

# 行政データ整備モデル事業 成果報告会

## 「データ」に対する取組について ～きれいなデータが何故必要？～

令和4年3月29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
- ケース2 画像、PDF
  - AI-OCRを利用した  
文字データ変換

## 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	医療法人社団貴理会 東京国際大塚病院	三鷹市下連雀4丁目8番40号
2	医療法人社団社友会 三鷹病院	三鷹市下連雀4丁目1番12号
診療所(R2.12.1 開設届)		
No	施設名称	施設所在地
3	篠原病院	
4	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 医療法人社団輝祥会 慶真整形外科	三鷹市井の頭2-1-17
5	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
6	8 医療法人社団研運会 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
7	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 医療法人社団尚隆医院	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
8	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

# 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

- 類似のデータがないか調査をしましょう。

推奨データセット（内閣官房）

<https://cio.go.jp/policy-opendata>

#	基本編/応用編	データ名	対象	作成にあたり準備すべきルールやフォーマット等（※1）		
1		AED設置箇所一覧			XLSX	CSV
2		介護サービス事業所一覧			XLSX	CSV
3		医療機関一覧			XLSX	CSV
4		文化財一覧			XLSX	CSV
5		観光施設一覧			XLSX	CSV
6		イベント一覧			XLSX	CSV
7		公衆無線LANアクセスポイント一覧			XLSX	CSV
8	基本編	公衆トイレ一覧	オープンデータに取り組み始める地方公共団体	データ項目定義書	XLSX	CSV
9		消防水利施設一覧			XLSX	CSV
10		指定緊急避難場所一覧			XLSX	CSV
11		地域・年齢別人口			XLSX	CSV
12		公共施設一覧			XLSX	CSV
13		子育て施設一覧			XLSX	CSV
14		オープンデータ一覧			XLSX	CSV
A-1		食品等営業許可・届出一覧	地方公共団体		全許可・届出一覧 XLSX	CSV
A-		学校給食献立			新規許可・届出一覧 XLSX	CSV

[データカタログサイト（国）](#)  
[データカタログサイト（東京都）](#)



他にもGoogle Dataset Searchなどで検索調査

<https://datasetsearch.research.google.com/>

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 x

1	2	3	4	5	町丁目表記 地域名	総人口	男性	女性	5歳刻み	
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				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	76
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		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。												
「住民基本台帳」		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				0歳		1歳		2歳		3歳		4歳		
「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	男	女	男	女	男	女	男	女	男	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									

1歳刻み

- ・縦横の表で表せる … NG
- ・目黒区データと比較して … NG
- ・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表記する … NG
- ・同一データは同一の表記で表す … NG

・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	
	<b>外国人の集計あり</b>		99,438	113,124	212,562	20	113
港区総数	外国人	—	10,374	9,187	19,561	—	65
<b>町丁目表記 支所管内</b>	合計	—	109,812	122,311	232,123	—	178
<b>芝地区総合支所管内</b>	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						

・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表記する … NG  
 ・同一データは同一の表記で表す … NG

区市町村名：東京都板橋区  
母集団区分：公営

集計区分2：板橋一丁目

年齢	男	女
0	37	35
1	33	28
2	29	25
3	19	26
4	24	25
5	17	21
6	22	16
7	11	18
8	16	25
9	17	12
10	18	9

町丁目表記 集計区分

指定区別年齢別男女別人口調

令和3年9月30日 現在(年齢)  
令和3年9月30日 現在(住民)  
令和3年10月1日 作成

年齢	男	女	計
25	72	105	177
26	62	81	143
27	69	103	172
28	78	85	163
29	76	90	166
30	78	98	176
31	74	75	149
32	56	77	133
33	74	77	151
34	76	75	151
35	81	69	150
36	80	65	145
37	57	64	121
38	80	53	133
39	68	70	138
40	62	64	126
41	58	66	124
42	70	53	123
43	69	53	122
44	65	67	132
45	64	56	120
46	66	59	125
47	48	69	117
48	72	67	139
49	63	57	120

別 ※※※

40~44	324	303	627
45~49	313	308	621
50~54	302	305	607
55~59	244	213	457
60~64	213	198	411
小計	2768	2880	5648
割合	36.4%	37.8%	74.2%
65~69	159	175	334
70~74	197	200	397
75~79	88	129	217

年齢	男	女	計
50	68	73	141
51	49	64	113
52	74	71	145
53	57	53	110
54	54	44	98
55	37	37	74
56	54	55	109
57	55	46	101
58	53	38	91
59	45	37	82
60	43	36	79
61	38	47	85
62	51	36	87
63	47	46	93
64	34	33	67
65	27	46	73
66	36	33	69
67	28	27	55
68	36	32	68
69	32	37	69
70	41	42	83
71	37	34	71
72	43	52	95
73	44	28	72
74	32	44	76

1歳刻み

5歳刻み

80~84	78	129	207
85~89	48	82	130
90~94	13	46	59
95~	2	12	14
小計	585	773	1358
割合	7.7%	10.2%	17.8%

合計	3655	3959	7614
世帯数			4922

平均年齢

※外国人を含めた集計です。  
※世帯数は複数国籍世帯も含まれます。

- ・縦横の表で表せる … NG
- ・目黒区データと比較して … NG
- ・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見せる … NG
- ・同一データは同一の見せ方で表す … NG

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

推奨データセットの

『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

行政基本情報データ連携モデル

<https://cio.go.jp/guides>

### 推奨データセット

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

最後にそのデータをどう使えばよいか、わかりやすく説明するための、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。

- 先に整備しておくことで問い合わせが減り、  
利用者が悩まずに利用できるようになります。

**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](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

#### • ケース2 画像、PDF

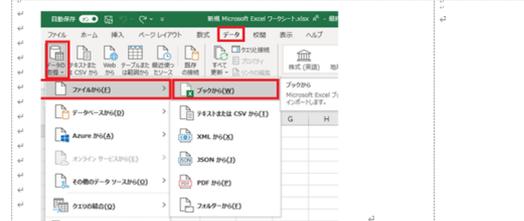
- AI-OCRを利用した文字データ変換

- 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。

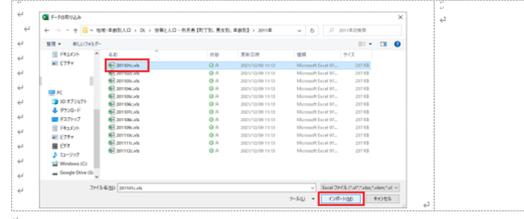
#### ① Excelで整形、加工したいファイルをインポートする

1. データ取得元（ソース）のファイルを開いている場合は、閉じておく

2. [データタブ]→[データの取得]→[ファイルから]→[バックから]をクリックする



3. Excelファイルをクリックして選択し、[インポート]をクリックする

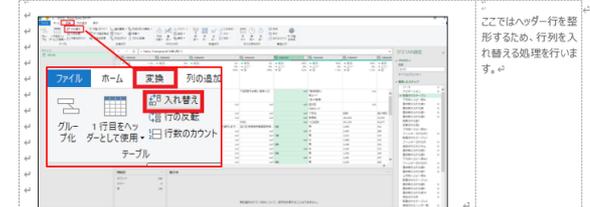


#### ② ExcelのPower Query機能でデータを整形する

1. 「クエリの設定」で、「変更された型」ステップをクリックして削除する



2. [変換]タブ→[入れ替え]をクリックする



3. 年籍の列を右をクリックし、[フィル]→[下]または[上]をクリックする



## 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

- ケース2 画像、PDF
  - AI-OCRを利用した文字データ変換

- 画像付きのPDFデータや、紙データのスキャン情報をデータ化しないといけない場合には、AI-OCRを利用します。

代表サンプルの修正

特定のテンプレートの分類活動やCLOVA OCRの認識活動の向上、テンプレートの作成方法についてサポートが必要な場合は、[テクニカルサポート](#)にお問い合わせください。

テンプレートの基本情報 \*必須入力項目です。\*

テンプレート名 \*

認識モデル  Basic  Premium

代表サンプルの設定

PDF / JPGなどの画像ベースの画像認識もサポート

20%  300% 85 初期値

フィールドの整列  並べ替え  上書き

代表サンプル名  編集管理

指定区別年齢別男女別人口調 修正

フィールド指定

フィールド名  フォーマット  削除

フィールド名  フォーマット  削除

フィールド名  フォーマット  削除

フィールド名  フォーマット  削除

キャンセル 実行

年齢	男	女	計												
0	25	27	52	25	27	54	50	26	27	53	75	25	27	52	47
1	26	28	54	26	28	54	51	28	29	57	76	26	28	54	57
2	27	29	56	27	29	56	52	29	30	59	77	27	29	56	58
3	28	30	58	28	30	58	53	30	31	61	78	28	30	58	48
4	29	31	60	29	31	60	54	31	32	63	79	29	31	60	53
5	30	32	62	30	32	62	55	32	33	65	80	30	32	62	47
6	31	33	64	31	33	64	56	33	34	67	81	31	33	64	46
7	32	34	66	32	34	66	57	34	35	69	82	32	34	66	35
8	33	35	68	33	35	68	58	35	36	71	83	33	35	68	33
9	34	36	70	34	36	70	59	36	37	73	84	34	36	70	36
10	35	37	72	35	37	72	60	37	38	75	85	35	37	72	29
11	36	38	74	36	38	74	61	38	39	77	86	36	38	74	24
12	37	39	76	37	39	76	62	39	40	79	87	37	39	76	34
13	38	40	78	38	40	78	63	40	41	81	88	38	40	78	17
14	39	41	80	39	41	80	64	41	42	83	89	39	41	80	13
15	40	42	82	40	42	82	65	42	43	85	90	40	42	82	10
16	41	43	84	41	43	84	66	43	44	87	91	41	43	84	18
17	42	44	86	42	44	86	67	44	45	89	92	42	44	86	12
18	43	45	88	43	45	88	68	45	46	91	93	43	45	88	11
19	44	46	90	44	46	90	69	46	47	93	94	44	46	90	7

- 機械可読性が低いデータの場合は、自動化が出来ない為、手作業が発生します。

# 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

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	合計
000111 台東1丁目	男	31	33	31	32	56	129	150	148	133	124	99	51	
		( 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)	( 15)	( 14)	( 11)	( 13)	( 4)		
	年 男	60~64	65~69	70~74	75~79	80~84	85~89	90~94	95~99	100~104	105~109	110~	合計	
		52	38	48	24	30	16	9	2	1	0	0	1,258	
	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)	( 7)	( 1)	( 0)	( 0)	( 0)	( 0)	( 0)	( 156)	

### HTML

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

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5

- データ形式の問題-データ構造が違う

### 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〕 <sup>↓</sup>
20210101	芝地区総合支所管内	芝一丁目	1288	968	1054	2022 <sup>↓</sup>
20210101	芝地区総合支所管内	芝二丁目	2218	1654	1874	3528 <sup>↓</sup>
20210101	芝地区総合支所管内	芝三丁目	2194	1705	1881	3586 <sup>↓</sup>
20210101	芝地区総合支所管内	芝四丁目	1397	1089	1220	2309 <sup>↓</sup>
20210101	芝地区総合支所管内	芝五丁目	1959	1155	1631	2786 <sup>↓</sup>
20210101	芝地区総合支所管内	海岸一丁目	1197	1014	980	1994 <sup>↓</sup>
20210101	芝地区総合支所管内	東新橋一丁目	782	677	781	1458 <sup>↓</sup>
20210101	芝地区総合支所管内	東新橋二丁目	406	336	224	560 <sup>↓</sup>

## 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-**項目の有無**が違う

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）

（令和3年4月1日 午前0時現在）

両国

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

- データ形式の問題-**情報提供頻度**が違う

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**

※不定期的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）](#)

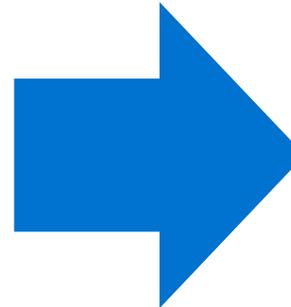
## 手順2 データ構造の統一 におけるデータ変換困難ケース5

### ・データ形式の問題-項目の表記ゆれ

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。

2019年（平成31年）3月

平成31年3月1日現在

年齢	総数	男	女
<b>総数</b>	<b>258,093</b>	<b>121,704</b>	<b>136,389</b>
<b>0歳～4歳</b>	<b>14,651</b>	<b>7,544</b>	<b>7,107</b>
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
<b>5歳～9歳</b>	<b>11,629</b>	<b>5,877</b>	<b>5,752</b>
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
<b>90歳～94歳</b>	<b>2,113</b>	<b>512</b>	<b>1,601</b>
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
<b>95歳～99歳</b>	<b>690</b>	<b>117</b>	<b>573</b>
<b>100歳以上</b>	<b>116</b>	<b>17</b>	<b>99</b>

2019年（平成31年）4月

平成31年4月1日現在

年齢	総数	男	女
<b>総数</b>	<b>258,696</b>	<b>121,989</b>	<b>136,707</b>
<b>0歳～4歳</b>	<b>14,516</b>	<b>7,473</b>	<b>7,043</b>
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
<b>5歳～9歳</b>	<b>11,622</b>	<b>5,866</b>	<b>5,756</b>
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
<b>95歳～99歳</b>	<b>690</b>	<b>118</b>	<b>572</b>
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
<b>100歳以上</b>	<b>127</b>	<b>19</b>	<b>108</b>
100歳	48	9	39
101歳	31	4	27
102歳	14	4	10
<b>103歳以上</b>	<b>34</b>	<b>2</b>	<b>32</b>

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

令和元年度特定健実施機関一覧表

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	社会医療法人協会 札幌里塚病院	004-0811	札幌市清田区 美しが丘3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

The screenshot displays four facility profiles from a website. Each profile has a header with the facility name and a '詳細情報を見る' (View Detailed Information) button. The profiles are:

- 鈴木内科リハビリセンター**: 所在地: 札幌市清田区清田4条2丁目10-25. サービス提供地域: 清田区. 営業時間: 平日 8:30~17:30, 土曜日 8:30~17:30, 日曜日 0:00~0:00, 祝日 0:00~0:00.
- 鈴木内科介護看護センター**: 所在地: 北海道札幌市清田区美しが丘2条4丁目17-1. サービス提供地域: 北海道札幌市清田区美しが丘全域.
- 鈴木内科訪問看護ステーション**: 所在地: 北海道札幌市清田区美しが丘2条4丁目17-1. サービス提供地域: 清田区全域および、近隣地域. 営業時間: 平日 9:00~17:00, 土曜日 9:00~17:00, 日曜日 9:00~17:00, 祝日 9:00~17:00.
- 医療法人社団 鈴木内科医院**: 所在地: 北海道札幌市清田区清田4条2丁目10-25. サービス提供地域: 清田区、白石区(南郷、栄通り)、豊平区(月寒、播住、西岡、羊が丘)、厚別区(厚別南、青葉町) 北広島市(大曲).

## 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-画像形式の地図データ



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гон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8

### ・データ形式の問題-データ構造外の情報の存在

	A	B	C	D	E	F	G	H	I
46	44	スマイル矯正歯科	三鷹市下運込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	三鷹市下運込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	三鷹市下運込3-44-5	希望ビル2階	歯	0422-29-8242			
49	47	運込の杜歯科	三鷹市下運込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	三鷹市下運込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	三鷹市下運込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	三鷹市下運込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	下運込デンタルクリニック	三鷹市下運込4-18-14	サンクマール三鷹下運込1階	歯・小歯・歯外・矯	0422-24-7850	以外		
54	52	角田歯科医院	三鷹市下運込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	三鷹市下運込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	あみやま歯科医院	三鷹市下運込5-8-1	シティコート下運込1F-B	歯・小歯・矯・歯外	0422-72-0220			
57	55	さくらい歯科医院	三鷹市下運込6-1-3		歯・小歯	0422-43-6303			
58	56	吉祥寺通り歯科クリニック	三鷹市下運込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	三鷹市下運込6-6-27		歯	0422-43-1324			
60	58	永江歯科医院	三鷹市下運込7-10-15	セドル三鷹201	歯・小歯	0422-42-7331			
61	59	ふくだ歯科	三鷹市下運込7-14-28	ベルムミタカ1F-2	歯・小歯・歯外・矯	0422-40-6480			
62	60	岩田歯科医院	三鷹市下運込7-16-13		歯・小歯・矯	0422-47-8860	開院 (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	三鷹市下運込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	三鷹市上運込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	岩田歯科医院	三鷹市上運込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。正しく入力される仕組みを作りましょう。**

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
  - 統一的な表記に変換

### 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

住所	郵便番号	正規化済み郵便番号
寺町20-47メフィ相沢1階	東京都八王子市寺町20-47	(NULL) 192-0073
打越町344-4北野ファイブ1階	東京都八王子市打越町344	(NULL) 192-0911
大塚425-4	東京都八王子市大塚425-4	(NULL) 192-0352
下柚木2-5-26	東京都八王子市下柚木2-5-26	(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4-27-1ビル2階	東京都八王子市明神町4丁目27-1ビル2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興伸ビル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) 192-0000
元横山町2-1-20	東京都八王子市元横山町2丁目1-20	(NULL) 192-0000
狭間町1450-1めじろ台コーポラス	東京都八王子市狭間町1450-1めじろ台コーポラス	(NULL) 192-0000
みなみ野3-9-7	東京都八王子市みなみ野3丁目9-7	(NULL) 192-0000
八日町4-9幸ビル3階	東京都八王子市八日町4-9幸ビル3階	(NULL) 192-0000
長沼町200-3旗野コーポラス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

市区町村を補完

半角英数字、カタカナ  
は全角に変換

郵便番号を追加付与

## 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

- データ形式の問題-**名前の変換**

### 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
- **一部の省略ケース**  
株) 自動処理 → 株式会社自動処理

### 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、高木祐介が正しかったとしても自動変換は出来ない。
- **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)
- **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

## 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

- 東京都豊島区東池袋3丁目1番1号  
サンシャイン60・43階・20号
- 東京都豊島区東池袋3丁目1番1号  
サンシャイン60 4320号
- 東京都豊島区東池袋3-1-1  
サンシャイン60 4320号
- 東京都豊島区東池袋  
サンシャイン60 (43階) 20号
- 東京都豊島区東池袋  
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39-176
- 宮城県塩竈市字杉ノ入裏39-176
- 宮城県塩竈市杉ノ入裏39-176

##### • 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

- 東京都中央区築地一丁目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390-2
- 兵庫県加古川市上荘町井野口390-2
- 兵庫県加古川市上荘町井乃口390-2  
(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91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390-2

(口の文字がカタカナ)

北海道北見市留辺蕊町旭公園95-4

(蕊の文字が書き間違いで本来は薬)

茨城県つくば市白井2090-20

(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13-21  
(熊本県熊本市が欠けている)  
静岡県浜松市浜北区貴布1200  
(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2-1-1  
(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  
愛知県名古屋市中川区西日置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

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

記入!

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) 本ぞまないかりけりせんたー  
 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

東京都                                  介護事業所・生活関連情報検索  
介護サービス情報公表システム                                  文字サイズの変更 中 大 最大

← 前のページに戻る 全国版トップ

**複数の情報が混ざっている**

2020年11月30日11:14 公表

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

画面を印刷する    ☆ お気に入りに追加する

お気に入り事業所一覧 0件

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

### マルマツ薬局 九段店

<b>マルマツ薬局 九段店</b>	飯田橋
住所	〒102-0072 千代田区飯田橋1-5-8 アクサンビル1階 <a href="#">地図を表示</a>
電話番号	(開店時間内)03-6808-5320 (開店時間外)03-3261-7224
施設情報	車椅子配慮  聴覚的配慮

### 参考) 店舗名がないケースが混在している

<b>マルマツ薬局</b>	飯田橋
住所	〒102-0072 千代田区飯田橋2-9-5 <a href="#">地図を表示</a>
電話番号	(開店時間内)03-3261-7224 (開店時間外)03-3261-7224
施設情報	車椅子配慮  視覚的配慮  聴覚的配慮

### 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 <a href="#">地図を開く</a>
連絡先	Tel : 0144-78-2180 ( / Fax : 0144-75-5228 <a href="#">ホームページを開く</a>

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)	複数電話番号	
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	

## 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 <a href="#">地図を開く</a>
連絡先	Tel : 04-7138-5345 / Fax : 04-7172-2055 <a href="#">ホームページを開く</a>

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	SAKURA 1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	夫沼ビル 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**

夫沼ビル KAISER三鷹フロント4階

## 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-名寄せをする為のID（番号）を判別する情報が存在しない

法人番号	商号又は名称	所在地	変更履歴情報等
6011501002115	シンニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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	ニホコウツウ 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
4010701030266	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・ YBEJL7F	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

延神社の情報

このページを印刷する

最新情報

法人番号  
1480005002414

商号又は名称  
延神社 外字

本店又は主たる事務所の所在地  
徳島県小松島市坂野町字 延50番地 外字

最終更新年月日  
平成27年11月27日

延神社

縮退先の  
ない外字

徳島県小松島市坂野町  
字 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	02	
	FAX番号	02	
	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

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

<Http://automation.jp>

<hhttp://automation.jp>

<http://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

## 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	29	32	18	14
計	29	32	18	14
京橋	63	72	34	38
	2	82	65	54
	3	56	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

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) (建物名・郵便番号等)	県の届でけ通り	

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 :

### 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

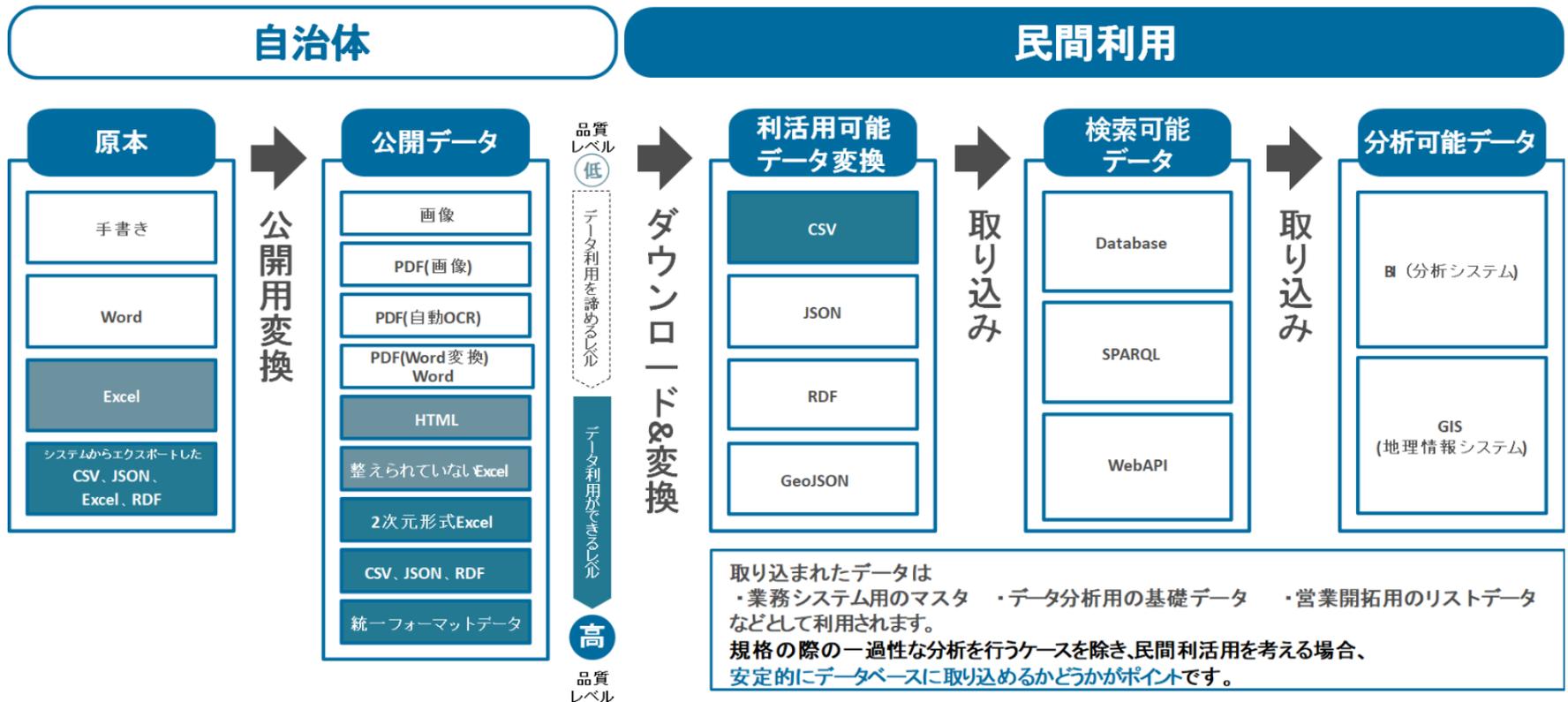
・ホームページURLを記入する個所に『あり』と記載されている

---

# データ整備を終えて

# データ整備を終えて

- 我々にとってはいつも通りではありましたが、やはり大変な作業でした。
  - 有識者の方からは専門家以外に出来るのかと指摘を受けるくらいの難しさはあったと思います。
  - 民間企業でも専門事業者でも、データサイエンティストやデータ整備業者でない限りはなかなか対応できないのデータが多かったです。
- ですが、データ利活用する側は必ずこの作業を行って、データ利用に進まなければいけません。
  - つまり、データ公開する側で作業を実施することが、データ活用の強い推進力となります。



- データを集めた後に、利活用するための作業を楽にするには

- **Excelを正しく使う**

- 1セルには1データ、セル結合はしない等。総務省統計局のガイダンス※を参考に

※統計表における機械判読可能なデータの表記方法の統一

[https://www.soumu.go.jp/menu\\_news/s-news/01toukatsu01\\_02000186.html](https://www.soumu.go.jp/menu_news/s-news/01toukatsu01_02000186.html)

- 現課の方に、機械可読性の高い形で、データを収集してもらう

- **デジタル・バイ・デフォルト**

- データ収集時点からデータとして情報を集める

- **申請時点で入力エラーに気づいてもらい、誤りを極力なくす**

- 誤ったデータに気づいてもらい、正しいデータを入力してもらう事で、クレンジング作業をなくす

- Google Form
- Excel Design Studio
- Kintone

# 最初から2次利用する事がしやすいデータを集める為の取り組みのご紹介

- Google Form Google
  - Google アンケートフォーム作成ツール
    - 正規表現によるデータチェックが可能

イベント参加申込書

開催日: 20XX年1月4~6日  
開催地: 郵便番号 都道府県 市区町村 番地  
お問い合わせ: (+81) 000-0000 または no\_reply@example.com

お名前

記述式

記述式テキスト (短文回答)

メールアドレス \*

記述式テキスト (短文回答)

所属組織 \*

記述式テキスト (短文回答)

参加日をお選びください。 \*

1日目

# 最初から2次利用しやすいデータを集める為の取り組みのご紹介

- Excel Design Studio 株式会社自動処理
  - Excelを利用してデータ入力やチェックが出来る仕組み。  
APIを利用したデータチェック、データアップロードする仕組みが作れる。

Excel Design Studio - 配布用アンケートの作成

説明にあるセルをクリックしてください。

質問 質問 回答

予算の状況 H38 ¥ 1,861 088

当初予算 J38

21年度 087

Step 1/2  
20/21  
前へ 次へ  
一覧表示  
設定完了

Q38 1861

配布用

事業番号 0033

平成24年行政事業レビューシート (経済産業省)

事業名	地域企業立地促進等共用施設整備費補助金		担当部署	経済産業政策局 地域経済産業グループ	作成責任者		
事業開始・終了(予定)年度	平成19年	平成22年	担当課室	産業施設課	課長 成瀬 茂夫		
会計区分	一般会計		施策名	1. 経済成長			
根拠法令 (具体的な 条項も記載)	-						
事業の目的 (目指す姿を 簡潔に。3行 程度以内)	企業立地の促進等による地域における産業集積の形成及び活性化に関する法律(平成19年法律第40号。以下「企業立地促進法」という。)に基づき、地域が主体となって産業集積の形成・活性化を進める際に、産業集積の中核を担う企業群の受け皿となる産業インフラを整備することを目的とする。						
事業概要 (5行程度以 内。別添 可)	企業立地促進法の規定により国の同意を得た基本計画の集積区域内において、企業立地促進及び産業集積の形成・活性化のために必要と認められ、当該計画に沿って行われる共用の施設等を整備する事業に対し、当該事業費の一部補助を行う。						
実施方法	<input type="checkbox"/> 直接実施 <input type="checkbox"/> 委託・請負 <input checked="" type="checkbox"/> 補助 <input type="checkbox"/> 負担 <input type="checkbox"/> 交付 <input type="checkbox"/> 貸付 <input type="checkbox"/> その他						
予算額・ 執行額 (単位:百万円)			21年度	22年度	23年度	24年度	25年度要求
	予算の状況	当初予算	¥ 1,861	¥ 1,911	¥ 1,376	¥ -	¥ -
		補正予算	¥ -20	¥ -	¥ -88	¥ -	¥ -
		繰越し等	¥ 89	¥ -265	¥ 280	¥ -	¥ -
		計	¥ 1,930	¥ 1,646	¥ 1,568	¥ -	¥ -
	執行額	¥ 1,770	¥ 1,245	¥ 1,308	¥ -	¥ -	
	執行率 (%)	91.7%	75.6%	83.4%	0.0%	0.0%	

説明一覧

- 0001 事業番号 0033
- 0002 事業名 地域企業立地促進等共用施設整備費
- 0003 担当部署 経済産業政策局地域経済産業グループ
- 0004 作成責任者 課長 成瀬 茂夫
- 0005 事業開始・終了(予定)年度 平成19年,平成22年
- 0006 担当課室 産業施設課
- 0007 会計区分 一般会計
- 0008 施策名 1. 経済成長
- 0009 根拠法令(具体的な条項も記載) -
- 0010 -
- 0011 事業の目的(目指す姿を簡潔に。3行程度以内)
- 0012 事業概要(5行程度以内。別添可) 企業立地促進法
- 0013 直接実施
- 0014 委託・請負
- 0015 補助
- 0016 負担
- 0017 交付

追加  
挿入  
削除

# 最初から2次利用しやすいデータを集める為の取り組みのご紹介

## • Kintone サイボウズ

— ウェブフォームを利用して、データ入力ができる仕組みを作れます。

様々な業務に利用できるテンプレートがある為、幅広い業務に利用できる。

- 正規表現によるデータチェックが可能

The screenshot shows the Kintone interface for configuring a survey form. The top navigation bar is yellow with the Kintone logo and a user profile for '相澤'. Below the navigation bar, there are icons for home, notifications, and search. The main content area is titled 'アンケート > アプリの設定' and shows the 'アンケート' (Survey) application settings. A green button 'アンケート' is visible, along with a note 'アプリ管理者用メモはありません (作成する)'. The 'フォーム' (Form) tab is selected, showing a list of form elements on the left and a configuration panel on the right. The configuration panel includes fields for '部署名 \*' (Department Name) and 'お名前 \*' (Name), a satisfaction scale for 'kintoneの満足度を教えてください \*' (Please tell us your satisfaction with Kintone) with radio buttons for '満足' (Satisfied), 'やや満足' (Somewhat Satisfied), 'やや不満' (Somewhat Dissatisfied), and '不満' (Dissatisfied), and a text area for 'ご意見・ご要望をお聞かせください。' (Please let us know your comments and requests).

---

ご清聴ありがとうございました。