

A Tag Management System for Cross Search among Data Repositories

メタデータ	言語: jpn 出版者: 公開日: 2020-11-02 キーワード (Ja): キーワード (En): 作成者: メールアドレス: 所属:
URL	https://doi.org/10.24517/00059779

This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 3.0
International License.



第 28 回年次大会予稿

データリポジトリの横断検索のためのタグ管理システム A Tag Management System for Cross Search among Data Repositories

河合秀明^{1*}, 笠原禎也¹, 高田良宏¹, 林正治²,
Hideaki KAWAI^{1*}, Yoshiya KASAHARA¹, Yoshihiro TAKATA¹,
Masaharu HAYASHI²,

1 金沢大学
Kanazawa University
〒 920-1192 金沢市角間町
E-Mail: kawai@cie.is.t.kanazawa-u.ac.jp

2 国立情報学研究所
National Institute of Informatics
〒 101-8430 東京都千代田区一ツ橋 2-1-2
*連絡先著者 Corresponding Author

近年, 世界では論文や研究データを始めとした研究成果などに対してアクセスを容易にし, データの発信, 共有, 再利用を促すことを目的としたオープンサイエンスと呼ばれる試みが活発になってきているこれに伴って国内でも既存のリポジトリソフトウェアを利用して, 論文だけでなく, 研究データを対象としたデータリポジトリの構築が行われている. しかし, データリポジトリとして利用することを考えた場合, メタデータを画一化できないことなどから, ユーザが所望したデータにたどりつけないなど, 現状では共有, 再利用は容易ではない. そこで, 本研究ではこの問題を解決しより利便性を向上させるインターフェースの開発を目的としている. 本インターフェースは, ユーザがより容易にアイテムを共有, 再利用できるよう, アイテムにタグを割り振ることによって, メタデータに欠けているものを補完し管理を行うシステムである.

Recent years, an attempt for the purpose of publishing, sharing, reusing data called "Open Science" has become active for easier access to papers and research data. Along with this, in Japan, academic repositories are built at many academic research institutions using a repository system named WEKO. When we consider using WEKO as a data repository, it is not easy for the users to search the data which they wish because metadata are not well standardized in many academic fields. In the present study, we developed a new user interface for the data repository using WEKO3, which is expected to be a mainstream of repository system in the future. This interface is a system that manages items by assigning tags to them, and aims to form collective knowledge by sharing tags among multiple repositories.

キーワード: WEKO, リポジトリ, オープンサイエンス, 研究データ
WEKO, repository, open science, research data

1 はじめに

近年, 世界では「オープンサイエンス」と呼ばれる試みが活発になっており, 論文だけでなく根拠となる研究データもリポジトリ化しようとする動きが加速している. それに伴って, 国内では WEKO^[2] をはじめとした様々なリポジトリソフトウェアを用いて, 多くの学術研究機関で

データリポジトリの構築が行われている. そもそもリポジトリは図書館が中心となり文献情報を蓄積・公開するために発展してきた経緯から, リポジトリシステムもその多くが文献リポジトリとして使われている. これをデータリポジトリとして利用することを考えたとき, 分類法やメタデータのつけ方が文献とは違い各分野で異なることが問題となる. 研究データを対象にす

る場合、分類やメタデータの付け方が学問分野によって多種多様な上に、こういった付け方をすることも十分に確立されていない分野の方が多いと考えられる。また、ユーザによってデータの使い方も異なることから、従来の分類法を拡張した多様性のある分類法が求められている。そこで本研究では分類を管理者だけでなく、ユーザが独自にデータを分類することができる仕組みを、プラグイン方式のアイテム管理インターフェースとして開発を進めてきた^[1]。

2 開発方針

本研究では、学問分野に大きく依存し、画一的な分類が難しいデータを柔軟に取り扱えるようにするため、多様な視点での分類を可能にする方法を提案する。具体的には、1. 横断的な分類、2. 一点に特化して細分化、3. 一つのデータに複数のカテゴリを付与する等の利用法を想定し、これらの方法のいずれにも対応可能な方式を検討する。そこで、本インターフェースでは、管理者による分類と複数のユーザによる独自かつ自由な分類の2つの視点からの分類が可能な、タグによるアイテム管理によってこの問題を解決した。

本研究で用いた WEKO は国立情報学研究所が開発した国内で広く使われているリポジトリシステムである。現在、大規模アップデート版の WEKO3^[3] が開発中で今後の主流になる見込みである。そこで本研究では WEKO3 を用いてアイテム管理インターフェースを開発した。

本研究では、WEKO に精通していないユーザを想定し、WEKO3 のプラグイン追加機能を利用したインターフェースをプラグイン方式で開発した。これによって、ユーザはサーバ上で本インターフェースのインストールを行うだけでそれぞれの環境に適した状態で導入を行うことができる。また、タグは多くのユーザに利用されることで洗練されていく仕組みであることを踏まえると単一のリポジトリでタグを保持するよりも複数のリポジトリ間で横断的にタグの共有を行った方が効率的である。そこで、今回はリポジトリのハーヴェスティングと呼ばれる仕組みを利用した複数リポジトリ間のタグ管理の検討を行った。さらに、本インターフェースのイン

ストール直後のまっさらな状態からのタギングをサポートするために、形態素解析を用いた自動タグ割り当てシステムの開発を行った。

本稿では、2章で本インターフェースの開発方針、3章で本インターフェースの概要を述べた後、今回検討、開発した複数リポジトリ間でのタグ管理、形態素解析によるタグ生成を4章、5章で述べる。

3 タグ管理システム

従来のリポジトリでは、所望したアイテムを探す際にはデータ管理者側が用意したメタデータをたよりに探す手段が主流である。ユーザが適切なキーワードを用意できない場合にそのアイテムにたどり着けないというケースが少なくない。さらに、適切なキーワードを用意できたとしても、そもそもリポジトリ毎にメタデータの付け方が統一されていない場合もあり、よりりたどり着きやすくなるなどの仕組みが望まれていた。アップロードされたときのメタデータに欠けている情報を後からタグを割り振ることによって補完し、ユーザがより容易にアイテムにたどり着けるようにするタグ管理機能を開発した。今回は、WEKO3 にプラグイン方式で導入できるタグ管理インターフェースとして実装した。本インターフェースには主に、タグ検索、タグ編集、タグ情報出力 API、タグ登録 API などの機能がある。図1に実際のタグ管理システムの UI を示す。以下に各機能の概要を示す。詳細は「WEKO3 に対応するアイテム管理インターフェースの開発」^[1]を参照されたい。

3.1 タグ編集

本インターフェースでは基本的にリポジトリ管理者でも一般ユーザでもリポジトリ内の各アイテムに対してタグを割りあてることができる。また、1つのデータに対して様々な分類を行うために役割が異なる3種類のタグを用意した。タグには一般ユーザ用の「一般タグ」、管理者用の「管理者タグ」、用意された語から作成できる「予約語タグ」の3種類がある。図2にタグの種類の UI を示す。

3.2 タグ検索

本インターフェースではユーザはインターフェース内の検索フォームでキーワードを入力することでキーワードとマッチしたアイテムにたどり着くことができる。

3.3 タグ情報出力 API

本インターフェースを参照するシステムや拡張させるようなシステムを作成することを想定した際、各アイテムのタグの情報を外部でも容易に取り出すことができるようなシステムが必要になる。そこで、本インターフェースでは、キーワードを入力することでそのキーワードに対応したアイテムに付与されているタグの情報を出力できる機能を実装した。また、出力するデータの形式は、一般的に WebAPI でデータを出力する際に用いられる JSON 形式を採用した。

これを用いることで図4のようにタグ情報を JSON 形式で出力することができる。



図 1: タグ管理システム UI



図 2: タグの種類

3.4 タグ登録 API

本インターフェースでアイテムに対して機械的に大量のタグを割り振りたいと考えたとき、UI 上以外でタグを登録できるような機能が必要になる。そこで、本インターフェースでは特定の情報を入力することで、タグを登録することができる機能を実装した。



図 3: リポジトリ管理者から見たタグと一般ユーザーからみたタグ

```

例:http://localhost:8001/tag_mng_api/export?keyword=tag
[
  {
    "itemNo": "2",
    "tags": [
      "titleJ_tag"
    ]
  },
  {
    "itemNo": "3",
    "tags": [
      "titleJ2_tag",
      "tag",
      "tag2",
      "tag3"
    ]
  },
  {
    "itemNo": "5",
    "tags": [
      "titleJ_tag",
      "titleJ2_tag",
      "titleJ3_tag"
    ]
  },
  {
    "itemNo": "7",
    "tags": [
      "titleJ_tag",
      "リポジトリ",
      "金沢大学",
      "通信情報研究室"
    ]
  }
]

```

図 4: JSON 出力結果

4 複数リポジトリ間でのタグ管理

4.1 概要

タグというシステムは利用するユーザ数やタグの数に比例して洗練されていくシステムであり、単一のリポジトリで運用する場合は非効率である。そこで、ハーヴェスティングによって複数のリポジトリ間でタグ管理を運用する仕組みの検討を行った。



図 5: ハーヴェスティング

4.2 ハーヴェスティング

ハーヴェスティングとは、リポジトリに登録されているアイテムのメタデータを他リポジトリが機械的に収集することができる仕組みのことである。この仕組みを利用し、本インターフェイスがインストールされているメタデータリポジトリを構築した。図 6 のように複数のリポジトリのメタデータを収集し、本インターフェイスでタグ管理することで、複数のリポジトリのアイテムに対するタグ管理を行えるようにした。これによって、リポジトリを複数集め、それらのアイテムに対してタグを割りあててことでより多くのアイテムに対して横断的にタグを割りあてることが可能になり、従来よりもタグによる恩恵を大きくすることが期待できる。

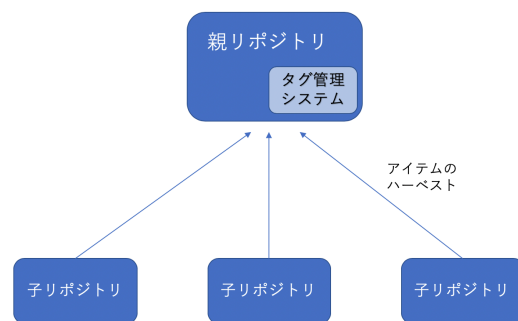


図 6: ハーヴェスティングによる複数のリポジトリ間でのタグ管理

5 形態素解析によるタグ生成

5.1 概要

本インターフェイスをインストールした環境構築直後は、タグが存在せずユーザがタグを付けづらいということが考えられる。そこで、本研究では形態素解析を用いることで自動的にタグを生成する手法の検討を行った。

5.2 形態素解析

形態素解析とは、メタデータ等が存在しない素のテキストデータを最小単位の単語で分割し解析する技術のことである。本研究では、MeCab^[4]

と呼ばれるオープンソースの形態素解析エンジンを利用した。これを用いて「WEKO3に対応するアイテム管理インターフェースの開発」というテキストを分割した場合は図7のようになる。

WEKO	名詞,一般,*,*,*,*,*
3	名詞,数,*,*,*,*,*
に	助詞,格助詞,一般,*,*,*,に,二,二
対応	名詞,サ変接続,*,*,*,*,対応,タイオウ,タイオー
する	動詞,自立,*,*,サ変・スル,基本形,する,スル,ス
アイテム	名詞,一般,*,*,*,*,アイテム,アイテム,
管理	名詞,サ変接続,*,*,*,*,管理,カンリ,カンリ
インターフェース	名詞,一般,*,*,*,*,インターフ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
開発	名詞,サ変接続,*,*,*,*,開発,カイハツ,カイハツ

図7: MeCabの形態素解析によるテキストの分割

この形態素解析をアイテムのタイトルに対して行うことで、細かい単語で分割したあとに名詞のみを抽出することでタグを自動的に生成することができる。しかし、形態素解析を用いると過剰に分割してしまい、所望しているタグとは違うものが生成されてしまうことがある。図7を例に挙げると、本来このテキストから抽出を期待する名詞は「WEKO3」、「アイテム管理インターフェース」、「インターフェース」などである。しかし、実際には「WEKO3」が「WEKO」、「3」の2つに分解されてしまったり、名詞が連なった「アイテム管理インターフェース」などの複合語をすべて分解してしまったり、あまり重要ではない「対応」、「開発」といった単語が抽出されてしまうという問題があった。本研究では、これを解決するために専門用語自動抽出というシステムを利用した。

5.3 専門用語自動抽出

専門用語自動抽出システム^[5]とは、東京大学情報基盤センター図書館電子化部門中川裕志教授および横浜国立大学環境情報研究院森辰則助教授が共同で開発したものである。これは、形態素解析したテキストデータから複合語からなる専門用語を抽出して重要度順に羅列するシステムである。これを用いることで、図8のように複数の名詞からなる複合語を抽出することや重要度が低い単語を除外することができる。

人工知能 7.937253933193772
 知能 5.291502622129181
 プログラミング言語 5.0743261995
 記号処理 3.5840246342157207
 記述 3.0
 アプローチ 3.0
 記号的明示性 2.96176521936472
 人工 2.6457513110645907
 家庭用電気機械器具 2.4082246852

図8: Wikipedia「人工知能」のページ^[6]のテキストからの専門用語自動抽出例の一部(右の数字が重要度)

5.4 自動タグ割り当てシステム

6.2.6.3章で述べたシステムを利用することでリポジトリ内のアイテムに対して、メタデータを自動的に生成するシステムを開発した。このシステムを用いることで、本インターフェース構築直後であっても即座にタグを割り当てることができる。タグ生成までの大まかな流れは図9のようになる。

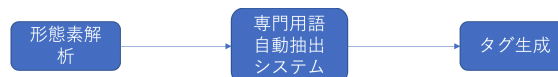


図9: タグ生成フロー

このシステムの主な機能としては以下の通りである。

- 選択したアイテム群の重要度を算出し羅列
- 設定したしきい値以上の重要度のアイテムに対してタグ生成
- 生成したタグがいくつのアイテムに割り当てるか表示

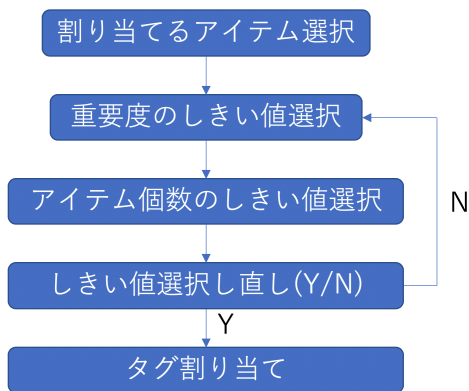


図 10: タグ生成システムフロー

6 まとめ

本研究では, WEKO3 を利用してデータリポジトリを構築し, それに適したタグ管理システムの開発を行っている. 今回は, ハーヴェスティングを用いた複数リポジトリ間のタグ共有の仕組みの検討と形態素解析を用いた自動タグ生成機能の実装などを行った.

今後は, ハーヴェスティング元のリポジトリからでも容易にタグを表示できる仕組みの検討や, 自然言語処理などを用いて最適なタグ候補を提案できるようなシステムの検討を考えている.

参考文献

- [1] 河合秀明; 笠原禎也; 高田良宏; 林正治: 「WEKO3 に対応するアイテム管理インターフェースの開発」, 情報知識学会誌, 29 巻 4 号, p. 352-355, 2020.
- [2] WEKO
<http://weko.at.nii.ac.jp/> (2020 年 4 月 7 日参照)
- [3] WEKO3
<https://rcos.nii.ac.jp/service/weko3/> (2020 年 4 月 7 日参照)
- [4] MeCab
<https://taku910.github.io/mecab/> (2020 年 4 月 7 日参照)
- [5] 専門用語 (キーワード) 自動抽出 Python モジュール termextract
<http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract/> (2020 年 4 月 7 日参照)
- [6] 人工知能 - Wikipedia
<https://ja.wikipedia.org/wiki/人工知能> (2020 年 4 月 7 日参照)