3N-05

既存メタデータに基づく記述対象を考慮したターム探索支援手法

二十歩 亮介† 西出 賴継‡ 本間 維‡ 永森 光晴†† ‡‡ 杉本 重雄††
筑波大学情報学群情報メディア創成学類† 筑波大学大学院図書館情報メディア研究科‡
筑波大学図書館情報メディア系†† 知的コミュニティ基盤研究センター‡‡

1. はじめに

メタデータの長期利用やコミュニティを超えた横断利用のためには、新たにメタデータスキーマを作成する際に既存のメタデータ語彙を再利用することが重要である。その一方で、既存メタデータ語彙の中から目的に応じたタームを探索する具体的な手法は確立されておらず[1]、ターム探索はメタデータスキーマ作成者の負担となっている。本研究は、既存メタデータにおいて実際にタームが使用されている例を確認することが適切なターム探索に有効であると考え、既存メタデータに基づくメタデータコーパスの作成、及びメタデータコーパスを用いたターム探索支援手法の提案を行う。

2. メタデータスキーマと語彙の再利用

図1は書籍を記述対象としたメタデータの記述例である。この例では書籍のタイトルや著者名といったメタデータ記述項目にdc:titleやfoaf:nameなどのタームを使用している。これらのタームの仕様はDCMI Metadata Terms[2]やFOAF[3]というメタデータ語彙の中で定義される。メタデータの記述において、各記述項目にどのタームを使用するか、構造やデータ値にはどのような制約を与えるか、といった規則を定義したものをメタデータスキーマという。

新たにメタデータスキーマを作成する場合、図1のようにFOAFなどの既存メタデータ語彙を再利用することで、メタデータの相互運用性を高めることができる。

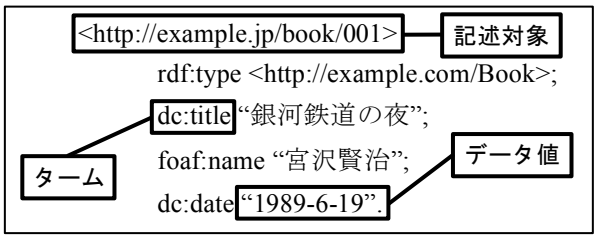

図 1 メタデータの記述例

“A Method for Searching Metadata Terms based on Metadata Instances”
†Ryosuke Nijubu. School of Infomatics. Univ of Tsukuba.
‡Yoritsugu Nishide. Tsunagu Honma. Graduate School of Library,Information and Media Studies. Univ of Tsukuba.
††Mitsuharu Nagamori. Shigeo Sugimoto. Faculty of Library, Information and Media Science. Univ of Tsukuba.
‡‡Research Ctr for Knowledge Communities. Univ of Tsukuba.

3. メタデータスキーマ作成時のターム探索における問題

現在、様々なメタデータ語彙が定義されているが、その中からメタデータスキーマ作成者の利用目的に応じたタームを発見、選択するための具体的な手法は確立されていない。実際にメタデータスキーマを作成したところ、ターム探索には多くの知識と時間を要することがわかった。その原因として、タームの定義文が曖昧であったり存在していなかったりする場合があることが挙げられる。そのため、あるタームが利用目的に適しているかの判断には、実際に当該タームが使用されている例を確認し使用傾向を知る必要がある。適切なタームの判断に有用だと考えられる情報として次のものがある。

(a) 当該タームが使用されている分野
(b) 当該タームを含むメタデータインスタンスにおける記述対象の特徴
(c) 当該タームと共に使用されているターム

しかし、現状タームに関するこれらの情報を得るための環境は十分に整備されていない。本研究はタームの使用傾向を分析するためにメタデータコーパスを作成し、そのコーパスを用いたターム探索支援手法を提案する。

4. メタデータコーパス

コーパスとは自然言語の分析のために作成されるデータベースである。様々な用途に使用され、例えば自然言語の辞書編纂の際、用例の検索や語の共起関係の把握に用いられる[4]。本研究はこの自然言語のコーパスの構成要素や作成方法を参考にメタデータのコーパスを設計、作成し、タームの使用傾向の分析を可能とする。

4.1 メタデータコーパスの設計

コーパスは分析対象を体系的に収集した基礎情報と、コーパスの用途に応じて作成される付加情報から構成される。メタデータコーパスにおいて基礎情報はメタデータインスタンスであり、付加情報は3章の(a)～(c)の情報を得るためのものである。

(a)の情報を得るためには、タームが使用されているメタデータインスタンスを含むデータセットの分野の情報が必要である。

表1　メタデータコーパスの要素

基礎情報	付加情報
記述対象	データセット名
記述対象のクラス	データセットの分野
プロパティ	記述対象のカテゴリ
データ値	プロパティのカテゴリ
データ値のクラス	データ値のカテゴリ

(b)の情報を得るためには、記述対象のクラスを参照すれば良い。クラスとはメタデータ語彙の中で定義されるタームの一種で、「人物」「文書」など共通の特徴を持つリソースをグループ化するために用いられる。図1ではhttp://example.com/Bookがクラスにあたる。また、類似する特徴を表すクラスが複数のメタデータ語彙で定義されている場合がある。例えばFOAFで「人物」を表すクラスとして定義されているfoaf:Personと、BBC Sport Ontology[5]で「アスリート」を表すクラスとして定義されているsport:Personなどが該当する。タームの使用傾向を分析する際、foaf:Personとsport:Personをどちらも「人物」として扱う方が都合が良い場合がある。そこで、クラスをいくつかのカテゴリに分類し、付加情報として記述する。

(c)の情報は、図1では「dc:titleはfoaf:nameと共に使用される」という情報である。この情報を得るためには、記述対象のクラスと、タームが使用されているメタデータインスタンスを含むデータセットを識別できる情報が必要である。

以上よりメタデータコーパスに記述する要素を決定し、それを列挙したものが表1である。

4.2　メタデータコーパスの作成

まず、基礎情報となるメタデータインスタンスの収集を行った。今回は約9000件のデータセットが登録されており、誰でも自由に閲覧、利用できるポータルサイトであるDataHub[6]で公開されているメタデータインスタンスを収集の対象とした。

次に付加情報を作成した。「データセットの分野」の作成にはDataHubのタグを利用した。DataHubではデータセット毎に複数のタグが付与されており、publications, geographicなどの分野を表すものが含まれている。そのうち10種のタグをデータセットの分野として採用した。

「記述対象のカテゴリ」の作成には、約300件の代表的なタイプのリソースを記述対象としたメタデータスキーマを公開しているschema.org[7]を利用した。schema.orgでメタデータの記述対象のタイプとされている”Person”, ”Book”などをカテゴリとし、収集したメタデータインスタンスに含まれるクラスをそのカテゴリへ分類した。

また、データセットを識別する情報としてDataHub内の各データセットページのURLを採用し、「データセット名」として記述した。

5.　メタデータコーパスを用いたターム探索支援手法

メタデータコーパスに含まれるタームを対象に探索を行う手法を示す。

5.1　分野や記述対象によるターム探索手法

分野や記述対象によるターム探索では、それぞれ付加情報の「データセットの分野」と「記述対象のカテゴリ」を利用する。例えば、”publications”の分野で使用されているタームを探索する場合、「データセットの分野」が”publications”であるタームが列挙される。また、キーワード検索と組み合わせることで検索結果の絞り込みを行うことができる。”name”をキーワードとし、記述対象を「人物」と指定した場合、「人の名前」を記述するタームが探索可能となる。

5.2　共起語によるターム探索手法

共起語によるターム探索では、「記述対象のクラス」と「データセット名」を利用する。例えば図1のメタデータインスタンスが”Book Resources”というデータセットに含まれている場合、入力として”dc:title”を受け取ると、「記述対象のクラス」として”http://example.com/Book”を、「データセット名」として”Book Resources”を取得する。取得した情報に基づきタームを検索することで、”dc:title”と共に使用されているタームとして”foaf:name”と”dc:date”を得る。

6.　おわりに

本研究では、タームの使用傾向を分析するためメタデータコーパスを作成し、それを用いたターム探索支援手法の提案を行った。これによりメタデータスキーマ作成者の適切なタームの探索コストの削減を図った。今後は本手法の評価や、メタデータコーパスを利用したターム推薦のスコア付けに取り組む。

参考文献

[1] Heath, Tom & Christian, Bizer. Linked Data: Evolving the Web into a Global Data Space. (2011). Morgan & Claypool.
[2] DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/
[3] FOAF Vocabulary Specification. http://xmlns.com/foaf/spec/
[4] 言語コーパスガイダンス コーパス開発センター. http://www.ninjal.ac.jp/corpus_center/guidance.html
[5] BCC – Ontologies – Sport Ontology. http://www.bbc.co.uk/ontologies/sport
[6] the DataHub. http://datahub.io/
[7] schema.org. http://schema.org/