データのクレンジングツール

◆無償利用可能なデータクレンジングツールもありますが、日本語への対応は不十分です。

- データクレンジングを行うための無償利用が可能な英語版ソフトウェアとしてOpenRefineが挙げられます。

【出所】OpenRefine http://openrefine.org/

- 日本語は英語に比べても、漢字表記や送り仮名の違い等の表記揺れが多く、標準化（名寄せ）は、より重要です。

住所表記・会社表記のデータ形式の標準化（名寄せ）例

住所の表記揺れ	
霞ヶ関1丁目1番地	霞が関1丁目1番地
霞ヶ関1丁目1	霞が関1丁目1
霞ヶ関1-1	霞が関1-1

住所表記の標準化の取り組み例

- 標準記載法の策定と公表
- 標準記載名データベースの公表
- 表記揺れの統一エンジンの公開

住所表記の標準化例

霞が関１丁目１
1-1 Kasumigaseki
〒100-0013
緯度: 35.675836 経度: 139.754734

- 住所の表記においては、「ヶ」と「が」の混在、丁番地の表記が不統一となっている事だけでも、一貫性が損なわれてしまいます。

ソニー株式会社の表記揺れ				
ソニー株式会社	Sony株式会社	SONY株式会社	Ｓｏｎｙ株式会社	ＳＯＮＹ株式会社
ソニー（株）	Sony（株）	SONY（株）	Ｓｏｎｙ（株）	ＳＯＮＹ（株）
ソニー(株)	Sony(株)	SONY(株)	Ｓｏｎｙ(株)	ＳＯＮＹ(株)
ソニー㈱	Sony㈱	SONY㈱	Ｓｏｎｙ㈱	ＳＯＮＹ㈱

会社表記の標準化例

ソニー株式会社
Sony Corporation
東証一部 6758（電気機器）
設立年月日　1946年5月7日

- 日本人が見れば、上記16種の企業表記は同一の企業だと分かりますが、文字列が異なるためデータ集計時には異なる企業として扱われてしまいます。
- 法人マイナンバー（法人番号）を利用すれば、正式な企業名を確認することができ、同じ企業名が複数ある場合でも企業を特定することができます。

- 日本語のデータクレンジングは、個々のケースに合わせてExcelやプログラミングで行っているケースが多くなっています。

□ 講座3-2では、Excelを用いて日本語の表記揺れの統一を含むデータクレンジングの実習を行います。