第35回　自然史情報発信に関する研究会

データクリーニングのポイント

2020.11.15.

国立科学博物館　標本資料センター
細矢　剛

GBIF Strategic Plan (2017-2021)

1. 科学および社会で必要とされているデータを提供する。
2. データの質を向上する。
3. データのギャップを埋める。
4. インフラ整備を推進する。
5. 国際的ネットワークへ注力する。

ファセット検索だと間違いが見えやすい

データクリーニング

参考文献：https://www.gbif.jp/v2/library/library_nov2017.html

Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data
データクリーニングの原理と方法－原生種および原生種分布のデータ
英語版/2005　和訳版/2005

エラーの種類：

不正確　　採集者：細谷　剛
不完全　　採集地：三国峠
不当　　採集日：2010年2月30日

「予防は治療に勝る」

チェックの視点：
形式のチェック、完全性のチェック、合理性のチェック、制限のチェック、異常値（地理的、統計的、時間的または環境的）あるいはその他のエラーを特定するためのデータ評価、および主題領域の専門家（例えば分類学の専門家）によるデータ評価など。

データ入力のお行儀を知る

1.データには、数値（字）型と文字型がある。

数値型→計算できる

文字型→計算できない；半角＝1バイト、全角＝2バイト

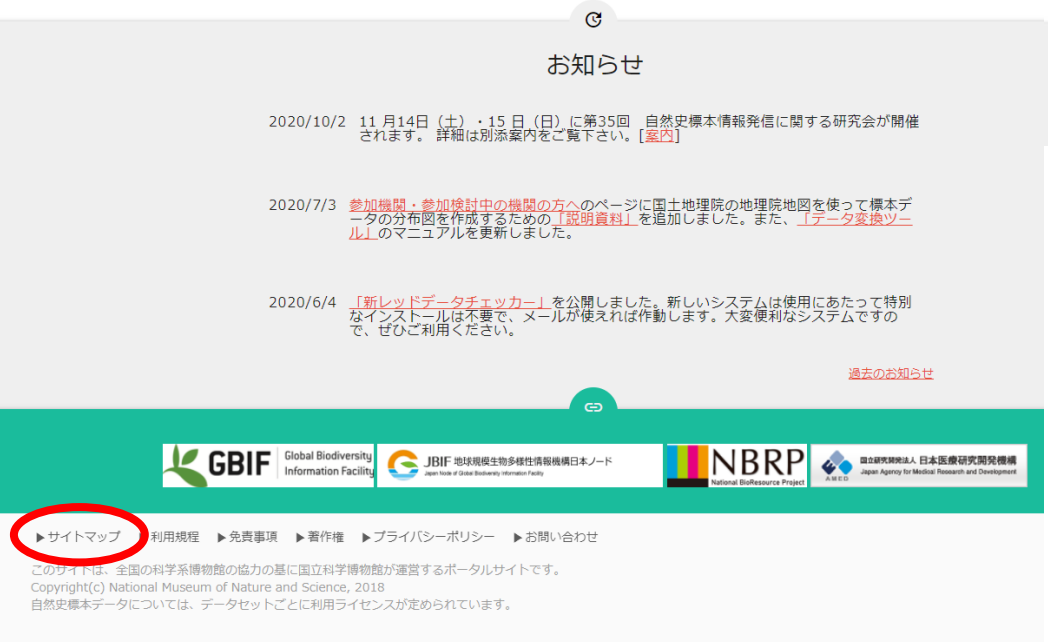
123	数字
123K	文字
123□（※□はスペース）	文字
65	文字
029-853-0000	文字

2.S-Netのデータ項目がどの形式かはマニュアル参照

URL:http://science-net.kahaku.go.jp/contents/tool/dataconv_manual_v1.10.pdf

http://science-net.kahaku.go.jp/?ln=ja

参加機関・参加検討中の機関の方へ



サイトマップ →

戻る

参加機関・参加検討中の機関の方へ

はじめに

データ提供について概略をご紹介します。

このページについて

このページには、すでに参加されている機関の方や、今後S-Net/GBIF活動へ参加を検討されている機関の方々に向けて、データ提供のプロセスや参考資料、ツールなどをリストしました。ここでは詳しい説明を省き、概略だけをご紹介します。

【ご注意】Internet Explorerなどブラウザの環境によっては、本ページ内のリンク先へのジャンプが動作しないことがあります。

S-Netへのデータ提出

S-Netへのデータ提出のためには、所定のデータが、所定のデータ項目名と形式で整理されている必要があります。どんなデータ項目が必要か、データ変換ツールのマニュアルの末尾の表（p.17～25）で、確認してみてください。もし、お手元のデータが所定の項目と合致していれば、データ変換ツールで、提出用のデータファイルを作成して、事務局（国立科学博物館S-Net/GBIF担当）まで、メールでお送りください。データが整っていない場合には、データ事前整形支援ツールを用いて、整形してから変換してください。なお、地名の統一などには自然史研究のための地名辞書・日本沿岸地名辞書などを利用することができます。また、レッドリスト種については、産地の公開に注意を払う必要があります。どの種が該当するかを調べるには新レッドデータチェッカーを利用することができます。

メタデータ情報の提供

提供されたデーター式は「データセット」として管理されています。これらがどんなデータセットかを説明するデータをメタデータ（たとえば、「○×博物館の昆虫コレクション」のようなものです）といい、提供データとは別にご用意頂く必要があります（詳しくは、メタデータ登録票の記載をご覧ください）。

データの公開とライセンス

ご提供いただいたデータは事務局におけるチェックを経て、S-Netから公開され、国内で利用されるとともに、GBIF（地球規模生物多様性情報機構、https://www.gbif.org/）やOBIS（海洋生物地理情報システム、http://www.iobis.org/）から公開され、世界的にも利用されます。データの二次利用については、事前に許諾（ライセンス）を与えることになっており、ライセンスの国際的な標準となっているクリエイティブコモンズのCC0、CC BY、あるいはCC BY-NCを指定していただきます（詳しくは、CCライセンスのご案内をご覧く

データクリーニングの重要なポイント

1.いつ
　日付の形式

2.どこで
　緯度経度を取得する
　測地系に配慮する

3.何を
　法規制種
　レッドリスト種
　学名のチェック

ホワイトスペース

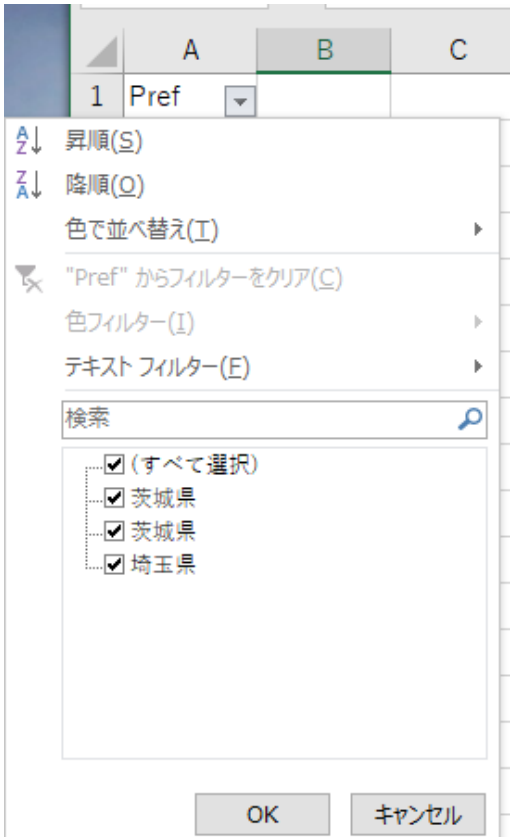
	A	B
1	Pref	
2	茨城県	
3	茨城県	
4	埼玉県	
5		

1）A2には「茨城県」の後に、全角のスペースが入っているが見えない

2）データタブからフィルタを選択

3）▼をクリック

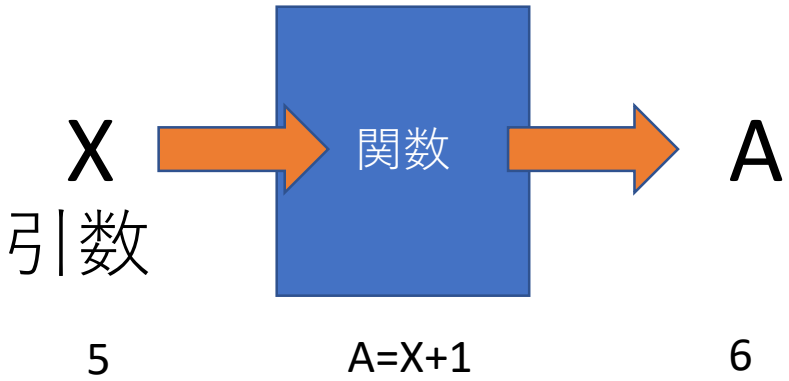
	A
1	Pref
2	茨城県
3	茨城県
4	埼玉県

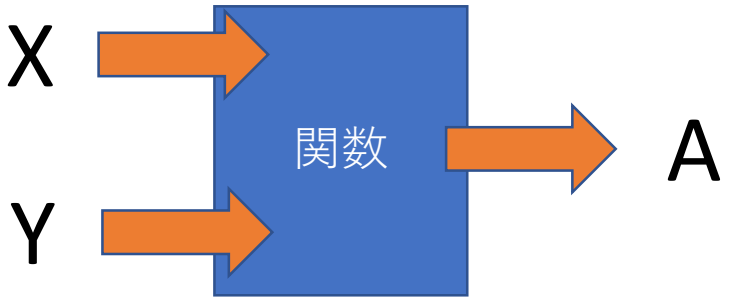


4）「茨城県」が2つ？？
→おかしい、と推論

エクセルの関数

関数：指定した値（1～複数）から、一つの値を誘導する





覚えておくと便利な関数

LEFT（文字列, 文字数）

Left(“ABCDE”, 2)=AB

RIGHT(文字列,文字数)

Right(“ABCDE”, 2)=DE

MID(文字列, 開始位置, 文字数)

Mid(“ABCDEF”, 3, 2)=CD

FIND(検索文字列, 対象, [開始位置])

[ ]はオプション（あってもなくてもよい）

Find(“ ”,“AB cdefg”, 1)=3

組み合わせ技で属と種を分ける

1） Find(“ ”, “Oryza sativa”,1)=6

2） Mid(“Oryza sativa”,7,100)=sativa

3) Left(“Oryza sativa”,5)=Oryza

4) 1)+2)でMid(“Oryza sativa”, Find(“ ”, “Oryza sativa”,1)+1,100)=sativa

※下線部が2）と同じになっている。

覚えておくと便利な関数

IF：関数の結果の真・偽で判断を分ける

書式：IF(A論理式, [B真の場合], [C偽の場合])

Aが正しければB、それ以外はC

B,Cのどちらかは必ず入れる。

IF(X-Y>0, “Xが大きい”, ”Yが大きい”)

X=3, Y=5のとき、 ”Yが大きい”

1-関数.xlsxにいくつかの関数を見本で示しました。