東京都

東京データプラット
フォームデータ整備事業事業報告会

データ整備

令和5年03月23日
株式会社自動処理

TDPF
Tokyo Data Platform

データ公開作業の流れ

• データ変換作業は『事前準備』、『データ構造の統一』、『データの精度の向上』の3つの手順で実施します。

手順1 事前準備

- 調査
- データフォーマット検討
  - データ項目検討
    ex) 病院名、住所、営業日
  - データ表記検討
    ex) 日付
    2021-09-27
    ex) 電話番号
    (03)9999-9999
- メタデータ検討
  - ライセンス
  - 問い合わせ先
  - 公開日
  - 更新頻度
    など

手順2 データ構造の統一

- ケース1 Excelデータ
  - 表データへ変換
    縦横多段クロス集計
    単純クロス集計
    単純表形式
    複数ファイル
    神Excel
    など

手順3 データの精度の向上

- データを合わせる
  - 診療所、住所、休日
    ↓
  - 病院名、住所、営業日
- クレンジング
  - 2021年09月27
    ↓ISO8601形式に変換
  - 20210927
  - 03-9999-9999
    ↓RFC3966形式に変換
  - (03)9999-9999

手順1 事前準備

手順1 事前準備　–調査

事前準備では、**自分たちが公開したい情報と、データ利用者が扱いやすいデータを見比べて**、
情報公開した際に、**データ利用者にとって利用しやすいデータフォーマットを調査・検討**する事になります。

手順1 事前準備

1.調査

2.メタデータ検討
- ライセンス
- 問い合わせ先
- 公開日
- 更新頻度

など

3.データフォーマット検討
1.データ項目検討
ex) 病院名、住所、営業日
2.データ表記検討
ex) 日付
2021-09-27
ex) 電話番号
(03)9999-9999

- まず**公開対象データを集め、**

病院(R2.12.1 開設届)

No	施設名称	施設所在地
1	医療法人社団実理会　東京国際大堀病院	三鷹市下連雀４丁目８番４０号
2	医療法人	
3	徳原病院	
4	医療法人	
5	公益財団	
6	医療法人	
7	杏林大学	
8	医療法人	

診療所(R2.12.1 開設届)

No	施設名称	施設所在地
1	医療法人社団健晶会　下川整形外科	三鷹市井の頭１－２４－１４
2	浜田耳鼻咽喉科	三鷹市井の頭１－３０－１３
3	笹本医院	三鷹市井の頭１－３１－２２
4	医療法人社団耀祥会　慶真整形外科	三鷹市井の頭２－１－１７
5	三鷹台ヒルズクリニック	三鷹市井の頭２－１－１７
6	松川内科クリニック	三鷹市井の頭２－１－１７
7	三鷹台眼科	三鷹市井の頭２－１－１７
8	医療法人社団研道会　高水クリニック	三鷹市井の頭２－１４－２
9	みこしばクリニック	三鷹市井の頭２－１９－２５
10	医療法人社団慈昭会　石井医院	三鷹市井の頭２－３２－３７
11	牟礼の里駅前クリニック	三鷹市井の頭２－７－９
12	藤林医院	三鷹市井の頭３－１２－１５
13	医療法人社団岡庭医院	三鷹市井の頭３－２１－１６
14	医療法人社団慈司会　若林医院	三鷹市井の頭４－１６－１０
15	武蔵野みどり診療所	三鷹市井の頭５－７－３６
16	ヨシコクリニック	三鷹市井口１－２２－２４
17	医療法人社団美々会　斉藤皮膚科	三鷹市井口２－１３－２６
18	医療法人社団加藤整形外科医院	三鷹市井口２－３－１
19	医療法人社団　かえでこどもクリニック	三鷹市井口３－６－１６

データフォーマット検討では、**公開データに適したデータ項目や表現方法は何かを検討する**事になります。

手順1 事前準備

1.調査

2.データフォーマット検討
1.データ項目検討
ex) 病院名、住所、営業日
2.データ表記検討
ex) 日付
2021-09-27
ex) 電話番号
(03)9999-9999

3.メタデータ検討
- ライセンス
- 問い合わせ先
- 公開日
- 更新頻度

など

- ここでポイントなのは

機械判読可能なデータ

にする事

ものすごくシンプルに表現すると

・縦横の表で表せる※
・同一ファイル形式である
・同じデータは同じデータ項目を持つ
・同じデータ項目は同じ名前で表記する
・同一データは同一の表記で表す
※上級者は必ずしも表でなくても良い。

という事になります。

手順1 事前準備

1.調査

2.データフォーマット検討
1.データ項目検討
ex) 病院名、住所、営業日
2.データ表記検討
ex) 日付
2021-09-27
ex) 電話番号
(03)9999-9999

3.メタデータ検討
- ライセンス
- 問い合わせ先
- 公開日
- 更新頻度

など

例えば

オープンデータを元に
複数市区町村の町丁目別人口データを比較する

という仕事をしなければいけない状況になったとします。

表データをコピペすれば、あとはExcelで
集計すれば何とかなる出来ると考えてませんか？

少し人口統計データのサンプルを少し挙げます。

人口データの項目としては
場所、性別、世帯、年齢、人口
項目はほとんど同じで、
比較的シンプルなデータを対象としています。

目黒区人口統計データ

- 単純表

131105_population_20110101.csv

町丁目表記　地域名

5歳刻み

	1	2	3	4	5	6	7	8	9	10
1	都道府県コード又は市区町村コード	地域コード	都道府県名	市区町村名	調査年月日	地域名	総人口	男性	女性	0-4歳の男性
2	131105	111	東京都	目黒区	2011-01-01	駒場一丁目	3698	1824	1874	70
3	131105	112	東京都	目黒区	2011-01-01	駒場二丁目	868	420	448	27
4	131105	113	東京都	目黒区	2011-01-01	駒場三丁目	742	350	392	12
5	131105	114	東京都	目黒区	2011-01-01	駒場四丁目	1239	576	663	37
6	131105	121	東京都	目黒区	2011-01-01	青葉台一丁目	2340	1078	1262	32
7	131105	122	東京都	目黒区	2011-01-01	青葉台二丁目	881	390	491	18
8	131105	123	東京都	目黒区	2011-01-01	青葉台三丁目	2689	1291	1398	55
9	131105	124	東京都	目黒区	2011-01-01	青葉台四丁目	1541	723	818	22
10	131105	131	東京都	目黒区	2011-01-01	東山一丁目	3376	1617	1759	59
11	131105	132	東京都	目黒区	2011-01-01	東山二丁目	6762	3234	3528	242
12	131105	133	東京都	目黒区	2011-01-01	東山三丁目	2868	1362	1506	71
13	131105	141	東京都	目黒区	2011-01-01	大橋一丁目	1538	738	800	32
14	131105	142	東京都	目黒区	2011-01-01	大橋二丁目	3829	1880	1949	70
15	131105	151	東京都	目黒区	2011-01-01	上目黒一丁目	1682	759	923	38
16	131105	152	東京都	目黒区	2011-01-01	上目黒二丁目	4170	1969	2201	67
17	131105	153	東京都	目黒区	2011-01-01	上目黒三丁目	4790	2219	2571	68
18	131105	154	東京都	目黒区	2011-01-01	上目黒四丁目	3725	1796	1929	56
19	131105	155	東京都	目黒区	2011-01-01	上目黒五丁目	3710	1777	1933	85
20	131105	161	東京都	目黒区	2011-01-01	中目黒一丁目	2606	1217	1389	38
21	131105	162	東京都	目黒区	2011-01-01	中目黒二丁目	2266	1157	1109	63
22	131105	163	東京都	目黒区	2011-01-01	中目黒三丁目	2501	1190	1311	62
23	131105	164	東京都	目黒区	2011-01-01	中目黒四丁目	2822	1367	1455	69
24	131105	165	東京都	目黒区	2011-01-01	中目黒五丁目	3139	1516	1623	65
25	131105	171	東京都	目黒区	2011-01-01	三田一丁目	2032	890	1142	33
26	131105	172	東京都	目黒区	2011-01-01	三田二丁目	3361	1533	1828	59
27	131105	181	東京都	目黒区	2011-01-01	目黒一丁目	3405	1548	1857	55
28	131105	182	東京都	目黒区	2011-01-01	目黒二丁目	2417	1095	1322	56
29	131105	183	東京都	目黒区	2011-01-01	目黒三丁目	2620	1249	1371	40
30	131105	184	東京都	目黒区	2011-01-01	目黒四丁目	3280	1577	1703	59
31	131105	191	東京都	目黒区	2011-01-01	下目黒一丁目	1239	548	691	20
32	131105	192	東京都	目黒区	2011-01-01	下目黒二丁目	4163	2022	2141	73
33	131105	193	東京都	目黒区	2011-01-01	下目黒三丁目	3417	1650	1767	47
34	131105	194	東京都	目黒区	2011-01-01	下目黒四丁目	2713	1290	1423	41
35	131105	195	東京都	目黒区	2011-01-01	下目黒五丁目	3224	1557	1667	89
36	131105	196	東京都	目黒区	2011-01-01	下目黒六丁目	2174	1038	1136	61
37	131105	201	東京都	目黒区	2011-01-01	中町一丁目	4637	2263	2374	84

品川区人口統計データ

- クロス集計

住民基本台帳による町丁目別および男女・年齢別人口

平成23年１月1日現在
「住民基本台帳」登録人口

ご注意：「印刷」をする場合、「すべて」を選択しますと大部数になりますのでご注意ください。
作成：品川区地域振興事業部地域活動課統計係

町丁目表記　丁目名

1歳刻み

「指定統計」等コード（五十音順）	品川区行政コード	丁目名	世帯数	人口総数	0歳 男	0歳 女	1歳 男	1歳 女	2歳 男	2歳 女	3歳 男	3歳 女	4歳 男	4歳 女
		総数	191,930	351,350	1,493	1,462	1,515	1,468	1,448	1,343	1,336	1,302	1,335	1,269
		品川地区	33,342	63,077	311	286	313	312	313	271	294	247	260	258
0060000100	1101	北品川1丁目	1,745	2,900	3	9	8	9	7	5	10	8	6	4
0060000200	1102	北品川2丁目	1,681	3,006	7	9	7	11	10	15	22	11	10	11
0060000300	1103	北品川3丁目	1,674	3,166	27	21	23	20	17	13	16	15	8	19
0060000400	1104	北品川4丁目	653	1,349	8	8	10	8	10	12	9	5	9	7
0060000500	1105	北品川5丁目	2,266	4,490	25	18	19	20	24	28	20	14	15	23
0060000600	1106	北品川6丁目	316	579	3	1	0	2	1	3	2	0	0	2
0180000100	1201	東品川1丁目	2,063	3,846	12	11	6	13	14	7	9	11	8	11
0180000200	1202	東品川2丁目	926	1,464	3	7	4	4	3	5	3	1	4	3
0180000300	1203	東品川3丁目	4,761	10,168	66	62	57	70	85	53	65	51	73	64
0180000400	1204	東品川4丁目	1,818	3,469	33	23	25	36	33	20	21	18	20	13
0180000500	1205	東品川5丁目	14	14	0	0	0	0	0	0	0	0	0	0
0240000100	1301	南品川1丁目	727	1,381	4	2	7	5	2	6	4	3	4	1
0240000200	1302	南品川2丁目	1,052	1,877	5	3	4	4	6	5	7	7	2	3
0240000300	1303	南品川3丁目	876	1,648	8	6	13	18	8	6	8	7	9	11
0240000400	1304	南品川4丁目	1,882	3,281	13	12	11	8	6	9	11	12	11	6
0240000500	1305	南品川5丁目	2,828	5,269	25	19	27	24	23	17	13	27	14	21
0240000600	1306	南品川6丁目	1,709	2,702										
0130000100	1401	西品川1丁目	2,068	4,055	1									
0130000200	1402	西品川2丁目	2,082	3,749										
0130000300	1403	西品川3丁目	1,558	2,821										
0210000100	1501	広町1丁目	37	52										
0210000200	1502	広町2丁目	606	1,791	3									

- 縦横の表で表せる　・・・ NG

目黒区データと比較して

- 同一ファイル形式である　・・・ NG
- 同じデータは同じデータ項目を持つ　・・・ NG
- 同じデータ項目は同じ名前で表記する　・・・ NG
- 同一データは同一の表記で表す　・・・ NG

港区人口データ

年齢区分無し

- 複雑なクロス集計

区分		世帯数	人口			前月比増減	
			男	女	計	世帯数	人口
港区総数			99,438	113,124	212,562	20	113
	外国人	—	10,374	9,187	19,561	—	65
	合計	—	109,812	122,311	232,123	—	178
芝地区総合支所管内	日本人	20,888	16,589	17,998	34,587	22	27
	外国人	—	1,551	1,240	2,791	—	10
	合計	—	18,140	19,238	37,378	—	37
麻布地区総合支所管内	日本人	27,638	21,556	24,974	46,530	△ 7	19
	外国人	—	3,980	3,577	7,557	—	74
	合計	—	25,536	28,551	54,087	—	93
赤坂地区総合支所管内	日本人	17,570	14,415	17,093	31,508	13	14
	外国人	—	1,742	1,577	3,319	—	17
	合計						
高輪地区総合支所管内	日本人	29,					
	外国人						

外国人の集計あり

町丁目表記　支所管内

・縦横の表で表せる　・・・ NG

目黒区データと比較して

・同一ファイル形式である　・・・ NG

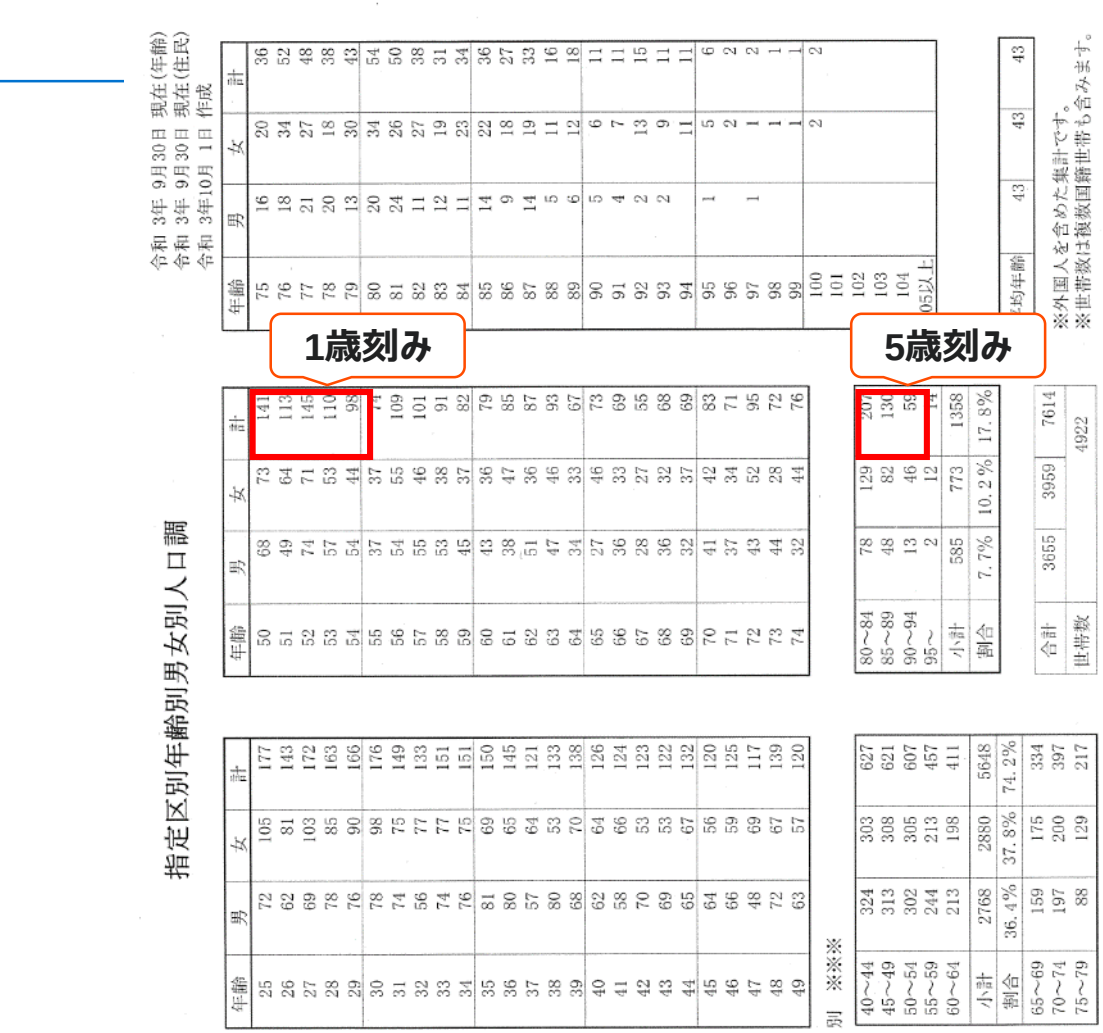
・同じデータは同じデータ項目を持つ　・・・ NG

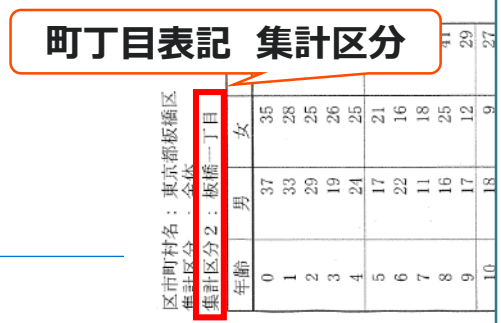
・同じデータ項目は同じ名前で表記する・・・ NG

・同一データは同一の表記で表す　・・・ NG

板橋区人口データ

- 横向きPDF





・縦横の表で表せる　・・・ NG
目黒区データと比較して
・同一ファイル形式である　・・・ NG
・同じデータは同じデータ項目を持つ　・・・ OK
・同じデータ項目は同じ名前で表記する　・・・ NG
・同一データは同一の表記で表す　・・・ NG

手順1 事前準備　–データフォーマット検討

追記

シンプルなデータですら、バラバラなフォーマットで公開されている。

事前準備

1.調査

2.データフォーマット検討
1.データ項目検討
ex) 病院名、住所、営業日
2.データ表記検討
ex) 日付
2021-09-27
ex) 電話番号
(03)9999-9999

3.メタデータ検討
- ライセンス
- 問い合わせ先
- 公開日
- 更新頻度

など

- **使い勝手の良いデータフォーマットが必要**です。

1.データ項目検討
自治体標準オープンデータセットの
『地域・年齢別人口』データを元に

町丁目表記は『地域名』と表記
『5歳刻み』データとして検討

2.データ表記検討
ex) 日付
2021-09-27
ex) 電話番号
(03)9999-9999

政府相互運用性フレームワーク(GIF)
440 コアデータパーツ
https://github.com/JDA-DM/GIF/tree/main/440_コアデータパーツ

自治体標準オープンデータセット

都道府県コード 又は 市区町村コード
地域コード
都道府県名
市区町村名
調査年月日
地域名
総人口
男性
女性
0-4歳の男性
0-4歳の女性
世帯数
備考

手順1 事前準備　–メタデータ検討

最後にそのデータをどう使えばよいのか、告知為、メタデータの検討を行います。

事前準備

1.調査

2.データフォーマット検討
  1.データ項目検討
    ex) 病院名、住所、営業日
  2.データ表記検討
    ex) 日付
    2021-09-27
    ex) 電話番号
     (03)9999-9999

3.メタデータ検討
- ライセンス
- 問い合わせ先
- 公開日
- 更新頻度
  など

- またデータを使う際に、

どう利用していいのか？　→　ライセンス
データに疑問があった場合には？　→　問い合わせ先
いつ更新されたデータか？　→　作成日
次の更新はいつか？　→　更新頻度

こういった**メタデータも整備も重要**です。

- 先に整備しておくと問い合わせが減り、
利用者が悩まずに利用する事が出来ます。

最初からメタデータを整理しておくことで、
問い合わせを減らし、
データ利用者が安心して使う事が出来るようになります。

手順1 事前準備　–メタデータ検討

- こういったメタデータは現在、デジタル庁内でメタデータの標準化の検討が進んでいます。
  データ項目に従って、整理すると過不足がなく情報を整理できます。

事前準備

1.調査

2.データフォーマット検討
  1.データ項目検討
    ex) 病院名、住所、営業日
  2.データ表記検討
    ex) 日付
    2021-09-27
    ex) 電話番号
    (03)9999-9999

3.メタデータ検討
- ライセンス
- 問い合わせ先
- 公開日
- 更新頻度

など

データセット	
管理ID	CA00001-DST001
タイトル	MJ文字情報一覧表
サブタイトル	
バージョン	Ver.006.01
説明	各文字に関するコード、読み、字母、画数等をまとめた情報。
キーワード	フォント、ヨミガナ、画数
対象地域	全国
対象期間	
分類	全ての業務
提供者	文字情報技術促進協議会
作成者	情報処理推進機構
連絡先情報	組織名：一般社団法人 文字情報技術促進協議会 メールアドレス：info@moji.or.jp フォームURL：https://moji.or.jp/about/contact/
タイプ	Strict Open XML
来歴情報	2020年10月に、情報処理推進機構から文字情報技術促進協議会に信託譲渡
品質評価	正確性、網羅性
品質測定結果	公務で使うのに十分な品質
公開日	2011/10/26
最終更新日	2020/8/26
更新頻度	不定期
言語	ja
公開範囲	公開
公開条件	
準拠する標準	
関連ドキュメント	
ランディングページ	

参考）メタデータルールと利用イメージの検討
https://cio.go.jp/dp2021_07

手順2　データ構造の統一

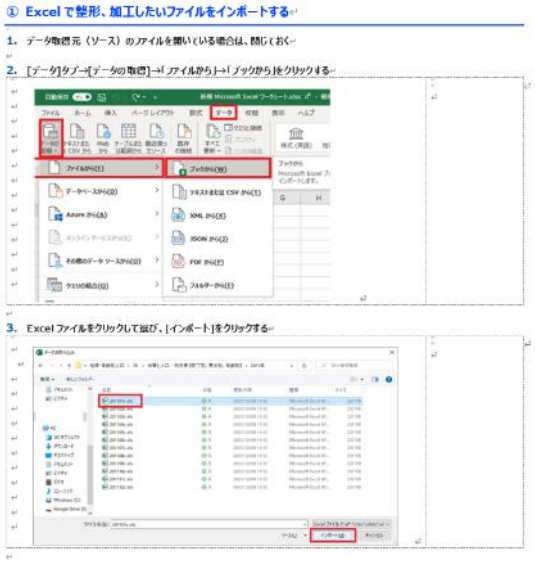
手順2　データ構造の統一

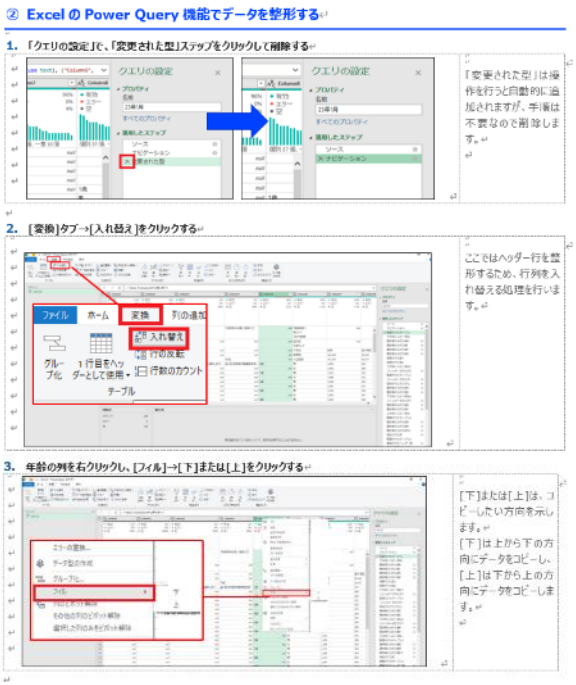
- 目指すべきデータの構造が定まったら、作業対象のそのデータの項目に合わせていく作業を実施します。

データ構造の統一

- **ケース1 Excelデータ**
  - 表データへ変換
    縦横多段クロス集計
    単純クロス集計
    単純表形式
    複数ファイル
    神Excel
    など

- 今回の事業の中で扱ったデータはほとんどがExcelで扱えるデータでした。
- Excelデータについては、PowerQueryを利用する事で、大量のデータを一気に変換する事が出来ます。





手順2　データ構造の統一

追記

- 目指すべきデータの構造が定まったら、作業対象のそのデータの項目に合わせていく作業を実施します。

データ構造の統一

- **ケース1 Excelデータ**
  - 表データへ変換
    縦横多段クロス集計
    単純クロス集計
    単純表形式
    複数ファイル
    神Excel
    など

- 専門知識があれば、データベースに取り込み、SQLにてデータ項目をマッピングして、データベースに取り込むことも可能です。

手順2 データ構造の統一　におけるデータ変換困難ケース１

- データ形式の問題-**データ形式が**違う

紙・PDF → **OCRもしくは専門業者に依頼**

RJkr540

行政区別・年齢別人口調べ

作成日付：令和 3年 4月 1日

1 頁

令和 3年 3月31日　現在

注：（）内の数字は、外国人住民の内数を示す。

行政区	年齢	0～4	5～9	10～14	15～19	20～24	25～29	30～34	35～39	40～44	45～49	50～54	55～59
000111	男	31	23	31	32	56	129	150	148	133	124	99	81
台東１丁目		(1)	(1)	(1)	(2)	(13)	(15)	(17)	(10)	(8)	(5)	(6)	(2)
	女	29	21	31	30	57	91	95	68	66	95	77	46
		(0)	(0)	(2)	(4)	(10)	(12)	(11)	(5)	(6)	(6)	(7)	(2)
	計	60	44	62	62	113	220	245	216	199	219	176	127
		(1)	(1)	(3)	(6)	(23)	(27)	(28)	(15)	(14)	(11)	(13)	(4)

年齢	60～64	65～69	70～74	75～79	80～84	85～89	90～94	95～99	100～104	105～109	110～	合計
男	52	38	49	24	30	16	9	2	1	0	0	1,258
	(1)	(4)	(0)	(0)	(0)	(1)	(0)	(0)	(0)	(0)	(0)	(87)
女	43	35	42	46	43	27	14	4	0	0	0	960
	(3)	(1)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(69)
計	95	73	91	70	73	43	23	6	1	0	0	2,218
	(4)	(5)	(0)	(0)	(0)	(1)	(0)	(0)	(0)	(0)	(0)	(156)

HTML → **専門業者に依頼**

2021年（令和3年）

令和3年５月1日現在

区分		世帯数	人口			前月比増減	
			男	女	計	世帯数	人口
港区総数	日本人	133,912	112,293	128,131	240,424	59	△ 106
	外国人	9,701	9,526	8,686	18,212	△ 7	△ 79
	複数国籍世帯	3,259	—	—	—	△ 13	—
	合計	146,872	121,819	136,817	258,636	39	△ 185
芝地区総合支所管内	日本人	24,142	18,462	20,509	38,971	13	△ 27
	外国人	1,607	1,451	1,197	2,648	28	15
	複数国籍世帯	459	—	—	—	△ 7	—
	合計	26,208	19,913	21,706	41,619	34	△ 12

CSV → **ダウンロード**

年月日［西暦］	地区	町丁目	世帯数	人口男［人］	人口女［人］	人口合計［人］
20210101	芝地区総合支所管内	芝一丁目	1288	968	1054	2022
20210101	芝地区総合支所管内	芝二丁目	2218	1654	1874	3528
20210101	芝地区総合支所管内	芝三丁目	2194	1705	1881	3586
20210101	芝地区総合支所管内	芝四丁目	1397	1089	1220	2309
20210101	芝地区総合支所管内	芝五丁目	1959	1155	1631	2786
20210101	芝地区総合支所管内	海岸一丁目	1197	1014	980	1994
20210101	芝地区総合支所管内	東新橋一丁目	782	677	781	1458
20210101	芝地区総合支所管内	東新橋二丁目	406	336	224	560

手順2 データ構造の統一　におけるデータ変換困難ケース２

- データ形式の問題-**データ構造**が違う　→　PowerQuery、**もしくはプログラム**

多段階クロス集計構造

2021年（令和3年）

令和3年５月1日現在

区分		世帯数	人口			前月比増減	
			男	女	計	世帯数	人口
港区総数	日本人	133,912	112,293	128,131	240,424	59	△ 106
	外国人	9,701	9,526	8,686	18,212	△ 7	△ 79
	複数国籍世帯	3,259	—	—	—	△ 13	—
	合計	146,872	121,819	136,817	258,636	39	△ 185
芝地区総合支所管内	日本人	24,142	18,462	20,509	38,971	13	△ 27
	外国人	1,607	1,451	1,197	2,648	28	15
	複数国籍世帯	459	—	—	—	△ 7	—
	合計	26,208	19,913	21,706	41,619	34	△ 12

クロス集計構造

年月日［西暦］	地区	町丁目	世帯数	人口男［人］	人口女［人］	人口合計［人］
20210101	芝地区総合支所管内	芝一丁目	1288	968	1054	2022
20210101	芝地区総合支所管内	芝二丁目	2218	1654	1874	3528
20210101	芝地区総合支所管内	芝三丁目	2194	1705	1881	3586
20210101	芝地区総合支所管内	芝四丁目	1397	1089	1220	2309
20210101	芝地区総合支所管内	芝五丁目	1959	1155	1631	2786
20210101	芝地区総合支所管内	海岸一丁目	1197	1014	980	1994
20210101	芝地区総合支所管内	東新橋一丁目	782	677	781	1458
20210101	芝地区総合支所管内	東新橋二丁目	406	336	224	560

手順2 データ構造の統一　におけるデータ変換困難ケース３

- データ形式の問題-**項目の有無**が違う　→　**現課に問い合わせ**

港区　　各総合支所管内別の町丁目別人口･世帯　年齢別情報なし
墨田区町丁別･年齢別人口　　年齢別情報あり

港区

各年１月１日現在の人口・世帯数（昭和29年～令和3年）

各年1月1日現在（ただし昭和29年は12月1日現在）

	1954年（昭和29年）	1955年（昭和30年）	1956年（昭和31年）	1957年（昭和32年）	1958年（昭和33年）
総人口	249,343	249,472	252,443	252,079	254,250
男	127,007	127,420	128,857	129,252	131,134
女	122,336	122,052	123,586	122,827	123,116
世帯数	63,901	64,343	65,843	65,974	67,575

墨田区

墨田区人口集計表（町丁別・年齢別）　両　国

（令和３年４月１日　午前０時現在）

	両国一丁目 男	女	計	両国二丁目 男	女	計	両国三丁目 男	女	計	両国四丁目 男	女	計	計 男	女	合計
0歳	9	7	16	6	8	14	8	11	19	12	6	18	35	32	67
1歳	9	11	20	5	6	11	6	5	11	12	8	20	32	30	62
2歳	3	5	8	6	9	15	11	3	14	9	5	14	29	22	51
3歳	7	3	10	6	5	11	6	6	12	8	10	18	27	24	51
4歳	3	6	9	8	6	14	3	5	8	8	12	20	22	29	51
5歳	7	3	10	5	10	15	5	4	9	8	3	11	25	20	45
6歳	2	2	4	5	8	13	5	2	7	12	5	17	24	17	41

手順2 データ構造の統一　におけるデータ変換困難ケース４

- データ形式の問題-**情報提供頻度**が違う　→　**集計、もしくは現課に問い合わせ**

墨田区町丁別年齢別人口 四半期
文京区町丁別世帯･人口　月次

※不定期なケースも存在する

文京区

町丁別世帯・人口（住民基本台帳）（毎月1日現在）

令和3年

- 3年1月(Excelファイル; 20KB)、 3年1月(PDFファイル; 353KB)
- 3年2月(Excelファイル; 20KB)、 3年2月(PDFファイル; 353KB)
- 3年3月(Excelファイル; 20KB)、 3年3月(PDFファイル; 353KB)
- 3年4月(Excelファイル; 20KB)、 3年4月(PDFファイル; 353KB)
- 3年5月(Excelファイル; 20KB)、 3年5月(PDFファイル; 353KB)
- 3年6月(Excelファイル; 33KB)、 3年6月(PDFファイル; 354KB)

墨田区

令和2年度

- 墨田区人口集計表（町丁別・年齢別）4月分（PDF：452KB）
- 墨田区人口集計表（町丁別・年齢別）7月分（PDF：452KB）
- 墨田区人口集計表（町丁別・年齢別）10月分（PDF：446KB）
- 墨田区人口集計表（町丁別・年齢別）1月分（PDF：453KB）

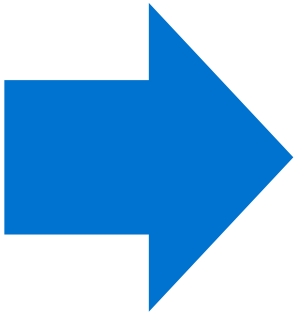
手順2 データ構造の統一　におけるデータ変換困難ケース５

- データ形式の問題-**項目の表記ゆれ**　→　**専門業者に依頼**

介護事業所における
営業時間項目

営業時間
窓口受付時間
受付営業時間
サービス提供時間
通いサービス提供時間
訪問可能時間帯
24時間対応
休業日
窓口定休日
受付休業日
通いサービスの定休日

マッピング難度が高い



推奨データセット
介護事業所営業時間項目

利用可能曜日
利用可能曜日特記事項

手順2 データ構造の統一　におけるデータ変換困難ケース６

- データ形式の問題-**経年でデータ構造が変わる**　→　PowerQuery、もしくは専門業者に依頼

2019年（平成31年）3月

平成31年3月1日現在

年齢	総数	男	女
総数	258,093	121,704	136,389
0歳～4歳	14,651	7,544	7,107
0歳	2,876	1,501	1,375
1歳	2,958	1,509	1,449
2歳	2,987	1,505	1,482
3歳	3,025	1,601	1,424
4歳	2,805	1,428	1,377
5歳～9歳	11,629	5,877	5,752
5歳	2,513	1,256	1,257
6歳	2,540	1,261	1,279
7歳	2,289	1,169	1,120
8歳	2,192	1,114	1,078
9歳	2,095	1,077	1,018
90歳～94歳	2,113	512	1,601
90歳	607	149	458
91歳	473	114	359
92歳	394	103	291
93歳	340	66	274
94歳	299	80	219
95歳～99歳	690	117	573
100歳以上	116	17	99

2019年（平成31年）4月

平成31年4月1日現在

年齢	総数	男	女
総数	258,696	121,989	136,707
0歳～4歳	14,516	7,473	7,043
0歳	2,802	1,457	1,345
1歳	2,954	1,519	1,435
2歳	2,976	1,498	1,478
3歳	2,968	1,567	1,401
4歳	2,816	1,432	1,384
5歳～9歳	11,622	5,866	5,756
5歳	2,533	1,263	1,270
6歳	2,514	1,240	1,274
7歳	2,268	1,154	1,114
8歳	2,216	1,132	1,084
9歳	2,091	1,077	1,014
95歳～99歳	690	118	572
95歳	242	44	198
96歳	162	21	141
97歳	131	29	102
98歳	97	12	85
99歳	58	12	46
100歳以上	127	19	108
100歳	48	9	39
101歳	31	4	27
102歳	14	4	10
103歳以上	34	2	32

2019年以降は100歳以上も1歳刻みで表示されている。

手順2 データ構造の統一　におけるデータ変換困難ケース7

- データ形式の問題-**情報の粒度**が違う　→　**現課に問い合わせ、もしくは専門業者に依頼**

令和元年度特定健実施機関一覧表

札幌市清田区

北海道医師会

医療機関コード	医療機関名	郵便番号	住所	
0110317658	社会医療法人　札幌清田病院	004-0831	札幌市清田区	真栄1条1丁目1−1
0110513892	医療法人北武会　美しが丘病院	004-0839	札幌市清田区	真栄61−1
0110513173	さかもと内科消化器クリニック	004-0841	札幌市清田区	清田1条4丁目4−30
0110513181	医療法人社団　サン内科外科医院	004-0842	札幌市清田区	清田2条1丁目8−1
0110317328	医療法人社団　鈴木内科医院	004-0844	札幌市清田区	清田4条2丁目10−25
0110316387	社会医療法人蘭友会　札幌里塚病院	004-0811	札幌市清田区	美しが丘1条6丁目1−5
0110511854	医療法人社団　美しが丘いとう内科	004-0813	札幌市清田区	美しが丘3条2丁目3−20
0110513066	医療法人社団群仁会　保坂内科クリニック	004-0814	札幌市清田区	美しが丘4条5丁目3−15
0110319928	医療法人　札幌平岡病院	004-0872	札幌市清田区	平岡2条1丁目15−20
0110319639	医療法人社団　ひらおか内科胃腸科	004-0876	札幌市清田区	平岡6条3丁目10−1
0110511243	医療法人社団　平岡公園整形外科・消化器科クリニック	004-0882	札幌市清田区	平岡公園東1丁目12−12
0110512589	ふじた内科循環器クリニック	004-0882	札幌市清田区	平岡公園東5丁目12−10メディカルビル平岡公園
0110314887	北海道医療生活協同組合　札幌緑愛病院	004-0861	札幌市清田区	北野1条1丁目6−30
0110316213	医療法人社団エス・エス・ジェイ　札幌整形循環器病院	004-0861	札幌市清田区	北野1条2丁目11−30
0110512274	医療法人　ほし内科消化器科クリニック	004-0863	札幌市清田区	北野3条2丁目13−57
0110317484	医療法人社団　小野内科医院	004-0865	札幌市清田区	北野5条5丁目15−27
0110515467	医療法人社団越仁会　北野循環器クリニック	004-0867	札幌市清田区	北野7条2丁目12−17
0110515178	みき内科消化器きたのクリニック	004-0867	札幌市清田区	北野7条5丁目12−20
0110515350	医療法人徳洲会　札幌南徳洲会病院	004-0801	札幌市清田区	里塚1条2丁目20−1

病院と介護施設でレコード単位が違う

手順2 データ構造の統一　におけるデータ変換困難ケース８

追記

- データ形式の問題-**集計単位**が政府標準と違う　→　**集計、もしくは現課に問い合わせ**
  - ― 総務省統計局　年齢コード　1歳、5歳刻み
  - ― 独立行政法人労働政策研究・研修機構　年齢階級コード　5歳刻み

"DateTime"	"Camera"	"入退場者"	"性別"	"年代"	"Count"
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"20-29"	3
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"30-39"	8
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"40-49"	10
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"50-59"	6
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"60-"	2
"2021-12-01 00:00:00"	" "	"Input"	"性別不明通行者"	"年齢不明"	77
"2021-12-01 00:00:00"	" "	"Input"	"女性"	"10-19"	1
"2021-12-01 00:00:00"	" "	"Input"	"女性"	"30-39"	7
"2021-12-01 00:00:00"	" "	"Input"	"女性"	"40-49"	12
"2021-12-01 00:00:00"	" "	"Input"	"女性"	"50-59"	11
"2021-12-01 00:00:00"	" "	"Input"	"女性"	"60-"	3
"2021-12-01 00:00:00"	" "	"Output"	"男性"	"20-29"	1
"2021-12-01 00:00:00"	" "	"Output"	"性別不明通行者"	"年齢不明"	82
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"20-29"	1
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"30-39"	5
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"40-49"	4
"2021-12-01 00:00:00"	" "	"Input"	"男性"	"50-59"	2

合わせないといけないわけではないが、10歳刻みの年齢区分コードがなかった為、国勢調査など、他の統計情報と組み合わせる際に、やや不便となる。

手順2 データ構造の統一　におけるデータ変換困難ケース９

- データ形式の問題-**画像形式の地図データ**　→　**専門業者に依頼**

一度画像になった地図から
緯度経度に変換するのは非常に難しい。

一度画像になった地図からポリゴンを起こすことは非常に難しい。
※例外的に市区町村、一部の町目については、国土地理院の情報からポリゴンを作成可能。

手順2 データ構造の統一　におけるデータ変換困難ケース１０

- データ形式の問題-**データ構造外**の情報の存在　→　**目視確認**

	A	B	C	D	E	F	G	H	I
46	44	スマイル矯正歯科	三鷹市下連雀３－４３－３２	松川ビル２階	矯	0422-71-6911			
47	45	医療法人社団林矯正歯科	三鷹市下連雀３－４４－１７	エルヴェ三鷹２階	矯	0422-76-8881			
48	46	医療法人社団永寿会　ふれあいデンタルクリニック	三鷹市下連雀３－４４－５	希望ビル２階	歯	0422-29-8242			
49	47	連雀の杜歯科	三鷹市下連雀４－１１－１５		歯科、歯科口腔外科、矯正歯科、小児歯科	0422-24-8822	以外		
50	48	えみた歯科・矯正歯科	三鷹市下連雀４－１６－１２	三鷹第一ビル１Ｆ	歯・歯外・小歯・矯	0422-26-5508	以外		
51	49	医療法人社団光成会　金子歯科医院	三鷹市下連雀４－１６－１８	長島ビル１階	歯・小歯	0422-41-4488			
52	50	鈴木デンタルクリニック	三鷹市下連雀４－１６－２	パークハウス三鷹セントレ１０１	歯・小歯・歯外	0422-70-1839	以外		
53	51	下連雀デンタルクリニック	三鷹市下連雀４－１８－１４	サンクマーレ三鷹下連雀１階	歯・小歯・歯外・矯	0422-24-7850	以外		
54	52	角田歯科医院	三鷹市下連雀４－９－１		歯・小歯・歯外	0422-48-7575			
55	53	三鷹ながえ歯科クリニック	三鷹市下連雀５－１－５	三鷹井上店舗１階	歯科、小児歯科、矯正歯科、歯科口腔外科	0422-26-6111	以外		
56	54	あきやま歯科医院	三鷹市下連雀５－８－１	シティコート下連雀１Ｆ－Ｂ	歯・小歯・矯・歯外	0422-72-0220			
57	55	さくらい歯科医院	三鷹市下連雀６－１－３		歯・小歯	0422-43-6303			
58	56	吉祥寺通り歯科クリニック	三鷹市下連雀６－２－１６	アトラス吉祥寺１階Ｄ区画	歯・小歯・矯・歯外	0422-24-8515			
59	57	鳥山歯科医院	三鷹市下連雀６－６－２７		歯	0422-43-1324			
60	58	永江歯科医院	三鷹市下連雀７－１０－１５	セドール三鷹２０１	歯・小歯	0422-42-7331			
61	59	ふくだ歯科	三鷹市下連雀７－１４－２８	ベルムミタカ１Ｆ－２	歯・小歯・歯外・矯	0422-40-6480			
62	60	~~若田歯科医院~~	~~三鷹市下連雀７－１６－１３~~		~~歯・小歯・矯~~	~~0422-47-8850~~	閉院（R3.2	4歯科医師会確認）	
63	61	医療法人社団まほろば　あかり歯科	三鷹市下連雀９－２－２０	エスポワール三鷹１０３・１０５号室	歯・小歯、矯正歯科、歯科口腔外科	0422-48-8524	以外		
64	62	三鷹駅前デンタルオフィス	三鷹市上連雀１－１－５	三鷹ロイヤルハイツ１１３号	歯・小歯・歯外・矯	0422-38-8641	以外		
65	63	榎石歯科医院	三鷹市上連雀１－８－２０	アサヒ三鷹ビル１０１	歯・小歯・矯	0422-54-7834			

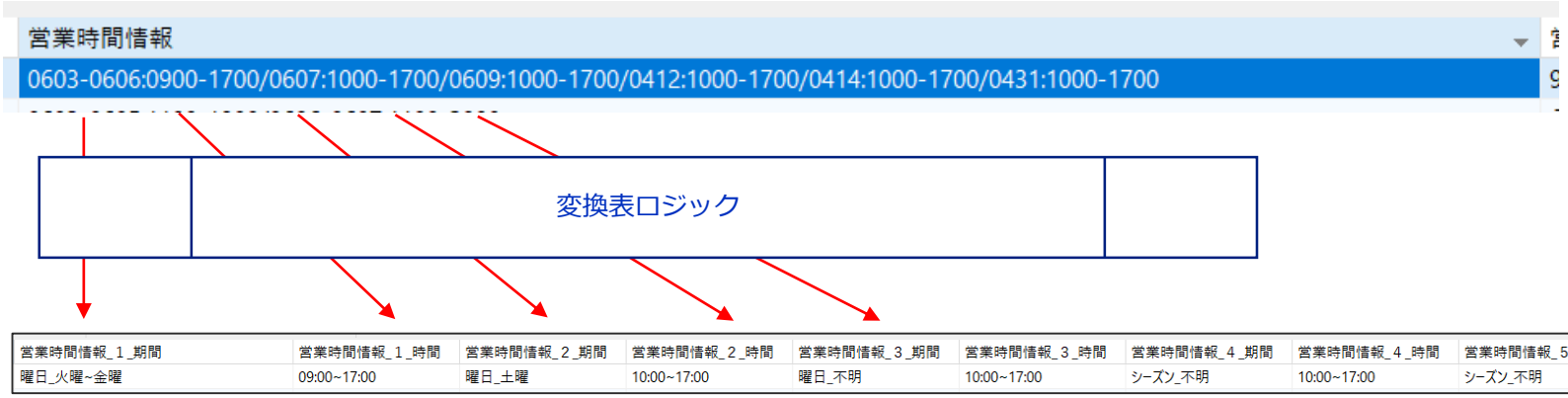
枠外表記有り

手順2 データ構造の統一　におけるデータ変換困難ケース１１

追記

- データ形式の問題-**1つの項目に複数データが定義**されている　→　**専門業者に依頼**

スラッシュ区切りで、1項目に複数値が設定さている。
営業時間情報1/営業時間情報２…営業時間情報５



営業時間情報_1_期間	営業時間情報_1_時間	営業時間情報_2_期間	営業時間情報_2_時間	営業時間情報_3_期間	営業時間情報_3_時間	営業時間情報_4_期間	営業時間情報_4_時間	営業時間情報_5_期間	営業時間情報_5_時間
曜日_火曜~金曜	09:00~17:00	曜日_土曜	10:00~17:00	曜日_不明	10:00~17:00	シーズン_不明	10:00~17:00	シーズン_不明	10:00~17:00

手順2 データ構造の統一　におけるデータ変換困難ケース１２

追記

- 自治体標準オープンデータセットの中に定義の仕方が存在しないケース　→　TDPFに相談
  ー キーとバリューを保存する場合の定義の方法が明確ではない

05 子育て施設

32	利用可能曜日	注7	文化財の利用可能曜日を記載。※記載方法について、「データ項目特記事項」シートの【共通ルール】を参照。	文字列	火水木金土日

17-2 小中学校通学区域情報

13	通学区域の住所		文字列で通学区域に含まれる住所を ";"（半角セミコロン）区切りで記載。市区町村名より後の住所を記載し、市区町村を越えた通学区域の場合は、市町村名（政令指定都市の場合は区名）から記載する。「○○」町の一部」「○○団地の○号棟まで」といった特殊な取り扱いも記載。	文字列	小石川1丁目;小石川2丁目;大滝の一部;北区山田町1丁目;北区山田町2丁目;北区山田町3丁目

金曜営業開始	金曜営業終了	土曜定休	土曜営業開始	土曜営業終了
7:00	23:00	0	8:00	18:00

出来ればこう定義したいが、どこにもそういった記述はない

金:7:00-23:00;土:8:00-18:00

手順２データ構造における民間データと自治体データの整備における特徴

追記

- 今年度は複数の民間企業様のデータを預かって整備させて頂きました。
  - ― 民間企業のデータはデータ整備に関する仕様書が定義されており、システムを利用して、継続的に整備されており、利活用前提でデータが整備されている事から、データに関しては仕様が一様であり、資料化もされている事から、データ構造を確認する作業については、比較的作業の進め方について見通しが立てやすい状況でした。
  - ― 但し、民間企業にデータ提供を依頼する場合、基本的には自社が開発しやすいデータであることも多そうでした。
    自治体が提供するデータはデジタル庁が、国際規格に基づき自治体オープンデータセットとして整えられていますので、そういった違いがありました。

民間企業データは各項目のデータがきっちりそろっている。

豊洲	ﾄﾖｽ	13	江東区	135-0061	江東区豊洲４－６－１		03-3533-2
牡丹	ﾎﾞﾀﾝ	13	江東区	135-0046	東京都江東区牡丹３－２０－		03-3641-2
森下	ﾓﾘｼﾀ	13	江東区	135-0004	東京都江東区森下３－１－１		03-3631-4
梅田	ｳﾒﾀﾞ	13	足立区	123-0851	東京都足立区梅田７丁目１６		03-3840-2
東陽	ﾄｳﾖｳ	13	江東区	135-0016	東京都江東区東陽３－２１－		03-3644-4
本郷	ﾎﾝｺﾞｳ	13	文京区	113-0033	東京都文京区本郷３－２３－		03-3814-5
善福寺	ｾﾞﾝﾌﾟｸｼﾞ	13	杉並区	167-0041	東京都杉並区善福寺１－２－		03-3399-0
高田馬場	ﾀｶﾀﾞﾉﾊﾞﾊﾞ	13	新宿区	169-0075	東京都新宿区高田馬場４－３		03-3368-4
練馬３丁目	ﾈﾘﾏ3ﾁﾖｳﾒ	13	練馬区	176-0001	東京都練馬区練馬３－１－１		03-3992-4
南篠崎	ﾐﾅﾐｼﾉｻﾞｷ	13	江戸川区	133-0065	東京都江戸川区南篠崎町５－	2	03-3679-6
中野中央	ﾅｶﾉﾁｭｳｵｳ	13	中野区	164-0011	東京都中野区中央４－２２－		03-3382-7
神宮前	ｼﾞﾝｸﾞｳﾏｴ	13	渋谷区	150-0001	渋谷区神宮前４－８－１		03-3401-1

"DateTime"	"Camera"	"入退場者"	"性別"	"年代"	"Count"
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"20-29"	3
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"30-39"	8
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"40-49"	10
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"50-59"	6
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"60-"	2
"2021-12-01 00:00:00"	"N	"Input"	"性別不明通行者"	"年齢不明"	77
"2021-12-01 00:00:00"	"N	"Input"	"女性"	"10-19"	1
"2021-12-01 00:00:00"	"N	"Input"	"女性"	"30-39"	7
"2021-12-01 00:00:00"	"N	"Input"	"女性"	"40-49"	12
"2021-12-01 00:00:00"	"N	"Input"	"女性"	"50-59"	11
"2021-12-01 00:00:00"	"N	"Input"	"女性"	"60-"	3
"2021-12-01 00:00:00"	"N	"Output"	"男性"	"20-29"	1
"2021-12-01 00:00:00"	"N	"Output"	"性別不明通行者"	"年齢不明"	82
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"20-29"	1
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"30-39"	5
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"40-49"	4
"2021-12-01 00:00:00"	"N	"Input"	"男性"	"50-59"	2

手順3 データの精度の向上

手順3 データの精度の向上

- データ構造が統一されたらデータを統一的な表記に変換・必要なデータを補完します。
  データのミスや文字化けを自動修正する事をクレンジングと言います。
- **機械可読性が高くともクレンジング作業は必ず発生します。**
  入力時点の確認が大事なので、正しく入力される仕組みを作りましょう。

手順3 データの精度の向上

- データ構造を合わせる
  - 診療所、住所、休日
    ↓
  - 病院名、住所、営業日

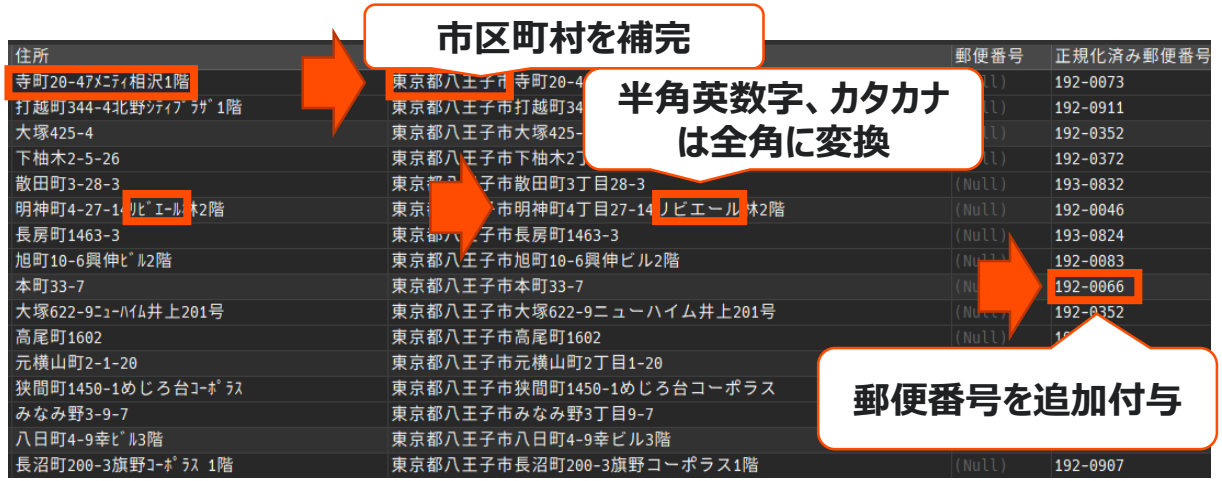
- クレンジング
  - 2021年09月27
    ↓ISO8601形式に変換
  - 20210927

  - 03-9999-9999
    ↓RFC3966形式に変換
  - (03)9999-9999

APIを利用して

- 文字列クレンジング
  - 英数字を半角
  - 記号を統一
  - カタカナを全角
- 住所クレンジング
  - 市区町村コード
  - 郵便番号
  - 緯度経度

  など 統一的な表記に変換



住所		郵便番号	正規化済み郵便番号
寺町20-4ｱﾒﾆﾃｨ相沢1階	東京都八王子市寺町20-4	(Null)	192-0073
打越町344-4北野ｼﾃｨﾌﾟﾗｻﾞ1階	東京都八王子市打越町34	(Null)	192-0911
大塚425-4	東京都八王子市大塚425-	(Null)	192-0352
下柚木2-5-26	東京都八王子市下柚木2丁	(Null)	192-0372
散田町3-28-3	東京都八王子市散田町3丁目28-3	(Null)	193-0832
明神町4-27-14ﾉﾋﾞｴｰﾙ 4階	東京都八王子市明神町4丁目27-14ノビエール 4階	(Null)	192-0046
長房町1463-3	東京都八王子市長房町1463-3	(Null)	193-0824
旭町10-6興伸ﾋﾞﾙ2階	東京都八王子市旭町10-6興伸ビル2階	(Null)	192-0083
本町33-7	東京都八王子市本町33-7	(Null)	192-0066
大塚622-9ﾆｭｰﾊｲﾑ井上201号	東京都八王子市大塚622-9ニューハイム井上201号	(Null)	192-0352
高尾町1602	東京都八王子市高尾町1602	(Null)	
元横山町2-1-20	東京都八王子市元横山町2丁目1-20		
狭間町1450-1めじろ台ｺｰﾎﾟﾗｽ	東京都八王子市狭間町1450-1めじろ台コーポラス		
みなみ野3-9-7	東京都八王子市みなみ野3丁目9-7		
八日町4-9幸ﾋﾞﾙ3階	東京都八王子市八日町4-9幸ビル3階		
長沼町200-3旗野ｺｰﾎﾟﾗｽ 1階	東京都八王子市長沼町200-3旗野コーポラス1階	(Null)	192-0907

手順3 データの精度の向上

追記

- データの精度向上は目視作業もしくは専門業者に依頼する事によって作業可能です。
  1000件以上になるようであれば、専門業者に依頼する事をお勧めします。

手順3 データの精度の向上

- データ構造を合わせる
  - 診療所、住所、休日
    ↓
  - 病院名、住所、営業日

- クレンジング
  - 2021年09月27
    ↓ISO8601形式に変換
  - 20210927

  - 03-9999-9999
    ↓RFC3966形式に変換
  - (03)9999-9999

APIを利用して、

- 電話番号
  - 住所から市外局番を補完
  - 統一的な表記に変換

市外局番を補完

電話番号	正規化済み電話番号
626-418	042-626-4188
648-3566	042-648-3566
677-0480	042-677-0480
679-1188	042-679-1188
661-0249	042-661-0249
645-3116	042-645-3116
661-6500	042-661-6500
644-3662	042-644-3662
623-2052	042-623-2052
676-8288	042-676-8288
661-0128	042-661-0128
642-1827	042-642-1827
664-5862	042-664-5862

- 三鷹市中原一丁は03
  それ以外は0422
- 八王子市は042
  ※2006/3/5までは0426

追記

- データ形式の問題-**文字の変換**

対応可能ケース

アルファベット(半角文字に統一)

スポーツオアシス　ＡＢＣＤＥＦ → スポーツオアシス　ABCDEF
ｘｘ-ｘｘｘ代官山 → xx-xxx代官山

カナ(全角文字に統一)

株式会社ｼﾞﾄﾞｳｼｮﾘ → 株式会社ジドウショリ
株式会社ﾄｳｷｮｳﾄ → 株式会社トウキョウト

数字(半角に統一)

株式会社サンプル　９Ｆ → 株式会社サンプル　9F
日本著作権協会　４Ｆ　カップ → 日本著作権協会　4F　カップ

無駄な余白の表記統一(全角空有白一つに統一)

abcde タカギ薬局原宿店 → abcde　タカギ薬局原宿店
学校法人IT学園　　新プラザ棟 → 学校法人IT学園　新プラザ棟

法人格表示揺れ統一

（株）自動処理 → 株式会社自動処理
かぶ）自動処理 → 株式会社自動処理

ハイフンや記号の異体字の統一

- → —
- → —

※上記全て同じ文字に見えますが、別の文字です。

追記

• データ形式の問題-**人物名の変換**

対応可能ケース

- **姓、名分割名前形式を、名前１項目に統合**
  田中　太郎　　→　田中太郎

対応困難ケース

- **正しい漢字が分からない。**

高木祐介という名前について、
髙木祐介が正しかったとしても
自動変換は出来ない。

- **ヨミガナの自動付与**

名前は利用する感じに制限はあるものの、名前の読みは自由につけて構わない事になっていますので、概ね問題ないものの正確に１００%付与する事は出来ない。

- **姓名分割の分割場所がわからない。**

田中太郎→　田中　太郎
（田中　　苗字ランキング 4位）
田中太郎→　田　　中太郎
（田　苗字ランキング 8,074位）

追記

• データ形式の問題-**事業所・法人名の変換**

対応可能ケース

- **法人格の統一**

（株）自動処理　→　株式会社自動処理
かぶ）自動処理　→　株式会社自動処理

対応困難ケース

- **複数情報が含まれる**

医療法人社団サンプル会 サンプルクリニック
↓
医療法人社団サンプル会　法人
サンプルクリニック　が事業所名

株式会社サンプル屋新宿店２ＦＪＲ口（公開空地）
「ふくしまプライド。」フェア
↓
株式会社サンプル屋新宿店が店舗名
２Ｆ場所の階数
ＪＲ口場所の情報
「ふくしまプライド。」フェア　イベント名

- **複数パターンある省略**

医）　→　医療法人、医療法人社団、
医療法人財団のどれか
山大　→　山口大学か山形大学のどちらか

• データ形式の問題-**電話番号の変換**

対応可能ケース

- **自動変換可能例**
  - **03−3543−021**
  - **033543021**
  - **(03)3543−021**
  - **033543021**
- **国際番号番号変換**
  - **+81−3−3543−021**
  - **+81(3)3543−021**

対応困難ケース

- **市外局番が欠落**
- 35430211
- **不正な文字が混入**
- 084－926－0139　**日曜日･祝日は担当者携帯電話**
- **複数電話番号が混入**
- 0974-75-2124、**もしくは、**0974-42-2270

- データ形式の問題-**住所の変換　その１**

対応可能ケース

- 自動変換可能例
  - 住所形式違い
    - 東京都A区東○○３丁目１番１号
      αビル･４３階･２０号
    - 東京都A区東○○３丁目１番１号
      サンシャイン６０ 4320号
    - 東京都A区東○○３−１−１
      サンシャイン６０ 4320号
    - 東京都A区東○○
      サンシャイン６０（４３階）20号
    - 東京都A区東○○
      サンシャイン６０ Ｆ４３ 20号
  - 異体字を統一する
    - 宮城県塩釜市字杉ノ入裏xxx−xxx
    - 宮城県塩竈市字杉ノ入裏xxx−xxx
    - 宮城県塩竈市杉ノ入裏xxx−xxx
  - 住所から郵便番号（90%前後)付与
  - 住所から緯度経度（70%前後)付与
  - 市区町村合併の対応

対応困難ケース

- 東京都B区A地区一丁目1番1号
  （一丁目の一がハイフンになっている）
- 兵庫県加古川市上荘町井ノ口XXX−X
- 兵庫県加古川市上荘町井野口XXX−X
- 兵庫県加古川市上荘町井乃口XXX−X
  （複数の書き方がある文字列）
- 法人登記に存在する平成以前の旧住所
  東京市小石川区久堅町XX番地
  （東京市小石川区は現在の東京都文京区）

- データ形式の問題-**住所の変換　その１**

対応困難ケース

別の文字が利用されている
兵庫県加古川市上荘町井ノ口XXX－X
（口の文字がカタカナ）
北海道北見市留辺蘂町旭公園XX-X
（蘂の文字が書き間違いで本来は蘂）
茨城県つくば市白井XXXX-XX
（白の文字が書き間違いで本来は臼）

対応困難ケース

- **文字が欠けている**
  南区御幸笛田７丁目XX-XX
  （熊本県熊本市が欠けている）
  静岡県浜松市浜北区貴布XXXX
  （貴布祢の祢が欠けている）

- **不正な文字が入力されている**
  〒305-0005　**茨城県**00000**つくば市天久保**XX-X
  （郵便番号や意味のない数字が含まれている）

- **住所に含まれる番地以降の建物名、方書の分離**
  愛知県名古屋市A**区**B**町**1**丁目**3-8**西一ビル**301

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題-データの**表記ゆれ**

令和元年度特定健実施機関一覧表

札幌市清田区　　19

北海道医師会

医療機関コード	医療機関名	郵便番号	住所		電話番号	実施体制			
						貧血	心電図	眼底	血清クレアチニン検査
0110317658	社会医療法人　札幌清田病院	004-0831	札幌市清田区	真栄1条1丁目1－1	011-883-6111	○	○	△	○
0110513892	医療法人北武会　美しが丘病院	004-0839	札幌市清田区	真栄61－1	011-883-8881	○	○	△	○
0110513173	さかもと内科消化器クリニック	004-0841	札幌市清田区						
0110513181	医療法人社団　サン内科外科医院	004-0842	札幌市清田区						
0110317328	医療法人社団　鈴木内科医院	004-0844	札幌市清田区						
0110316387	社会医療法人蘭友会　札幌里塚病院	004-0811	札幌市清田区						
0110511854	医療法人社団　美しが丘いとう内科	004-0813	札幌市清田区						
0110513066	医療法人社団群仁会　保坂内科クリニック	004-0814	札幌市清田区						
0110319928	医療法人　札幌平岡病院	004-0872	札幌市清田区						
0110319639	医療法人社団　ひらおか内科胃腸科	004-0876	札幌市清田区						
0110511243	医療法人社団　平岡公園整形外科・消化器科クリニック	004-0882	札幌市清田区						
0110512589	ふじた内科循環器クリニック	004-0882	札幌市清田区						
0110314887	北海道医療生活協同組合　札幌緑愛病院	004-0861	札幌市清田区						

鈴木内科リハビリセンター

事業所の概要　事業所の特色　事業所の詳細　運営状況　その他

介護サービスの種類	通所リハビリテーション
所在地	〒004-0844　札幌市清田区清田4条2丁目10-25　地図を開く
連絡先	Tel：011-882-5608／Fax：011-885-2720　ホームページを開く

法人情報　所在地等　従業者　サービス内容　利用料等

● 2．介護サービス（予防を含む）を

表記ゆれ

事業所の名称、所在地及び電話番号その他の連絡先

事業所の名称	（ふりがな）	すずきないかりはびりせんたー
	鈴木内科リハビリセンター	
事業所の所在地	〒004-0844	市区町村コード　札幌市清田区
	（都道府県から番地まで）	札幌市清田区清田4条2丁目10-25
	（建物名・部屋番号等）	

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題-**データの意味とデータが適切に分離されていない**

IMSグループ	医療法人社団明芳会	イムス板橋リハビリテーション病院	訪問リハビリテーション事業所
グループ名	法人名	事業所名	サービス名業

マルマツ薬局　　九段店

参考）店舗名がないケースが混在している

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題- １つの項目に**複数のデータが格納**されている

記入日：2021年03月10日

介護サービスの種類	認知症対応型共同生活介護
所在地	〒053-0816　北海道苫小牧市日吉町1丁目2番23号 地図を開く
連絡先	Tel：0144-78-2180（／Fax：0144-75-5228 ホームページを開く

法人情報 | 所在地等 | 従業者 | サービス内容 | 利用料等

● 2．介護サービス（予防を含む）を提供し、又は提供しようとする事業所に関する事項

事業所の名称、所在地及び電話番号その他の連絡先			
事業所の名称	(ふりがな)	ふれあいのさとぐるーぷほーむやまぼうし	
	ふれあいの里グループホーム山法師		
事業所の所在地	〒053-0816	市区町村コード	苫小牧市
	（都道府県から番地まで）	北海道苫小牧市日吉町1丁目2番23号	
	（建物名・部屋番号等）		
事業所の連絡先	電話番号	0144-78-2180（1F）　0144-78-2181（2F）	
	FAX番号	0144-75-5228	
	ホームページ	あり	
介護保険事業所番号	0193600343		
事業所の管理者の氏名及び職名	氏名	山川　喜代美	

複数電話番号

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題-**入力項目誤り**

千葉県柏市今谷上町51-2

事業所の概要 | 事業所の特色 | 事業所の詳細 | 運営状況 | その他

記入日：2018年11月14日

介護サービスの種類	居宅介護支援
所在地	〒277-0074　千葉県柏市今谷上町51-2　リアンレーヴ南柏 地図を開く
連絡先	Tel：04-7138-5345／Fax：04-7172-2055 ホームページを開く

法人情報 | 所在地等 | 従業者 | サービス内容 | 利用料等

● 2．介護サービスを提供し、又は提供しようとする事業

ふりながなに漢字表記

事業所の名称、所在地及び電話番号その他の連絡先				
事業所の名称	(ふりがな)	木下の介護　南柏		
	千葉県柏市今谷上町51-2			
事業所の所在地	〒277-0074		市区町村コード	柏市
	(都道府県から番地まで)	千葉県柏市今谷上町51-2		
	(建物名・部屋番号等)	リアンレーヴ南柏		
	電話番号	04-7138-5345		

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題-**打消し線**データ

15	武蔵野みどり診療所	三鷹市井の頭５－７－３６	ＳＡＫＵＲＡ１階
16	ヨシコクリニック	三鷹市井口１－２２－２４	
17	医療法人社団美々会　斉藤皮膚科	三鷹市井口２－１３－２６	１階
18	医療法人社団加藤整形外科医院	三鷹市井口２－３－１	１階
19	医療法人社団　かえでこどもクリニック	三鷹市井口３－６－１６	アップルかえで通りビル　１Ｆ－Ａ
20	井の頭公園前ヒフ科	三鷹市下連雀１－１２－５	
21	医療法人社団鎌田医院	三鷹市下連雀１－３０－１２	
22	医療法人社団東京清心会　吉祥寺通り花岡クリニック	三鷹市下連雀１－９－２４	１階
23	花岡心療内科クリニック	三鷹市下連雀１－９－２４	
24	下田医院	三鷹市下連雀２－１８－１	
25	佐竹耳鼻咽喉科気管食道内科医院	三鷹市下連雀３－１４－２８	
26	松崎眼科クリニック	三鷹市下連雀３－１５－１	
27	サンクリニック三鷹	三鷹市下連雀３－１５－１８	~~大沼ビル~~　ＫＡＩＳＥＲ三鷹フロント４階
28	のぞみクリニック三鷹	三鷹市下連雀３－１７－１９	パレスマンション３０２号

打消し線データ

手順3 データの精度の向上　–データクレンジングケース

追記

- データ精度の問題-名寄せをする為の**ID（番号）を判別する情報が存在しない**
  ― 本事業の中で名寄せを行った為、こういったケースで問題が発生する事となった。

法人番号	商号又は名称	所在地	変更履歴情報等
6011501002115	シンニホンコウウン 新日本交通株式会社	東京都北区浮間2丁目24番13号	履歴等
9011501015898	第十日本交通株式会社	東京都北区浮間5丁目4番51号	履歴等 閉鎖等
1011501015889	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号2階102号	履歴等
2011501015888	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号2階101号	履歴等
3011501015895	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号4階102号	履歴等
3011501022512	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号3階108号	履歴等
4011501015894	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号4階101号	履歴等
4011501018187	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号3階104号	履歴等
4011501021983	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号3階107号	履歴等
5011501015893	ニホンコウウン 日本交通株式会社	東京都北区浮間5丁目4番51号3階103号	履歴等

名称・住所が
ほぼ同じで別法人番号

法人番号	商号又は名称	所在地	変更履歴情報等
4010701030286	株式会社NDR JAPAN	東京都品川区北品川1丁目9番2号	履歴等
9010401091323	株式会社NDR JAPAN	東京都品川区北品川1丁目9番2号TOKYO・YBビル7F	履歴等

建物ありなしで別法人番号

法人番号	商号又は名称	所在地	変更履歴情報等
8040001106000	ショウヤ 株式会社SHoYA	千葉県船橋市印内町599番地3	履歴等
1040001094474	株式会社SHOya	千葉県船橋市印内町599番地3	履歴等
2040001094473	shoya株式会社	千葉県船橋市印内町599番地3	履歴等
3040001022986	株式会社SHOYA	千葉県船橋市印内町599番地3	履歴等
3040001094472	SHOYA株式会社	千葉県船橋市印内町599番地3	履歴等
4040001094471	株式会社shoya	千葉県船橋市印内町599番地3	履歴等

法人名の大文字小文字違いで
別法人番号

法人番号	商号又は名称	所在地	変更履歴情報等
4300005002643	熊野社	佐賀県多久市西多久町大字板屋無番地	履歴等
5300005002642	熊野社	佐賀県多久市北多久町大字多久原2176番地	履歴等
6300005002641	熊野社	佐賀県多久市北多久町大字多久原2176番地	履歴等

名称・住所が同一で別法人番号

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題-**外字、情報が欠けている**

橿延神社

**縮退先の
ない外字**

**徳島県小松島市坂野町
字橿 延50番地**

● 2．介護サービス（予防を含む）を提供し、又は提供しようとする事業所に関する事項

事業所の名称、所在地及び電話番号その他の連絡先				
事業所の名称	(ふりがな)	しょうじゅえんりはびりけあせんたー		
	松寿園リハビリケアセンター			
事業所の所在地	〒311-2206	市区町村コード	鹿嶋市	
	(都道府県から番地まで)	茨城県		
	(建物名・部屋番号等)			
事業所の連絡先	電話番号	0		
	FAX番号	0		
	ホームページ	あり		
介護保険事業所番号	0852280023			

**途中までしか住所が
格納されていない**

- データ精度の問題-**誤情報**が含まれている

日本赤十字社の代表として公開されている代表者名　8パターン

近藤忠煇	←	誤字
近藤忠輝	←	誤字
近衛忠	←	誤字文字欠け
近衛忠煇	←	誤字
近衛忠輝	←	誤字
近衛忠	←	文字欠け
近衛忠煇	←	正解
近衛忠輝	←	誤字

URL　プロトコル記入ミス 13パターンパターン

文字欠け
http://automation.jp
ttp://automation.jp
htt://automation.jp
http//automation.jp
http:/automation.jp
文字増
Htttp://automation.jp
hhttp://automation.jp
httpp://automation.jp
http::/automation.jp
http:///automation.jp
文字誤入力
http;///automation.jp
http:///automation.jp
http://http://automation.jp

手順3 データの精度の向上　–データクレンジングケース

追記

- データ精度の問題-**複数のマスタの存在**
  — **食品営業許可届出一覧の営業の種類と、行政基本情報データ連携モデルのPOIコードのマスタが存在**

【15.食品等営業許可・届出一覧】

営業の種類(令和3年11月17日までは下記より選択)

食品衛生法施行令第35条にて定める34の営業の分類と、各地方公共団体における条例で定めた分類の中から該当するものを記載。

食品衛生法施行令第35条にて定める34の営業の分類	備考
飲食店営業	
喫茶店営業	
菓子製造業	
あん類製造業	
アイスクリーム類製造業	
乳処理業	
特別牛乳搾取処理業	
乳製品製造業	
集乳業	
乳類販売業	
食肉処理業	
食肉販売業	
食肉製品製造業	
魚介類販売業	
魚介類せり売営業	
魚肉ねり製品製造業	
食品の冷凍又は冷蔵業	
食品の放射線照射業	
清涼飲料水製造業	
乳酸菌飲料製造業	
氷雪製造業	
氷雪販売業	
食用油脂製造業	
マーガリン又はショートニング製造業	
みそ製造業	
醤油製造業	
ソース類製造業	
酒類製造業	
豆腐製造業	
納豆製造業	
めん類製造業	
そうざい製造業	
缶詰又は瓶詰食品製造業	
添加物製造業	

飲食				0503	食・グルメ	食をテーマとした観光利用の拠点。	観光庁　観光入込客統計に関する共通基準　観光地点等分類コード	7611	食堂、レストラン
飲食	2401	レストラン	レストラン	5	レストラン		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対訳表】		
飲食	2402	和食	和食	6	和食		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	2621	日本料理店
飲食	2403	イタリアン料理	イタリアン料理	7	イタリアン料理		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2404	フランス料理	フランス料理	8	フランス料理		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2405	中華料理	中華料理	9	中華料理		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7623	中華料理店
飲食	2406	居酒屋	居酒屋	10	居酒屋		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	765	酒場、ビヤホール
飲食	2407	ビアガーデン	ビアガーデン		ビアガーデン		OpenStreetMap Wiki, JA:Map Features(2017-11-28)	765	酒場、ビヤホール
飲食	2408	バー	バー、スナックバー、キャバレー、ナイトクラブ	11	バー		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	766	バー、キャバレー、ナイトクラブ
飲食	2409	ラーメン	ラーメン	12	ラーメン		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7624	ラーメン店
飲食	2410	寿司	寿司	13	寿司		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	764	すし店
飲食	2411	そば	そば	14	そば		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	763	そば・うどん店
飲食	2412	うどん	うどん	15	うどん		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	763	そば・うどん店
飲食	2413	ファーストフード	ファーストフード	16	ファーストフード		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	769	その他の飲食店
飲食	2414	焼肉	焼肉	17	焼肉		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7625	焼き肉店
飲食	2415	とんかつ	とんかつ	18	とんかつ		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2416	お好み焼き	お好み焼き	19	お好み焼き		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7692	お好み焼き、焼きそば、たこ焼き店
飲食	2417	牛丼	牛丼	20	牛丼		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2418	すき焼き	すき焼き	21	すき焼き		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2419	しゃぶしゃぶ	しゃぶしゃぶ	22	しゃぶしゃぶ		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2420	焼き鳥	焼き鳥	23	焼き鳥		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2421	カレー	カレー	24	カレー		東京都　国内外旅行者のためのわかりやすい案内サイン標準化指針【東京都版対	7629	その他専門料理店
飲食	2422	喫茶店、カフェ	喫茶店、カフェ		喫茶店、カフェ		OpenStreetMap Wiki, JA:Map Features(2017-11-28)	767	喫茶店
飲食					喫茶店、カフェ		schema.org(2017-11-28)	767	喫茶店
飲食	2423	アイスクリーム店	アイスクリーム店		アイスクリーム店		OpenStreetMap Wiki, JA:Map Features(2017-11-28)	5862	菓子小売業(製造小売でないもの)
飲食					アイスクリーム店		schema.org(2017-11-28)	5862	菓子小売業(製造小売でないもの)
飲食	2424	インターネットカフェ	インターネットカフェ		インターネットカフェ		OpenStreetMap Wiki, JA:Map Features(2017-11-28)	767	喫茶店
飲食	2499	その他飲食施設							

手順3 データの精度の向上　–データクレンジングケース

- データ精度の問題-その他

データの意味を表す**表側が空白**のレコードが存在する。

地域（町丁名）		世帯数	人口 総数	男	女
八重洲	2	29	32	18	14
	計	29	32	18	14
京橋	1	63	72	34	38
	2	82	119	65	54
	3	56	74	41	33
	計	201	265	140	125

地域（町丁名）	世帯数	人口 総数	男	女
総数	347,846	693,586	349,479	344,107
小松川 1丁目	2,524	5,319	2,569	2,750
2丁目	2,098	4,621	2,106	2,515
3丁目	1,865	4,635	2,246	2,389
4丁目	655	1,300	671	629

同上や〃などを利用して別セルを指示する繰り返し文字がある

町丁目	世帯
総数	101,894
後楽一丁目	230
〃 二丁目	676

土地区画整理事業内など**住所が未確定**な場合がある

ケアサポート　COCOLO

事業所の概要 | 事業所の特色 | 事業所の詳細 | 運営状況 | その他

介護サービスの種類	訪問介護
所在地	〒501-6254 岐阜県羽島市福寿町本郷駅北本郷土地区画整理事業内34街区6-1 地図を開く
連絡先	Tel：058-393-2301／Fax：

項目に**対応しない文字列**が格納されている。

事業所の所在地	〒311-4153	市区町村コード	水戸市
	（都道府県から番地まで）	県の届でけ通り	
	（建物名・部屋番号等）		

その他

メタデータ

・データの主幹の記載がない為、どこに問い合わせしていいかわからない。

ライセンス

・ライセンスの記述がない為、どこまで自由に使っていいのかわからない。

あり・なしで回答されている

・ホームページURLを記入する個所に『あり』と記載されている

手順3 データの精度の向上　–データ補完ケース

追記

- データ精度の問題-POIコードの補完を行う。
  — POIコードについて、出典データを元に学習を行い名称からPOIコードを推測する処理を実施しました。

出典データ

POIコード、POIを
学習データ用に手動付与

営業所名称	営業の種類	POIコード	POI
ファミリーマート　代々木駅西	コンビニエンスストア	2710	コンビニエンスストア
CAFE：MONOCHROME	その他の食料・飲料販売業	2713	コーヒー店
十番右京　恵比寿店	その他の食料・飲料販売業	2410	レストラン
ローソン神宮前六丁目店	コンビニエンスストア	2710	コンビニエンスストア
JPローソン代々木郵便局店	コンビニエンスストア	2710	コンビニエンスストア
ローソン渋谷道玄坂上店	コンビニエンスストア	2710	コンビニエンスストア
Cygames　ガーデン　19F	コップ式自動販売機	2106	自動販売機
すき家　幡ヶ谷駅前店	その他の食料・飲料販売業	2410	レストラン

1,000レコード

JSON

```
{
  処理対象データ : "株式会社高島屋新宿店11開催会場　大北海道展"
  大分類 : "飲食"
  POI : "レストラン"
  POIコード : 2401
}
```

手順3 データの精度の向上　–データ補完ケース

追記

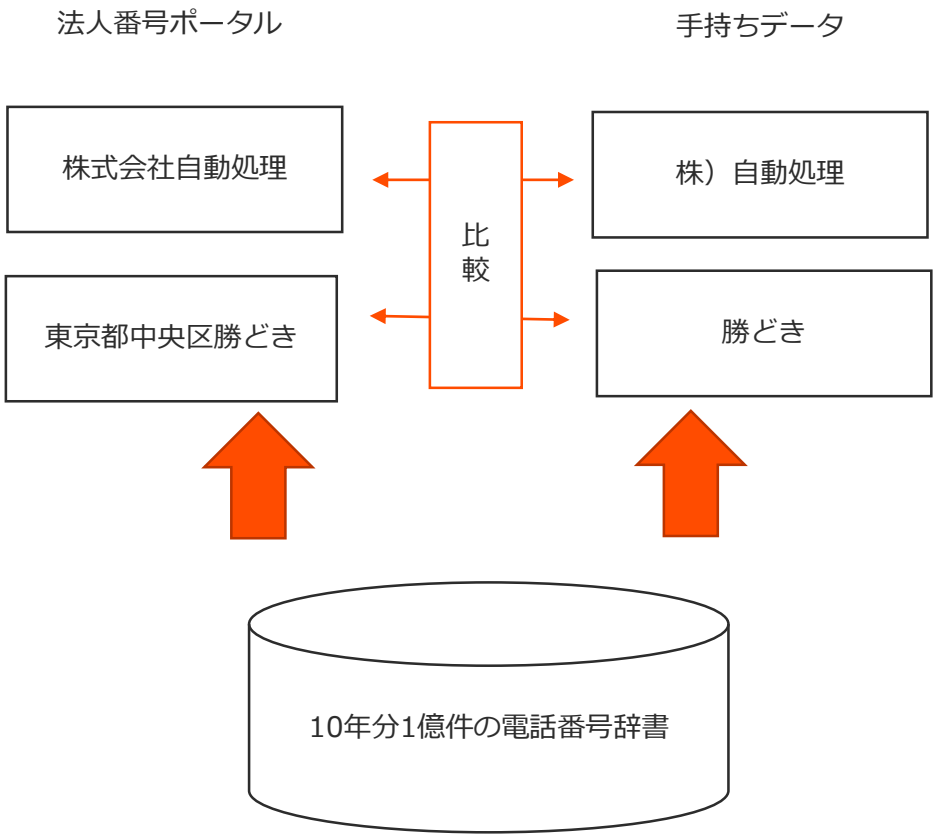
- データ精度の問題-緯度経度の追加付与
  — 住所情報から、住所情報を元に緯度経度を付与しました。

営業所所在地	緯度	経度
東京都台東区秋葉原1番5	35.7020182292	139.7752284071
東京都台東区秋葉原3番7	35.7022222222	139.7748787435
東京都台東区秋葉原3番1	35.7020610894	139.7750165473
東京都台東区秋葉原3番1	35.7020610894	139.7750165473
東京都台東区浅草1丁目1	35.7109556749	139.7974820964
東京都台東区浅草1丁目1	35.7109556749	139.7974820964
東京都台東区浅草1丁目1	35.7109556749	139.7974820964
東京都台東区浅草1丁目1	35.7109754774	139.7973019748
東京都台東区浅草1丁目1	35.7109754774	139.7973019748
東京都台東区浅草1丁目1	35.7110061306	139.7971964518
東京都台東区浅草1丁目1	35.7111564128	139.7971199544
東京都台東区浅草1丁目1	35.7112689887	139.7971397569
東京都台東区浅草1丁目1	35.7113864475	139.7971411133
東京都台東区浅草1丁目1	35.7113864475	139.7971411133
東京都台東区浅草1丁目1	35.7115386285	139.7971568468
東京都台東区浅草1丁目1	35.7116238064	139.7973258464
東京都台東区浅草1丁目1	35.7116238064	139.7973258464
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459
東京都台東区浅草1丁目1	35.7116080729	139.7974948459

手順3 データの精度の向上　–データ補完ケース

追記

- データ精度の問題-法人番号の追加付与
  ― 事業所の名称、住所、電話番号情報から、法人番号を付与しました。



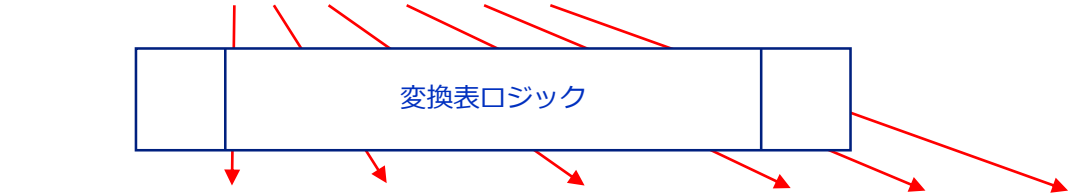
手順3 データ構造における民間データと自治体データの整備における特徴

追記

- 今年度は民間企業様のデータを預かって整備させて頂きました。
  - ー 民間企業のデータはデータ整備に関する仕様書が定義されており、システムを利用して、継続的に整備されており、利活用前提でデータが整備されている事から、データに関しては仕様が一様であり、資料化もされている事から、データの品質を上げる作業については、比較的作業の進め方について見通しが立てやすい状況でした。
  - ー データ整備に関する仕様書が定義されており、システムを利用して、継続的に整備されている事から、データ品質が一定であることが利活用のしやすさにつながっていると思われます。
  - ー 但し、民間企業にデータ提供を依頼する場合、基本的には自社が開発しやすいデータであることも多そうでした。
    自治体が提供するデータはデジタル庁が、国際規格に基づき自治体オープンデータセットとして整えられていますので、そういった違いがありました。

スラッシュ区切りで、1項目に複数値が設定さている。
　　定休日情報1/定休日情報２/…定休日情報9
　　※定休日は0個～9個と格納数は様々（実データより）。
・変換表を参照しないと、意味のわかるデータとならない。
　値フォーマット：「大区分_小区分(-大区分_小区分)」のデータフォーマットとなっている。

定休日情報	定
01010202/01010203/01010201/01120231/0607/0609	土



定休日情報_1_期間	定休日情報_2_期間	定休日情報_3_期間	定休日情報_4_期間	定休日情報_5_期間	定休日情報_6_期間	定休日情報_7_期間	定休日情報_8_期間	定休日情報_9_期間
月_1月02日	月_1月03日	月_1月01日	月_12月31日	曜日_土曜	曜日_日・祝祭日			

データクレンジングについて調達のポイントをまとめています。

追記

- 昨年と違いデータクレンジングが中心の業務となりました。
- データクレンジング作業は目視で実施するか、事業者に依頼するかのどちらかになります。
- 調達の際には参考にしていただければと思っております。

【文字列正規化における委託事業者へ依頼時の仕様イメージ】

項目	内容	備考
作業の目的・概要	各データ項目に対して、文字単位の基本的な表記揺れを判別し、想定している文字列の型に合わせて修正する	
アウトプット	基本的な機械的なチェックにより、文字単位の表記揺れが解消された状態	・定量的にアウトプットの品質を定義しにくいため、作業内容を明確に指示することが望ましい
作業内容の想定例	・アルファベット、カナ、数字、漢字、記号等の、文字列種別毎の表記揺れのチェック（半角/全角、空白の長さ、常用漢字/常用外漢字など） ・MJ 縮退マップを利用して、異体字を変換 ・記号の異体字を統一 ・省略表現による表記揺れのチェック（法人格の表記など） ・その他の表記揺れのチェック（読み仮名、ローマ字など） ・表記揺れの修正方針の決定、統一化作業	・委託事業者別に独自ノウハウが存在している部分であり、委託事業者に相談した上で、作業内容や統一化方針等を具体化することが望ましい

【データの値チェックおける委託事業者へ依頼時の仕様イメージ】

項目	内容	備考
作業の目的・概要	データのクレンジングや正規化の実施前に、各データの荒れ傾向や特性を把握する	
アウトプット	各データの荒れ傾向や特性が分かり、データのクレンジングや正規化の方針を検討できる状態	
作業内容の想定例	・目視によるデータ項目別のデータ値の制御状況の把握（自由記載か否か、電話番号、氏名、施設名などの表記が統一されているか）	・対象データの質や量に応じ、作業量が変化するため、サンプルデータや課題感を事前に提示した上で、作業内容・範囲を決める形が望ましい

【データのマッピングにおける委託事業者へ依頼時の仕様イメージ】

項目	内容	備考
作業の目的・概要	各データのフォーマットから、想定する共通フォーマットへ統一化する	
アウトプット	想定する共通フォーマットのデータ項目に、データ値が正しく格納された状態	
作業内容の想定例	・統一化する共通フォーマットの検討・決定（事前準備） ・統一化する共通フォーマットのデータ項目と、各データの様式・フォーマットのデータ項目の対応関係の整理 ・統一化する共通フォーマットのデータ項目に対応するデータ値の格納（マッピング） ・データ項目別のデータ正規化、不足するデータ項目に対するデータ補完の並行実施	・統一化するフォーマットの検討方法は、手順 1 の「データ構造の検討」「データ構造の調査」をご参照 ・データ正規化とデータ補完については、後述の「データの正規化」「複雑なデータ補完」をご参照

【データ構造に関するクレンジングにおける委託事業者へ依頼時の仕様イメージ】

項目	内容	備考
作業の目的・概要	目視で確認を行い、機械判読性の高いデータ構造に変換する	
アウトプット	機械判読性の高いデータ構造に変換されたデータ	・対象となるデータの構造により、作業量や実現可能なアウトプットの品質が異なるため、サンプルデータを提示した上で、作業内容・アウトプットを決める形が望ましい
作業内容の想定例（例）	・1 セル 1 データへの変換 ・データ項目の表記ゆれの統一 ・経年でデータ粒度が異なる場合の粒度統一 ・枠外にデータが追加されている場合のデータの取込	

【データ構造に関するクレンジングにおける委託事業者へ依頼時の仕様イメージ】

項目	内容	備考
作業の目的・概要	目視で確認を行い、データの内容を機械判読性の高いデータに変換する	
アウトプット	機械判読性の高いデータに変換されたデータ	・対象となるデータの形式や値により、作業量や実現可能なアウトプットの品質が異なるため、サンプルデータを提示した上で、作業内容・アウトプットを決める形が望ましい
作業内容の想定例（例）	・枠外にデータが追加されている場合のデータの取込 ・打消し線で修正されているデータへの対応（打消し線部分のデータの削除） ・異表記なデータ形式の修正	

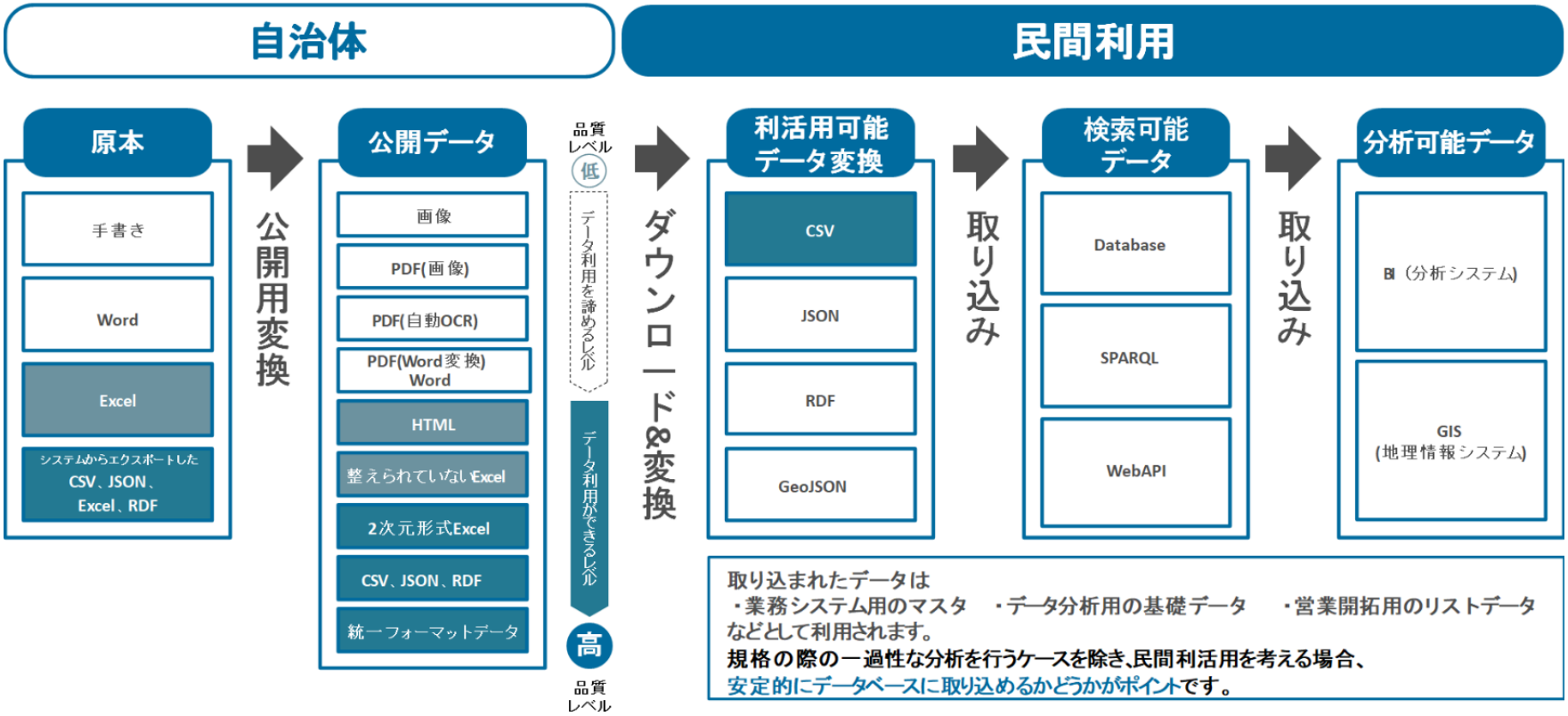
おわりに

データ整備を終えて

追記

- データ整備は2年目になりましたが、大変な作業でした。
  — 新しいデータが出れば、その分新しいケースが増える事になります。
  — みんなで同じデータを整備すれば、データ利活用の前に数カ月作業をしないといけない状況は回避できます。

— 昨年、今年と知見が貯まってきておりますので、知見をもっと皆さまにフィードバックできるようにしたいと考えております。



ご清聴ありがとうございました。