第36回　自然史情報発信に関する研究会

# エクセル（関数など）を使ったクリーニング

2021.2.6.

国立科学博物館　標本資料センター
細矢　剛

以下の資料をダウンロードしてください。

http://science-net.kahaku.go.jp/app/page/activity.html##studygroup

「第36回　GBIF関連サイトの使い方とより品質の高いデータ提供のためのテクニック」

にあるファイル

# ソートとフィルタ

# エラー発見の基本原理

変なデータを見つける

1. ソートする
2. フィルタを利用する

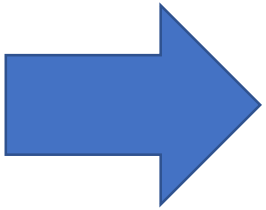
注意！

段ズレに注意する

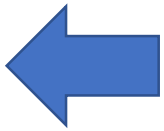
多様すぎるデータには対応困難

| 県名 |
| --- |
| 茨城県 |
| 埼玉県 |
| 沖縄県 |
| 茨木県 |
| 大阪 |
| 大阪府 |
| 埼玉県 |
| 中央区 |
| 東京都 |
| 茨城県 |
| 埼玉県 |
| 埼玉県 |

ソート



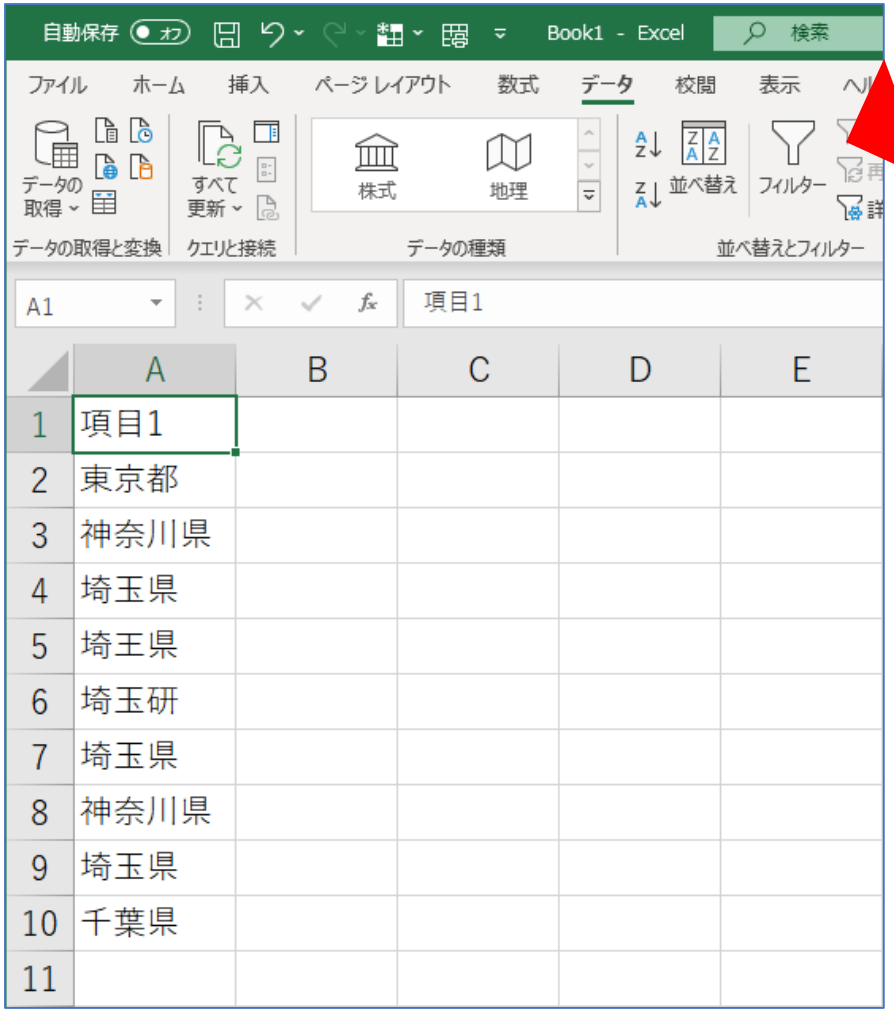
| 県名 |
| --- |
| 茨城県 |
| 茨城県 |
| 茨木県 |
| 大阪 |
| 大阪府 |
| 沖縄県 |
| 埼玉県 |
| 埼玉県 |
| 埼玉県 |
| 埼玉県 |
| 中央区 |
| 東京都 |

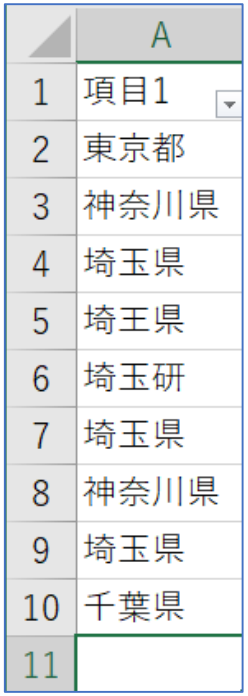


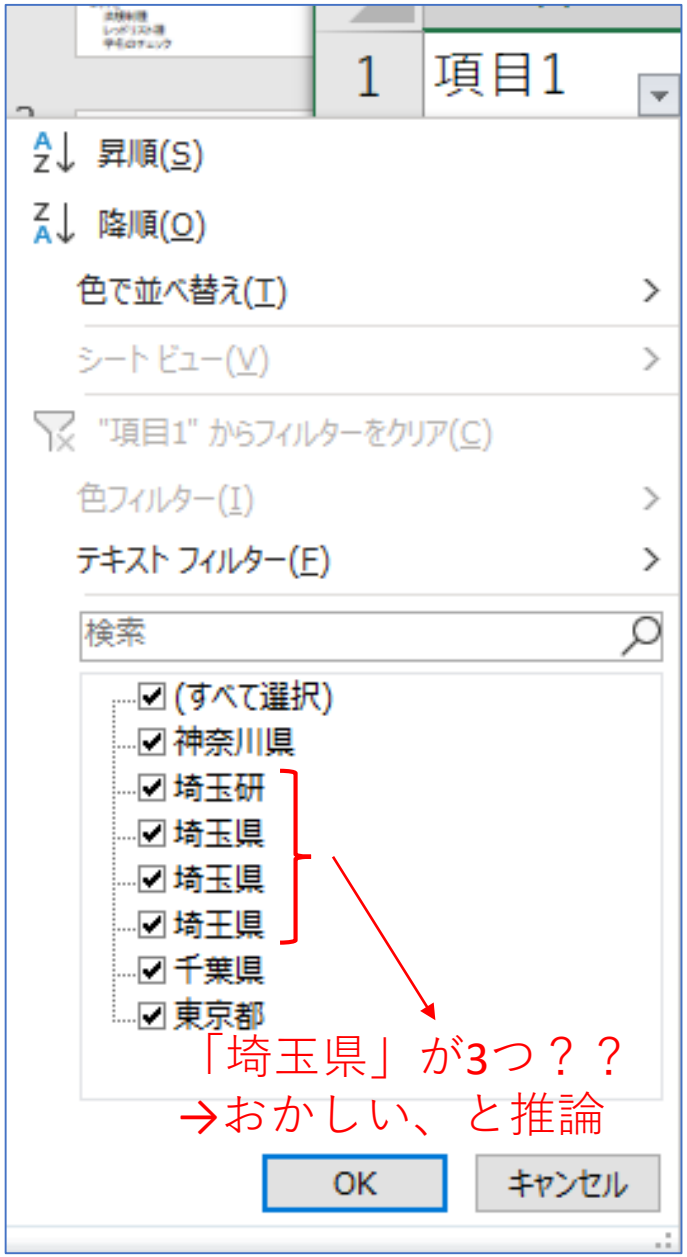
埼玉県□

見えないスペース
（ホワイトスペース）

# フィルターの利用


自動保存
Book1 - Excel
検索
ファイル
ホーム
挿入
ページ レイアウト
数式
データ
校閲
表示
データの取得
すべて更新
株式
地理
並べ替え
フィルター
データの取得と変換
クエリと接続
データの種類
並べ替えとフィルター
A1
項目1
A B C D E
1 項目1
2 東京都
3 神奈川県
4 埼玉県
5 埼王県
6 埼玉研
7 埼玉県
8 神奈川県
9 埼玉県
10 千葉県
11


A
1 項目1
2 東京都
3 神奈川県
4 埼玉県
5 埼王県
6 埼玉研
7 埼玉県
8 神奈川県
9 埼玉県
10 千葉県
11


1 項目1
昇順(S)
降順(O)
色で並べ替え(T)
シート ビュー(V)
"項目1" からフィルターをクリア(C)
色フィルター(I)
テキスト フィルター(F)
検索
(すべて選択)
神奈川県
埼玉研
埼玉県
埼玉県
埼王県
千葉県
東京都
「埼玉県」が3つ？？
→おかしい、と推論
OK
キャンセル

# 検索だけでも見つけられる

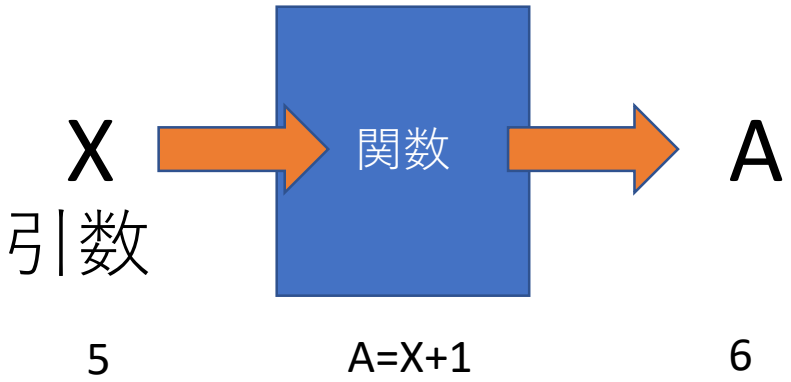
## Ctrl+F（検索）

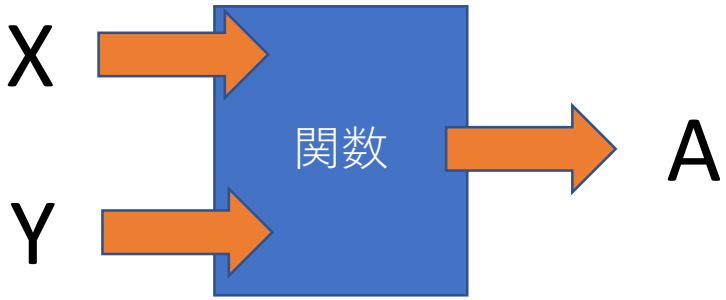

cf. Ctrl+R （置換）

# 関数

# 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。

TRIM(文字列)　無駄なスペースを消す

Trim(“ABC__”)=AB　　※“__”はホワイトスペース

# 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。

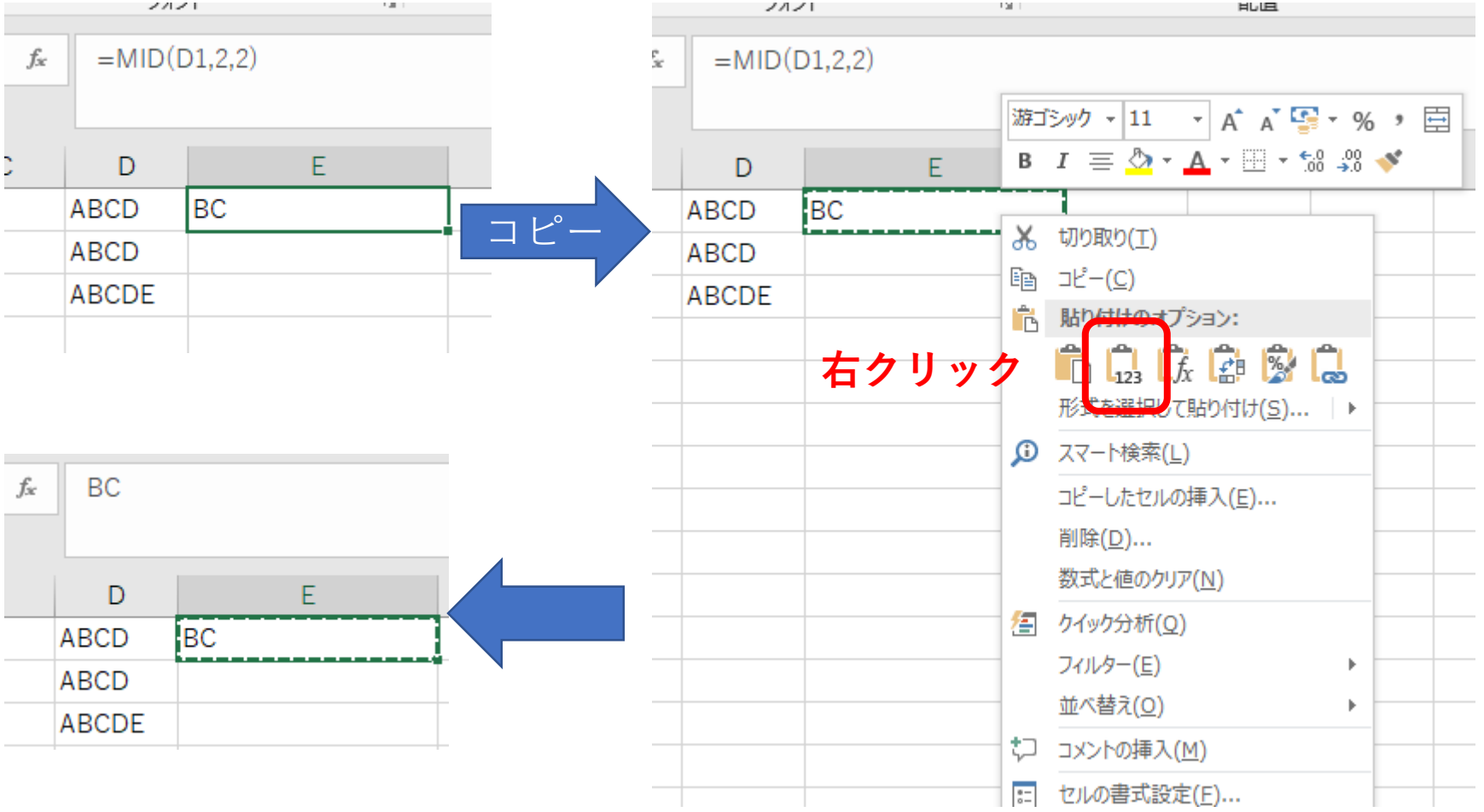
IF(X-Y>0, “Xが大きい”, ”Yが大きい”)

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。

# 関数を消す(“値複写”)

コピー→ペースト時に右クリック、「形式を選択して貼り付け」で「値（左から2番め）」を選ぶ。



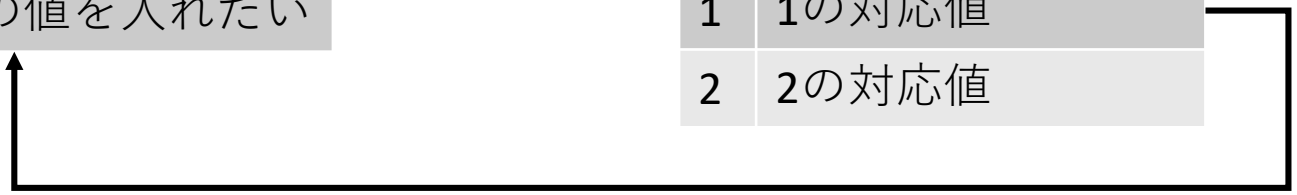
# VLOOKUP

Sheet1

| A | B |
|---|---|
| 1 | ここにYの値を入れたい |

Sheet2 (Name List)

| X | Y |
|---|---|
| 1 | 1の対応値 |
| 2 | 2の対応値 |



VLOOKUP(B2,'Name List'!$A$1:$B$10,2,FALSE)

①　②　③　④条件

元になる値　参照先範囲　何番目の列

①の値を覚え、
②の範囲の一番左の列からその値を見つけ
その列から数えて、③で与えられた番号の列の値を返しなさい。
④（オプション）ただし、完全一致の場合に限る

※ポイント：参照先を動かさないようにA1:B10ではなく$A$1:$B$

# 具体例の紹介

1. 単純ミス・文字化け
2. 数値項目にありがちなミス
3. 一貫性に関するミス

1. 単純な入力ミスと文字化け

1）単純な入力ミス。修正して提出してください。

例 1：”鳥網”（網（こう）が　網（あみ）になっている）→”鳥綱”に修正

例 2：”#N/A”や”#VALUE!”が残っている（作業途中のゴミの消し忘れ）→削除する

解説：綱（こう）が　網（あみ）になっている間違いは、しばしば見られます。注意してください。
エクセルで関数（VLOOKUP、HLOOKUP、LOOKUP、MATCH など）を使用した際、参照値が見つからずエラー値 ”#N/A”が示される場合があります。また、同じくエクセルで数式に文字列が含まれていると”#VALUE”が示されます。これらの値は忘れずに削除してください。

ちょっと恥ずかしい・・・
フィルター機能を使えば、簡単に見破れるはず

2) ウムラウト付きの特殊文字が「?」などに置き換えられている。修正して提出してください。

例：[学名]が"Pidonia (Pidonia) shikokensis shikokensis Ch?j? et Hayashi, 1951"

→"Pidonia (Pidonia) shikokensis shikokensis Chûjô et Hayashi, 1951"または

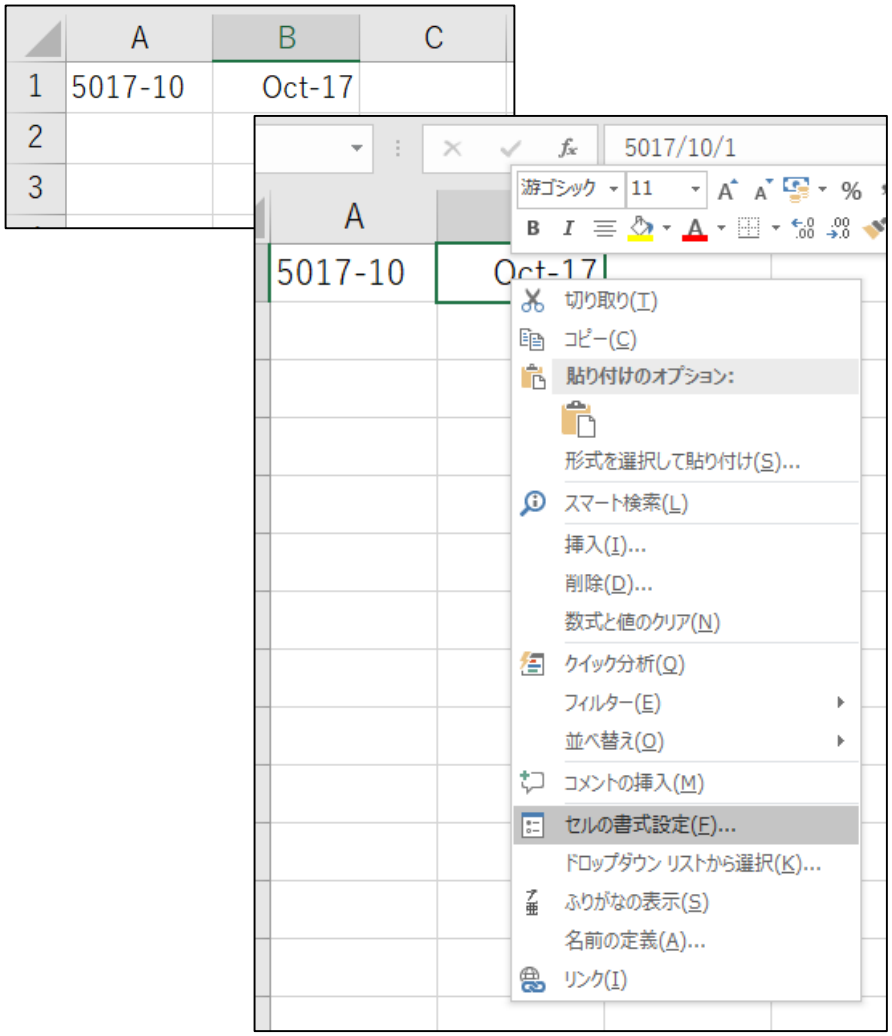
"Pidonia (Pidonia) shikokensis shikokensis Chujo et Hayashi, 1951"に修正

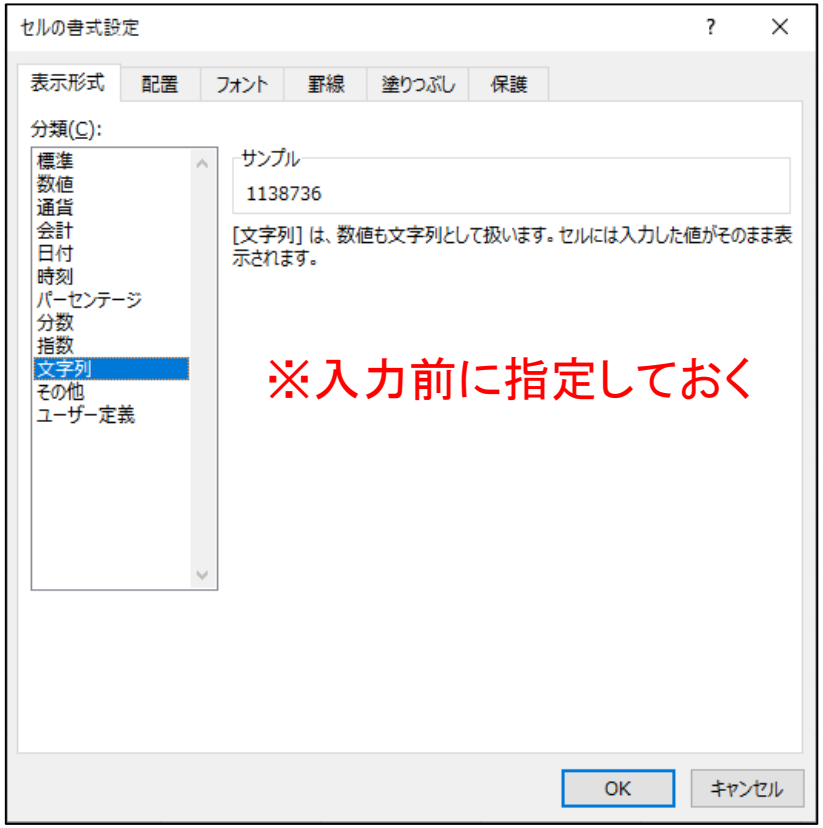
解説：[学名][学名の著者]に見られます。これは、エクセルの標準形式からシフト JIS 形式の CSV ファイルに変換するときによく生じる事象です。Excel2019 からは Unicode を使えるようになりましたので、保存のときに Unicode を指定しておけば、特殊文字が維持されます。

3) ハイフン付きの数字が日付型に置き換えられている。修正して提出してください。

例：[カタログ番号]"5017-10"が"Oct 5017"になっている→"5017-10"に修正

解説：[カタログ番号]などのハイフン付きの数字が日付型に置き換えられているときは、エクセルで自動変換が行われるので、列の分類を「文字列」に変更して再入力します。




※入力前に指定しておく

※「‘5017-10」でも、表示上はOKだが、余計な文字が入るので使うべきではない。

# 具体例の紹介

1. 単純ミス・文字化け
2. 数値項目にありがちなミス
3. 一貫性に関するミス

参考資料D02参照

2. 数値項目にありがちなミス

1）値が不明な場合に 0 値が入っている。空白にして提出してください。

例：[メッシュコード]、[緯度（十進数表記)]、[経度（十進数表記)] に“0”→空白とする

解説：”0”は「値がない」という意味ではありません。不明な場合は空白としてください。特に緯度や経度における”0”は、赤道を意味する位置情報になってしまいます。エクセルなどでの入力であれば、「フィルタ」機能を利用するなどして”0”を見つけることができます。

2)海抜、水深、記録年月日の範囲値が逆転して入っている。修正して提出してください。

例 1：[最低海抜] > [最高海抜] → [最低海抜] ≦ [最高海抜]

例 2：[記録年月日（始め)] > [記録年月日（終わり)] → [記録年月日（始め)] ≦ [記録年月日（終わり)]

解説：値が範囲でない場合は、先頭末尾両方の項目、または先頭項目のほうに値を入れてください。
記録年月日は数値項目ではありませんが同様に注意してください。アスタリスク（*）を含む場合もあり機械的なチェックは難しいですが、完全な年月日であれば、入力時に注意する他、「終わり」から「初始め」を引いて、正の値が得られるか、などを基準にしてチェックすると良いでしょう。

# 日付の形式

1. S-Netでは、「始まり」「終わり」がある。
2. 片方しかない時には、どちらか（はじめ）に一貫して入力。
3. 入力の形式はYYYYMMDD。
4. 不明の箇所は**で埋める。ただし、より上位が不明の場合、下位は**とする。

◯19750219
◯197502**
◯1975****
×1975**19 → ◯1975****
×1975*219 → ◯1975****
×197**219 → ◯********
×197002** → ◯197002**

# 採集年月日情報をS-Net型に変更(1)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | No | Yr | Mo | Da | ① | ② | ③ | ④ | ⑤ |
| 2 | 1 | **1963** | **12** | **19** | **1963** | **12** | **19** | **19631219** | **19631219** |
| 3 | 2 | 1982 | 4 | 1 | 1982 | 04 | 01 | 19820401 | 19820401 |
| 4 | 3 | 1955 | 12 | 15 | 1955 | 12 | 15 | 19551215 | 19551215 |
| 5 | 4 | 2014 | 10 | | 2014 | 10 | ** | 201410** | 201410** |
| 6 | 5 | 2011 | | 9 | 2011 | ** | ** | 2011**** | 2011**** |
| 7 | 6 | | | | **** | ** | ** | ******** | ******** |

①IF( ISNUMBER(B2),B2,"****")
Yrが数字かどうか調べ、空欄ならば"****"を返す

②IF(ISBLANK(C2),"**",IF(C2<10,"0"&C2,C2))
Moが空欄かどうか調べ、空欄ならば"**"を返す。空欄でなければ、10より小さいかどうか調べ、小さければ、0で2桁にした文字を返す。それ以外は、そのままの数字（文字）を返す。

③IF(ISBLANK(C2),"**",IF(ISBLANK(D2),"**",REPT("0",2-LEN(D2))&D2))
Dayが空欄かどうか調べ、空欄ならば"**"を返す。空欄でなければ、Moが空欄かどうか調べ、空欄ならば"**"を返す。以上に該当しなければ、0で2桁にした文字を返す。それ以外は、そのままの数字（文字）を返す。

# 採集年月日情報をS-Net型に変更(2)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | No | Yr | Mo | Da | ① | ② | ③ | ④ | ⑤ |
| 2 | 1 | **1963** | **12** | **19** | **1963** | **12** | **19** | **19631219** | **19631219** |
| 3 | 2 | 1982 | 4 | 1 | 1982 | 04 | 01 | 19820401 | 19820401 |
| 4 | 3 | 1955 | 12 | 15 | 1955 | 12 | 15 | 19551215 | 19551215 |
| 5 | 4 | 2014 | 10 | | 2014 | 10 | ** | 201410** | 201410** |
| 6 | 5 | 2011 | | 9 | 2011 | ** | ** | 2011**** | 2011**** |
| 7 | 6 | | | | **** | ** | ** | ******** | ******** |

①IF( ISNUMBER(B2),B2,"****")

②IF(ISBLANK(C2),"**",IF(C2<10,"0"&C2,C2))

③IF(ISBLANK(C2),"**",IF(ISBLANK(D2),"**",REPT("0",2-LEN(D2))&D2))

④E2&F2&G2

IF( ISNUMBER(B2),B2,"****")&IF(ISBLANK(C2),"**",IF(C2<10,"0"&C2,C2))&IF(ISBLANK(C2),"**",IF(ISBLANK(D2),"**",REPT("0",2-LEN(D2))&D2))

3）緯度と経度が逆転している。提出前に確認をお願いします。

例：北海道石狩市の［緯度（十進数表記）］が”141.3155”、［経度（十進数表記）］が”43.1713”

解説：これも致命的な間違った情報になってしまいますが、機械的なチェックは難しいです。しかし、国内であれば、緯度、経度のそれぞれを「フィルタ」や「ソート」を利用して逆転した数値を比較的簡単に見つけることができます。

# 具体例の紹介

1. 単純ミス・文字化け
2. 数値項目にありがちなミス
3. 一貫性に関するミス
4. レッドデータチェックに必要な項目
5. 変換ツールで治るミス

参考資料D02参照

3.データセットを通じて一貫しているべき項目に不備がある

1）メタデータの通りになっていない。メタデータに合わせてください。

例：国立科学博物館（植物）維管束植物コレクションの場合

［機関名］　”National museum of nature and science”→ ”National Museum of Nature and Science”

［機関名（日本語）］　”国立科学博物館[植物]”→“国立科学博物館（植物）”

［機関コード］　”TSN”→ ”TNS”

［コレクションコード］　”vs”→ ”VS”

解説：大変多く見られる誤りです。これらの値は、全データベースを通じて一貫していることが重要です。不安なときは、過去提出したデータを検索して、確認しましょう。掲載済みのメタデータの情報はS-Netサイトの「機関・データセット一覧」（http://science-net.kahaku.go.jp/app/k）で確認できます。

メタデータの方を修正されたい場合はご相談ください。

2) [カタログ番号]の形式が統一されていない。極力統一してください。

例：以前が”AAA-BBBB-0001”で今回が”0010”→　今回も”AAA-BBBB-0010”で統一する。

解説：カタログ番号の形式が統一されていないと、データの重複が起こりやすくなります。特に理由がある場合を除き、データセット通じて統一した形式にしましょう。不安なときは、過去提出したデータを検索して、確認しましょう。掲載済みのデータは S-Net サイトの「機関・データセット一覧」（http://science-net.kahaku.go.jp/app/k）でデータセットを選択し、「データを見る」で表示できます。

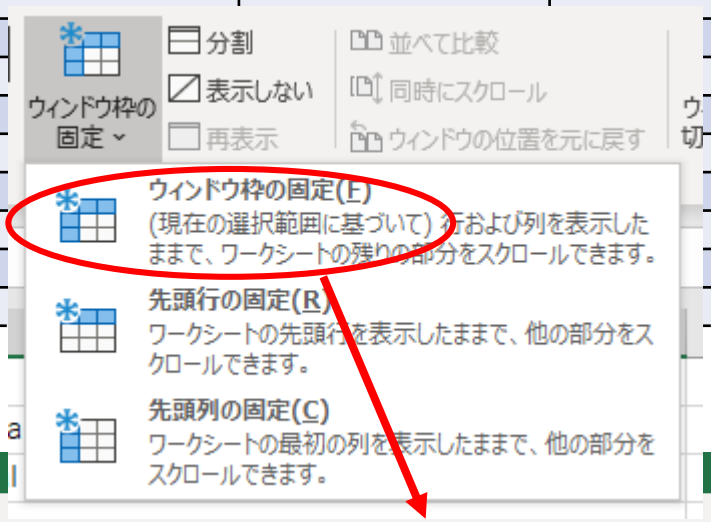
また、ハイフンの半角、全角のチェックも重要なポイントです。通常は半角のハイフンを使います。

前回提出から、何年かを経て提出する
複数の担当者がデータ作成に関わる

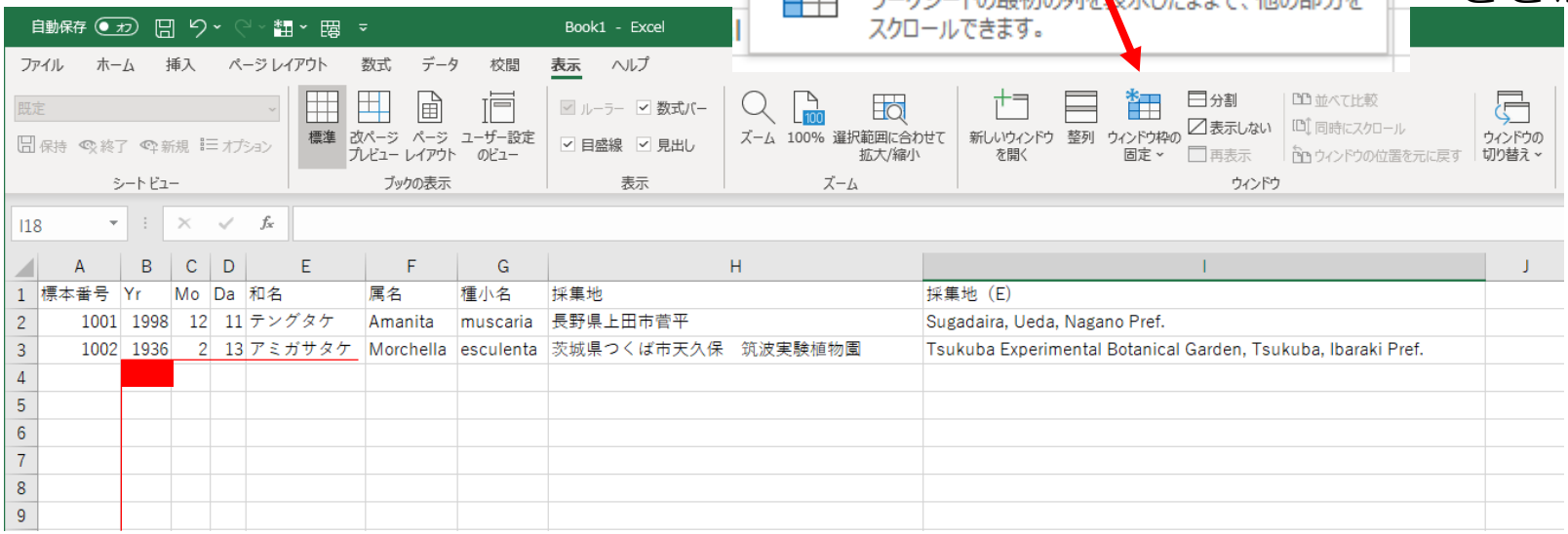
注意！

# 入力時に「見本」を示す

| | | | | |
|---|---|---|---|---|
| 1 | Sample data | Sample data | Sample data | Sample data |
| 2 | Sample data | Sample data | Sample data | Sample data |
| 3 | Sample data | Sample data | Sample data | Sample data |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |
| 11 | | | | |
| 12 | | | | |




ここから入力




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 標本番号 | Yr | Mo | Da | 和名 | 属名 | 種小名 | 採集地 | 採集地（E） | |
| 2 | 1001 | 1998 | 12 | 11 | テングタケ | Amanita | muscaria | 長野県上田市菅平 | Sugadaira, Ueda, Nagano Pref. | |
| 3 | 1002 | 1936 | 2 | 13 | アミガサタケ | Morchella | esculenta | 茨城県つくば市天久保　筑波実験植物園 | Tsukuba Experimental Botanical Garden, Tsukuba, Ibaraki Pref. | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |
| 8 | | | | | | | | | | |
| 9 | | | | | | | | | | |

3）[カタログ番号]は重複がないようにチェックしてください。

解説：エクセルの「ピボットテーブル」機能（挿入＞ピボットテーブル）で、対象列のデータごとに出現する個数を知ることができます（行に目的の項目を指定し、Σ値にその項目の「個数」を指定）。
または、カタログ番号の列(登録データファイルではM列)を選択し[条件付き書式]>[セルの協調表示ルール]>[重複する値]を行い、[フィルター]を設定して[色フィルター]＞[セルの色でフィルター]で絞り込むことでも確認することができます。
「重複の削除」機能を使うと、重複している行の一方が確認なしにまとめて削除されるのでご注意ください。

ピボットテーブルは、とても便利な機能です。
この機会にぜひマスターしましょう。

# 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;大小関係がある

文字型→計算できない;半角＝1バイト、全角＝2バイト

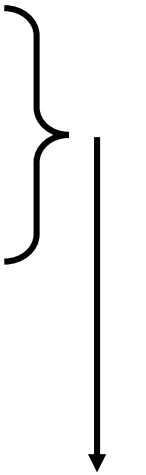
| | |
|---|---|
| 123 | 数字 |
| 123K | 文字 |
| 123□(※□はスペース) | 文字 |
| 65 | 文字 |
| 029-853-0000 | 文字 |

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

# ホワイト行問題

| | | | | |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |
| 11 | | | | |
| 12 | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

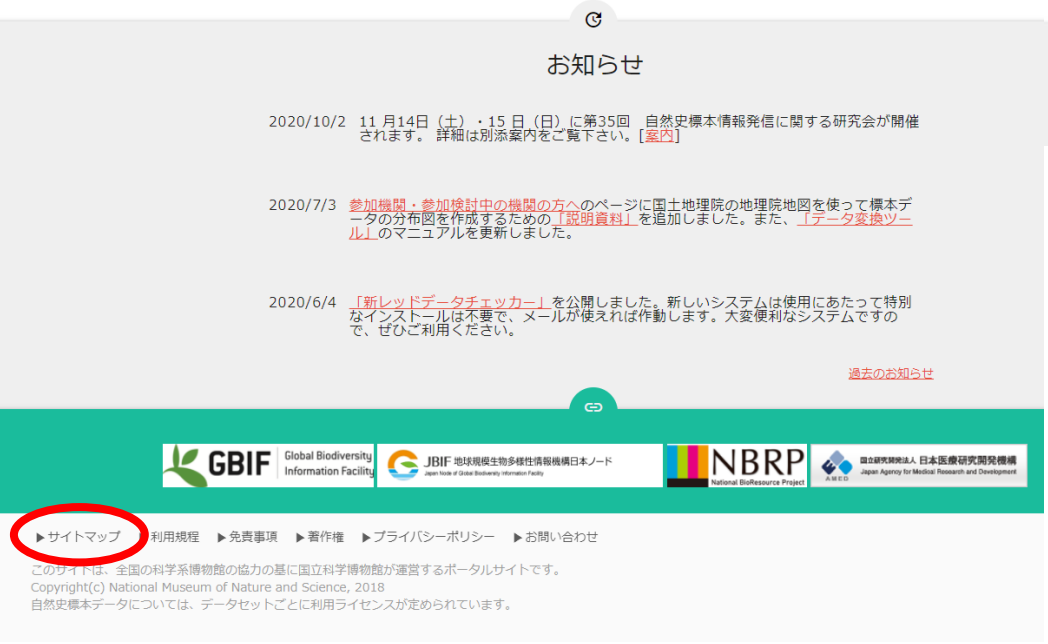


データがないはずの行に見えないデータが入っていることがある

念の為に、最初に削除

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

# レッドデータチェックのために
# ご協力をお願いします

## 和名

1. レッドデータチェックのため、不明でない場合は入力してください。
2. カタカナ表記で、1個だけ（レッドリスト掲載名）を入れてください。標準和名以外の情報は各備考欄に入力してください。

× 「ベニテングタケ（ベニテングダケ、紅天狗茸）」「コムラサキ　黒色型」「ニホンイイズナ（本州亜種）」「コレラタケ＝ドクアジロガサ」

○ 「ベニテングタケ」「コムラサキ」「ニホンイイズナ」「コレラタケ」

×［和名］に「ヒメネズミ（頭蓋骨）」

○［和名］に「ヒメネズミ」、［オカレンス備考（日本語）］に「ヒメネズミ（頭蓋骨）」

※［和名］における雑種式は、全角の”×”(掛け算記号)をお使いください。
（例）「ミヤマチョウジザクラ×タカネザクラ」

# レッドデータチェックのために
# ご協力をお願いします

## 都道府県（日本語）

都道府県名を必ず入れてください（県をまたぐ場合は”愛知県／静岡県”のように入れて下さい）。

http://science-net.kahaku.go.jp/app/page/activity.html
に過去の資料あり