第36回　自然史情報発信に関する研究会

S-Netデータ提供時のデータチェックのプロセス

2021.2.6.

国立科学博物館　標本資料センター
細矢　剛

背景：日本ノードからのデータ出版の流れ

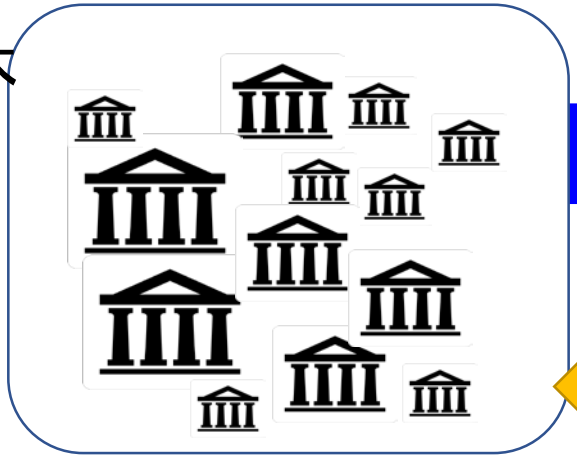
文献
観察情報
学術誌など

東大

遺伝研

GBIF

標本



103の研究機関・
自然史博物館・大学

国立科学博物館

英語

日本語

変換手数料
研修・実習・
情報交換

S-Net／JBIFサイト
／ジャパンサーチ

集める　から　使う　の時代へ

1. GBIFサイトは日本語化されてさらに便利に。
2. 利用例（お手本となる論文）
   - Science review→日本語版も出版
3. S-Netもリニューアル（2018年4月）され、便利に。

データを使う

よいデータを集める

ファセット検索だと間違いが見えやすい

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。

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

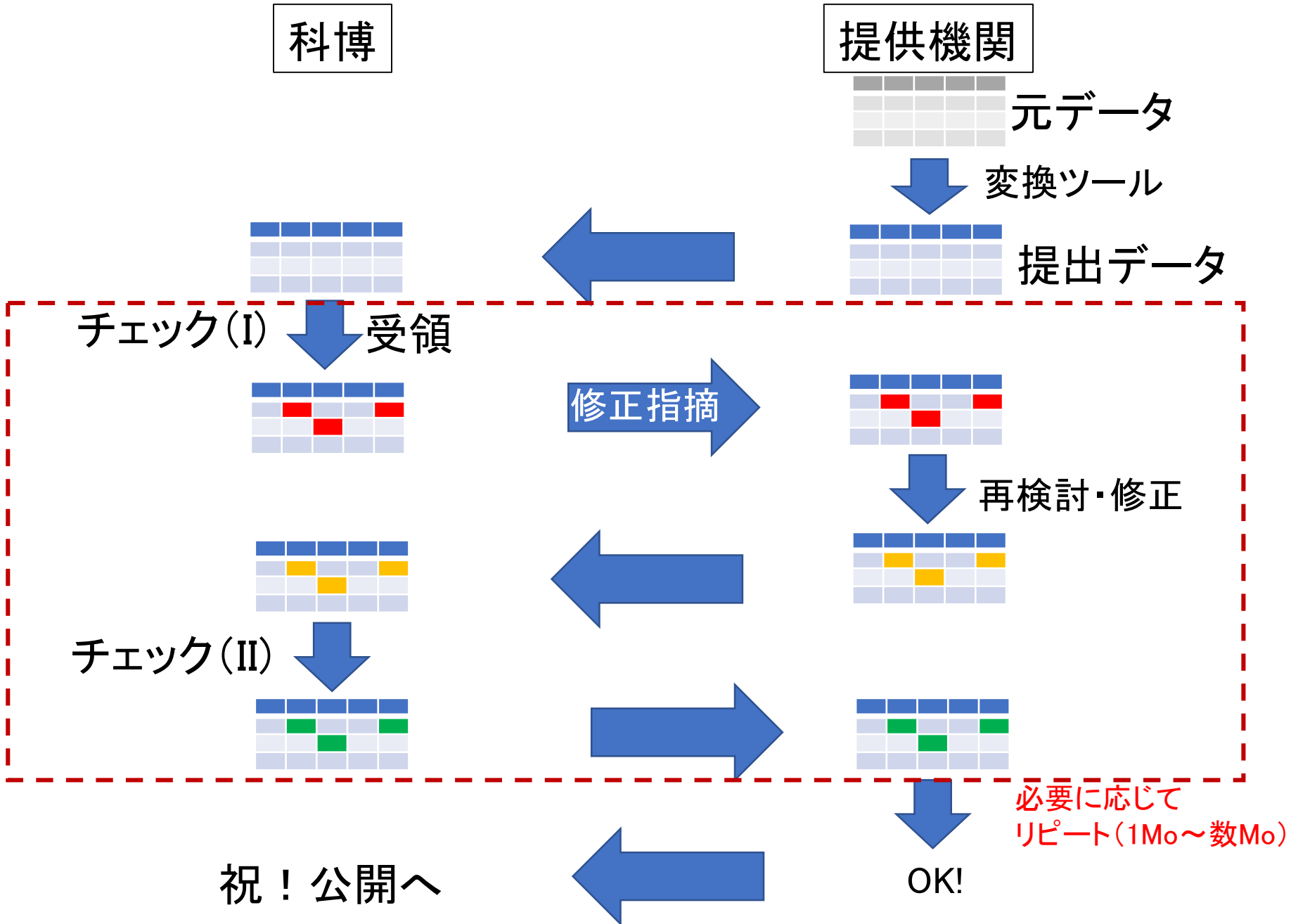
「予防は治療に勝る」

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。

本研究会のアウトライン

1. 諸注意（細矢）
2. S-Netデータ提供時のデータチェック過程【細矢】
3. 導入：データクリーニングとは（座学）【神保】
4. 初心者向け：エクセル（関数など）を使ったクリーニング【細矢】
5. 中級～上級者向け：エクセル以外のツールを使ったクリーニング【神保】

データ受領～公開までのプロセス



チェックには膨大な作業と時間がかかる

作業	1データセット	全DS（118）
変換ツール	0.5h	25d
内容チェック	0.5h	25d
レッドデータチェック	1.5h	60d
修正確認ファイル作成	1h	60d
修正案内作成	1h	60d
返答待ち期間	(30d)	
返答ファイルから最終データ作成	1.25h	35d
最終確認案内作成	1.5h	24d
返答待ち期間	(1wk)	
公開作業	2h	60d
Total（待ち時間除く）	9.25h	

※R1年度の118データセットをもとにした経験値。作業者は5hr/day。

ここを短縮することがすべての節約につながる

こんなデータはいやだっ!!　ワースト10-6

10.「データ登録日時」に値が入っている

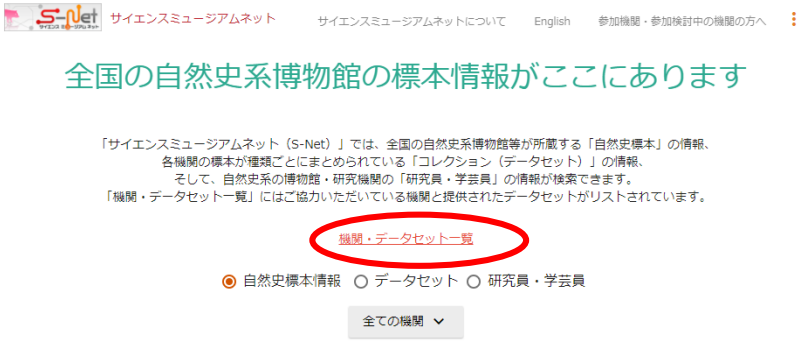
事務局使用欄です。

9.「最低海抜」と「最高海抜」逆（先頭値が空白）

念の為確認する場合もあります。

8.機関コードに誤り

過去の例を確認して、
一貫したコードを使用してください。



7.「綱」が「網」

気持ちは分かりますが・・・。

6.十進の緯度経度の値が“0”

緯度経度の0は「ない」ではなく大海原。

こんなデータはいやだっ!!　ワースト5-1

5.機関名（日本語）の誤り

例）科博、国立科学博物館、etc.。登録された機関データ参照。

4.学名の著者が文字化け

例）“G?nther”→”Günther“。ファイルの“文字コード”に注意

3.「記録年月日（始め）」「記録年月日（終わり）」が1900年以前

ホントの場合がないわけではありませんが、稀有です。

2.「地名公開レベル」が ”0” 以外

地名公開レベル」は”0”とします。レッドリストとは無関係に詳細地名非表示を希望される場合、「非公開情報に関する備考（日本語）」に”採集地保護のため詳細地名非表示“と記述し、非公開としたい地名情報を「備考２（非公開）」か「備考３（非公開）」に移動してください。

1.学名が文字化け

全角が入っている事が多い。

データの誤りワースト選

1. ●学名が文字化け
2. ●「地名公開レベル」が ”0” 以外
3. ●「記録年月日（始め）」「記録年月日（終わり）」が1900年以前
4. ●学名の著者が文字化け
5. ●機関名（日本語）の誤り
6. ●十進の緯度経度の値が“0”
7. ×「綱」が「網」
8. ●機関コードに誤り
9. ●「最低海抜」と「最高海抜」逆（先頭値が空白）
10. ●「データ登録日時」に値が入っている
11. ●「コレクションコード」誤り
12. ●「記録年月日（始め）」が未来日付
13. ●「GBIF公開フラグ」”1”とあるべきが”0”
14. ●「記録年月日（始め）」と「記録年月日（終わり）」逆
15. ●「メッシュコード」”544070**”の”**”
16. ●「最浅水深」と「最深水深」が逆

●文字化け；●コード入力に注意；●一貫性に注意；●範囲に注意　×論外