第 36 回 自然史標本情報の発信に関する研究会　2021.2.6（神保）

「標本データのチェックとクリーニングの技法」

中級編：エクセル以外のツール

本実習では、エクセル以外の、標本データのチェックやクリーニングに有用な、以下のようなツールを紹介します。

(1) ウェブツール
(2) エディタ（基本編）
(3) OpenRefine
(4) エディタ（応用編：正規表現）

(1) ウェブツール

ここでは、学名をチェックするウェブツールを 2 つ紹介します。
サンプルとしてサンプルとして UJsample1.csv, UJsample1.xslx を用意しました。

Interim Register of Marine and Nonmarine Genera (IRMNG) のスペルチェッカー
・学名をあいまい検索し、よく似た学名を修正候補として提示する。
・単純に学名の誤記を探すときに有用

1. https://www.irmng.org/ へアクセス、メニューから「Taxa Match」を選ぶ。
2. 学名の入っているファイル（エクセルも可）をアップロードする。
3. 界（kingdom)、学名（scientificName）に相当する列を指定し、「Match」させる。
4. 学名の一覧が表示される。綴りの誤りがある種の場合、種名の候補が単数あるいは複数表示される。
5. 複数の正しい種名綴り候補がある場合、選択メニューから今回採用する学名を選び出す。
エクセルファイル等でダウンロード可能。

GBIF species name matching
・GBIF の分類体系（GBIF backbone taxonomy）と照合し、綴りの間違い、シノニム等の情報を追加し、最も妥当な名前とその分類情報を返す。
・GBIF の体系との齟齬を見る際に有用。

1. 空のエクセルファイルを用意し、「kingdom」列と「scientificName」列を作る。scientificName には学名を、kingdom には所属する界の学名を入れる。UTF-8 の csv 形式

で保存する。
2. https://www.gbif.org/tools/species-lookup にアクセスし、1. で作成したファイルをドラッグアンドドロップでアップロードする。
3. 出てきた対応表を確認する。
4. 画面右下「GENERATE CSV」をクリックすると、結果全体を CSV ファイルでダウンロードできる。数が多い場合は時間がかかる。

verbatimScientificName	preferedKingdom	matchType	confidence	scientificName (editable)	status
Papilio xuthus	Animalia	EXACT	100	Papilio xuthus Linnaeus, 1767	ACCEPTED
Papilio pretenor	Animalia	FUZZY	96	Papilio protenor Cramer, 1775	ACCEPTED
Papilio machon	Animalia	FUZZY	95	Papilio machaon Linnaeus, 1758	ACCEPTED
Papilio hippocrates	Animalia	EXACT	99	Papilio hippocrates Felder & Felder, 1864	SYNONYM
Pieris rapae	Animalia	EXACT	100	Pieris rapae (Linnaeus, 1758)	ACCEPTED
Pieris japonica	Plantae	EXACT	99	Pieris japonica D.Don ex G.Don	ACCEPTED
Hypenagonia aokii	Animalia	HIGHERRANK	96	Hypenagonia Hampson, 1893	ACCEPTED

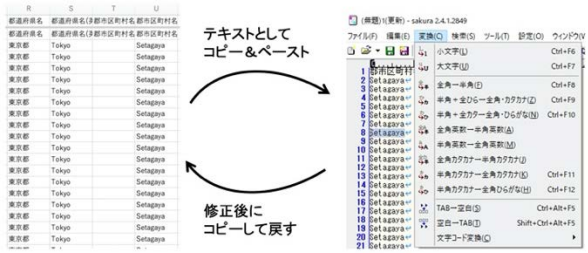
status	rank	kingdom	phylum	class	order	family	genus	species
ACCEPTED	Species	Animalia	Arthropoda	Insecta	Lepidoptera	Papilionidae	Papilio	Papilio xuthus
ACCEPTED	Species	Animalia	Arthropoda	Insecta	Lepidoptera	Papilionidae	Papilio	Papilio protenor
ACCEPTED	Species	Animalia	Arthropoda	Insecta	Lepidoptera	Papilionidae	Papilio	Papilio machaon
SYNONYM	Species	Animalia	Arthropoda	Insecta	Lepidoptera	Papilionidae	Papilio	Papilio machaon
ACCEPTED	Species	Animalia	Arthropoda	Insecta	Lepidoptera	Pieridae	Pieris	Pieris rapae
ACCEPTED	Species	Plantae	Tracheophyta	Magnoliopsida	Ericales	Ericaceae	Pieris	Pieris japonica
ACCEPTED	Genus	Animalia	Arthropoda	Insecta	Lepidoptera	Noctuidae	Hypenagonia	

補足：EXACT は完全に一致、FUZZY はスペルが完全一致しない場合で、候補となる学名も表示される。HIGHERRANK は、種名は無いが属まではある場合等。それぞれ、あわせて、上位分類群も補足される。後ろには GBIF backbone taxonomy での分類情報がつき、シノニムの場合は、現在の学名で表示される。

kingdom が指定されている場合、指定した界のみで照合作業が行われる。特に、異なる命名規約で同じ学名がつけられている場合、それぞれに対応して上位分類群が補足される。たとえば、*Pieris* 属は、動物界ならモンシロチョウの仲間、植物界ならアセビの仲間だが、それぞれ正しくシロチョウ科とツツジ科が紐付けられる。

(2) エディタ（導入編）

・テキストデータを専門で扱うソフトウェア
　文字情報だけのデータ：csv ファイル⇔文字以外も含むデータ：オフィス、画像…
・テキストに特化した編集機能：検索・変換
・様々な用途のエディタがリリース　例：文章書き用、プログラミング用…
・有償・無償問わずたくさんの選択肢
　主なエディタ　Windows：秀丸エディタ・サクラエディタ；Mac：mi
・エクセルデータ編集に「ちょい足し」すると効率よくクリーニングできる時がある。
　全角・半角の変換、余計な空白の存在チェック、等



1. 気に入ったエディタをダウンロードして起動させる。
2. きれいにしたいエクセル表の列を、エディタにコピーする。
3. 半角・全角の変換等を実施して、アルファベットや英数字等は半角、カナやかなは全角で統一する。

※スペースやタブ文字など、特殊な記号も表示するような設定にすると、編集がしやすい。

(3) OpenRefine

・データクリーニングや変換に特化した、オープンソースのフリーソフトウェア
「乱雑なデータを処理するためのパワーツール」

・以前は Google Refine と呼ばれていた

・ウェブブラウザ上で動作し（ネットに接続しなくても利用か可)、Windows/Mac どちらでも使える。Windows 版は Java も必要

インストール・言語設定

1. 公式ページのダウンロードページ（https://openrefine.org/download.html）等から、Win 版（Java 同梱版有り）および Mac 版をダウンロード、解凍して起動する。

2. 起動すると、ふつう Chrome が起動し、ウェブページのように起動画面が表示される。URL は特に気にすることはないが、特定のデータを開いたページをブックマークすることもできる。

3. メニューを日本語にする。起動画面の右メニューから「Language Settings」をクリック、言語で「日本語」を選択する。

※Mac の場合、「製造元が未確認」でダブルクリックでは起動しないときがある。その場合右クリックメニューから「開く」をクリック、確認画面でも「開く」を選択すれば起動する。

ファイルを開く

ここでは UJsample2.xlsx を使って実習を行う。各ファイルと進行状況は、「プロジェクト」というかたまりで管理される。

＜実習＞

1. UJsample2.xlsx を OpenRefine で開く。起動画面に戻り（左上のダイヤのロゴマークをクリック)、「プロジェクトをつくる」を選択する。ファイル選択をクリックして、開きたいファイルを選択する。「次へ」をクリック。

2. 読み込むときの詳細オプションが表示される。名前など必要に応じて変更し、今回は特に何もせず「プロジェクトを作成」をクリック、しばらく待つとメインの画面が表示される。

※セーブは自動的に行われるので、自分でセーブする必要は無い。普通は見えない作業フォルダに保存されるが、エクセルファイルなどにエクスポートしてダウンロードもできる。

※二回目以降は、「プロジェクトを開く」から作業したいプロジェクトを選んで読み出せる。
※ファイルを開いているときも、左上のロゴ（ダイヤのマーク）をクリックすると、起動画面に戻る。

基本的な操作

データは一覧形式で表示され、一画面には最大 50 行まで表示できる。先頭には各行を区別する番号がついており、並び替えても最初の並び順が保存されることになる。
ヒストリー機能がある。左列の上「取り消す／やり直す」をクリックすると、それまでにやった操作が一覧になっており、任意の位置をクリックすると、そこまで作業を戻すことができる。

＜実習＞ 「機関コード」に大文字小文字があるので、大文字にしたい
1. 画面の要素やメニュー等の位置を確認する。
2. データを直接修正する。「機関コード」列で小文字のセルの上にマウスを持っていくと「edit」表示が出るのでクリック、出てきたダイアログでデータを修正する。修正の際に、「同じ内容の全セルに適用」をクリックすると、同じ値のデータが全て新しい値で上書きされる。
3. ヒストリ機能を確認する。「取り消す／やり直す」をクリック、一覧で「Mass Edit」よりも前にあわせてクリックし、元に戻す。
4. 小文字を大文字に変換。「機関コード」見出し左の下向き矢印をクリック、「セル編集」→「よく使う変換」→「大文字に」の順にもっていく。

ファセットによる作業

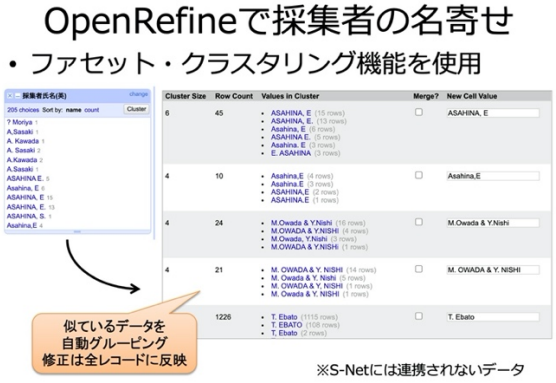
OpenRefine では、ファセット（値を修正してまとめて表示）を使って様々な編集作業を行う。絞り込みやまとめて置換、さらにはファセットをグループ化して編集することもできる。

＜実習＞ファセットによるクリーニング
1. 「コレクションコード」見出し左の下向き矢印をクリック、ファセット→テキストファセットと順に選択すると、ファセットは画面の左に表示される。
2. ファセットの各値をクリックすると、選択された値のデータのみがフィルタリングされて表示される。また、各値の右に表示される「include」をクリックすると、複数の値のデータをまとめて表示できる。フィルターによる絞り込みは「exclude」で解除できる。
3. 一括変換する。修正したい値の右に表示される「編集」をクリックし、新しい値を入力して「適用」すると、同じ値のデータが全て新しい値で上書きされるとともに、ファセットも再集計される。
4. グループ化。「詳細地名」見出し左の下向き矢印をクリック、ファセット→テキストファ

セットと順に選択しファセットを表示させる。ファセットの右上の「クラスタ」をクリックすると、よく似た値のデータがグループ化された状態で表示される。データを統一させたい場合「マージしますか？」にチェックをして新しい値を入力、「マージして閉じる」にすればいい。
5. 以上の作業を各行について繰り返すことで、データをかなりの部分でクリーニングできる。

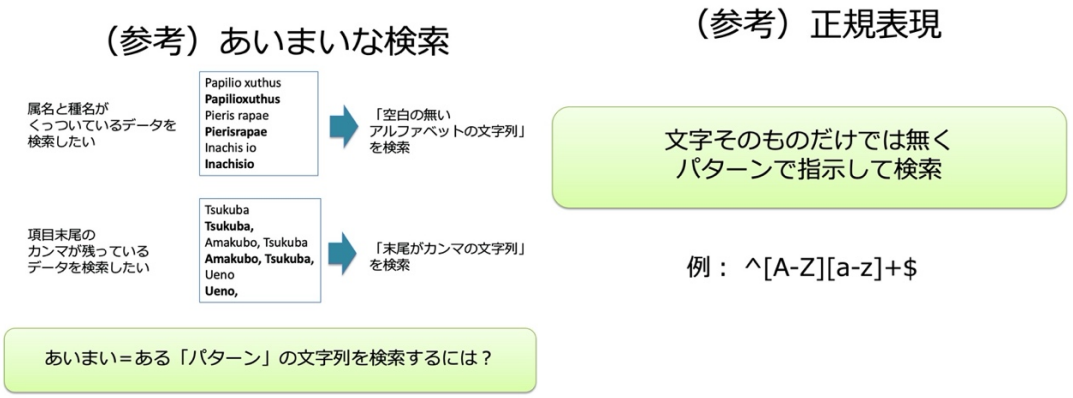
※数値の場合は「数値ファセット」、空白かどうかで集計したいときは「空白ファセット」、データの重複をチェックしたいときは「重複ファセット」を使う。また、各行冒頭のスターやフラグでの絞り込みも可能。

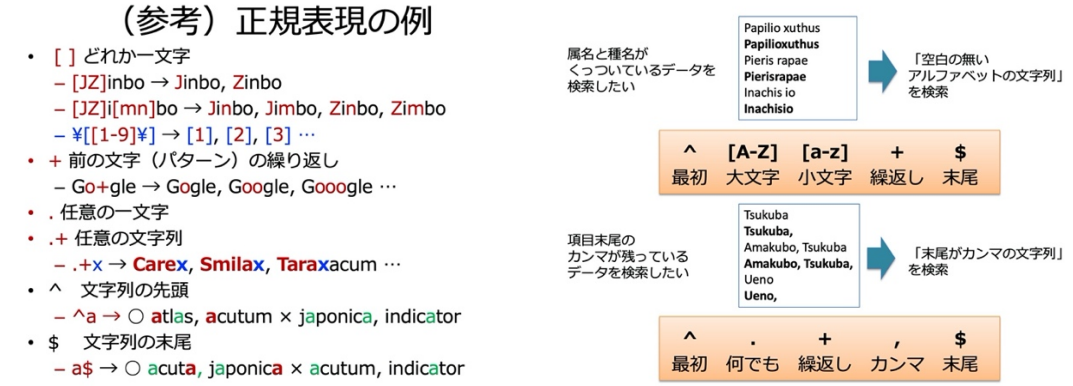
ファイルでの保存
データは自動保存されるので保存操作は必要ないが、エクセルファイルなどにエクスポートすることもできる。
1. 画面右上の「出力」から、欲しいファイル形式を選択すると、その形式でファイルがダウンロードされる。



(4) エディタ（応用編：正規表現）

文字そのものでは無く「パターン」に一致するかどうかで検索をする「正規表現」を使うことで、複雑な検索や置換ができるようになる。上級者向け。対応しているエディタのほか、Google スプレッドシートや OpenRefine などでも使える。





正規表現による置換例

やりたいこと	検索	置換
神保のローマ字を Jinbo で統一したい	[JZ]i[mn]bo	Jinbo
項目末尾のカンマを削除したい	,$	空白文字列
空白行を消したい	¥n+	¥n
1, 2. … を [1], [2] … にしたい	([0-9]+)¥.	[$1]
文字の間にタブを挟みたい	(.)	$1¥t

※正規表現で特別な意味を持つ文字自体を検索したいときは、前に「¥（円マーク、もしくはバックスラッシュ）」をつける。改行は¥n、タブ文字は¥t と書く。

※置換するときは、正規表現を()で囲った部分にマッチした文字列を「$1」で指定できる。

※似た機能に、ワードの「ワイルドカード」がある。エクセルの検索でも、?が任意の一文字、*が任意の文字列を表す（これらの記号自体の検索には、その前に~をつける）。

最後に：データクリーニングのコツ

データクリーニングのコツ

- 入力時の揺れが起きにくいようにする
  - 項目をできるだけ分ける
  - 入力制限等も活用する

	A	B	C	D	E	F
1	雑誌名	巻	開始ページ	終了ページ		
2	Lepidoptera Science	70	99	148		Lepidoptera Science 70: 99–148

<表記揺れ例>

Lepidoptera Science 70: 99–148
Lepidoptera Science 70:99–148
Lepidoptera Science70 :99 - 148
Lepidoptera Science, 70: 99–148

=A2&" "&B2&": "&C2&"–"&D2

結合は簡単、分割は困難

データクリーニングのコツ

- ツールを使い分ける
  - エクセル　何でも　かゆい所に手が届きにくい
  - ウェブツール　ある目的に特化
  - OpenRefine　中級者　データの扱い楽
  - エディタ　中級者　ちょっとした作業を追加
  - スクリプト　上級者　大量データ処理

- できるだけ作業とその結果を残す
  - エクセル：行番号をつけて戻せるようにする
  - OpenRefine：修正履歴を活用する
  - 作業・レシピをメモしておき再利用