社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DD－63 (1)
2007／9／28

情報技術テキストコンテンツのためのOWL記述ツールの開発

都原 安貴[†] 塚本 享治[†]
東京工科大学[†]

ネットワーク上のリソース増加に対して検索時間が増大する問題がある。その 1 つの解決方法としてセマンティックウェブが提唱されている。セマンティックウェブの現在の状況は、RDF/OWL というメタデータが標準化され、メタデータの構築の段階に入りつつあり、そのためのエディタが提供されてきている。

しかし、現実にはメタデータの構築が進んでいるとは言いがたい。理由はいくつか考えられるが、次の要因が大きな理由として挙げられる。実際のデータに対するメタデータとして OWL を記述するためには、クラス、プロパティ、インスタンスが必要になるが、インスタンスを構築するまでのプロセスが複雑で難しく、時間がかかってしまうという理由である。

本研究では、情報技術テキストコンテンツのための OWL を記述するためのツールを開発し、OWL の構築を行った。

The development of the OWL description tool for information technology text contents

Yasutaka Tohara[†] Michiharu Tsukamoto[†]
Tokyo University of Technology[†]

There is the problem that increases of search time for resource increase on the network. Semantic Web is proposed as its solution. Meta data called RDF/OWL is standardized, and the present of Semantic Web is a stage of the construction of Meta data, and the editor has been provided.

However, it is hard to say that the construction of Meta data actually advances. There are some reasons, but the next factor is given for a main reason. Class, property and instance are necessary to describe OWL as Meta data for actual data and the processes before building instance is complicated and difficult and takes time.

In this study, we developed the tool to describe OWL for information technology text contents and built OWL.

1．はじめに

近年、インターネット上のリソースの増加に伴う検索時間の増加が問題として指摘されている。その 1 つの解決方法としてセマンティックウェブが提唱されている。セマンティックウェブの現在の状況は、RDF/OWL というメタデータの仕組みが標準化され、メタデータの構築の段階に入りつつあり、そのためのエディタも提供されてきている。

セマンティックウェブの実現によって期待されているのは、ユーザに対する検索支援である。OWL によって表現されたメタデータを解析することで、ユーザは検索に必要な知識を事前に手にすることができる。

しかし、OWL が提供する知識表現をコンピュータに分かりやすい XML などの形式で記述することは複雑で難しく、時間がかかる。そこで、本研究では、情報技術テキストコンテンツに対する OWL 記述をシンプルわかりやすく行うためのツールを開発し、OWL の構築を行った。

2．OWL 記述の現状と課題

OWL 記述の現状の課題は 2 つある。1 つ目は、記述方法自体が複雑な点である。OWL はそもそもの意味がわかりにくく、また、OWL の意味がわかっているような人でも、既存のツールを利用した記述に時間がかかってしまう。そのため、OWL を記述できる人が少なく、メタデータが普及していない。2 つ目は、OWL を記述する人が、ナビゲートしたい分野に関して知識を持っていなければならない点である。OWL はユーザをナビゲートする知識表現であるため、OWL を記述する人にはナビゲートする分野に関してユーザ以上に知識が必要になるのである。

これを解決するためには、OWL の記述プロセスを分割したツールが必要になる。

OWL の記述は大まかに、次のような 4 つのプロセスになる。

(1) クラス設計プロセス
(2) データタイププロパティ設計プロセス
(3) オブジェクトプロパティ設計プロセス
(4) インスタンス入力プロセス

(1) から (3) までの作業は、ナビゲートしたい分野の知識だけでなく、OWL を記述するための知識が求められる。1 つ 1 つのクラス、オブジェクトプロパティ、データタイププロパティの作成に時間がかかり、それぞれの作業を分割して行うことが難しいが、量が必要になる作業ではなく、1 度決定すれば大きな変更が発生しない設計に近い作業になる。一方で、(4) の作業は、ナビゲートしたい分野の知識だけで行える。1 つ 1 つのインスタンス作成には時間はかからず、作業も分割することが可能であるが、量が必要になる作業である。

既存の OWL 記述ツールは (4) の機能に特化して OWL を記述することができないので、本研究では、(4) に特化した OWL 記述が可能なツールの開発を行い、OWL の知識に乏しいユーザが記述可能な環境を提供する。

3．本研究で利用する OWL の設計

3・1　本研究がナビゲートする分野

本研究では、情報技術に関するテキストに対する OWL を記述するツールの開発を行う。筆者の知識分野が情報技術に強いので、ツールの利用と評価を行うことができるからである。

本研究では、クラスの設計、データタイププロパティの設計、オブジェクトプロパティの設計を事前に行い、ツールに組み込むことで、ユーザがインスタンスを作成することに集中できるようにする。

3・2　情報技術テキストコンテンツの OWL

情報技術テキストコンテンツの OWL クラスは図 1 のように設計した。テキストとコンテンツは、目次とページのような関係である。1 つ 1 つの内容はコンテンツの中に入っていて、それらを統合するのがテキストである。

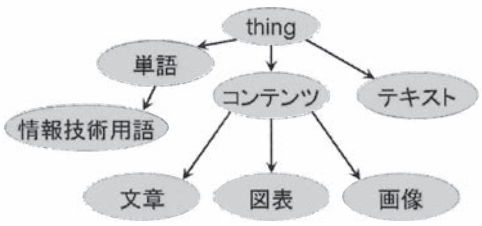

図1　設計した OWL クラス

OWL データタイププロパティはコンテンツ、テキストに対して設計を行った。コンテンツに対するデータタイププロパティは次の 7 つである。制約はそのデータタイプが何個なのかを示す。

・コンテンツ名(制約：1)
・作成者名(制約：1 以上)
・作成日付(制約：1)
・修正日付(制約：0 以上)
・URL(制約：1)
・修正 URL(制約：0 以上)
・オリジナルコンテンツ(制約：1)

オリジナルコンテンツは、コンテンツが何度か修正された時に、どのコンテンツが最初に作成されたのかを明示するための項目である。修正コンテンツの差分を保存していくことで、検索に引っかかるコンテンツが増加しすぎることを抑えるためのである。それ自身がオリジナルコンテンツであれば、URL と同じ値を指定する。

コンテンツの内容をデータタイププロパティとして保存することも検討を行った。しかし、OWL にコンテンツ内容を保存した場合、1. OWL の複雑な構造から取り出さなければならないこと、2. コンテンツの再利用性を考えた場合、XML 階層の一貫性が保障されていない OWL は XLST などでの変換が非常に困難になること、の 2 点からコンテンツ内容は柔軟でシンプルな形式で保存した。

テキストに対するデータタイププロパティは次の 7 つである。構造はコンテンツの場合と全く同じである。

・テキスト名(制約：1)
・作成者名(制約：1 以上)
・作成日付(制約：1)
・修正日付(制約：0 以上)
・URL(制約：1)
・修正 URL(制約：0 以上)
・オリジナルテキスト(制約：1)

OWL オブジェクトプロパティは、次の 4 種類を作成した。

・単語と単語の関係
・テキストとコンテンツの関係
・テキストと情報技術用語の関係
・コンテンツと情報技術用語の関係

単語と単語の関係が、従来までの OWL 記述で注目されてきた点であり、現在でも様々な研究がなされている。現時点で、莫大な量の単語と単語の関係を自動的に記述する明確な解決策はまだ研究段階である。そこで手作業によって解決する方法があるが、そのためにはコンテンツ作成者の協力が必要になる。しかし、コンテンツ作成者が単語と単語の関係を作成することは難しい。それは次の 2 つの理由による。

(1) コンテンツ作成者が OWL に熟知しているかどうか不明である。

(2) コンテンツ作成者にとって、メリットが見えない。

(1) に関しては、本ツールによって解消を目指す。問題は (2) である。コンテンツと単語の関係を記述することで、作成したコンテンツの利用が促進されることは明確なため、コンテンツ作成者をメタデータ記述に協力させやすいが、単語間の関係の記述では、必ずしもコンテンツ利用が促進されるわけではないため、コンテンツ作成者にとって単語間の関係記述がメリットになりにくい。

そこで、コンテンツ作成者に単語間の OWL 記述に協力させるためには、コンテンツと単語の関係の OWL 記述の過程で、単語間の関係を記述可能にするオブジェクトプロパティを考え出すことが不可欠であると考えた。それを前提として、作成した OWL は図 2、図 3、図 4 のように設計した。（これは現在でも発展させる予定である。）

図 2 のオブジェクトプロパティを基本として考えている。これをコンテンツ側から記述するために、図 3、図 4 のようなオブジェクトプロパティを考えた。図 3、図 4 の太字のプロパティは、図 2 のプロパティと関係のあるプロパティである。

システム側からこの 2 つのプロパティを関係付けて記述することで、コンテンツと単語の関係から単語と単語の関係を作成できるようにする。

この関係は、実際のテキストにメタデータを付与しながら試行錯誤して決定していった。

同意語である
古い用語である
反意語である
関係がある
属している
所属クラスが同じである
具体例である
構造である
機能である

利用例である
利用技術である
部品である
　入力部品である
　出力部品である
内部の表現である
内部の動作である
利用している単位である
仕組みである
手法である
体系である
部品である
分類である
人物の関係である
　発明者である
　発見者である
　開発者である
　提案者である
種類である
下位概念の関係である
下位概念の比較である
性能を表す

図2　単語と単語の関係

説明である
　概念の説明である
　機能の説明である
　仕組みの説明である
　構成の説明である
　利用例の説明である
　目的の説明である
　背景の説明である
　評価の説明である
　性能の説明である
　同一の説明である
　現状の説明である
　具体例である
　利用技術の説明である
　部品の説明である
　　入力部品の説明である
　　出力部品の説明である
　内部表現の説明である
　内部動作の説明である
　単位の説明である
　体系の説明である
　分類の説明である
方法である
方式である

比較である
　同等概念との比較をしている
　反対概念との比較をしている
　下位概念同士の比較をしている
人物のコンテンツである
　発見者のコンテンツである
　作成者のコンテンツである
　発明者のコンテンツである
　提案者である
古いコンテンツである
歴史である
関係がある
　下位概念または部分概念に関するコンテンツである
　上位概念または属する概念に関するコンテンツである
　同等概念または同等クラスに関係するコンテンツである
　反対概念に関係するコンテンツである

図3　コンテンツと単語の関係

主題である
古いテキストである

図4　テキストと単語の関係

テキストとコンテンツの関係は、図 5 のようになっている。コンテンツ側から見た OWL である。テキスト側から見た OWL は図 6 のようになる。

コンテンツ管理に関するデータは、所属関係のみに絞った。OWL によってコンテンツを管理するのは複雑になりすぎるからである。コンテンツとテキストの細かい階層関係などに関しては、テキストのデータタイププロパティの URL で参照できる XML に記述を行った。

属している
テキストの中で重要である

図5　テキストとコンテンツの関係

一部である
重要なコンテンツである

図6　テキストとコンテンツの逆関係

上記のように設計した OWL を元に、インスタンスを作成する OWL 記述ツールの開発を行った。

４．OWL 記述ツールの開発

4・1　OWL 記述ツールの全体構成

OWL 記述ツールの全体構成は図 7 のようになっている。OWL のインスタンスを作成し、インスタンスを元にテキスト、コンテンツの検索を行っている。

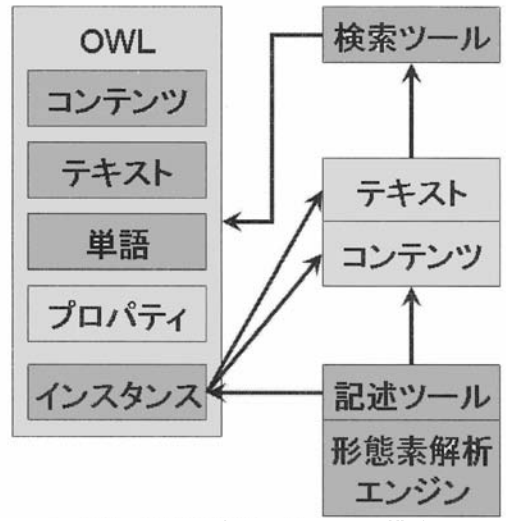

図7　OWL 記述ツールの全体構成

4・2　記述ツールの開発

4・2・1　記述プロセスの軽減

テキストコンテンツのインスタンスを記述するプロセスは既存のツールの場合、4 つになる。

(1) テキスト、コンテンツのインスタンスを作成する
(2) テキスト、コンテンツからキーワードを選び、キーワードのインスタンスを作成、または選択する
(3) テキスト、コンテンツとキーワード間のオブジェクトプロパティを選択する
(4) キーワード間のオブジェクトプロパティを選択する
(5) テキスト、コンテンツのデータタイププロパティの値を入力する

本研究で開発するツールでは、これらのプロセスを簡略化、支援する。

テキスト、コンテンツのインスタンスはオートナンバリングで処理を行う。本来であればインスタンス名を利用するのが一番シンプルであるが、OWL ではインスタンスなどで同名の存在が許されていないため、競合が発生する可能性のあるコンテンツ、テキストのインスタンス名はオートナンバリングによって処理を行う。

テキスト、コンテンツからキーワードを選択するために、形態素解析による名詞抽出を行う。それにより、従来までの手入力による負担を軽減できる。

データタイププロパティは、システム側でデータが全て管理できるので、ユーザが特別に入力する必要がないようにしている。

4・2・2　実際の記述画面

OWL の記述は図 8、図 9 の画面から行う。コンテンツの入力を行った後、形態素解析を行い、名詞の中からプロパティを選択する流れになる。図 9 のプロパティは、あらかじめ設計した OWL のオブジェクトプロパティを読み込み、ユーザに選択させる。

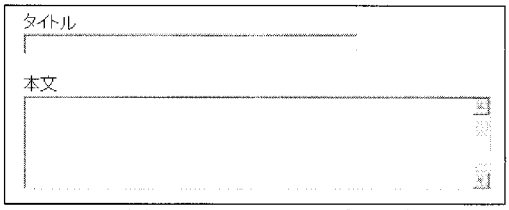

図8　コンテンツ入力画面

単語とコンテンツの関係は、プロパティを選択するだけで可能となる。

また、単語と単語の関係を作成する場合は、図 9 の枠で囲まれた場所に、もう 1 つのキーワードを入力することで可能になる。例えば、図 9 のようにメモリの SRAM の解説をしている文書において、キーワード SRAM は DRAM と「**所属クラスが同じ概念**」である。ここで、キーワード SRAM のプロパティを「**同等概念または同等クラスに関係するコンテンツである**」に設定し、右側の空欄に DRAM と入力することで、SRAM と DRAM のインスタンスおよびその関係が作成される仕組みになっている。

これらの入力を経て作成された OWL のインスタンスは、あらかじめ設計された OWL に追加される。

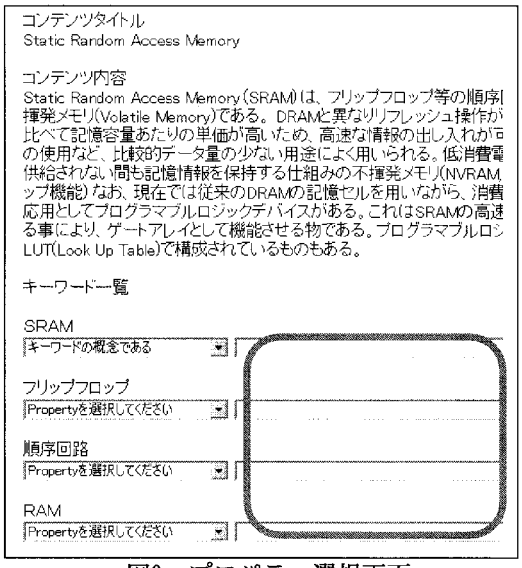

図9　プロパティ選択画面

4・3　検索ツールの開発

検索ツールは、ユーザに対してプロパティの選択を支援させることができる。検索のプロセスは次のようになる。

(1) 入力された単語と関係のある単語があるかどうかを探す

(2) 関係のある単語がある場合、その単語も含めて、OWL の中から選択された関係のあるコンテンツ、テキストを探し出す

(3) 絞られたコンテンツから、ユーザに選択をさせる。

上記のようにして、検索が実現される。

5．情報技術 OWL の作成

このツールを元に、情報技術テキスト、コンテンツ、OWL の作成を行った。情報技術テキストの題材として、ハードウェアとソフトウェア（基本情報図解テキスト）を元に、OWL の記述を行った。

ハードウェアとソフトウェアは 300 ページ強の書籍である。内容は、情報処理技術者試験問題のための学習教材であり、ハードウェア、ソフトウェアの内容を中心にしている。

作成したコンテンツ数は 1150 になる。コンテンツは、1 つの節程度の区切りである。

6．実験と評価

本研究で作成した OWL 記述ツールの評価や課題を述べる。

6・1　実験

本研究によって OWL を利用したコンテンツの検索利便性が向上したかどうかを確認するため、「コンピュータ」という言葉を知るために検索した時にどのくらい利便性が高くなっているかを評価する。

6・1・2　全文検索による結果

「コンピュータ」という言葉で全文検索を行うと、166 のコンテンツ（ページ数で言うと、120 ページ程度）が検索対象として引っかかる。内容に関しての精査は自分で行わなければならず、また、「コンピュータグラフィックス」のように、本来の意味とかけ離れてしまうような複合語や、何気なくコンピュータと記述してあるコンテンツが引っかかってしまう。

6・1・3　OWL を利用した検索による結果

166 のコンテンツを上記のプロパティによって分類することができた。また、コンピュータグラフィックスなどの全く別の単語や意味のない記述を 20 コンテンツ省くことができ、検索性能を向上させることができた。また、「コンピュータ」が書かれているコンテンツ内で、直接関係する単語と「コンピュータ」との関係を作成することができ、それを元に「コンピュータ」という単語をより知ることができた。

6・2　既存の OWL 記述ツールとの比較

6・2・1　Protégé との比較

Protégé はスタンフォード大学が開発している有名な OWL 記述ツールである。

Protégé と比較した場合、インスタンス作成時間において、本研究の OWL 記述ツールが勝っている。本研究のツールでは、コンテンツマネージメントの部分が記述ツールに組み

込まれているので、インスタンスを別途選択したり入力したりする必要がなく、データタイププロパティを入力する必要がない、コンテンツ、テキストと単語の関係を作成することによって単語間の関係も作成されるため、作業効率が高い、キーワードが形態素解析により選択されるため、ユーザがキーワードを改めて入力する必要がない、などの利点がある。

一方、本研究ツールでは OWL 記述の柔軟性にかけている。クラス、データタイププロパティをシステム作成後に変更を行った場合、システムを大幅に変更しなければならない。Protégé は、OWL を余すことなく記述することに特化しているので、柔軟に OWL を記述することができる。

クラスやデータタイププロパティを頻繁に変えながら OWL 記述を考える必要がある、設計などの場合には Protégé を採用し、インスタンスを多く作成することに特化する場合には本研究でのツールを使用することが望ましいだろう。

6・2・2 その他ツールとの比較

・SWOOP

基本的な OWL 作成手順は Protégé と同様になる。Protégé との比較内容は変わらない。

・OILed

これも、基本的な OWL 作成手順は Protégé と同様になる。Protégé との比較内容は変わらない。

・法造

基本的なオントロジ作成手順はこれも変わらない。しかし、オントロジの構造に特徴があり、構築がしやすい GUI ツールである。現在のところ、OWL には対応していないが、すばらしいツールである。

6・3 課題

最終的に、OWL の記述を効率化することができ、テキストからの OWL の記述を行うことができたが、いくつかの課題が残った。

(1) Protégé に代表されるツールに比べ、OWL を差し替えられる可能性が非常に低いのは問題であった。設計した OWL に応じてシステムを新たにジェネレートする仕組みが必要であると考える。一番の問題は、データタイププロパティの変更である。データタイププロパティは、いうなれば、データベースで言うテーブルの変更に近い。それだけ、インスタンス作成の根幹に関わる部分である。ここに、クリティカルなデータを入力するような OWL にする場合は、データタイププロパティの部分を専門に入力するツールが必要になる。

(2) コンテンツ内に記述されていない単語の扱いを考える必要があること、また、細かいプロパティの分類をせざる得ない一方、細かいプロパティの設定に悩む場面に多々遭遇した。プロパティ選択のためのナビゲーションおよび、そのための仕組みをツールに実装する必要がある。

(3) コンテンツの分け方に関して、本研究では書籍の節に当たる部分を元に分割を行った。もう少し大きい単位でコンテンツを分割した場合の実験を行い、コンテンツの分割が結果に影響を与えるのかどうか調べる必要がある。

7．おわりに

本研究において、OWL のインスタンス部分を効率的に作成するツールを開発し、情報技術テキストコンテンツとその OWL を記述することができた。

参考文献

[1] Resource Description Framework (RDF), W3C, "http://www.w3.org/RDF/"
[2] RDF Vocabulary Description Language 1.0:RDF Schema, W3C, "http://www.w3.org/TR/rdf-schema/"
[3] Web Ontology Language (OWL), W3C, "http://www.w3.org/TR/OWL/"
[4] ハードウェアとソフトウェア (情報処理技術者試験基本情報図解テキスト), NEC E ラーニング事業部
[5] CD-ROM で始めるセマンティック Web, Grigoris Antoniou, Frank van Harmelen
[6] セマンティック・ウェブのための RDF/OWL 入門, 神崎正英
[7] オントロジ技術入門-ウェブオントロジ OWL (セマンティック技術シリーズ), AIDOS
[8] オントロジー構築入門, 古崎晃司, 笹島

宗彦, 來村徳信
[9] オントロジー工学(知の科学), 溝口理一郎, 人工知能学会
[10] コンテキスト依存性に基づくロール概念組織化の枠組み, 砂川 英一, 古崎 晃司, 來村 徳信, 溝口 理一郎, 人工知能学会論文誌 Vol. 20 (2005)
[11] セマンティック Web 推論と議論エージェント推論の統合, 若木 利子, 沢村一, 福本 太郎, 向井 孝徳, 新田 克己, 人工知能学会論文誌 Vol. 22 (2007)
[12] オントロジーに基づく機能的知識の体系的記述とその機能構造設計支援における利用, 來村 徳信, 笠井 俊信, 吉川 真理子, 高橋 賢, 古崎 晃司, 溝口 理一郎, 人工知能学会論文誌 Vol. 17 (2002)
[13] オントロジー工学に基づく機能的知識体系化の枠組み, 來村 徳信, 溝口理一郎, 人工知能学会論文誌 Vol. 17 (2002)
[14] 「ロール」および「関係」に関する基礎的考察に基づくオントロジー記述環境の開発, 古崎 晃司, 來村 徳信, 池田 満, 溝口 理一郎, 人工知能学会論文誌 Vol. 17 (2002)
[15] オントロジー構築・利用環境「法造」の開発と利用-実規模プラントのオントロジーを例として-, 古崎 晃司, 來村 徳信, 佐野 年伸, 本松 慎一郎, 石川 誠一, 溝口 理一郎, 人工知能学会論文誌 Vol. 17 (2002)
[16] タスク・ドメインロールに基づくオントロジー構築ガイドシステムの設計と開発-石油精製プラントを例として-, 石川 誠一, 久保 成毅, 古崎 晃司, 來村 徳信, 溝口 理一郎, 人工知能学会論文誌 Vol. 17 (2002)
[17] 学術分野動向把握のためのオントロジー構築, 荒木次郎, 人工知能学会発表資料, (2007)
[18] 学習・教授理論オントロジーの構築と利用～Theory-aware なオーサリングツールの試作～, 林雄介, Jacqueline Bourdeau, 溝口理一郎, 人工知能学会発表資料, (2007)