国立情報学研究所教育研修事業
「大学図書館員のための IT 総合研修」2020 年度
「Web API を使ったデータの入手とその整備」講義資料

東京大学の機関リポジトリデータクレンジング実例

2020 年 9 月 14 日
東京大学情報システム部　前田　朗

東京大学では 2017 年度に機関リポジトリシステムを DSpace から JAIRO Cloud に変更した。そのリプレイス作業の機会を使い、以下のとおりデータクレンジングを実施した。これを実例として提示する。また、補記に国立国会図書館の学位論文メタデータとの照合の例を示す。

１．言語コード

メタデータ中に ISO639-1 と ISO639-2 の言語コードをそれぞれ項目としてもっており、手当を行った。Excel のピボッドテーブル機能で、コードごとの件数集計表を作成し目視確認。実際に見つかった問題となる言語コードは以下のとおり（「→」の先は本来あるべき言語コードを示す)。

・ISO639-1

- em → en
- eneng → en
- eng → en
- lat → la
- pt → es
- jpn → ja
- ip → ja　※ 43 件
- jp → ja　※ 395 件

・ISO639-2

- en → eng
- get → lat
- ja → jpn

・言語コードとして ISO639-1 と ISO639-2 のそれぞれをセットしているが、片方のコードが決まればもう片方も決まるといった関係にある。そこで ISO639-2 の値が空欄の場合、ISO639-1 の値を基にセットした。ISO639-1 と ISO639-2 とも空値の 60 件についてはアイテム詳細を確認してセットした。

ポイント

・コードはグループ化して集計すると、誤記載の確認がしやすい。今回は Excel のピボットテーブルで行ったが、データベースであれば SQL で確認できる。

・コード表があれば実際に照合してみるのもよい

２．出版者（Publisher）

同じ出版者でもアイテムによって表記ゆれが散見された。そこで OpenRefine により表記ゆれ調査を実施した。表記ゆれの確認の際のパラメータ設定によりどこまで表記ゆれとみなすかの結果が異なる。パラメータをいくつか試した中で、以下が多く結果が出たので、その結果をもとに表記ゆれの精査を行った。

・Method → key collision

・Distance Function → RPM

・Radius → 1.0

・Block Chars → 6

３．雑誌名(jtitle)

・「出版者」と同様に同じ雑誌名でも表記ゆれ調査を実施した。

・全角スペースをすべて半角に

・雑誌名を junii2 準拠にするには繰り返しなしにする

- タイトル別名とヨミが混在　→　ヨミ削除
- 和英は和英併記に　→　件数が多いので一括返還

・雑誌名中の年・年度表記

- 書誌記述にはそぐわないので、年・年度は巻(volume)に記載を移す

4. ページ(page)

・一部 ()付きのページ No あり、手作業で修正

例: (48)-(1)

5. NII 書誌 ID(NCID)の連番付与

Excel のフィル機能で連番を振ってしまったと思しき NCID を修正

> **ポイント**
>
> ・Excel でセルのコピーのつもりが連番を振るのはありがちなため注意したい

6. NII 書誌 ID(NCID)と雑誌名(jtitle)

同じ NII 書誌 ID でも、それに対応する雑誌名(jtitle)の表記がまちまち。最終的には CiNii Books の Web API により、NII 書誌 ID に対応する CiNii Books の雑誌名を取り出すことで、表記を統一した。

> **ポイント**
>
> ・Web API でアイテム情報を取り出し、それを手持ちのデータに適用する

7. NCID と NII 資源タイプ

図書の NCID にも関わらず、NII 資源タイプの指定が「Journal Article」(学術雑誌論文)になっていたものを確認した。確認したところ、雑誌とも図書ともとれるアイテムであったため、そのままにしている。

8. 「7 桁」の ISSN

ISSN は「8 桁」のコードであるが、ISSN のデータ項目に値として「7 桁」になっていたものが見つかりチェックを行った。実例は以下のとおり。すべて、先頭の「0」が消えていた。件数が多いことからプログラムで一括処理を実施した。

2896400

2898527

3873307

5638089

9129731

9133801

9138277

9158758

9190473

ポイント

・数値のみのコード値はExcelでコード値の先頭の[0]が削除されることがある

9. NII資源タイプ(NIIType)

NII資源タイプはアイテムについて、そのアイテムが雑誌論文なのか、書評なのかなどを記載する。しかし、その適用にゆれがあったため適用の厳密化をはかった。以下は、実際に東京大学で行った例である。

以下の語がタイトル中に含まれるアイテムを抽出し、アイテムを精査することで、NII資源タイプを決め直した。

・「書評」

・「表紙」

・「巻頭言」

・「イントロダクション」

・「書評」「review」

10. 関連(relation)のURI記述

・DOIのURL表記に書き換えられるものは、すべて書き換える。

・本来がURIを記載すべきデータ項目であるにも関わらず、文章による説明が入っていた。URL形式以外について抽出し、手作業で修正した

11. DOIの記法の統一

次の3パターンを info.doi/... に統一。プログラムにより一括変換により実施した。

・info.doi/prefix/suffix

・doi:prefix/suffix

・prefix/suffix

ポイント

・記法の統一は正規表現を使えば簡単に処理できる

12. 言語の推定

記録が残っていなかったが以下のツールなどで言語の推定を試していたかもしれない。

[Perl モジュール Lingua::LanguageGuesser のお試しページ]

http://gensen.dl.itc.u-tokyo.ac.jp/LanguageGuesser/LanguageGuesser_demo_ja.html

追記　国立国会図書館の DOI 付与済み博士論文リストとの照合

JAIRO Cloud 移行後に、国立国会図書館が DOI を付与した学位論文について、機関リポジトリでも同じ DOI を設定するため、国立国会図書館の DOI 付与済み学位論文リストとの照合も行った。これにより、機関リポジトリの学位論文メタデータについて、タイプミスなどを確認することができた。

１．国立国会図書館の DOI 付与済み博士論文リストを入手

（１）以下から全国分のデータをダウンロード
http://www.ndl.go.jp/jp/aboutus/dlib/cooperation/doi.html#anchor05

（２）Excel でデータを取り込み→「博士授与大学名」を東京大学で絞りこみ
(11,796 件)

２．機関リポジトリのメタデータとの照合

（１）学位授与番号でマッチングさせた一覧を作成し確認