データクレンジングとは

松野 渉（国立情報学研究所 研究データ基盤整備チーム）

データクレンジングとは

データベース（データの集合体）内のデータの品質を高める

- ✓ 誤記や標記揺れの修正
- ✓ 重複の削除

➡ “汚いデータ”の修正・削除

データクレンジングはスムーズで正確なデータ処理の前処理として必要不可欠！

“汚いデータ”（1）：不要なスペース・改行

状態	例
先頭にスペース	␣夏目漱石
末尾にスペース	夏目漱石␣
途中改行	夏目↵ 漱石

※利用しているアプリケーション・ソフトウェアによっては目視で発見するのが困難

“汚いデータ”(2)：文字種の揺れ

状態	例
ひらがな カタカナ	なつめそうせき ナツメソウセキ
全角 半角	１８６７ 1867
大文字 小文字 先頭のみ大文字	HEMINGWAY hemingway Hemingway

“汚いデータ”(3)：表記の揺れ

状態	例
用語の揺れ	国立大学法人 東京大学 東京大学 東大
	コンピューター コンピュータ
誤字	夏目漱易

“汚いデータ”(4)：その他

状態	例
言語コードのルール	jp jpn
データ入力のルール	[夏目] [漱石] 夏目␣漱石 夏目,漱石
その他	Info.doi/10.11501/8695992 et al. [円城塔,伊藤計劃]

なぜ発生するのか？

- システムの民主化
  - ✓データ作成は基本的に分業制
  - ✓データ入力の自由度が上がるとデータは汚くなる

- 移行時の変換で瑕疵が見える化
  - ✓ システムごとにデータの入力規則やスキーマは変わる
  - ✓ データを別のシステムで扱った時にそれまで問題にならなかった部分が大きな傷に

なぜ必要なのか？

- システムのため
  - ✓“汚いデータ”はシステムを止める
    - ex(1):一行＝一件の書誌として処理するプログラムに読み込ませるデータの氏名の中にうっかり改行が紛れ込んでいると…？
    - ex(2):ISSNを「8ケタの数字(or英字)」として処理するプログラムにうっかりハイフンが紛れ込んだデータを処理させると…？

なぜ必要なのか？

- ユーザのため
  - ✓“汚いデータ”は検索性や利便性を下げる
    - ex(1):姓・名に分けて検索する時、姓のフィールドにフルネームが入っていると…？
    - ex(2):特定の言語に絞って資料検索をしたい時に、ルールを無視した言語コードを持つデータが含まれていると…？

※いずれも実際のシステムや業務上の操作の際に回避が可能な事象だが、その分、**コストが上がる**

おわりに

「Web API完全に理解した！」
…だけでは十分ではないかもしれない

➡ 世の中のデータセットは意外と”汚い”
（残念ながら）図書館のデータも例外ではない…

データクレンジングで正確・高速なデータ処理を！