

研究・調査データの標準化管理手法に関する研究

佐藤 和孝

1. はじめに

今日、計量的な研究では、IT 技術の利用が不可欠なものとなってきた。また、新たな統計手法の利用やデータ処理において、IT 技術の発展がそれらを支えてきたことも事実である。さらに、時間が経過するにつれ、データ量の増加と分析方法の多様化が進み、分析データの加工方法も複雑化し、ますます IT 技術なしでは分析処理が難しい状況になっている。そのような動向を受けて、データの管理手法についても工夫がこらされるようになってきた。

実際に、業務や情報システムについて標準化、最適化する方法を考え、組織の効率化を図っていくエンタープライズ・アーキテクチャ（以下、EA）¹と呼ばれる概念が注目されてきており、それを導入する企業も増えてきている。現実的に、組織内に存在するデータをすべて整理したとき、組織内で日常的に必要なデータや、重複していて不必要なデータが存在する可能性もあり、その適用は、その組織において研究効率を上げる結果となることが予想される。

実際の研究・調査の元となるデータは、研究結果に直結する極めて重要なものであるが、研究が終わると研究者個人で管理し、分析時のままの状態で放置されている場合が多い。そのデータは、他の研究で再利用できる可能性もあり、どのような研究機関でも、研究・調査データの有効利用を考えているはずである。しかし、研究者独自の手法で管理されているため、特有のデータ形式になっていることが多く、第三者が閲覧または利用しようとした場合に、大変取り扱いにくいのが現状である。また、EA の実施により、実際にどのような効果があるのかが形として見えないため、組織の中でそれらを管理する必要性を感じていない現実もある。

そこで、その複雑化したデータのフォーマットを標準化し、様々なデータを統括して扱えるような管理手法を考案した。また、その手法を実際にシステム上で実現させるために、データの条件抽出や並べ替え処理などの機能を備えたデータベースシステムを構築し、その実用性を検証した。

2. 方法

1) 利用データ

サンプルデータとして、①「国民健康保険の実態」、②「全国市区町村別の面積」、③「平成 17 年度医療機器データベース」の 3 種類を利用した。これらを詳細に表したものを表 1 に示す。①のデータは、国民健康保険中央会と都道府県国民健康保険団体連合会が毎年出版している、2003 年度および 2004 年度の「国民健康保険の実態」^{2,3}を参照したもので、細かく分類すると、データ表が年度単位で 9 つあり、計 18 のデータ表が存在する。また、②のデータは、毎年 10 月に国土地理院のホームページで公表される面積情報である。③のデータは、毎年、医療情報システム開発センターが有料で公表している医療機器情報である。

表1 データ表リスト

No.	データ表名	件数	列数
①	平成15年度一般状況	3,390 件	54 列
	平成15年度一般状況都道府県別計	240 件	52 列
	平成15年度保険税（料）の賦課徴収状況	3,244 件	41 列
	平成15年度療養の給付（診療費）および療養諸費諸率（一般被保険者分）	3,390 件	31 列
	平成15年度療養の給付（診療費）および療養諸費諸率（退職者医療分）	3,390 件	31 列
	平成15年度療養の給付（診療費）および療養諸費諸率（老人保健分）	3,390 件	31 列
	平成15年度療養の給付（診療費）および療養諸費諸率都道府県別計	720 件	30 列
	平成15年度保険税（料）収納状況および経理関係諸率	3,390 件	41 列
	平成15年度保険税（料）収納状況および経理関係諸率都道府県別計	240 件	40 列
	平成16年度一般状況	3,310 件	54 列
	平成16年度一般状況都道府県別計	240 件	52 列
	平成16年度保険税（料）の賦課徴収状況	3,229 件	41 列
	平成16年度療養の給付（診療費）および療養諸費諸率（一般被保険者分）	3,310 件	31 列
	平成16年度療養の給付（診療費）および療養諸費諸率（退職者医療分）	3,310 件	31 列
	平成16年度療養の給付（診療費）および療養諸費諸率（老人保健分）	3,310 件	31 列
	平成16年度療養の給付（診療費）および療養諸費諸率都道府県別計	720 件	30 列
	平成16年度保険税（料）収納状況および経理関係諸率	3,310 件	45 列
	平成16年度保険税（料）収納状況および経理関係諸率都道府県別計	240 件	44 列
②	平成15年10月公表全国市区町村別面積	3,397 件	13 列
	平成16年10月公表全国市区町村別面積	3,392 件	12 列
③	平成17年度医療機器データベース	256,617 件	256 列

2) システムの開発

システムの開発には、Web サーバの OS に Debian GNU/Linux 3.0 を採用し、データベースには PostgreSQL 7.2.1 を利用し、Apache 1.3.26 の環境上で、php（PHP: Hypertext Preprocessor）4.1.2 と JavaScript を利用して開発を行った。また、サーバのスペックは、PentiumⅢ800MHz、メモリ 512MB で行った。

3. 結果

1) 標準化手法の実装

(1) 標準化定義表の作成

標準化手法の実現のため、データ表の各列の属性情報を抜き出して、それを表形式にまとめた標準化定義表を作成し、この表を利用してデータを管理することとした。具体的な属性情報の例を表2に示す。データベースのテーブルを構成するために必要な基本情報に加えて、データを表示するときの表示設定に必要な項目を考慮して導出したものである。

表2 各データ表から抜き出す列の属性情報群

番号	列名	説明
1	取得年度	データ取得年度
2	データ表名	データ表の表題
3	列名 (英語)	列の表題 (英語表記)
4	列番号	列の通し番号
5	見出し1	列の大見出し
6	見出し2	列の中見出し
7	見出し3	列の小見出し
8	見出し4	列の見出し予備
9	見出し5	列の見出し予備
10	タイトル	列の表示される列名
11	単位	列の単位名
12	データ型	数値列、文字列、コード化列判定
13	表示幅	列幅をピクセル値で指定
14	公開可否	列の表示・非表示判定
15	小数点以下桁数	数値列の場合の小数点以下桁数
16	配置方法	列の左寄せ、中央配置、右寄せなどの配置

(2) 標準化定義表の作成手順

作成手順は、最初に1つのデータ表の1列目に注目し、その列に対する表2の属性値を設定し、その属性値の間にカンマ記号やTAB記号などの文字を挿入して値を連結させた1行の列属性情報データを作成する。続けて、次の列以降にも同様の作業を行い、作成した列属性情報データを行単位に追加していく。つまり、あるデータ表の1列が、標準化定義表の1行となる。この作業を、1つのデータ表の全ての列に対して行えば、新たに1つの表が出来上がる。さらに、この作業を表2の全てのデータ表に適用して作成した、データベース全体の標準化定義表を図1に示す。

具体的には、次の手順に従って、列の属性情報を整理した。

手順 1.整理するデータ表の数を確認し、各データ表にタイトルを付ける。

手順 2.それらのデータ表をまとめて表現するタイトルを付ける (データベース名の作成)。

手順 3.各データ表に取得年度を付ける。

手順 4.各データ表から列を抜き出し、列番号を付ける。

手順 5.各列に英語タイトル、日本語タイトルを付ける。

手順 6.各列の単位名を付ける。

手順 7.各列のデータ型（数値列、文字列、カテゴリ（コード化）列など）を調べる。

数値列（合計や平均などの計算が可能なデータ：実数と整数に分けて考える）

文字列（名称や、コメントなどの数値ではない文字列データ）

カテゴリ（県コード（1～47）や性別（M、F）など、コード化されたデータ）

手順 8.カテゴリ項目がある場合は、そのコードと名称を別のカテゴリデータ表にまとめる。

例）データ表：性別（コード=“M”：名称=“男”、コード=“F”：名称=“女”）

カテゴリデータ表のタイトルと列名を決め、元のデータ表との連結列を確認する。

手順 9.各列の表示幅と配置方法（左寄せ、中央配置、右寄せなど）を決める。

手順 10. 整理したものを表形式にまとめる（コード化可能なものはコード変換して登録）。

図1 データを標準化するための定義例

取得年度	データ表名	列名(英語)	列番号	見出し1	見出し2	データ型	タイトル	単位	表示幅	公開可否(0否、1可)	配置方法(0左寄せ、1センタリング、2右寄せ)		
3	2003	HI5_KH0110H	K	2		1、実数	レポート番号	2	0	0	1		
4	2003	HI5_KH0110H	M0011	3		2、整数	レポート区分	2	0	0	1		
5	2003	HI5_KH0110H	M0012	4		3、コード整数	都道府県コード	3	1	1	1		
6	2003	HI5_KH0110H	M0013	5		4、コード文字	保険者コード	3	2	1	1		
7	2003	HI5_KH0110H	KEN_NAME	6		5、文字列	保険者区分	3	3	1	1		
8	2003	HI5_KH0110H	KEN_NAME	7		6、日付	都道府県名	5	10	1	1		
9	2003	HI5_KH0110H	RA207	8			保険者名	5	50	1	1		
10	2003	HI5_KH0110H	RA208	9			総世帯数(年度末)	世帯	2	0	1	1	
11	2003	HI5_KH0110H	RA209	10	世帯数	年度末現在	総人口(年度末)	人	2	0	1	1	
12	2003	HI5_KH0110H	RA210	11	世帯数	年度末現在	総数	世帯	2	0	1	1	
13	2003	HI5_KH0110H	RA211	12	世帯数	年度末現在	総数(加入率)	%	1	1	1	1	
14	2003	HI5_KH0110H	RA212	13	世帯数	年度末現在	退職者単独世帯	世帯	2	0	1	0	1
15	2003	HI5_KH0110H	RA213	14	世帯数	年間平均	退職者混合世帯	世帯	2	0	1	0	1
16	2003	HI5_KH0110H	RA214	15	世帯数	年間平均	総数	世帯	2	0	1	0	1
17	2003	HI5_KH0110H	RA215	16	世帯数	年間平均	退職者単独世帯	世帯	2	0	1	0	1
18	2003	HI5_KH0110H	RA223	17	被保険者数	年間平均	退職者混合世帯	世帯	2	0	1	0	1
19	2003	HI5_KH0110H	RA245	18	被保険者数	年間平均	総数	人	2	0	1	0	1
20	2003	HI5_KH0110H	RA246	19	被保険者数	年間平均	総数(3歳未満)	人	2	0	1	0	1
21	2003	HI5_KH0110H	RA247	20	被保険者数	年間平均	総数(70歳以上一般)	人	2	0	1	0	1
22	2003	HI5_KH0110H	RA248	21	被保険者数	年間平均	総数(一定以上所得者)	人	2	0	1	0	1
23	2003	HI5_KH0110H	RA249	22	被保険者数	年間平均	一般	人	2	0	1	0	1
24	2003	HI5_KH0110H	RA249	23	被保険者数	年間平均	一般(3歳未満)	人	2	0	1	0	1
25	2003	HI5_KH0110H	RA250	24	被保険者数	年間平均	一般(70歳以上一般)	人	2	0	1	0	1
26	2003	HI5_KH0110H	RA251	25	被保険者数	年間平均	一般(一定以上所得者)	人	2	0	1	0	1
27	2003	HI5_KH0110H	RA252	26	被保険者数	年間平均	一般(構成率)	%	1	1	1	1	1
28	2003	HI5_KH0110H	RA227	27	被保険者数	年間平均	退職被保険者	人	2	0	1	0	1
29	2003	HI5_KH0110H	RA228	28	被保険者数	年間平均	退職被扶養者	人	2	0	1	0	1
30	2003	HI5_KH0110H	RA251	29	被保険者数	年間平均	退職者合計	人	2	0	1	0	1
31	2003	HI5_KH0110H	RA252	30	被保険者数	年間平均	退職者合計(3歳未満)	人	2	0	1	0	1
32	2003	HI5_KH0110H	RA253	31	被保険者数	年間平均	退職者合計(70歳以上一般)	人	2	0	1	0	1
33	2003	HI5_KH0110H	RA253	32	被保険者数	年間平均	退職者合計(一定以上所得者)	人	2	0	1	0	1
34	2003	HI5_KH0110H	RA229	32	被保険者数	年間平均	退職者合計(構成率)	%	1	1	1	1	1
34	2003	HI5_KH0110H	RA230	33	被保険者数	年間平均	老人保健医療給付対象者	人	2	0	1	0	1
35	2003	HI5_KH0110H	RA231	34	被保険者数	年間平均	老人保健医療給付対象者(構成率)	%	1	1	1	1	1
36	2003	HI5_KH0110H	RA235	35	被保険者数	年間平均	介護保険第2号	人	2	0	1	0	1
37	2003	HI5_KH0110H	RA216	36			不明1	2	0	0	0	1	1
38	2003	HI5_KH0110H	RA236	37			不明2	2	0	0	0	1	1
39	2003	HI5_KH0110H	RA237	38			不明3	2	0	0	0	1	1
40	2003	HI5_KH0110H	RA238	39			不明4	2	0	0	0	1	1
41	2003	HI5_KH0110H	RA217	40			不明5	1	1	0	1	1	1
42	2003	HI5_KH0110H	RA218	41			不明6	2	0	0	0	1	1
43	2003	HI5_KH0110H	RA239	42			不明7	2	0	0	0	1	1
44	2003	HI5_KH0110H	RA240	43			不明8	2	0	0	0	1	1
45	2003	HI5_KH0110H	RA241	44			不明9	2	0	0	0	1	1
46	2003	HI5_KH0110H	RA219	45			不明10	2	0	0	0	1	1
47	2003	HI5_KH0110H	RA220	46			不明11	2	0	0	0	1	1

(3) カテゴリデータ連結定義表の作成

前記の手順 8 により、カテゴリデータ表を作成した場合、元のデータ表と連結させるための情報が必要となるため、カテゴリデータ連結定義表を作成した。この表には、カテゴリデータ表名、カテゴリデータ表のコード列名、連結させる元のデータ表名、連結させる元のデータ表のコード列名を登録する。

今回のサンプルデータの中で、表 1 の①のデータについては、「都道府県コード」「市町村区分コード」「地域コード」「合計グループコード」、②のデータについては、「都道府県コード」「地域コード」、③のデータについては、「商品分類コード」が含まれるデータ表について、それらに対応するデータ表をカテゴリデータ表として作成し、連結定義情報をカテゴリデータ連結定義表に登録した。

2) システムの動作検証

システムの実行画面は、上下にフレームを分割し、上側フレームでデータの選択や、表示条件の設定を行い、下側フレームにその結果を表示する構成とした。また、システムの基本機能として、①特定のコードデータだけを表示させるようにする限定表示機能、②表示させる行数を指定できる表示行数設定機能、③ある特定条件を設定し、その条件を満たすデータを抽出する条件抽出機能、④並べ替え列を指定できる並べ替え機能、⑤表示させたい列を選択できる表示列選択機能、⑥表示した結果を CSV 形式（カンマ区切り）でダウンロードできるテキストファイル出力機能を実装させた。さらに、合計値や平均値などの単純集計機能にも対応できるように設計した。

実際に、データベースに「国民健康保険の実態」、閲覧データに「平成 15 年度療養の給付（診療費）および療養諸費諸率（老人保健分）」を指定した場合の実行画面を図 2 に示す。ここで、基本機能の設定箇所は、図に数字で示したとおりで、③～⑤の機能については、機能の説明文をクリックすると設定画面を表示する仕組みとなっている。

図2 実行画面例（2003年度療養の給付（診療費）および療養諸費諸率（老人保健分））

日医総研共有データベース - Microsoft Internet Explorer

アドレス http://sato/jmari_db.php

データベース一覧 国民健康保険の実態 年度選択 選択して下さい

データ選択 平成15年度療養の給付(診療費)及び療養諸費諸率(老人保健分)

表示行数の設定 先頭行 から 100行 表示 ②

結果表示

抽出条件の設定 ③

並び替え列の設定 ④

表示列の設定 ⑤

【限定表示】複数選択: [Ctrl]+クリック

都道府県 市町村 地方

北海道 青森県 岩手県 宮城県 秋田県 再表示 全選択 全解除

市町村 組合 再表示 全選択 全解除

地方 北海道 東北 関東 甲信越 北陸 再表示 全選択 全解除

平成15年度療養の給付(診療費)及び療養諸費諸率(老人保健分)

3390件中、1～100件目を表示 次へ

このデータをCSV形式で保存する ⑥ 全データをCSV形式で保存する

療養の給付(診療費)																				療養諸費の老人以外(一般+退職)の比率															
都道府県コード		保険者コード		保険者区分		保険者名		入院				入院外				歯科				計				1人当り療養諸費用額		療養諸費の老人以外(一般+退職)の比率									
								受診率		1件当り日数		費用額		受診率		1件当り日数		費用額		受診率		1件当り日数		費用額		受診率		1件当り日数		費用額		1人当り療養諸費用額		療養諸費の老人以外(一般+退職)の比率	
								日		円		円		円		日		円		円		円		日		円		円		円		円		%	
1	1	1	北海道	札幌市	113.269	20.27	458466	22613	519298	1577.086	2.22	17109	7697	269830	153.064	2.91	24705	8496	37815	1843.418	3.39	44859	13237	826943	992412	58.9	3.7	14							
1	2	1	北海道	函館市	90.747	19.69	451768	22939	409965	1763.23	2.54	15795	6228	278507	132.796	2.78	22484	8089	29858	1986.772	3.34	36156	10837	718330	893285	61.01	3.11	32							

ページが表示されました

インターネット

この画面では、データ列の見出しが段階的であったり、他列で共通の見出しとして利用されていたりする項目があったが、それらの表示条件にも対応した形で結果を表示できた。それは、標準化定義表の見出し列の表示定義から、列の段組や連結処理を施して、動的に列タイトルの構成を組み換えるアルゴリズムを実装させていたため、このアルゴリズムが正常に動作していることが確認できた。また、表示列の単位も正常に表示されていることがわかる。

他の画面例として、データベースに「全国市区町村別面積」、閲覧データに「平成16年10月公表全国市区町村別面積」を選択した画面を図3に示す。この例では、限定表示列が図2の例とは異なることがわかる。また、単位表示で“km²”のように上付き文字にも対応させた。

図3 実行画面例（2004年10月公表全国市区町村別面積）

日医総研共有データベース - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

アドレス http://www.kato-jmari-db.php

データベース一覧 全国市区町村別面積 年度選択 選択して下さい

データ選択 2004年10月全国市区町村別面積公表データ

表示行数の設定 先頭行 から 100行 表示

結果表示

【限定表示】複数選択: Ctrl+クリック

都道府県 北海道 青森県 岩手県 宮城県 秋田県

地方 北海道 東北 関東 甲信越 北陸

再表示 全選択 全解除

2004年10月全国市区町村別面積公表データ

3392件中、1～100件目を表示 次へ

このデータをCSV形式で保存する 全データをCSV形式で保存する

都道府県コード	市区町村通番	市支庁郡名	区町村名	保険者コード	市区町村かな	平成16年面積	平成15年面積	備考	増減面積	摘要	参考値
						km ²	km ²		km ²		km ²
1	1	釧路市		6	くしろし	222.1	222.1		0	埋立等	
1	2	帯広市		7	おびひろし	618.94	618.94				
1	3	北見市		8	きたみし	421.08	421.08				
1	4	網走市		11	あひし	470.92	470.92				
1	5	紋別市		19	もんべつし	境界未定	境界未定				*830.36
1	6	根室市		23	ねむろし	512.59	512.59	AB			
1	7	網走支庁	東藻琴村	139	ひかしもことむら	184.38	184.38				
1	8	網走支庁	女満別町	140	めまんべつちよう	159.24	159.24				
1	9	網走支庁	美幌町	141	びほろちよう	438.36	438.36				
1	10	網走支庁	津別町	142	つべつちよう	716.6	716.6				
1	11	網走支庁	斜里町	143	しゃりちよう	736.97	736.97				
1	12	網走支庁	清里町	144	ぎよさとちよう	402.73	402.73				
1	13	網走支庁	小清水町	145	こしみずちよう	287.04	287.04				
1	14	網走支庁	端野町	146	たんのちよう	163.5	163.5				
1	15	網走支庁	訓子府町	147	くねっぶちよう	190.89	190.89				
1	16	網走支庁	置戸町	148	おけとちよう	527.54	527.54				
1	17	網走支庁	留辺蘂町	149	るべしべちよう	564.69	564.69				
1	18	網走支庁	佐呂間町	150	さろまちよう	404.99	404.99				
1	19	網走支庁	常呂町	151	ところちよう	278.29	278.29				
1	20	網走支庁	生田原町	152	いきたはらちよう	境界未定	境界未定				*269.10

ページが表示されました

インターネット

さらに、データベースに「医療機器データベース」、閲覧データに「平成 17 年度医療機器」を選択した画面を図 4 に示す。

図 4 実行画面例（2005 年度医療機器）

日医総研共有データベース - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

戻る 進む 検索 お気に入り

アドレス http://www.nihon-iseigaku-kenkyukai.co.jp/ksato/jmari_db.php

データベース一覧 医療機器データベース 年度選択 選択して下さい

データ選択 平成17年度医療機器

表示行数の設定 先頭から 100行 表示

結果表示

【限定表示】複数選択: Ctrl+クリック

機器大分類

検査・診断用関係
処置・手術用: 縫合類
処置・手術用: 注射・穿刺類
処置・手術用: 輸液類
処置・手術用: 洗浄器

再表示 全選択 全解除

平成17年度医療機器

256617件中、1～100件目を表示 次へ

このデータをCSV形式で保存する 全データをCSV形式で保存する

データ状態ステータス	商分類名	登録企業略称名	データ区分	初期登録日	最終更新日	国内総販売元業者: 企業略称名	JMDNコード	一般的名称	特定保守管理区分	修理区分	JANコード	薬事申請書上の販売名	商品名	商品名略称	規格	製品番号	材料価格調査用保険償還単位換算値	薬事法承認(認証)番号または許可番号	添付文書情報管理コード	材料償還請求区分	生物由来製品分類	償還価格合計	PIOの単位の保険償還単位数	商品画像ページリンク先	医療機関最小購入単位梱包インジケータ
0 10000	アイエスケ	アイエスケ	1	2005-06-01 00:00:00	2005-06-01 00:00:00	アイエスケ	35473000	血小板凝集測定装置	1	0	8 4517850000822	WBA-Neo (全血血小板凝集能測定装置)	WBA-Neo (全血血小板凝集能測定装置)	ISK-100		0 13BZ2368			0	0	0	0	0	0	0
0 10000	アイエスケ	アイエスケ	2	2005-06-01 00:00:00	2005-06-01 00:00:00	アイエスケ	33936000	染色体分析装置	1	0	8 4517850000716	ブラビー	ブラビー	コラーゲン	1 10038		0 13BZ2368			0	0	0	1	0	0

ページが表示されました インターネット

ページが表示されました インターネット

このデータは、列数が 256 列、行数が 256,617 件という大規模なデータであり、表示されるまで多少の時間を要するが、それ以外は問題なく動作していることが確認できた。ここでも、限定表示項目が動的に変更されており、カテゴリ定義表の連結機能も正常に動作していることが確認された。

(2) システムの基本機能

ここで、システムの基本機能が正常に動作するかを確認した結果を図5に示す。この画面では、図2で示したデータを利用し、東京都の市町村に限定し、表示列を[都道府県名][保険者名][療養の給付の計]と[1人当り療養諸費費用額]に設定して、[1人当り療養諸費費用額]に対して、降順に並べ替えかつ、抽出条件にそれが800,000円以上または490,000円以下という条件を設定して実行した画面である。その結果、東京都の市区町村の「1人当り療養諸費費用額」のトップ3とワースト3を同時に表示した結果となっている。この実行例1つで、基本機能の①～⑤までが正常に動作していることが確認できた。また、⑥の機能についても動作確認したところ、表示結果の内容を、CSV形式でダウンロードでき、全ての基本機能が正常に動作したことを確認できた。

図5 実行画面例（基本機能の動作確認）

日医総研共有データベース - Microsoft Internet Explorer

アドレス http://www.knsato/jmari_db.php

データベース一覧 国民健康保険の実態 年度選択 選択して下さい

データ選択 平成15年度療養の給付(診療費)及び療養諸費諸率(老人保健分)

表示行数の設定 先頭行 から 100行 表示

結果表示

抽出条件の設定

抽出条件1 1人当り療養諸費費用額 が 800000 以上 かつ または 再表示 条件1クリア

抽出条件2 1人当り療養諸費費用額 が 490000 以下 かつ または 再表示 条件2クリア

抽出条件3 指定しない が 指定しない かつ または 再表示 条件3クリア

抽出条件4 指定しない が 指定しない かつ または 再表示 条件4クリア

抽出条件5 指定しない が 指定しない かつ または 再表示 条件5クリア

並べ替え列の設定

並べ替え列1 1人当り療養諸費費用額 昇順 降順 再表示

平成15年度療養の給付(診療費)及び療養諸費諸率(老人保健分)
条件検索結果

[1人当り療養諸費費用額]が800000以上
または
[1人当り療養諸費費用額]が490000以下

6件中、1～6件目を表示

このデータをCSV形式で保存する 全データをCSV形式で保存する

都道府県名	保険者名	療養の給付(診療費)					1人当り療養諸費費用額
		受診率	1件当り日数	費用額			
				1件当り	1日当り	1人当り	
		日	円	円	円	円	
東京都	台東区	1954.63	3.19	34377	10772	671949	850460
東京都	文京区	2113.019	2.86	30406	10647	642489	810647
東京都	荒川区	1961.091	3.39	31570	9323	619120	807089
東京都	新島村	1256.902	2.43	35227	14513	442772	489370
東京都	檜原村	1381.414	2.59	28689	11060	396307	464322
東京都	御蔵島村	1224	2.31	35982	15574	440422	460694

国民健康保険中央会
都道府県国民健康保険団体連合会

ページが表示されました インターネット

4. 考 察

IT技術の発展により、総務省や厚生労働省などの国の機関をはじめ、様々な研究機関より、調査結果や統計報告がインターネットを通じて発信されている現在において、それらの情報をどう受け止め、どう活用できるかを考えることも、研究者にとっては必要である。また、長期的な研究においては、分析するデータ量も、年を重ねるにつれ、必然的に肥大化し、作業も複雑になっていくが、データの管理方法については、現段階では問題となっていないため、真剣に取り組まれていないのが現状である。

例えば、今回利用した全国の市区町村別の面積データの場合を考えると、国土地理院のホームページ内に、そのデータが公開されており、早速そのデータを利用しようとする。しかし、そのホームページでは、データが都道府県ごとに別々に表示されるようになっているため、全ての都道府県のページを閲覧し、それらのデータをまとめて保存できる標準化フォーマットを考えてから、データを保存していかなければ、全てのデータを標準的に扱えない。とくに市町村のデータについては、市町村合併のため、年度によって市町村自体が変化することがあり、データの整理の仕方が大変重要となる。従って、年月が経過すればするほど、データは複雑化し、肥大化することとなる。

一般的に、複数の人の目で、データを確認することは、そのデータの間違いの発見や、妥当性が考慮されるということであり、そのデータの信頼性が向上することにつながる。とくに、研究者によっては、統計ソフトに依存し、結果のみを信用して報告することも見受けられるが、本当に重要なのは、統計処理を行うまでの過程である。つまり、データの収集方法や、クレンジング処理であり、その過程が妥当でなければ、統計処理は無意味なものとなる。

実際にデータの管理を行うとなると、データ量、利用方法および、利用可能なIT技術を考えれば、有効な手法の1つとしてデータベース化を考えることとなる。本研究は、このようなデータベースの構築を行う場合に、そのデータの種類やフォーマットに関係なく、それらを共通に管理できるような手法を考案するものである。

通常、データベースの構築は、データ表ごとに定義情報を作成してから、テーブルを作成するのが基本であるが、利用するデータベースの種類によりその定義情報が微妙に異なるため、その実現には、データベースアプリケーション特有の規約に基づいて、それを作成しなければならないという制限がある。今回考案した手法は、その定義情報そのものを標準化した定義表を利用して、その規約を独自に制御し、データの統合管理を実現しようというものである。

また、カテゴリデータ連結定義表を作成し、データベースのリレーション機能を実現させたことにより、データ表間での結合処理も可能となるため、年度間でのデータ表の結合処理や、市町村別の面積データ表との結合処理など、必要性に応じて、自由なデータ表の作成が可能となるように設計している。

これらの定義表が存在しない状態で、システム開発を業者に委託した場合、単純にデータを読み込んで表示するシステムは出来上がる。また、「表示幅」や「公開可否」などの設定情報は、その仕様を明確に伝えておけば、その対応も可能である。しかし、そのような設定は、発注者から要求しない限り、プログラムソースで対応することが多いため、仕様変更があった場合、プログラムソースを直接修正することとなり、結局は業者に依存したシステムとなってしまう。さらに、追加したいデータがある場合、追加用に新たなプログラムが必要となるため、その都度、開発費が積み重なる。

今回考案した定義表を利用すれば、データの追加や細かな設定が必要となったときに、データベ

ースに登録している定義表を直接修正するだけで、それらの対応が可能となり、システム内のプログラムソースを修正する必要がなくなるため、業者に依存せず、研究機関内で、それらの処理に対応することが可能となる。

実際、サンプルデータを利用してシステムの動作を検証したところ、限定表示や、並べ替え機能など、一通りの基本機能の動作確認ができ、本定義表を利用すれば、汎用性の高いシステムが構築可能であることが示された。また、結果に示すとおり、データを実際に条件で絞り込んで閲覧することで、分析結果を確認することも可能であり、さらに本格的な分析をしたい場合は、閲覧している結果を CSV 形式でダウンロードし、専用の分析ソフトで活用できるため、研究を支援するツールとして利用できると考えられる。また、Windows の基本機能であるコピー、ペースト機能も利用できるため、結果の画面を他アプリケーションにも簡易に移植することも可能である。

サンプルデータで利用した医療機器データについては、Windows で主要な表計算ソフトである Excel で扱えるデータ数の制限を大きく超えており、分析するには、データベースソフトや、統計処理ソフトなどの特殊なアプリケーションを利用しなければならない。しかし、どのようなデータであっても、本システムに登録することができれば、システムの基本機能を利用して、必要なデータを即座に指定し、取得することが可能となる。とくに重要なのは、Web ブラウザのみで本システムの利用を実現した点であり、利用者が多ければ、新しい要望やデータの種類なども増え、システムの改良に役立つと考えられる。

本システムを開発するにあたり、次の 2 つの条件を考慮して開発を行っている。第一の条件として、特別なソフトを必要としない利用できるシステムの構築である。この条件を満たすため、世界で最も利用人口の多い Windows ユーザをシステムの利用者として想定し、利用するアプリケーションは、OS インストール時に組み込まれる Web ブラウザのインターネットエクスプローラで動作するシステムを構築した。また、第二の条件として、標準的なデータベースを扱うことができ、少しでも安価に開発ができることを実現するために、Web サーバの OS を Linux、データベースには PostgreSQL を利用し、コストパフォーマンスも考慮したシステムに仕上がった。

本定義表は、まだ発展途上であり、テーブル間での動的な連結処理は実現しておらず、今後、定義表の改良が必要である。また、複数のデータベースサーバにも対応するためには、標準化定義表にサーバへのアクセス定義情報が必要である。これからさらに新しいデータを追加していき、本定義表を完全なものに仕上げることをこれからの課題としたい。

また、あらゆる研究分野における IT 化の発展を考えると、データの意味の理解とデータベースシステム技術の両方を兼ね備えた専門の知識が必要である。このような知識を持った人材を育成することも、これから先の重要な課題である。

過去に、XML (eXtensible Markup Language) を利用してデータベースの標準化を行う手法が報告されているが⁴、本研究の手法は、技術的にもそれほど高度でなく、データベースシステムの基幹的な部分の手法であるので、一般的に広く活用できると考えられる。

さらに、本システムを公開用システムとして拡張すれば、研究機関専用の公開情報データベースシステムとして活用することが可能であり、研究で利用したデータを公開すれば、研究に興味を持った利用者が独自に結果を確認できるようになる。その研究の健全性を証明でき、その結果に説得力を持たせる 1 つの手段として活用できる。

本研究で考案した定義表を利用することにより、様々な種類のデータの標準化を実現し、異種のデータを統括して管理することが可能となった。また、定義表に属性情報を登録することで、数値

情報や、カテゴリ情報の判別ができ、単純な集計機能だけでなく、クロス集計機能も実現可能となり、拡張性の高いシステムの基本的な標準化規準として利用できると考えられる。さらに機能を拡張し、相関係数や、多変量解析などの統計機能に対応した分析ツールの作成も可能であり、あらゆる研究分野で広く利用できると思われる。

文 献

- 1 J.A.Zachman : A framework for information systems architecture. IBM System Journal 26 (3) 276-292,1987
- 2 国民健康保険中央会・都道府県国民健康保険団体連合会：平成 15 年度国民健康保険の実態、2004
- 3 国民健康保険中央会・都道府県国民健康保険団体連合会：平成 16 年度国民健康保険の実態、2005
- 4 細川卓哉・高坂貴弘・遠里由佳子・伊達進・下條真司・松田秀雄：データグリッド技術を用いた異種分子生物学データベースの連携手法、情報処理学会研究報告 (91)、37-40、2003