creation and accumulation of content underlies the activities in the other two directions. The CDL description and markup of the digital content and XML tagged content will be described and marked up by CDL, and will be accumulated for sharing.

Research of semantic representation and knowledge representation includes deeper semantics processing and meaning understanding, knowledge extraction from content, more sophisticated inference engine, etc.

The Semantic ICT makes ICT fields more intelligent, more familiar, more reliable based on the Semantic Computing and is expected to progress into the following, for example:

Semantic CAL (Computer Assisted Learning)

Semantic CAD (Computer Aided Design)

Semantic CASE (Computer Aided Software Engineering)

Semantic CACE (Computer Aided Content Engineering)

Semantic CASEE (Computer Aided Service Engineering)

Semantic KM (Knowledge Management)

7. ACKNOWLEDGEMENTS

We express our sincere thanks to Prof. Mitsuru Ishizuka, the chairperson of WF04, for his helpful supports.

8. REFERENCES

[1] Minsky, M. L. (ed.). Semantic Information Processing, The MIT Press, 1969.

[2] CDL Specification Manual (version-1), ISeC, 2005.

[3] SemanticWeb.
http://www.semanticweb.org

[4] UNL.
http://www.undl.org/

[5] OWL.
http://www.w3.org/2004/OWL/

[6] RDF.
http://www.w3.org/RDF/

[7] MPEG7.

http://ipsi.fraunhofer.de/delite/Projects/MPEG7/