

統制語彙を利用した SBM 用ブラウザ機能の試作 (Prototype of the browser function using controlled vocaburary for Social Bookmark)

長屋 俊・日本原子力研究開発機構[†] NAGAYA Shun・Japan Atomic Energy Agency[‡]

板橋 慶造・日本原子力研究開発機構[†] ITABASHI Keizo・Japan Atomic Energy Agency[‡]

1 はじめに

ウェブの普及および学術情報流通の電子化とともに、ウェブ上の情報資源の量が爆発的に増えている。原子力分野でも図書・雑誌・レポートを中心とした書誌情報データベースを作成し学術情報流通の改善を行ってきたが、昨今では従来の書誌情報だけではなく新たに登場したウェブ情報資源の収集とその組織化が課題となっている¹。そこで、原子力分野におけるウェブ情報資源の組織化の一手法として Social BookMark (以下「SBM」) に焦点をあてた。SBM の特徴である「自由語によるタグ付け」は、同時に SBM において「タグの表記揺れ」という欠点になっている。今回はこの課題を解決するために、原子力分野で使われている統制語彙である INIS² シソーラスを WebAPI (Application Programming Interface) 化し、それを用いて SBM 上で統制語によるタグの付与を行うためのブラウザ機能の補助システムの試作を行った。

2 SBM と統制語彙

SBM は以前からウェブ情報資源の組織化の一つの手法として注目を浴びている。SBM とはウェブサービス上で利用者同士がブックマークを共有するという仕組みで、日本でもウェブサービスとしていくつかの SBM が提供されている。SBM はブックマークを共有する際にタグと呼ばれるキーワードを付与することで、ブックマークそのものに索引付けを行い組織化する、という機能を持っている。このように利用者同士が自由にタグ付けを行い組織化を行う仕組みは「フォークソノミー (folksonomy)」と呼ばれ、専門家が統制語によるタグ付けをするよりもコストがかからず、なおかつ自由語を付与できる、というメリットがある。しかし、自由語を付与できるということは、意味として同じタグであっても利用者によって違うタグがつけられてしまう、という表記揺れの問題も同時に引き起こしている³。

一方、図書館情報学の分野では統制語彙という技術が使われてきた。伝統的な書誌情報デー

タベースでは書誌情報毎に統制語を付与し、情報検索の精度を高めることで利用者の支援を行っている。しかし、専門家が内容分析を行いキーワードを付与する従来のモデルは前述したとおりコストがかかる⁴。

SBM で統制語彙が利用されない理由としては語彙の作成および管理にかかるコストの問題、そして階層構造から統制語を選択する作業などがあり利用のハードルが高い、加えて統制語彙の掲載メディアが書籍、PDF などとで実用に向かないという 3 点が考えられる。1 つ目の課題は既存の統制語彙を利用することで解決でき、2 つ目と 3 つ目の課題はシステム側で利用支援を行うことで解決できると考えた。今回は原子力分野の統制語彙である INIS シソーラス (国際原子力機関 IAEA⁵ の INIS 本部から提供されている INIS データベース⁶ のためのシソーラス) を利用した。また、単純にデータベース化するだけでなく将来的な他システムからの利用も考え Web API 機能を実装することで相互運用性を高めた。

2.1 原子力分野における統制語彙

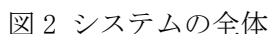
INIS シソーラス自体は INIS への参加国が新規ディスクリプタ (シソーラス中の階層構造を持つキーワード) を提案し、INIS 本部にて調整を行い、更新が行われている。現在、INIS シソーラスは PDF 形式で配布されている⁷。また機械可読可能なデータとしては IAEA が INIS 加盟国で公開しているシソーラスの固定長データがあるので今回はそれを用いた。そのデータは図 1 のようになっている。1 レコード 167 byte で固定長データ全体で 301,445 行からなる。シソーラスに登録された統制語は 30,550 語である (2010 年 12 月現在)。必ずしも固定長データの 1 レコード = 1 ディスクリプタとして構成されているわけではなく階層関係 (BT, NT, RT, etc) の表現や固定長に収まりきれないデータの都合上、複数行に展開されているためファイルの行数に比べ統制語の数は少なくなる。この固定長データを MySQL 上に展開し、PHP にて INIS シソーラスの WebAPI 化を行った。

図 1 INIS シソーラスの 167byte 固定長のデータ (一部抜粋)

SBMサービスとし

試作をした本システムの全体像は図 2 の通りである。各利用者はブラウザから bookmarklet を用いて本システムを呼び出す方法を用いた。利用者は事前に bookmarklet をブラウザのブックマークに登録しておき、はてなブックマークへウェブ情報資源を登録したいときにはブラウザ上で bookmarklet を起動することで本システムを呼び出すことができる。bookmarklet を使うと、各利用者のブラウザにポップアップで登録画面が表示される。表示された登録画面上では自由語だけではなく INIS シソーラスをもとにした統制語を用いてタグ付をし、はてなブックマークに登録することができる仕組みになっている。また、システムからはてなブックマークへの登録にははてなが正式に提供している「はてなブックマーク AtomAPI」を利用した。

The diagram illustrates the system architecture. A user (represented by a person at a computer) interacts with a 'ブラウザ「登録画面」' (Browser 'Registration Screen'). The user can perform a 'ブックマーク' (Bookmark) action, which is stored in the 'はてなブックマーク' (Hatena Bookmark) database. The user can also perform a 'ポップアップ' (Popup) action, which is handled by the 'システム' (System) component. The 'システム' component interacts with the 'はてなブックマーク' database via a 'シソーラス' (Thesaurus) interface. The 'システム' component also interacts with a 'シソーラス (Web API)' (Thesaurus (Web API)) component. The 'シソーラス (Web API)' component interacts with the 'システム' component via a 'bookmarklet' interface.



今後の課題としては自由語によるタグの付与ではなく統制語を用いてタグを付与していくことでウェブ情報資源の組織化にどういった変化を及ぼすのか、といった視点での利用分析が挙げられる。

[参考文献・参考 URL・注記]

- Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.