健康診断データ後利用システムの開発と評価
―ユニケージ®開発手法による結合／抽出ロジックの実装―

大貫　亮**　藤井　香*　松本　可愛*
高橋　綾*　清　奈帆美*　櫻井　勉*
金子　康樹**　押見　淳**　河邊　博史*
齊藤　郁夫*

当大学における健康診断システム（IDST：Information and Database of health care Service Tools）は，毎年度見直される可能性のある検査項目について追加や変更などが生じた場合でも対応できる，汎用的かつ単年度管理可能なデータ構造にて開発され，教職員健診，学生健診，特定業務従事者健診，雇入時健診の4健診で全面利用している。

しかし，IDSTの汎用検索機能の処理が非常に遅く，本来開発目的の一つであるイレギュラーな統計処理や疫学研究などで利用するデータ出力が難しく，データの利活用ができていない。そこで，これら一連のパフォーマンス問題における業務への影響を解消するために，ユニケージ開発手法[1)注1)]を用いてデータ後利用システムを構築し，結合ロジック[注2)]，抽出ロジック[注3)]を新たに実装したので評価したい。

注1）ユニケージ開発手法（Unicage software development method）；UNIX系オペレーティングシステム上（特にLinux上）において，シェルスクリプトのみでシステムを開発する手法。通常のテキストファイル形式でデータを持たせ，独自開発されたフィルタコマンド群をパイプラインで接続して並列処理する技術を用いる。

注2）結合ロジック；出力されている複数の検査データから年度及び健診種類ごとに一受付一レコード（≒一人一レコード）にマージする機能。

注3）抽出ロジック；結合ロジックにてマージしたCSVファイルを対象に，WEBクライアントからの任意の条件によるリクエストに対して，ヒットしたデータをCSVファイル形式で出力する機能。

対象と方法

業務用ネットワーク上での健康診断システムにおけるデータ後利用システムとして，結合／抽出ロジックのパフォーマンスを中心としたシステム全体の評価を行った。

IDSTを利用した全項目CSV出力と，今回開発した抽出ロジックの機能を用いた同一データのCSV出力における処理時間を計測し，パフォーマンスの比較を行った。結合／抽出ロジックにお

* 慶應義塾大学保健管理センター　** 慶應義塾インフォメーションテクノロジーセンター

いては，2006年度～2011年度各年度の教職員健診，学生健診，特定業務従事者健診，雇入時健診の全項目出力時間，2 ～ 6年度分の複数年度での全項目出力時間（秒）を計測した。

なお，IDST，結合／抽出ロジックそれぞれの処理に用いたサーバはその用途やシステムの特性上の観点から仕様上異なるが，IDSTで利用した基幹サーバのマシンスペックは，2012年度学生健康診断実施時におけるパフォーマンス監視結果でみると，CPU使用率は最大約20used%であり，十分なリソースの余力があったことを確認している（図1）。このことから，処理時間の差は，データ構造上の問題，及びそれに付随したアプリケーションの実装上の差であり，基幹サーバのマシンスペック上の問題とは考えにくい。

システム概要

1．後利用システム

1）システム形態（図2）

システム全体を，ユーザーインターフェイス層（クライアント側），ビジネスロジック層（サーバ側），データベース層（サーバ側）の3階層に処理を分割して構成するWEB3階層型システムとした。

本システムでは，①ユーザーインターフェイス層をCSV出力用画面（クライアントPC上のWEBブラウザ），②ビジネスロジック層を結合／抽出ロジック（サーバ上のシェルスクリプト），③データ層を各検査項目データ（サーバ上のCSVファイル）の3つの要素で3階層とした。

2）結合／抽出ロジック　サーバ構成

結合／抽出ロジック専用環境として独立したマシン上に構築した。

a）ハードウェア

IBM System x3550 M3 （1台）

CPU: インテル（R）Xeon（R）プロセッサーE5620（4コア/8スレッド）2.40GHz x2

メモリ：48GB（4GB x12）

ディスク：300GB（内蔵型SAS）x4，実効容量600GB（RAID10構成）

b）OS／ミドルウェア

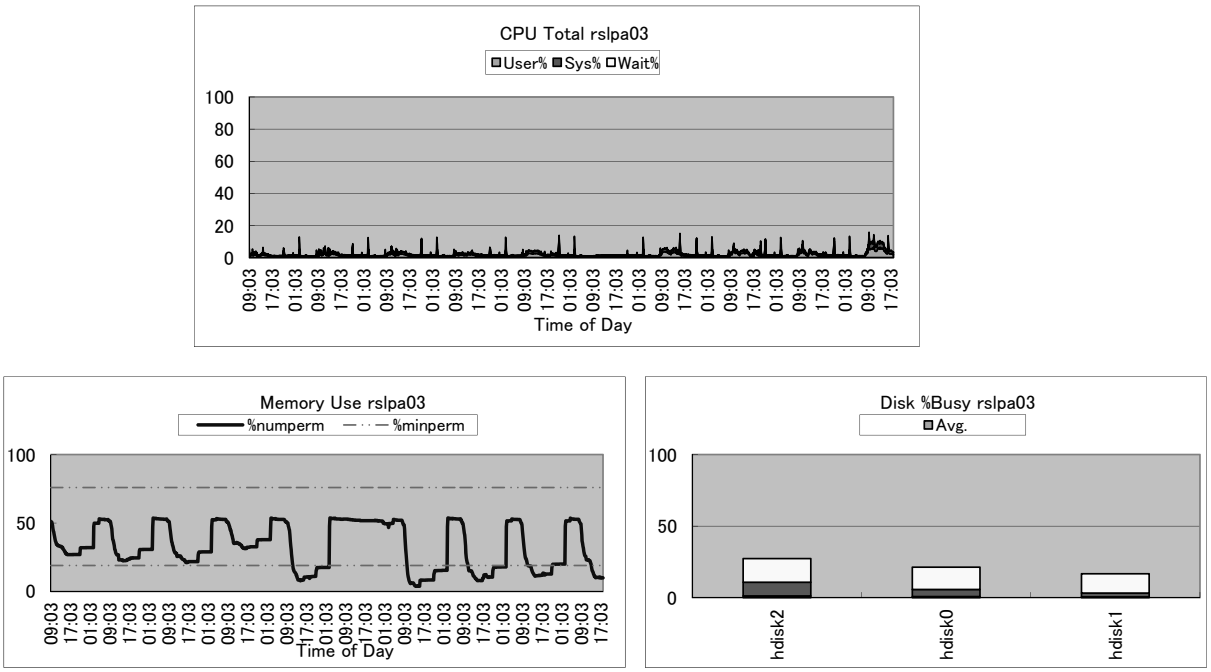

図1　2012年度学生健診実施時の基幹サーバ（rslpa03）パフォーマンス監視結果

サーバ OS として CentOS（フリーの Linux ディストリビューション）バージョン 6 を利用した。さらに，本システムのコアとなる usp Tukubai® コマンド群（有限会社ユニバーサル・シェル・プログラミング研究所）をインストールした。なお，リレーショナルデータベース管理用のミドルウェアは特に利用していない。クライアント PC からの WEB アクセス用には，OS 標準の Apache HTTP Server（オープンソース・ソフトウェア）を利用した。

c）パフォーマンス要件

全データの結合処理および各種抽出処理について，複数年度にまたがる処理であっても 5 分以内に終了することを目標値とした。

3）データ構造と結合／抽出ロジック（図 3）

プログラムは，usp Tukubai コマンド群を利用したシェルスクリプトのみで記述した。

IDST のデータ構造は，健診種類ごとにテーブルが存在し，同一年度，同一健診種類でも，一人に紐付くべきデータが不定数レコード存在しており，検査ごとに CSV 出力する機能を持っている。そこで，結合ロジックは，IDST から出力された各検査項目の CSV を対象に，年度及び健診種類ごとに一受付一レコード（≒一人一レコード）にマージする処理を実装した。

また，抽出ロジックは，結合ロジックにてマージした CSV ファイルを対象に，WEB クライアントからの任意の条件によるリクエストに対して，ヒットしたデータを CSV ファイル形式で出力する機能を実装した。データ結合ロジックの仕様として，同一健診内における複数回検査による列ずれが発生しないように，また，結合対象となる各検査項目データについては，個人属性など同じ内容の列があった場合に，それらも重複しないように制御した。

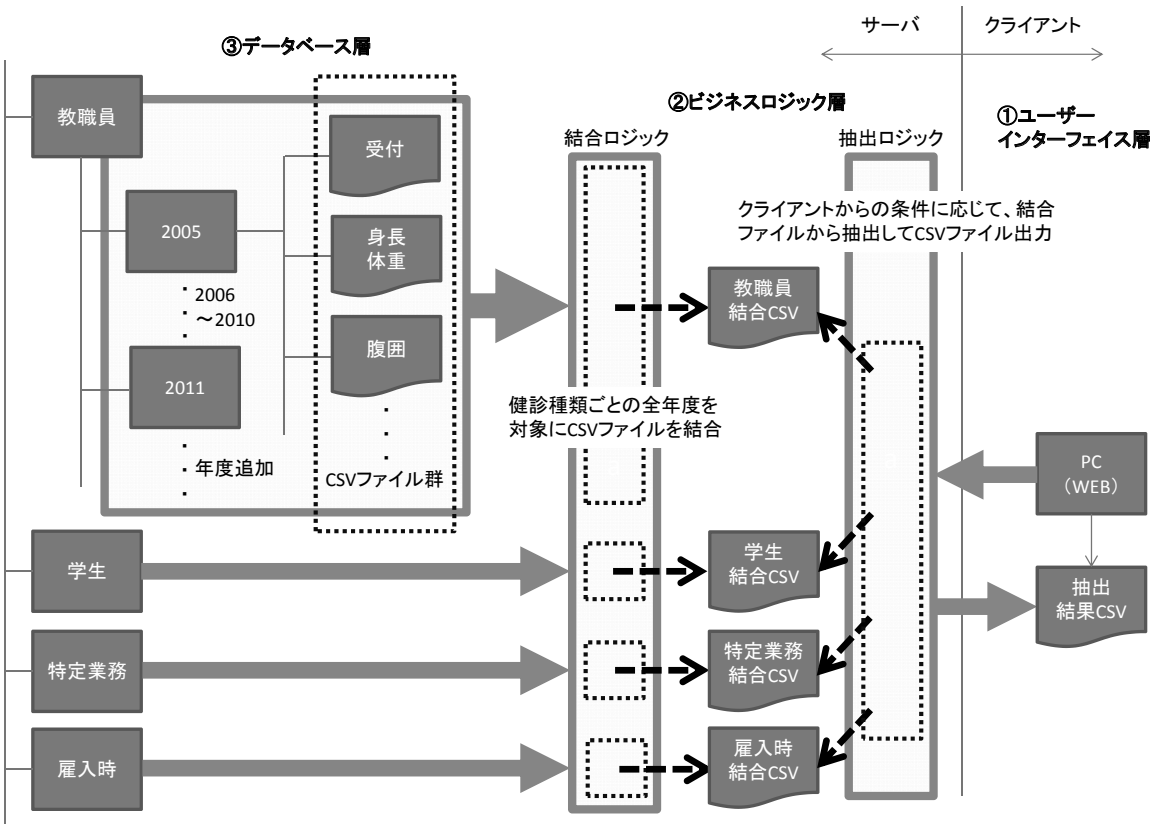

図２　WEB３階層と処理フロー

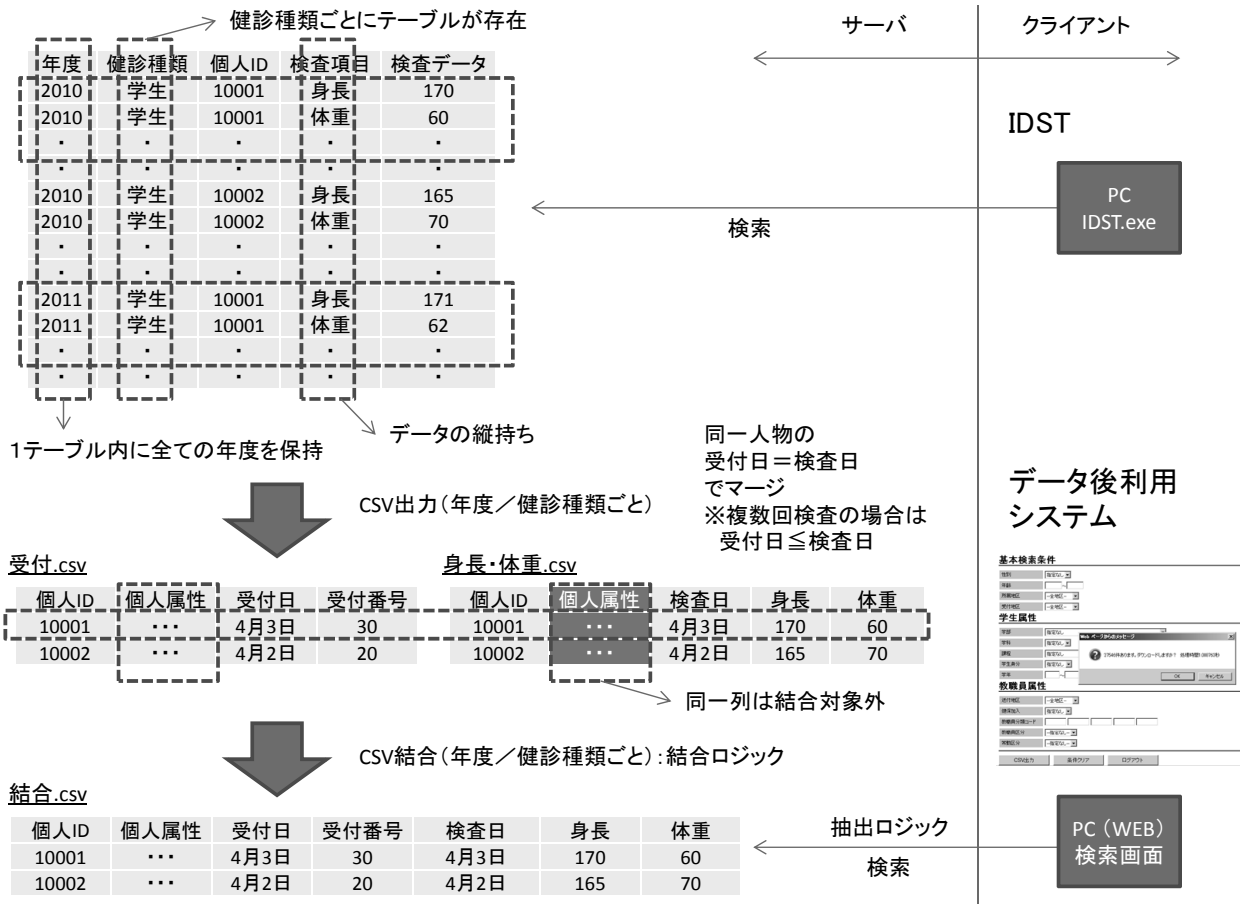

図３　データ構造と結合／抽出ロジック

結合のキーとなるのは健診受付日および各検査日であり，同一人物が同一種類の健診を二度受診したとしても，複数回検査や外部受診についても直近の健診に紐付くように設定した。受付日のないデータは不要データの位置付けで精査され，元となる健康診断システムへのフィードバックを可能とした。

4）クライアント利用

業務用PC標準のWEBブラウザであるInternet Explorer 9（日本マイクロソフト株式会社）での利用を想定した。

5）ネットワーク・セキュリティ

ファイアウォールによる制限が設けられている業務用LAN（有線）を利用した。また，サーバ側でのアカウント管理，およびシステム利用時の認証機構と，使用端末登録による利用制限を設けた。

成　　績

1．IDSTでのCSV出力，抽出ロジックによるCSV出力のパフォーマンス評価（表１）

IDSTでの全項目CSV出力の処理時間をみると，教職員健診では，2008年度　5,575件24分（1,440秒），2009年度　5,780件25分（1,500秒）であった。一方，抽出ロジックでのCSV出力の処理時間は，同条件で2008年度0.3秒，2009年度0.3～0.4秒であった。

2．複数年度を対象とした，抽出ロジックによるCSV出力のパフォーマンス評価（表２）

抽出ロジックを用いたCSV出力の処理時間を２年度分～６年度分の複数年度で計測した。教職員健診では，2006～2011年度６年度分33,847件1.9秒，学生健診では，2006～2011年度６年度分173,958件6.6～6.7秒であった。

表 1　IDST での CSV 出力，抽出ロジックによる CSV 出力のパフォーマンス評価

健康診断システム（IDST ＊1 ）における CSV 処理時間の計測

＊1 Information and Database of health care Service Tools（慶應義塾にて開発した大規模大学向け健康診断システム）

教職員健診	件　数	1 回目	2 回目
2008年度	5,575	1,440.000	1,380.000
2009年度	5,780	1,500.000	1,500.000

（秒）

学生健診	件　数	1 回目	2 回目
2008年度	28,687	5,100.000	5,040.000
2009年度	29,709	5,700.000	5,700.000

（秒）

抽出ロジックにおける CSV 処理時間の計測

教職員健診	件　数	1 回目	2 回目
2006年度	5,130	0.272	0.272
2007年度	5,302	0.295	0.309
2008年度	5,575	0.329	0.332
2009年度	5,780	0.341	0.360
2010年度	5,834	0.335	0.354
2011年度	6,228	0.382	0.388

（秒）

学生健診	件　数	1 回目	2 回目
2006年度	27,546	1.006	1.083
2007年度	27,761	1.015	1.029
2008年度	28,687	1.124	1.172
2009年度	29,709	1.187	1.096
2010年度	30,124	1.180	1.244
2011年度	30,131	1.217	1.362

（秒）

特定業務健診	件　数	1 回目	2 回目
2006年度	1,450	0.068	0.069
2007年度	1,480	0.067	0.071
2008年度	1,674	0.086	0.082
2009年度	1,762	0.074	0.082
2010年度	1,775	0.077	0.078
2011年度	1,941	0.079	0.080

（秒）

雇入時健診	件　数	1 回目	2 回目
2006年度	373	0.050	0.048
2007年度	349	0.049	0.049
2008年度	341	0.049	0.048
2009年度	325	0.050	0.048
2010年度	372	0.050	0.052
2011年度	479	0.055	0.055

（秒）

結合ロジックにおける CSV 処理時間の計測（参考*）

教職員健診	件　数	1 回目	2 回目
2006年度	5,130	2.404	2.383
2007年度	5,302	2.460	2.520
2008年度	5,575	3.190	3.256
2009年度	5,780	3.244	3.334
2010年度	5,834	3.032	3.111
2011年度	6,228	3.441	3.385

（秒）

学生健診	件　数	1 回目	2 回目
2006年度	27,546	8.167	8.317
2007年度	27,761	8.242	8.232
2008年度	28,687	8.521	8.512
2009年度	29,709	8.809	8.758
2010年度	30,124	10.208	10.262
2011年度	30,131	10.668	10.302

（秒）

特定業務健診	件　数	1 回目	2 回目
2006年度	1,450	0.553	0.561
2007年度	1,480	0.620	0.581
2008年度	1,674	0.587	0.596
2009年度	1,763	0.653	0.661
2010年度	1,775	0.662	0.681
2011年度	1,941	0.743	0.675

（秒）

雇入時健診	件　数	1 回目	2 回目
2006年度	373	0.472	0.471
2007年度	349	0.477	0.466
2008年度	341	0.466	0.465
2009年度	325	0.543	0.477
2010年度	372	0.500	0.499
2011年度	479	0.530	0.528

（秒）

＊結合ロジックは IDST における CSV 処理時間との比較が出来ないため参考値

表２　複数年度を対象とした，抽出ロジックによる CSV 出力のパフォーマンス評価

教職員健診	件　数	1 回目	2 回目
2006-2007 2 年度分	10,430	0.508	0.533
2006-2008 3 年度分	16,005	0.815	0.843
2006-2009 4 年度分	21,785	1.185	1.188
2006-2010 5 年度分	27,619	1.547	1.519
2006-2011 6 年度分	33,847	1.859	1.856

（秒）

学生健診	件　数	1 回目	2 回目
2006-2007 2 年度分	55,307	1.998	2.009
2006-2008 3 年度分	83,994	3.079	3.060
2006-2009 4 年度分	113,703	4.392	4.052
2006-2010 5 年度分	143,827	5.464	5.370
2006-2011 6 年度分	173,958	6.659	6.583

（秒）

教職員健診	件　数	1 回目	2 回目
2007-2008 2 年度分	10,876	0.619	0.585
2007-2009 3 年度分	16,656	0.899	0.905
2007-2010 4 年度分	22,490	1.416	1.211
2007-2011 5 年度分	28,718	1.577	1.623

（秒）

学生健診	件　数	1 回目	2 回目
2007-2008 2 年度分	56,448	2.099	2.086
2007-2009 3 年度分	86,157	3.273	3.158
2007-2010 4 年度分	116,281	4.553	4.419
2007-2011 5 年度分	146,412	5.650	5.398

（秒）

教職員健診	件　数	1 回目	2 回目
2008-2009 2 年度分	11,355	0.653	0.686
2008-2010 3 年度分	17,189	1.031	0.972
2008-2011 4 年度分	23,417	1.382	1.389

（秒）

学生健診	件　数	1 回目	2 回目
2008-2009 2 年度分	58,396	2.230	2.169
2008-2010 3 年度分	88,520	3.377	3.378
2008-2011 4 年度分	118,651	4.661	4.571

（秒）

教職員健診	件　数	1 回目	2 回目
2009-2010 2 年度分	11,614	0.733	0.666
2009-2011 3 年度分	17,842	1.289	1.024

（秒）

学生健診	件　数	1 回目	2 回目
2009-2010 2 年度分	59,833	2.329	2.316
2009-2011 3 年度分	89,964	3.562	3.809

（秒）

教職員健診	件　数	1 回目	2 回目
2010-2011 2 年度分	12,062	0.680	0.712

（秒）

学生健診	件　数	1 回目	2 回目
2010-2011 2 年度分	60,255	2.455	2.402

（秒）

考　察

今回実装したユニケージ開発手法を用いた結合／抽出ロジックは，IDST におけるデータ管理方法とデータ構造の違いはあるが，両システムのパフォーマンス比較より，圧倒的に処理時間が短く，優位であることが示された。ユニケージ開発手法のシステム構成上の特徴と優れた点として，データ管理におけるミドルウェア＝データベースを必要とせず，プログラムにおける複雑な記述が発生しないことが挙げられる。プログラムは usp Tukubai コマンド群を用いたシェルスクリプトのみ，データはテキストファイルのみで構成されるという非常にシンプルで無駄の無い構成であるため，プログラムの実行処理は著しく速く，開発当初の処理時間目標値である 5 分をはるかに上回る結果が達成できた。職員の業務負担になることなく，さらなる業務面での効率化や蓄積されたデータの後利用における研究面での発展が期待できた。

また，システムの特性上，CSV ファイルを非常に容易に扱えることで，他システムとの連携が柔軟であるため，今後の診療系システム，画像系システム，医事会計システムなどの導入に応じてそれらシステムとの連携を通じ，保健管理センターとしての健康診断結果への利活用だけでなく，診療所の観点からのビッグデータ分析にも有用となることであろう。

ユニケージ開発手法はミドルウェアを利用しないため，不要なランニングコストやアップデート作業などが発生せず，データもプログラムもテキストファイル形式なので，今後のハードウェアの入れ替えなどにおいても単純なファイルコピーだけでシステム移行が済むことから，将来的な管理面での負荷も劇的に軽減されることが期待された。

一方で，データ自体がデータベースではなくファイルで管理されているという特徴から，サーバ自体へのアクセス制限と権限設定については，他のシステム同様に十分な配慮が必要である。

総　括

1．現健診システムである IDST の汎用検索機能の処理が非常に遅く，データの利活用ができていない。そこで，これら一連のパフォーマンス問題における業務への影響を解消するために，データ後利用システムを構築し，結合／抽出ロジックを実装した。

2．ユニケージ開発手法を用い，プログラムは，usp Tukubai コマンド群を用いたシェルスクリプトのみで記述した。

3．IDST での全項目 CSV 出力の処理時間をみると，教職員健診では，2008 年度　5,575 件 24 分（1,440 秒），2009 年度　5,780 件 25 分（1,500 秒）であった。一方，抽出ロジックでの CSV 出力の処理時間は，同条件で 2008 年度 0.3 秒，2009 年度 0.3 ～ 0.4 秒であった。

4．両システムのパフォーマンス比較より，圧倒的に処理時間が短く，優位であることが示された。さらなる業務面での効率化や蓄積されたデータの後利用における研究面での発展が期待できた。

文　献

1）ユニケージ開発手法：http://www.usp-lab.com/methodology.html