第36回自然史標本データ整備事業による標本情報の発信に関する研究会
2021.2.6　オンライン

データクリーニングとは？

国立科学博物館　動物研究部
神保　宇嗣

「データクリーニング」とは？

データを「きれいにする」ことで
品質を高め利用しやすくすること

「きれいにする」とは？

誤字・表記や形式の揺れを統一させること

形式的な表記揺れ

Lepidoptera Science 70: 99–148
Lepidoptera Science 70:99–148
Lepidoptera Science 70: 99 – 148
Lepidoptera Science 70: 99–148.
Lepidoptera Science, 70: 99–148
Lepidoptera Science 70: 99–148
Lepidptera Science 70: 99–148
Lepidoptera Science 70: 99-148
Lepidoptera Science 70: 99–148
Lepid. Sci. 70: 99-148
Lepid Sci 70: 99-148
Lep. Sci. 70: 99-148

形式的な表記揺れ

Lepidoptera Science 70: 99–148
Lepidoptera Science 70:99–148
Lepidoptera Science 70: 99 – 148
Lepidoptera Science 70: 99–148.
Lepidoptera Science, 70: 99–148
Lepidoptera Science 70: 99–148
Lepidptera Science 70: 99–148
Lepidoptera Science 70: 99-148
Lepidoptera Science 70: 99–148
Lepid. Sci. 70: 99-148
Lepid Sci 70: 99-148
Lep. Sci. 70: 99-148

- 記号の使い方
- 空白の入れ方
- 誤記
- 略称

等々…

ファセット表示で見る表記揺れ

国（日本語）

日本 (4,756)

日本国 (96)

もっと見る

「日本」「日本国」「Japan」
「中国」「中華人民共和国」

国（日本語）

台湾 (84) Japan (14) 中国 (13) ロシア (7) 韓国 (7) タイ (6) マレーシア (6) ミャンマー (5) ネパール (4) タイ王国 (3) アフガニスタン (2) アメリカ (2) イギリス (2) キルギス (2) スリランカ (2) 中華人民共和国 (2) インド (1) カナダ (1) フィリピン共和国 (1) ベトナム (1) ラオス (1)

都道府県（日本語）

岩手県 (585)

沖縄県 (450)

もっと見る

「東京」「東京都」
「兵庫」「兵庫県」

都道府県（日本語）

北海道 (391) 栃木県 (327) 茨城県 (315) 宮城県 (309) 鹿児島県 (259) 滋賀県 (198) 神奈川県 (197) 長崎県 (197) 兵庫県 (177) 愛知県 (165) 千葉県 (163) 埼玉県 (150) 群馬県 (135) 石川県 (124) 静岡県 (102) 富山県 (89) 東京 (83) 三重県 (70) 徳島県 (61) 東京都 (59) 山形県 (49) 大阪府 (40) 長野県 (32) 山梨県 (31) 福井県 (31) 福島県 (28) 青森県 (28) 千葉 (25) 新潟県 (25) 岐阜県 (22) 秋田県 (22) 南投県 (15) 山口県 (14) 奈良県 (11) 鳥取県 (10) 兵庫 (8) 四川省 (8) 神奈川 (7) カチン州 (5) 京都府 (5) 河南省 (5) 北京市 (4) 群馬 (4) 佐賀県 (3) 慶尚北道 (3) 高雄県 (3) ハワイ州 (2) バンコク (2) 和歌山県 (2) 基隆市 (2) 宮崎県 (2) 島根県 (2) 広島県 (2) 彰化県 (2) 愛媛県 (2) 新北市 (2) 福岡県 (2) 香港 (2) 香港特別行政区 (2) 高知県 (2) no data (1) パラワン州 (1) ブリティッシュコロンビア州 (1) ラムドン省 (1) 台中県 (1) 台湾 (1) 岡山県 (1) 広東省 (1) 新竹県 (1) 熊本県 (1) 西ベンガル州 (1) 香川県 (1) 鹿児島 (1)

S-Net 「モンシロチョウ」の検索結果

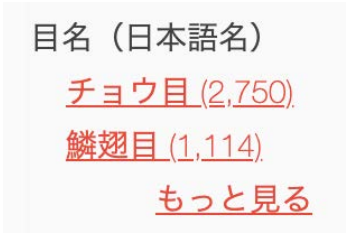
なぜ困る？

S-Netには、何カ国・地域のモンシロチョウの標本が登録されているのか？

- 同じ国・地域の表記が統一していなければ、正しい答えに行き着けない。検索もうまくいかない
- S-Netでできるだけ統一されていれば、利用者の手間を減らせ、検索時の漏れも減らせる

表記揺れあれこれ

- 形式的な違い　「日本」「日本国」
  「チョウ目」「チョウ」
- 正式名と通称　「中華人民共和国」「中国」
- 典拠の違い
  「チョウ目」「鱗翅目」「チョウ目（鱗翅目）」
  「チョウ目チョウ類」
- 誤記「チヨウ目」



S-Net 「モンシロチョウ」の検索結果

（参考）形式的以上の表現揺れ

- 台湾の扱いは？
  - 「国名コードの国際標準」に従って別にする(ISO-3166-1）
- 慣用の漢字目名と科学用語集のカタカナ目名

<複雑な「表記揺れ」パターン>

目名（日本語名）

齧歯目 (15,013)

げっ歯目 (422)

もっと見る

目名（日本語名）

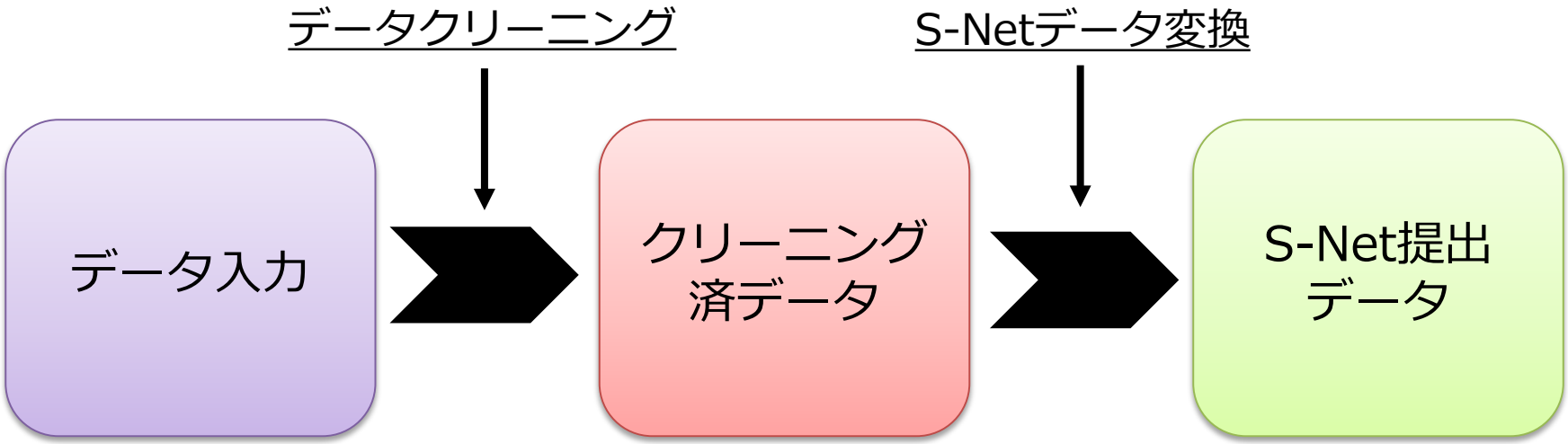
ネズミ目 (160)　齧歯目（ネズミ目） (54)　ゲッ歯目 (3)　ネズミ目（齧歯目） (3)　げっ歯 (1)

S-Net「Rodentia」（目の学名）での検索結果

S-Netでも「おすすめの入力方法」ガイドラインを整備する必要があり

いつクリーニングするのか？

- S-Net変換前（マスターデータ）作成時
- あとから見つかった修正が必要な箇所は、クリーニング済データに反映させたあと、さらにクリーニングを実施する。



どんなことをするのか？

- 形式の統一
  - 全角・半角
  - ひらがな・カタカナ、大文字・小文字
  - 余計な空白や記号の除去
- 表記の統一
  - 同じデータの書き方を統一する＝名寄せ
  - 対応するデータを同じように入れる
    （例：和名から学名を入力）

データクリーニングのツール

- 表計算ソフトウェア
  - Microsoft Excel, Google Spreadsheetなど
- 専用のプログラム・ウェブ上のツール
  - 事前整形支援ツール・変換ツール
  - レッドデータチェッカー
  - GBIFの学名チェッカー
- データクリーニングソフトウェア
  - Open Refine
- エディタ
- スクリプトによる自動化