B-021

博物館用メタデータスキーマ間における類似性発見方式の基礎的検討

A Basic Consideration on Similarity Discovery Between Metadata Schema of Museum Information

村田 智紘† 秋元 良仁† 亀山 渉†
Tomohiro Murata Ryoji Akimoto Wataru Kameyama

1．はじめに

近年，博物館では多くのデータベースが公開され，様々な用途で活用されつつある．しかしながら，それぞれのデータベースは異なるメタデータスキーマで構成されるため，横断的かつ複合的なデータの利用は困難な状況にある．データの相互利用を実現するためには，スキーマ間に見られる項目の意味的な類似性を発見し，それを活用することが必要不可欠となる．本稿では，スキーマを構成する概念間の関係構造記述を，シソーラスを基に，スキーマ間の概念関係を自動的に組織化し，スキーマ間における類似性を発見する方式の基礎的検討を行う．

以下，2 章で博物館におけるデータの相互利用の現状と問題点を整理し，3 章でシソーラスとオンライン百科事典 Wikipedia を用いたスキーマ間の概念類似性の発見方式を検討する．次に 4 章でシステム設計を行い，5 章でまとめと考察を行う．

2．博物館におけるデータの相互利用

2.1 博物館におけるデータの相互利用の現状

博物館におけるデータの相互利用の実現は，展覧会における所蔵品の貸借，歴史研究における文化財情報の収集等，実務・研究両面から急務とされてきた．近年では，情報技術の進展に伴い，各博物館の文化財情報はデータベース化され，インターネットを介して内外に発信されつつある．このような状況で，様々な形式で管理される情報資源を横断的かつ複合的に取り扱う試みがある．

大学共同利用機関法人人間文化研究機構では，データベースの拡充，高次化と研究資源の共有を目的として，2008 年 4 月から「研究資源共有化システム」を公開している[1]．このシステムは(1)複数のデータベースの横断・統合検索システム，(2)データベース作成支援システム，(3)GIS 情報分析システムからなる．特に(1)は複数存在するメタデータスキーマを Dublin Core と拡張要素を用いて共通のスキーマにマッピングする規則を適用することでデータベース間の横断・統合検索を実現している．

独立行政法人国立美術館では，文化庁が運営する文化財情報ポータルサイト「文化遺産オンライン」に対し，自館の文化財メタデータを XSLT で文化遺産オンライン用メタデータに変換し，OAI-PMH を介して提供する試みや，複数館の展覧会情報や美術図書の目録情報等の美術関連メタデータを，文書間の類似度を計算する連想検索エンジン GETA に登録することで横断的に検索できるシステムの構築を行っている[2]．

† 早稲田大学大学院国際情報通信研究科

```
<?xml version="1.0" encoding="UTF-8"?>
<noun xml:id="5455460">
 <gloss>
  the person or thing chosen or selected; "he was my pick for mayor"
 </gloss>
 <word-form>choice</word-form>
 <word-form>pick</word-form>
 <word-form>selection</word-form>
 <hypernym part-of-speech="noun" target="5453619"/>
 <frames part-of-speech="verb" target="653781"/>
 <hyponym part-of-speech="noun" target="5455670"/>
 <hyponym part-of-speech="noun" target="5455968"/>
 <hyponym part-of-speech="noun" target="5456920"/>
</noun>
```

図 1 WordNet を用いた XML 記述例

2.2 博物館におけるデータの相互利用の問題点

このような取り組みは，ユーザにとって，(1)利便性：検索システム毎に検索方法を覚えなくて良い，(2)網羅性：一度の検索で複数のシステムを検索できる，(3)品質の確保：Web 上のデータと異なり博物館の判断に基づいたデータ品質が確保されている，という観点から有益な取り組みであると言える．しかし，各データベースのスキーマを共通のスキーマにマッピングする際には，どの項目をどの項目にマッピングするのか，専門家による高度な知識と経験が必要となる．ここで，異論のない意味的に整合の取れたマッピングを実現することは困難である．

3．スキーマ間の概念類似性発見方式の提案

3.1 シソーラス

概念間の関係構造を記述する方式の一つにシソーラスがある．シソーラスは，語彙を等価関係，階層関係，関連等に分類し，辞書あるいはデータベースとして利用する．

WordNet[3]は，英語の大規模概念辞書であり，シソーラスに見られる関係を用いて構築されている．近年では，WordNet から XML 形式で概念関係を抽出し，セマンティック XML アプリケーションに応用する試みも行われている[4]．図 1 に WordNet を用いた XML 記述例を示す．WordNet は英語の概念辞書であるため，日本語処理を目的としたアプリケーションへの応用を考える際には，語の翻訳が必要となる．これまでにも，和英辞書を用いて機械翻訳する試み[5]等がなされている．

SKOS(Simple Knowledge Organization System)[6]は，シソーラスの構造内容表現に RDF を用いたモデルである．2008 年 6 月現在，SKOS の仕様ステータスは W3C の Working Draft で，現在も議論が行われている．

3.2 Wikipedia

Wikipedia[5]は，Wiki を利用して構築された大規模 Web 事典であり，2008 年 6 月現在，50 万項目を超える様々な分野の記事(日本語のみ)をカバーしている．市販の百科事

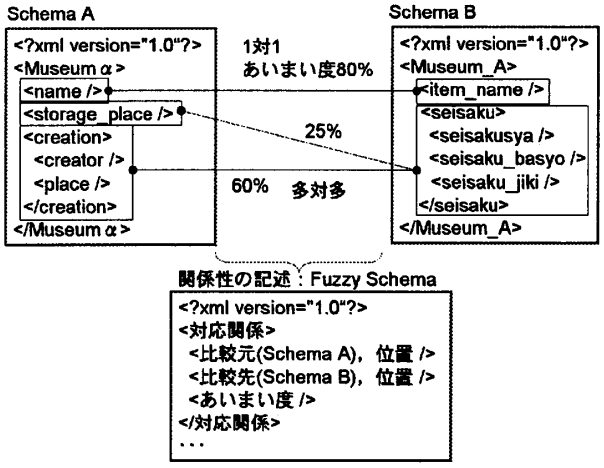

図 2　Fuzzy Schema

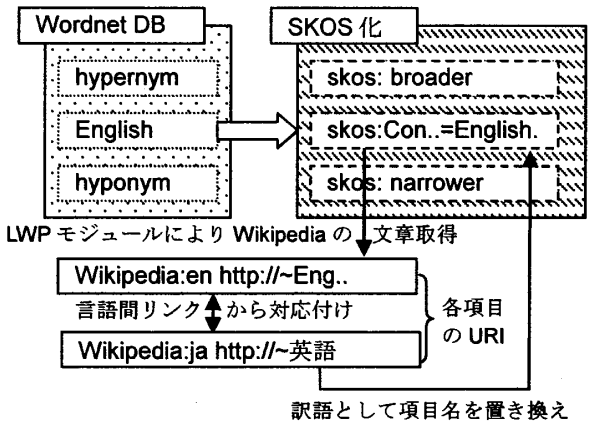

図 3　システム構成

典は記事数が約 10 万であることから比較しても膨大な記事数をカバーしていることが分かる．また，Wikipedia は記事(概念)同士がハイパーリンクで互いに参照されており，かつ，1 つの記事には 1 つの URL が割り当てられているため，インターネット上において記事を一意に参照できるという特徴を持つ．

Wikipedia を用いた概念形成や多言語化には，リンクの共起性を解析することでシソーラス辞書を構築した例[7]や，Wikipedia の各国版ハイパーリンクを利用した例[8]がある．

3.3　提案手法

本研究では，スキーマの項目間に見られる類似性を発見するため，シソーラスの持つ概念関係記述方式と Wikipedia の持つ広範な記事量を利用することを考える．WordNet には，概念に対して一意に ID が付与されている．そこで，(1)WordNet から概念 ID 及び関連する概念関係性を抽出し，(2)概念 ID は Wikipedia の 1 記事(日本語ページ URL)と対応付ける．また，(3)関連する概念関係性は SKOS に変換する．更に，(4)Wikipedia の記事と対応付けた概念 ID と変換した SKOS のマージを行う．

更に，SKOS を拡張し，筆者らが提案している Fuzzy Schema[4]への適用を考える．Fuzzy Schema とは，複数の異なるスキーマ間の関係を，項目の対応パターンと類似性を表すあいまい度を用いて表現する言語である．図 2 に Fuzzy Schema の例を示す．本研究では，Fuzzy Schema 内に記述するあいまい度の参照先に拡張した SKOS を指定し，指定された SKOS とその関連 SKOS からあいまい度を計算することを考える．

4．システム構成

システム構成図を図 3 に示す．まず，Wordnet のデータベースファイルを XML に出力するプログラムを用い，各単語間の関係を SKOS に変換する．その後得られた項目名(skos:concept の項)を直接英語版 Wikipedia の URI に充て，LWP module を使用しページを取得する．取得したファイルに対して HTML パーサを用いて HTML 構文を解析，言語間リンクの URI を得る．この事により英語版，日本語版 Wikipedia の URI と対応関係が得られる．また，Wikipedia においては単語が多義性を持ちえず，URI に単語が直接使用されているため，URI の単語を訳語として使用する．上下関係のある他項目に対しても同様の事を行い，各項目を置き換える．得られた Wikipedia の記事については skos:note とし，参照先とする．これにより英語シソーラスと日本語シソーラスの対応が可能となり，また Wikipedia とも関連付けられることにより記事の自然言語解析等から，有用なオントロジの自動生成が可能になると思われる．

5．まとめと今後の課題

本稿では，博物館用メタデータスキーマ間における類似性発見方式について基礎的な検討を行った．博物館におけるデータの相互利用では，スキーマ間の意味的マッピングが問題であることがわかった．そこで，WordNet と Wikioedia を用いたスキーマ間の意味的な類似性の発見方式について検討を行った．今後は，検討方式を拡張し，具体的なシステムの構築に取り組んでいく予定である．

参考文献

[1] 研究資源共有化シンポジウム講演予稿集, 2008-03-14, 人間文化研究機構.

[2] 水谷ほか: “独立行政法人国立美術館における情報（連携）の試み－美術館情報資源の利活用試案ならびに他関連機構との連携について”, 近美紀要, No.12, pp.5-26, 2008.

[3] WordNet
http://wordnet.princeton.edu/ (accessed 2008-07-02)

[4] Thinking XML: Querying WordNet as XML
http://www.ibm.com/developerworks/xml/library/x-think29.html
(accessed 2008-07-02)

[5] The Web KANZAKI
http://www.kanzaki.com/ (accessed 2008-07-02)

[6] W3C Semantic Web Activity SKOS
http://www.w3.org/2004/02/skos/ (accessed 2008-07-02)

[7] 伊藤他: “Wikipedia のリンク共起性解析によるシソーラス辞書構築”, 情処論, Vol.48, No.SIG-20, pp. 39-49 (2007-12)

[8] Maike Erdmanm. et al.,: “Wikipedia Link Structure Analysis for Extracting Bilingual Terminology”, 情処研報, Vol. 2007, No.65, pp.551-556 (2007-07)