第35回　自然史情報発信に関する研究会

問題あるデータのチェックポイント

2020.11.15.

国立科学博物館　標本資料センター
細矢　剛

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35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

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

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

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。

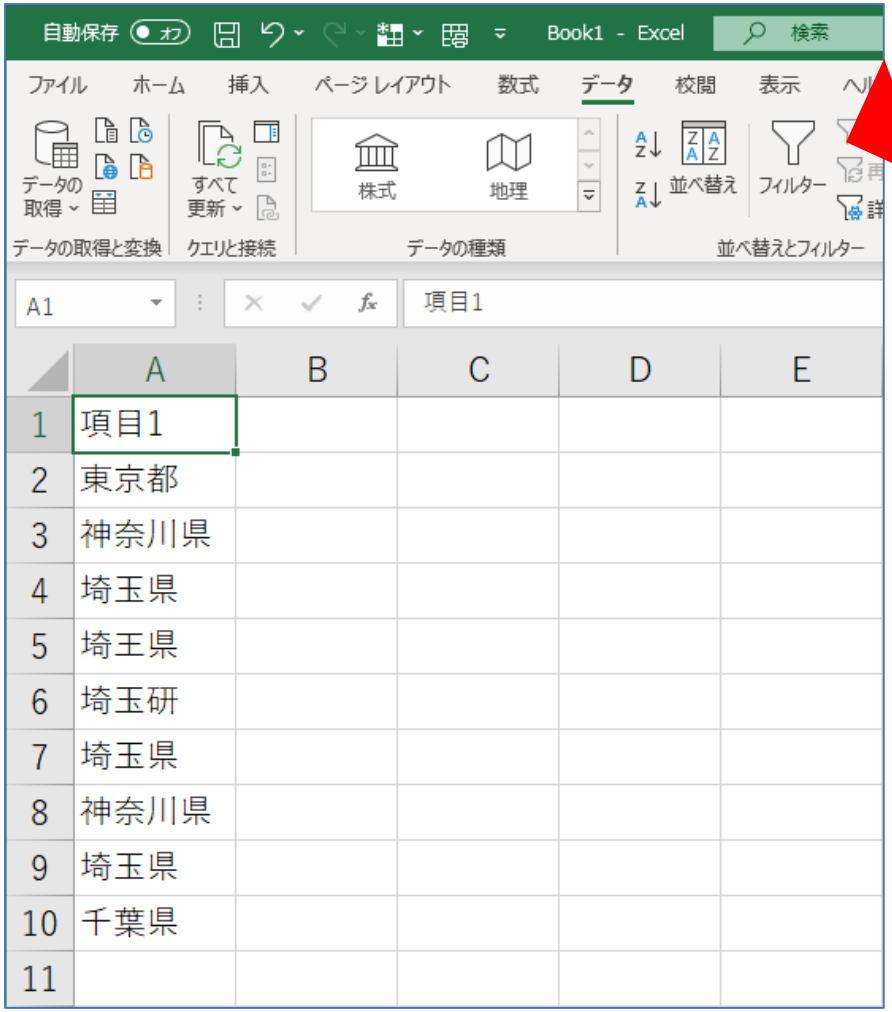
例 1：”鳥網”（網（こう）が　網（あみ）になっている）→”鳥綱”に修正

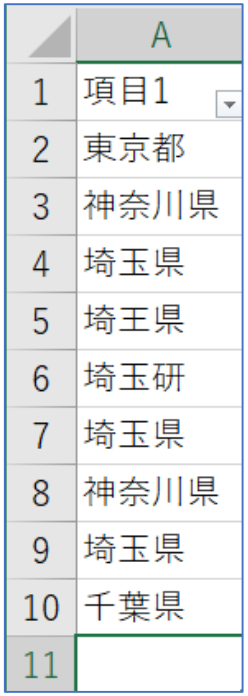
例 2：”#N/A”や”#VALUE!”が残っている（作業途中のゴミの消し忘れ）→削除する

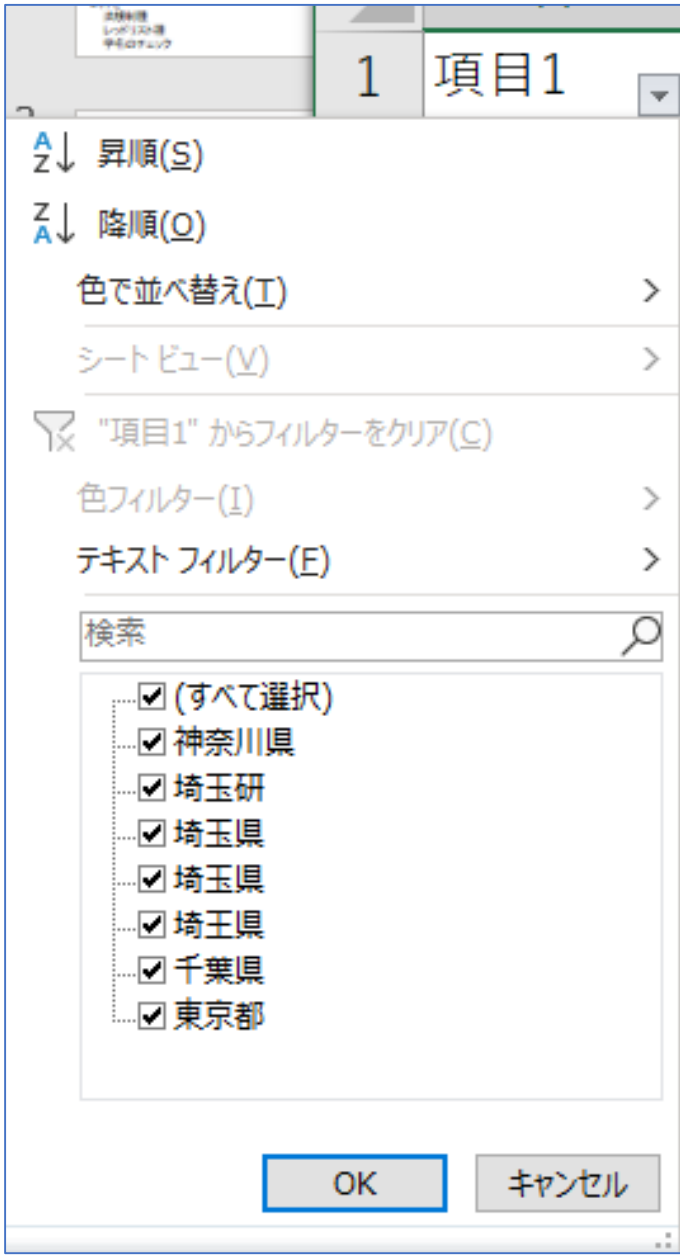
解説：綱（こう）が　網（あみ）になっている間違いは、しばしば見られます。注意してください。
エクセルで関数（VLOOKUP、HLOOKUP、LOOKUP、MATCH など）を使用した際、参照値が見つからずエラー値 ”#N/A”が示される場合があります。また、同じくエクセルで数式に文字列が含まれていると”#VALUE”が示されます。これらの値は忘れずに削除してください。

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

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 埼玉県
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 埼玉県
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
埼玉県
千葉県
東京都
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）

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

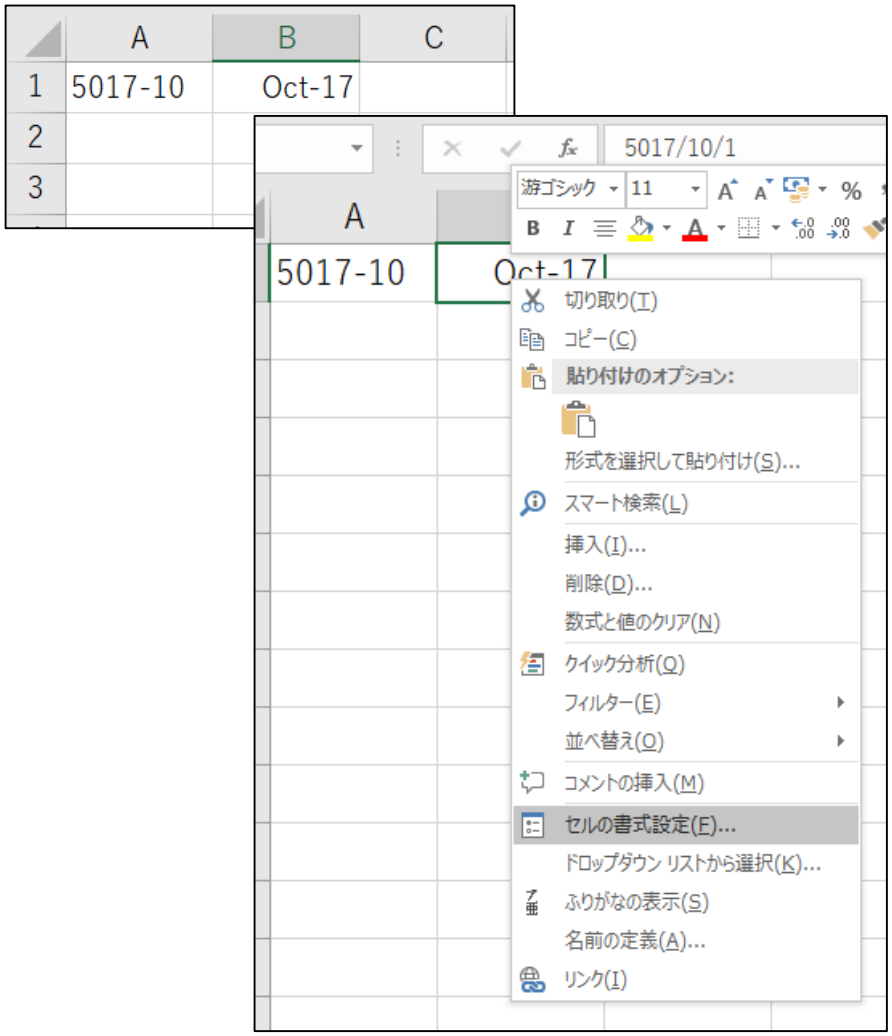
"Pidonia (Pidonia) shikokensis shikokensis Chujo et Hayashi, 1951"に修正

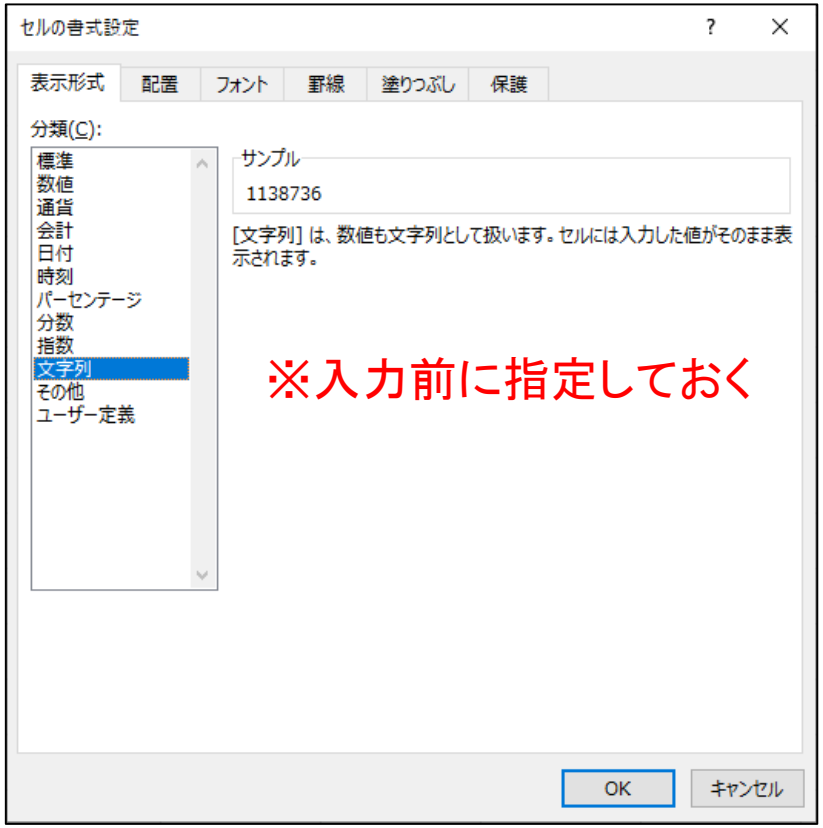
解説：[学名][学名の著者]に見られます。これは、エクセルの標準形式からシフト JIS 形式の CSV ファイルに変換するときによく生じる事象です。Excel2019 からは Unicode を使えるようになりましたので、保存のときに Unicode を指定しておけば、特殊文字が維持されます。

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
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

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 → ◯********

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！

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。

4）マッピング時の項目選択が適切でない。提出前に確認をお願いします。

例 1：[綱名（日本語名）]に“Magnoliopsida“→正しくは[綱名（学名）]にマッピングされているべき

例 2：[国（日本語）]に“Japan”→正しくは[国]にマッピングされているべき

解説：特に日本語項目に英語項目をマッピングしてしまう誤りが多く見られます。これは変換ツールでエラーとならないため気づきにくいパターンですが、掲載済みのデータと内容が一貫しない原因にもなりますのでご注意ください。

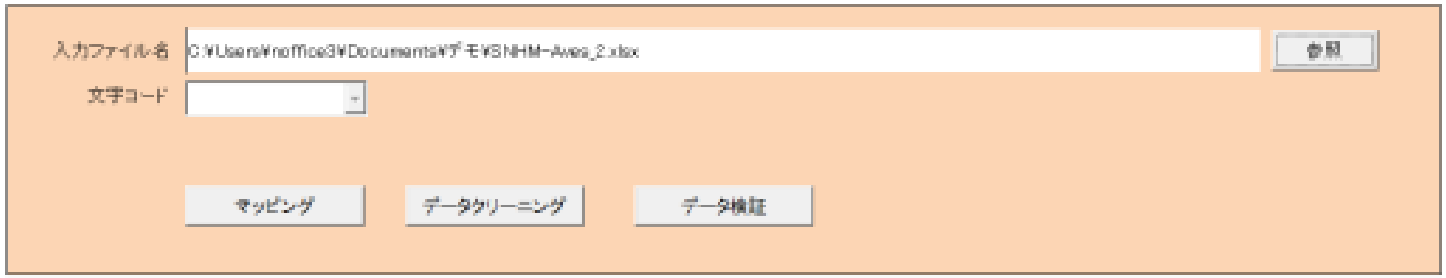
マッピング：自分のデータベース項目に適当なS-Netのデータ項目を参照させること

1. S-Netにデータを提出するためには、S-Netのデータ形式に変換する必要があります。

2. 変換のための詳細は、http://science-net.kahaku.go.jp/app/page/tool_download.html　S-Netのホームページの右上の「参加機関・参加検討中の機関の方へ」の「S-Netへのデータ提出」を御覧ください。

3. 変換を行うために「データ変換ツール」を配布しています（無料）。

標本データ変換



④[入力ファイル項目名]欄のドロップダウンリストから、事前整形ファイルの項目を選択するか、「直接入力」を選択して [値の直接入力]欄に値を入力し、S-Netの項目を対応付け（マッピング）していきます。

項目マッピング

	情報群	S-Net項目名	データ型	項目説明	入力ファイル項目名	値の直接入力	表示項目
6	基本情報	データ登録日時	日時	S-Net/GBIF事務局管理用項目			○
7	基本情報	GBIF公開フラグ	論理(真偽)型	0:GBIFで公開しない、1:GBIFで公開する(既定値)	直接入力	☑	○
8	基本情報	レコード種別	文字列(半角英数字)	PreservedSpecimen, FossilSpecimen, LivingSpecimen, HumanObservation, MachineObservation, MaterialSample, Occurrenceのいずれか	直接入力	PreservedSpecimen	○
9	基本情報	機関名	文字列(半角英数字)	機関名の英文名称	直接入力	Saito Natural History I	○
10	基本情報	機関名(日本語)	文字列(日本語可)	機関名の和文名称	直接入力	斎藤報恩会自然史博物	○
11	基本情報	機関コード	文字列(半角英数字)	標本の機関コード(例:TNS, NMST, KPM)	直接入力	SNHM	○
12	基本情報	コレクションコード	文字列(半角英数字)	標本のコレクションコード(例:VS, F, AL)。コレクションコードが未設定の機関では、機関コードを入れる。	直接入力	Aves	○
13	基本情報	カタログ番号	文字列(半角英数字)	標本番号	カタログ番号		○
14	オカレンス情報	採集者番号	文字列(半角英数字)	採集者によるオリジナルの標本番号			○
15	オカレンス情報	オカレンス備考	文字列(半角英数字)	標本等に関する補足説明(例:found dead on the road)			○
16	オカレンス情報	オカレンス備考(日本語)	文字列(日本語可)	標本等に関する補足説明(例:路上轢死体)			○
17	オカレンス情報	性別	文字列(半角英数字)	例:male, female	性別		○
18	オカレンス情報	性別(日本語)	文字列(日本語可)	例:オス、メス	性別(日本語)		○
19	オカレンス情報	生活型・世代型	文字列(半角英数字)	個体のライフステージ(例:juvenile, adult, sporophyte)			○
20	オカレンス情報	生活型・世代型(日本語)	文字列(日本語可)	例:幼虫、成虫、胞子体			○
21	オカレンス情報	成熟状況	文字列(半角英数字)	例:pregnant, in bloom, fruit-bearing			○
22	オカレンス情報	成熟状況(日本語)	文字列(日本語可)	例:妊娠中、開花中、結実中			○
23	オカレンス情報	行動	文字列(半角英数字)	採集時の個体の振る舞い(例:roosting, foraging, running)			○
24	オカレンス情報	行動(日本語)	文字列(日本語可)	例:ねぐらに滞在、摂食中、走っていた			○
25	オカレンス情報	生息環境	文字列(半角英数字)	例:oak savanna, pre-cordilleran steppe			○

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

次演題で説明します

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

5. 変換ツールにかけることで無くなるデータ不備

1)値に垂直タブ（Mac 版ファイルメーカーセル内改行など）等の制御コードが入っているケースがある。

解説：これは、トラブルの原因としては非常に頻度高く出てきますが、エクセルでは目視できないことが多く、始末が悪いです。
最新の変換ツール(データ変換ツール.xlsm)でデータクリーニングを行うと、制御コードが自動的に削除されます。変換ツールの最新版は、S-Net サイトの「ツール・辞書」データ変換ツール（http://science-net.kahaku.go.jp/app/page/tool_download.html##dataconv）からダウンロードしてください。
変換ツールを使用せずに見つける一つの方法としては、サクラエディタ（https://sakura-editor.github.io/download.html からダウンロードできる無料のエディタ）で「検索」メニューを選び、「正規表現」にチェックを入れて、条件を[^¥t¥r¥n[:print:]]とすることで検索が出来ます。なお、垂直タブの除去は、最新変換ツールのクリーニング機能で対応済みですが、過去のデータを扱う際には要注意です。可能な限り、消去し、代わりに「|」（半角パイプ；¥と同じキーにあることが多い）を使って、同一フィールド内部に質的に異なる情報が入っていることを示すようにするとよいでしょう。

なんだ、エクセル講座かよ

その通りです。でも、エクセルは優れたソフトです。簡単な機能を知るだけで、十分役に立ちます。

もっと上級者向けには・・・

Google Open Refine

https://www.gbif.org/tool/81727/open-refine-google-refine

https://www.slideshare.net/arosawa/open-refine-237216272

起こりやすいエラーの実例

1. 「自然史標本データサンプル_20200831_コメント非表示.xlsx」を御覧ください。これは、起こりやすいエラーをまとめたものです。
2. 2つのシートからなっており、「1−自然史標本データサンプル」のなかでエラーがあるセルは、「2−自然史標本データサンプル_答え」で赤くハイライトされています。各セルを選択すると、コメントが表示されます