2004年2月3日

公開講座「医療情報システム工学」第三回

臨床情報システムでのデータウェアハウスの設計

松谷 司郎

東京大学大学院医学系研究科

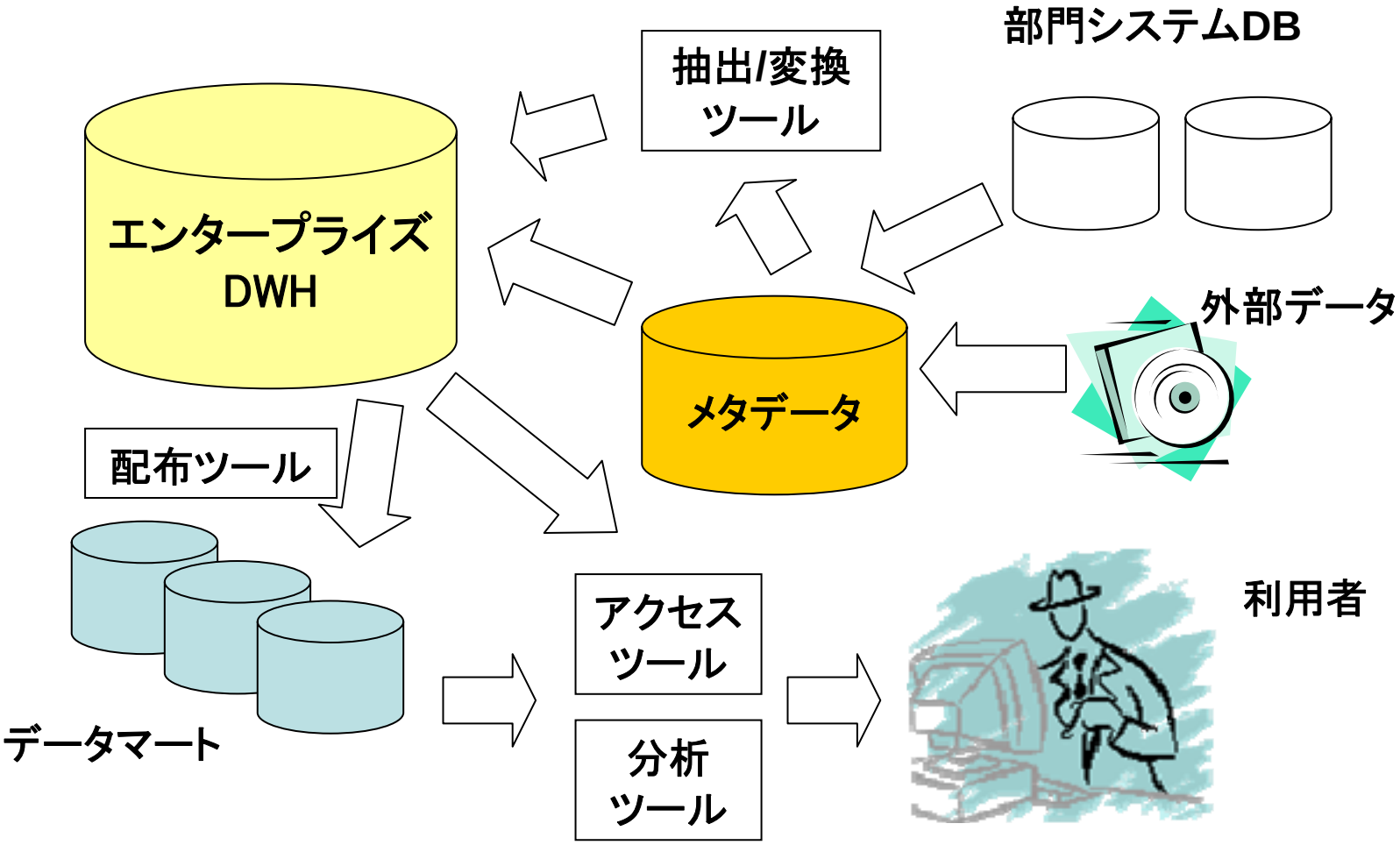
クリニカルバイオインフォマティクス
研究ユニット(CBI)

内　容

- データウェアハウスのイメージ
- 臨床系のデータウェアハウス例
- データウェアハウスの簡単な歴史
- 臨床研究と病院内の情報
- 臨床研究支援のためのDWH設計
- まとめ

- データウェアハウスのイメージ

データウェアハウスのイメージ



臨床研究支援のためのDWHの必要性

- 研究目的に応じたデータベースを構築しないと研究が進まなくなってきたが、できるだけ集中管理をすれば効率的である
- 病院内で発生する情報が急激に電子化されてきており、医療の質を高めたり、病院経営の効率化のためには、組織全体あるいは他組織との情報共有・交換が必要となってきた
- 文献や経験に頼った研究から情報処理を用いた科学的医療（臨床）の推進が必要となってきた

- 臨床系データウェアハウスの例

臨床系のデータウェアハウス例

- PubMedで ”clinical data warehouse” で検索
- 主な利用目的
  - 症例分類
  - Decision Support System (DSS)
  - 診療ガイドラインへの適合度チェック
  - 院内感染制御
  - 医学教育
  - 薬物副作用の頻度・コスト分析

例１：症例分類

- 小児患者の”呼吸困難”という症状を「細気管支炎」、「細菌性肺炎」、「喘息」に分類するアルゴリズムを開発
- Clinical Data Warehouse (CDW)　上のフリーテキストのデータ項目を対象
- Feasibility study だが分類は自動化できると考えている
  - 「Feasibility of using a large clinical data warehouse to automate the selection of diagnostic cohorts.」*(R.Stephen et al.; Proc AMIA Symp. 2003)*

例2：Case-based Reasoning DSS

- ガイドラインに適合しない新患が来院した場合、熟練の医者は治療方針の判断基礎として過去に自分が扱った類似の症例を持った患者の診療例での経験を使う
- このDSSシステムは新患の症例をガイドラインと比較、次に他の症例と比較、さらに類似症例の検索を自動的に行う予定
- 今後は”類似“の指標の定義などをしていく予定
  - 「Modelling of a case-based retrieval system for oncology.」 *(D.Rossille,Stud Health Technol Inform. 2003; 95: 565-70.)*

例3：ガイドラインへの適応度解析

- ガイドラインへの適合度を測る研究や商用製品は多くない。大量のデータを入力させて、evidence-basedなガイドラインへの適合度のプロファイルをシステマティックで詳細に作成する製品へのニーズがある
- まだ開発中であるがmedical and pharmacy claimsを入力としてガイドライン適合度プロファイルを自動作成するツールを開発した
- プロファイルの詳細を興味のある観点から見ることができるようにするため、DWHを構築した
  - 「An automated tool for an analysis of compliance to evidence-based clinical guidelines.」*（B. A. Metfessel; Medinfo. 2001; 10(Pt 1): 226-30.）*

例4：院内感染制御

- 3つの関連病院の病院情報システムからデータを収集してCDWを構築した
- 抗菌剤の耐性のモニタリング、抗菌剤の使用の計測、院内血液感染の検知、感染のコストの計測、抗菌剤処方のミスを検知することが目的
- 病院内のトランザクション（あるまとまった処理単位）を保管しているサーバ上のデータを使用して、研究や品質管理のサポートをすることは、病院のシステムにとって発展的なテーマである
  - 「Development of a clinical data warehouse for hospital infection control.」*（M. F. Wisniewski ; J Am Med Inform Assoc. 2003 Sep-Oct; 10(5): 454-62.）*

例５：医学教育

- 患者情報を格納した教育用DWHを構築し、医学部２年生の既存の「地域医療」の教育コースの中で使ってみた（１回/週、選択式問題形式）
- 学生に対する課題は風土病の罹患率、診療のパターン、患者特性についての検索・解釈をさせること
- 教官と学生の間には、この科目に教育用DWHを使った課題を取り入れる、という了解がとれつつある（DWHを使った課題は概ね好評である）

  – 「Introducing an academic data warehouse into the undergraduate medical curriculum.」(*J.A. Lyman; Proc AMIA Symp. 2002; : 474-8*)

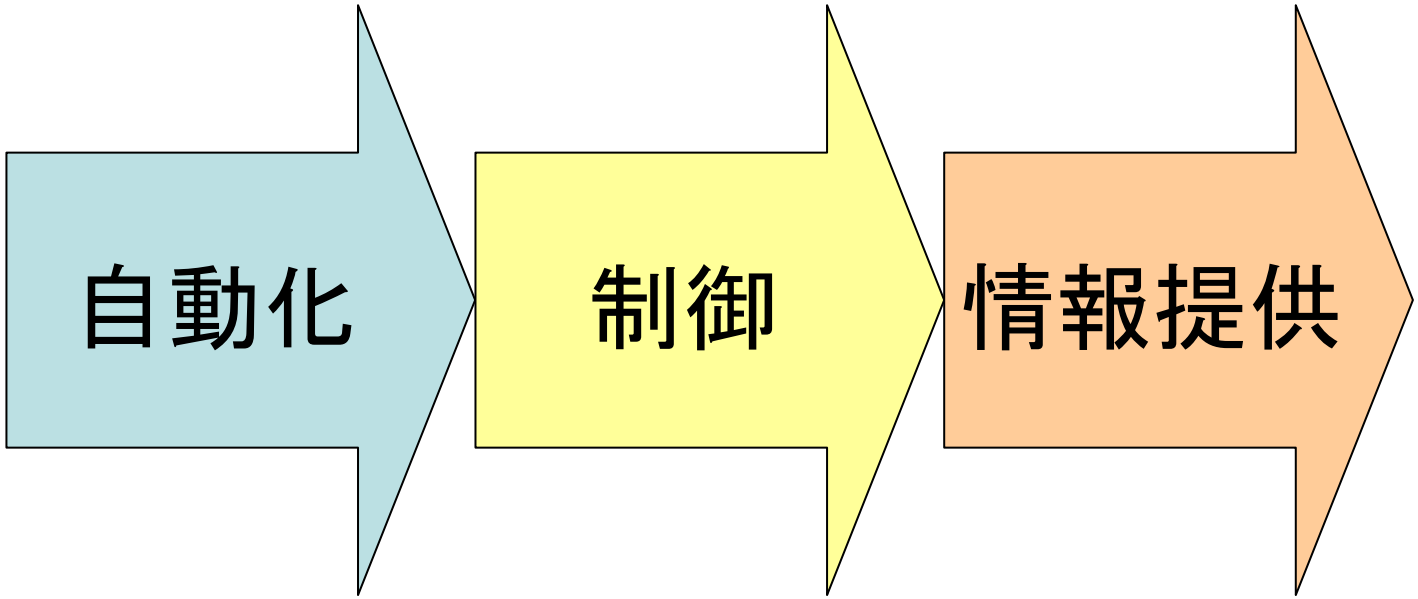
例６：薬物副作用の頻度・コスト分析

- 入院患者の薬物副作用とコストのretrospective analysisによる見積り
- 公開されている評価基準から潜在的な副作用を見つけ出し、CDWを用いて副作用の可能性がある入院患者と症例を検出
- 評価基準から、副作用イベントと予防可能だった副作用イベントの数、および副作用のコストと余分な入院日数の見積り（１００入院あたり予想より多い10.4～11.5件発生の可能性）
- 副作用検知と副作用の影響度を見積もるために使われているルールと基準を評価するのに、今回の方法がかなり役に立つ

  – 「Using a clinical data repository to estimate the frequency and costs of adverse drug events.」(*J.S. Einbinder; Proc AMIA Symp. 2001; : 154-8.*)

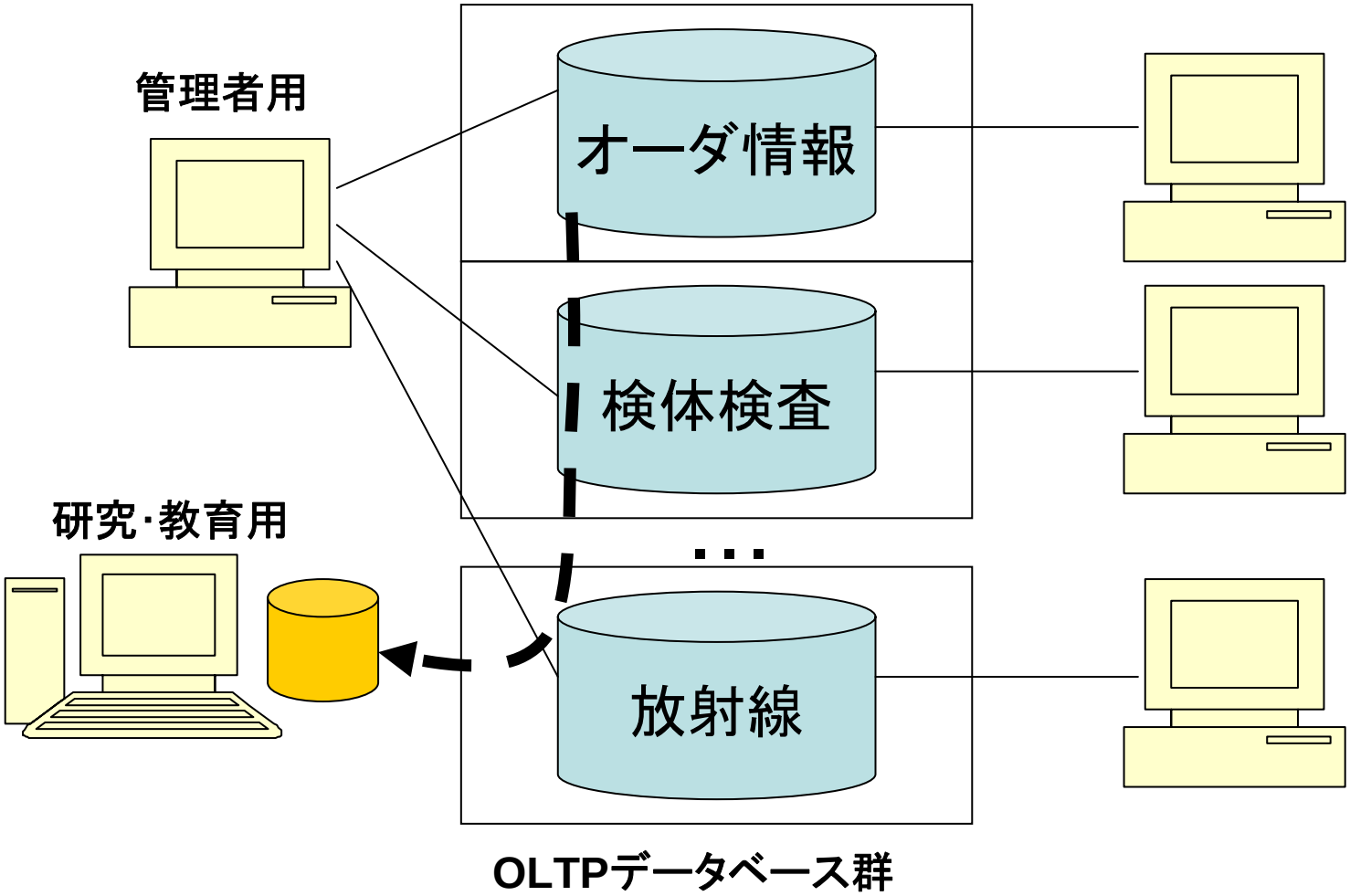
- データウェアハウスの簡単な歴史

データウェアハウス以前（1）



（「ウォルマートに学ぶデータ・ウェアハウジング」, Paul Westerman著, 翔泳社, 一部改）

データウェアハウス以前（2）

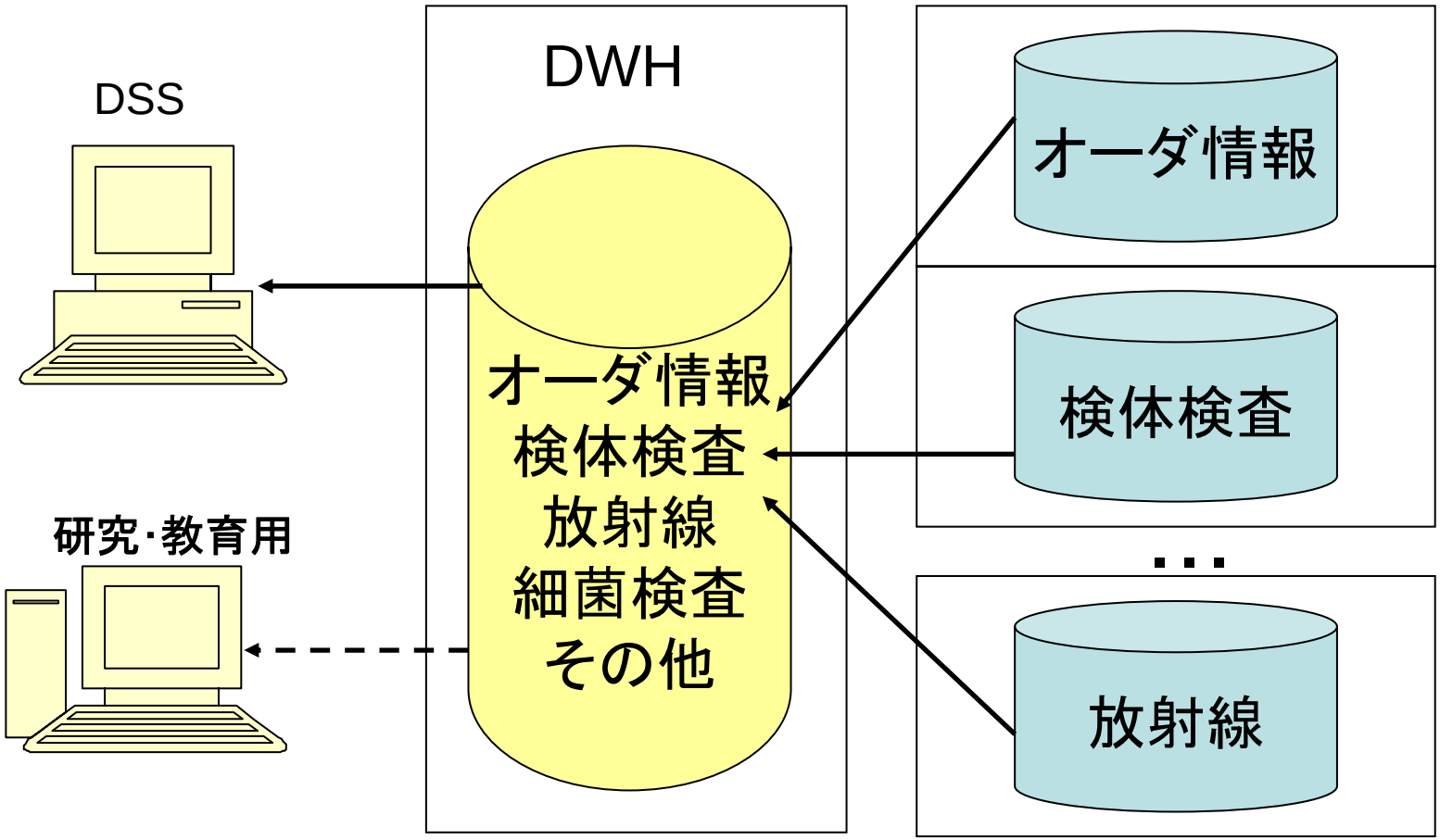


（「ウォルマートに学ぶデータ･ウェアハウジング」, Paul Westerman著, 翔泳社, 一部改）

データウェアハウス誕生の契機

- 自部門にない他部門のデータを取り込んで使いたいが、データ項目の長さや単位、日付型が異なっていたり、異なるDB間でデータの一貫性がない
- 一貫性のないDB間のデータを利用者側が整理・変換するのは煩雑である
- 基幹システム（オーダエントリなど）とは別に研究目的でシステム（DB)を構築することが増えてきた
- 本来業務とは違ったデータの利用・活用ができる（履歴を管理する、など）ため、有用性が認められてきた

データウェアハウスの構成



（「ウォルマートに学ぶデータ･ウェアハウジング」, Paul Westerman著, 翔泳社, 一部改）

データウェアハウスの主目的

「情報化された証拠を基にした意思決定を支援すること」

（Evidence-based Decision Making Support)

- 臨床研究と病院内の情報

臨床研究の定義

- 医療における疾病の予防方法、診断方法及び治療方法の改善、疾病原因及び病態の理解並びに患者の生活の質の向上を目的として実施される医学系研究であって、人を対象とするもの（個人を特定できる人由来の材料及びデータに関する研究を含む）。
  - 診断及び治療のみを目的とした医療行為並びに他の法令及び指針の適用範囲に含まれる研究を除く。

(「臨床研究に関する倫理指針の概要」、厚生労働省
http://www.mhlw.go.jp/shingi/2003/07/s0729-7h.html)

コンピュータを使った臨床研究

- 臨床研究の倫理規定に則る
- 患者データの扱い
  - 個人情報保護法が適用される施設（5000件以上のカルテ保有）は，診療情報を研究に利用する場合は「目的外利用」に該当するので原則として本人の同意を得なければならない．独立行政法人で1,000人以上の個人情報を持つ場合、届出が必要
  - 学術機関における研究利用は個人情報保護法の除外規定とされているのが現状であるが、2004年度より国立大学の独立行政法人化に伴い、上記の義務が発生

個人識別情報の管理

- 個人識別情報
  - 個人を識別できる情報を含む情報
- 個人情報
  - 個人識別情報以外の情報
  - 「個人医療情報」も個人情報

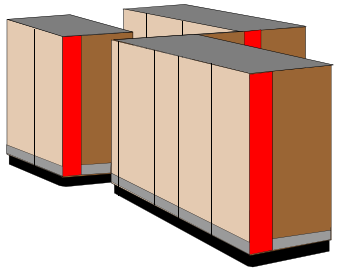
匿名化

- 連結情報
  1. 個人識別情報 ⇔ ID番号（対応表）
  2. ID番号 ⇔ 個人医療情報
  3. ID番号 ⇔ 研究結果（データ）
- 連結不可能情報
  4. 匿名化コード ⇔ 個人医療情報
  5. 匿名化コード ⇔ 研究結果
- 通常、＃1の対応表（DB）を厳重管理

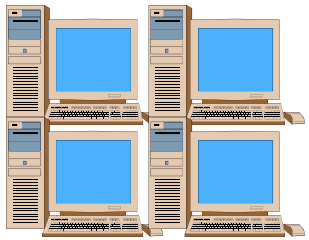
病院情報システム

診療部門

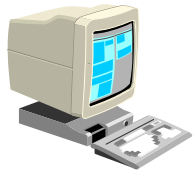
UNIX Servers



PC Servers



医事会計システム



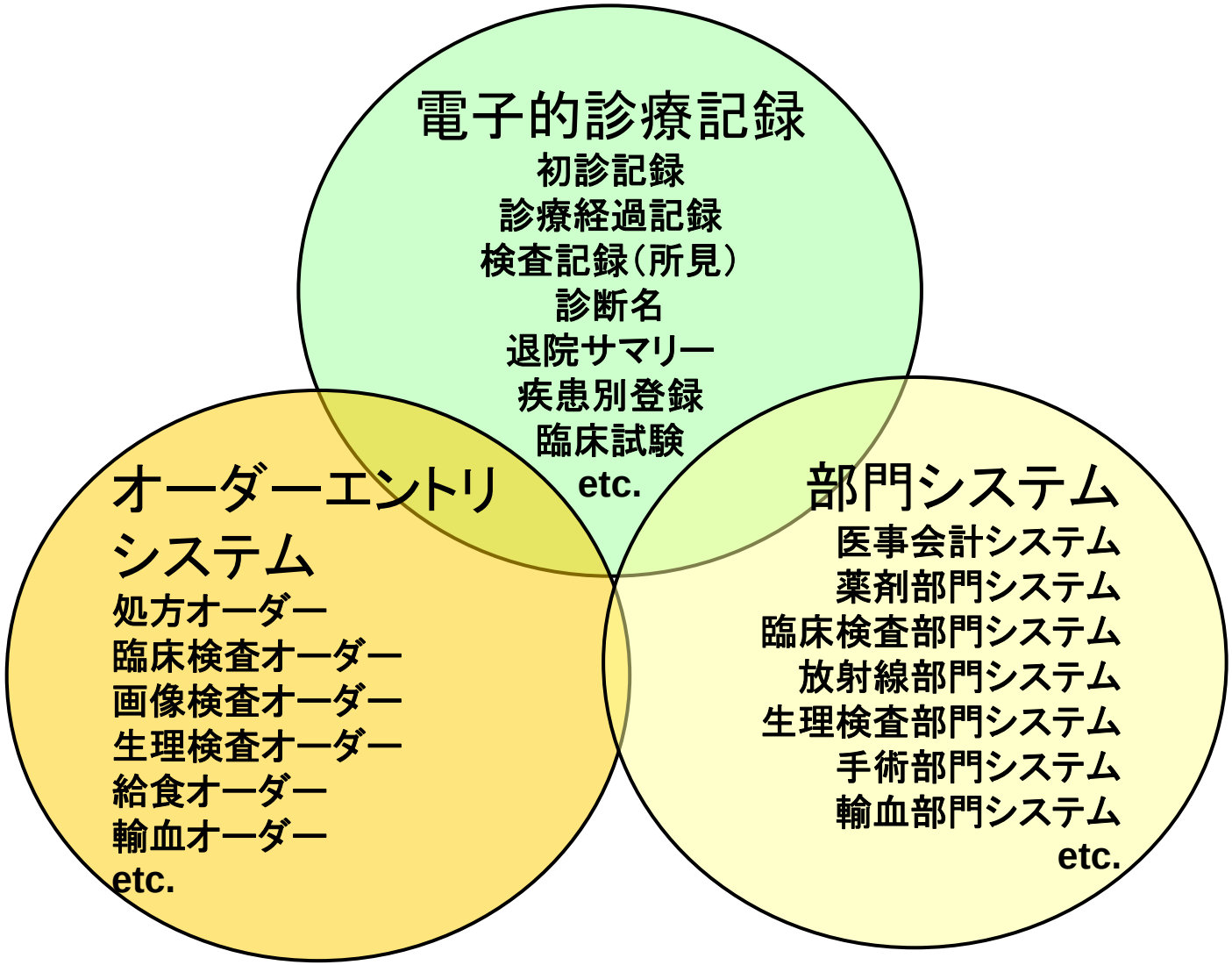
物品管理システム

臨床研究支援システム

院内ネットワーク

薬剤部門システム
検体検査システム
放射線診断システム
栄養管理システム
病歴システム
生理検査システム
放射線治療システム
病理部門システム
RI検査システム
細菌検査システム
輸血部門システム

病院にはどんな情報があるか（１）



病院にはどんな情報があるか（2）

臨床上使用されている帳票からみた分類

予約・指示箋	表形式	文章形式	チェックリスト形式	説明書類
再来予約票	指示表	1号記録用紙	申し送り簿	オリエンテーション手引き
他科受診予約票	温度表	2号記録用紙	麻酔前問診票	手術を受けられる方へ
内服処方箋	検温表	看護目標	術前チェックリスト	
注射処方箋	注射指示表	転科サマリー	検査チェックリスト	検査記録台紙
画像検査予約票	CCU看護記録表	退院サマリー	看護チェックリスト	カルテ台紙
生理検査予約票	水分バランス納表	看護要約	承諾書	
食事箋	手術前指示表	個人記録	術式略図	
	業務分担表	病棟日誌	入院手術申込票	
予約結果参照	他科依頼表	看護日誌	添付許可申請書	
処方内容参照	点滴表	麻酔記録		
検査結果参照		手術記録		
		日報		

公開情報にはどんなものがあるか

- PubMedなどの文献データベース
- ゲノム関連データベース
- たんぱく質関連データベース
- パスウェイデータベース
- 薬剤副作用情報
- 診療ガイドライン

・・・

Windber Research Instituteの例

- The Windber Research Institute ・・・ to help the facility create the first central data warehouse for molecular and clinical information by combining the information of five organizations that generate data ranging from protein interactions to metabolic pathways. (9/24/2003 5:00:00 PM)

(http://www.itbusiness.ca/index.asp?theaction=61&lid=1&sid=53533&adBanner=dat)

臨床情報工学部門のDWH構想

医師
看護師
検査技師
・・・

検査・診断・処方

意思決定

各種分析

適切な治療法は？
治療の効果は？
治療の副作用の可能性は？

患者臨床薬物治療プロファイルデータ

患者SNPsプロファイルデータ

患者薬物代謝遺伝子プロファイルデータ

疾患（がん）プロファイルデータ

⊕

患者治療効果データ

患者治療副作用データ

推論エンジン
テキスト/データマイニング

HIS
電子カルテ
がん登録

臨床試験
CRF DB

SNPs検査

ヒト遺伝子発現
検査(24,500)

がん遺伝子発現
DB

蛋白質DB
(130,000)

Pathway DB

- 臨床研究支援のためのDWH設計

業務系システムとDWHのデータの違い

	業務系システム	DWH
指向性	現在のオペレーション	分析・予見
対象データ	業務運営のための全データ	情報を組み立てるための素材
データ構造	必要なデータを網羅	部分的かつ統合的
エンティティ	マスター 短期データ 目的別サマリー	分類基準 時系列データ 基本サマリー

臨床研究用のDWHの特徴

流通業などのDWHと比較すると・・・

- 情報に恣意性有り（カルテなど）
- テキスト/イメージが重要
- 測定誤差を含む（検査値など）
- 大量の外部データベースの取り込み
- ギガバイトからテラバイト/ペタバイトへ

臨床研究用のDWHの方針

- 柔軟で臨床のニーズ変化に対応する強固な基礎として構築
- 質の高い臨床データの確保
- データ、利用者、クエリーの増加に対する容易な対応
- リーズナブルなTCO (Total Cost of Ownership)

データウェアハウス技術の分類（1）

- プラットホーム
  - –データ・サーバ
  - –オペレーティング・システム
  - –DBMS
  - –アプリケーション・サーバ
  - –ネットワーク・プロトコル
  - –クライアントPC

（ バーキン他、「データウェアハウスの戦略と設計」、日経BP社）

データウェアハウス技術の分類（2）

- データウェアハウス管理
  - データ・モデリングとデータベース定義
  - データ・リエンジニアリングとデータ洗浄
  - データ抽出と変換
  - 操作性と性能

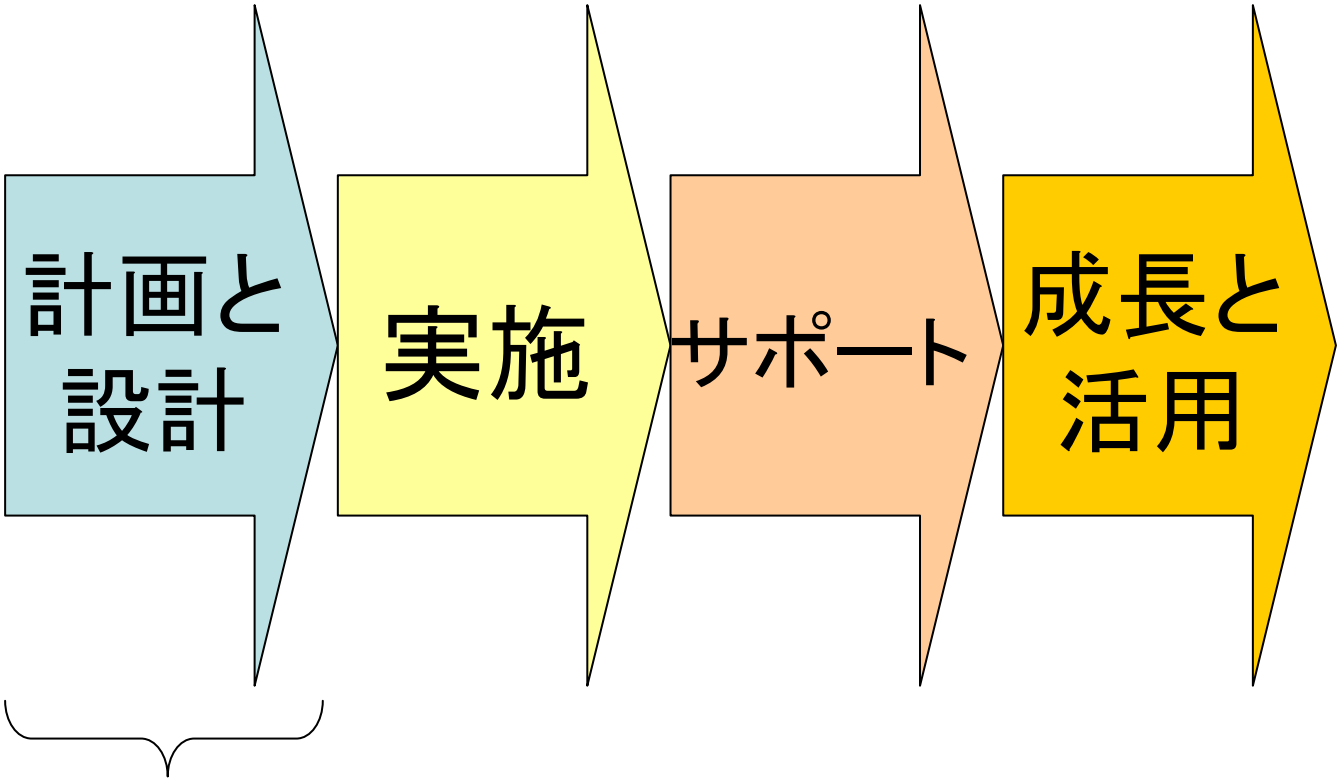
（ バーキン他、「データウェアハウスの戦略と設計」、日経BP社）

データウェアハウス技術の分類（3）

- クライアントツール
  - –アドホックな問合せ
  - –定型的な問合せ
  - –多次元分析・データマイニング・DSS
- メタデータ
  - –メタデータ・リポジトリ

（ バーキン他、「データウェアハウスの戦略と設計」、日経BP社）

全体の流れ



計画（1）

- 導入の目的
- 予算・スケジュール・体制
- 業務要件・システム要件
- システム概要（DWHの3要素：DB、アプリ、U/I）

以上を、基本計画書としてまとめる

計画（2）

- 基本計画書（中長期計画も）
- 業務分析、モデル化、仕様書作成
- システム選定基準（できるだけ信用のおける外部機関のデータや、業種を問わずDWHの成功事例を研究する）

（導入するまでわからないことも多いのが現実！）

システム選定基準の例

- 性能のベンチマーキング
  - ある程度大きなDBで評価する（オモチャはダメ）
  - テストで使うプログラムは現実な課題を解決するものを使う（簡単なサンプルプログラムではダメ）
  - 実運用に近い利用者数、クエリーの多重度・複雑度を設定して評価する（1人が一度に1回しかクエリーを出さないようなテストはダメ）
- TCO (Total Cost of Ownership)

ベンチマーキングの例（1）：TPC-H

- TPC-Hは、独立したベンチマーク監査機関である米国トランザクション処理性能評議会（TPC）によって開発された意思決定支援のためのベンチマーク。このベンチマークは、複数の同時ユーザーによる複雑なビジネス上のクエリーに対するデータベースの性能を測定
- TPC-Hテストは、22種類の意思決定支援向けクエリーと2種類のデータベース更新処理を実行し、パフォーマンス（処理性能）を測定します。測定基準単位にQphH（Query-per-Hour Performance Metric：1時間あたりに実行可能な検索処理数）を用い、プライス・パフォーマンス（価格性能比）、利用可能日を加えた3つを測定
- 投資する際は、同じ分野での導入事例を充分に調査し、必要に応じて実世界のデータウェアハスウ環境に即したベンチマークテストを実施することを推奨

（http://www.teradata-j.com/product/pro1_02.html）

ベンチマーキングの例（2）

TPC-H：100GB

Rank	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	IBM eServer 325	12,216	71 US $	2011/8/3	IBM DB2 UDB 8.1	Suse Linux Enterprise Server 8	07/29/03
2	IBM eServer xSeries 445 8P	5,602	73 US $	12/31/03	IBM DB2 UDB 8.1	Microsoft Windows Server 2003 Enterprise Edition	06/30/03
3	MAXDATA Platinum 9000-4R	4,307	70 Euros	2011/8/3	IBM DB2 UDB 8.1	Microsoft Windows Server 2003 Enterprise Edition	2007/3/3
4	HP ProLiant DL 760 G2 8P	4,224	43 US $	08/15/03	Microsoft SQL Server 2000 Enterprise Edition	Microsoft Windows Server 2003 Enterprise Edition	08/15/03
5	IBM eServer xSeries 440	3,861	102 US $	07/15/03	IBM DB2 UDB 8.1	Microsoft Windows Server 2003 Enterprise Edition	04/22/03

（日付が間違っているものがありますが修正せず、そのままにしています）（http://www.tpc.org/tpch/results/tpch_perf_results.asp）

ベンチマーキングの例（3）

TPC-H：1,000GB

Rank	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	Fujitsu (Siemens) PRIMEPOWER 2500	34,492	156 Euros	2003/8/4	Oracle Database 10g Enterprise Edition	Sun Solaris 9	2009/8/3
2	Fujitsu PRIMEPOWER 2500	34,492	141 US $	2003/8/4	Oracle Database 10g Enterprise Edition	Sun Solaris 9	11/13/03
3	HP 9000 Superdome Enterprise Server	25,805	203 US $	10/30/02	Oracle 9i Database Enterprise Edition v9.2.0.2.0	HP UX 11.i 64-bit	10/29/02
4	HPProLiant DL760 X900-128P	22,361	253 US $	06/20/02	IBM DB2 UDB 7.2	Microsoft Windows 2000 Advanced Server	2002/6/2
5	IBM eServer p655 with DB2 UDB	20,221	69 US $	2006/8/4	IBM DB2 UDB 8.1	IBM AIX 5L V5.2	2012/8/3

（日付が間違っているものがありますが修正せず、そのままにしています）（http://www.tpc.org/tpch/results/tpch_perf_results.asp）

ベンチマーキングの例（４）

TPC-H：3,000GBのDB

Rank	System	QphH	Price/QphH	System Availability	Database	Operating System	Date Submitted
1	NCR 5350	79,528	213 US $	12/20/02	Teradata V2R5.0	MP-RAS 3.02.00	2001/6/3
2	HP Integrity Superdome Enterprise Server	45,247	109 US $	03/25/04	Oracle Database 10g Enterprise Edition	HP UX 11.i 64-bit	09/25/03
3	Fujitsu (Siemens) PRIMEPOWER 2500	34,345	161 Euros	02/22/04	Oracle Database 10g Enterprise Edition	Sun Solaris 9	08/22/03
4	Fujitsu PRIMEPOWER 2500	34,345	147 US $	02/22/04	Oracle Database 10g Enterprise Edition	Sun Solaris 9	08/26/03
5	Sun Fire[TM] 15K server	28,948	184 US $	04/30/03	Oracle 9i R2 Enterprise Edition	Sun Solaris 9	2004/7/3

（http://www.tpc.org/tpch/results/tpch_perf_results.asp）

TCOによる評価（1）

- TCO (Total Cost of Ownership)
  - 「IT資産の購入および維持に要する直接的支出のみならず、技術の習得、維持管理、利用を可能にするための人件費も視野にいれた、何年間かのライフサイクルにまたがって積算された総合的な保有コスト」 **-日本ガートナーグループのTCO(Total Cost of Ownership)の定義-**
- TCOの要素
  - ①資産　②管理　③エンドユーザオペレーション　④テクニカルサポート（⑤トラブルによる業務上の機会損失）

TCOによる評価（2）

- 従来、コンピュータシステムのコストは製品価格(導入費用)で評価されることが多かったが、近年のコンピュータシステムの複雑化や製品価格の下落などにより、コンピュータシステムの維持・管理やアップグレード、ユーザの教育、システムダウンによる損失など、導入後にかかる費用(ランニングコスト)が相対的に大きな存在となったため注目されるようになった。

TCOによる評価（3）

- DWHのTCO削減のための指標例
  - 生データ量と実際に必要なディスクスペースの比
  - 論理モデルが物理モデルに直接利用できる割合
  - データローディングやデータ配置に要する時間
  - DBを管理するのに何人のDB管理者がいるか
  - スペース管理がどの程度自動化できるか
  - 性能向上のチューニングが簡単か
  - 利用者や開発スタッフへの教育が簡単か

設　計

- マスタの整備・コードの標準化
- 部門サーバからのデータ取得
- データの洗浄
- データベース
- 業務ロジック
- クライアントツール
- 運用設計

DWH構築の大部分の時間を要するので以下、これを中心に説明

データベースとクライアントツールはどのようなプラットホームを選択するかで何ができるかおよそ決まってしまう

コード・用語の標準化

分野	分類コード
疾患分類	ICD-9,ICD-10, ICD-9CM, SNOP, SNOMED, DSM-III, MeSH, JICST
放射線診断	放射線診断コード集JRSC, IRDコードなど
ＰＡＣＳ	DICOM-3, MIPS, IS&CA ACR-NEMA300など
看護情報	看護診断（北米看護協会）など
臨床検査	臨床検査項目分類コード, 臨床検査自動化機器用検査項目集
薬剤	日本標準商品分類コード
輸血	輸血製剤コード, 血液型マスター
給食	日本食品標準成分表

部門サーバからのデータ取得

- メタデータ（データウェアハウスの標準化）
  - MDCのOIMとOMGとCWM の統合（2000/9）
  - 統一の標準仕様を用いることで、ユーザーは異なるベンダーの異なる製品間であっても、自由にメタデータの交換が可能
- メッセージ交換形式
  - HL7、XML、・・・
- 通信プロトコル
  - CORBA、SOAP、・・・

MDC: Meta Data Coalition

OIM：Open Information Model

OMG: Open Management Group

CWM：Common Warehouse Metamodel

メタデータとは

- メタデータとは、「データに関する記述データ（データ辞書）」。データ構造、データ定義、データの意味、データの出所、発生場所、保管場所、入手時期、データ間の関連性等の情報を保持している。
- データ・ウェアハウスの管理やアプリケーション開発に必要な“データに関する技術的な青写真”を提供するもので、データベース内の「リポジトリ」と呼ばれるテーブルに保存。

（http://www.ncr.co.jp/products/software/terasoft/mds.html）

部門間データ交換プロジェクト

- 平成12、13年のMEDISからJAHIS、JIRAに委託された病院情報システムのマルチベンダーによる開発プロジェクト
- 診療部門と医事会計を含む6部門システム間をHL7メッセージをベースにCORBAネットワーク上で接続された研究
- 開発目標：XML表記されたHL7メッセージをCORBAによって部門間で共有すること

HL7とは

- 包括的な電子的医療情報の交換のための規約
- 医療情報交換のための標準規約（プロトコル）で、患者管理、オーダー、照会、財務、検査報告、マスターファイル、情報管理、予約、患者紹介、患者ケアなどの情報交換を取り扱う
- HL7の名前は、医療情報システム間のISOーOSI第7層アプリケーション層上での抽象メッセージ」に由来

（http://www.hl7.jp/）

HL7メッセージの構成

メッセージ：MSHセグメント<cr>
xxxセグメント<cr>
yyyセグメント<cr>
セグメント：セグメントID｜フィールド1｜・・・<cr>
フィールド：エレメント1^エレメント2^｜・・・<cr>
[メッセージ記述の主要素]
- セグメント属性テーブル
- 区切り文字（<cr>, | , ^ ）
- データ型（フィールドおよびエレメントの記述様式）

細菌検査結果データのHL7化（１）

- ORUメッセージの規格はGCP97を参考したが、異なる点はGCP97に省略されたPV1（患者所在情報セグメント）を採用し、入院・外来患者、病棟、診療科、入・退院日などの情報を表現。
- 培養同定検査結果と感受性検査結果は親子のような関係があり、検体検査より関連付けが複雑である。そのような関連付けはHL7で表現するのは難しいため、GCP97のHL7インターフェースは細菌検査結果をサポートしていない。

（平成12年東大健康科学・看護学専攻　陳俊成 修士論文）

細菌検査結果データのHL7化（2）

- HL7規格に準拠する複数表現の可能性がある。例えば、培養同定検査結果には菌名と菌量のペアがあり、東大病院では両方ともコードで表現するため、以下のような2つの表現ができる。
- １個のOBXでの表現（OBXの5番目フィールドに繰返し）
  - OBX|1|CE|検査結果コード||菌名～菌量|A|||F
- ２個のOBXでの表現
  - OBX|1|CE|検査結果コード||菌名|A|||F
  - OBX|1|CE|検査結果コード||菌量|A|||F
  - 注：CE（Code Element）はHL7に定義されたデータ型でコード体系を意味する。OBXの2番目の標記は5番目の結果のデータ型を示す。
- JAHISに相談の結果、２個のOBXでの表現を推奨された。

（平成12年東大健康科学・看護学専攻　陳俊成 修士論文)

細菌検査結果データのHL7化（3）

- 標準検査項目コード
  - JLAC10は当院検査項目コードと完全に対応できないため、HL7規格のCE型（コード^コード文書^コード体系名^代替コード^代替コード文書^代替コード体系名）に準拠しながら代替コードとして導入（対応できない項目はJLAC10を省略）。
- 標準細菌コード
  - 現在、日本では標準化細菌コードがないため、当面当院の細菌コードで検出菌の結果を表現する。
  - 標準化菌コードがあれば、ORUメッセージが共通化となり、DWHのスキーマなどの標準化実現の可能性も高まる。

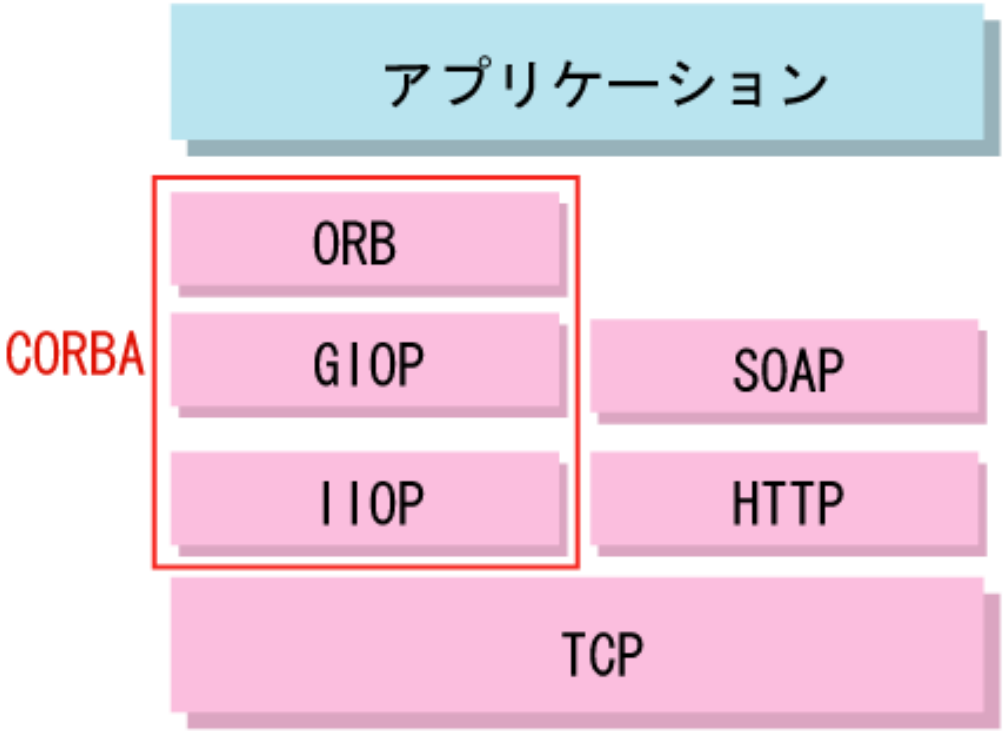
（平成12年東大健康科学・看護学専攻　陳俊成 修士論文）

通信プロトコル（１）

- CORBA
  - Common Object Request Broker Architecture
  - 異機種分散システムを統合するための共通基盤としての利用
  - クライアントが、ネットワーク上に存在するオブジェクト（分散オブジェクト）を呼び出すための基盤を提供する
  - ポータビリティと相互運用性に優れる
  - OMGがCORBAのSOAPマッピングの標準化作業進行中
- SOAP
  - Simple Object Access Protocol
  - XMLドキュメントをHTTPなどのトランスポート上で送受信するためのプロトコル
  - オブジェクト呼び出しの要求と応答メッセージをXMLで記述することで、分散オブジェクト呼び出しのプロトコルとして使用可能

通信プロトコル(2)

- CORBA とSOAP



ORB (Object Request Broker)

GIOP（General Inter-ORB Protocol）

IIOP (Internet Inter-ORB Protocol)

（http://mikilab.doshisha.ac.jp/dia/research/report/2002/0606/015/report20020606015.html）

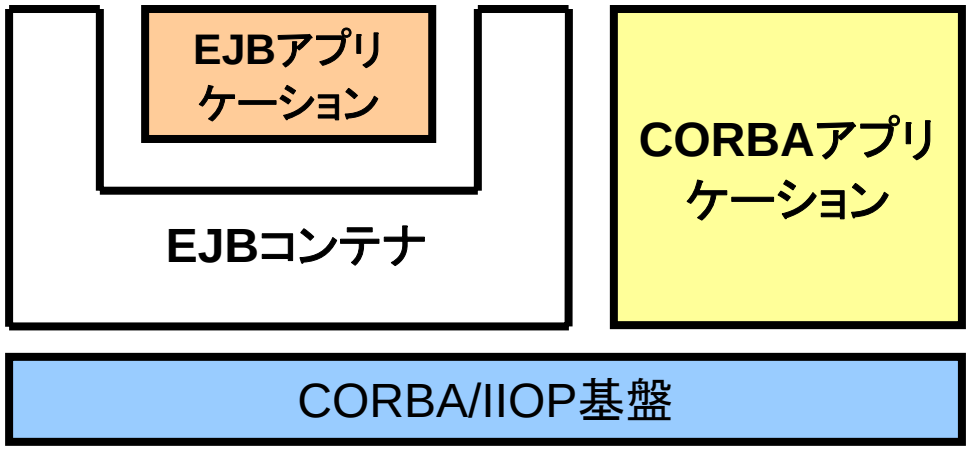
SOAPエンジンとORBの例

名前	SOAP/CORBA	対応言語	フリー/シェア
orbix	CORBA	JAVA, C++, C	シェア
VisiBroker	CORBA	JAVA, C++	シェア
HORB	CORBA	JAVA	フリー
PEAR	SOAP	PHP	フリー
Apache AXIS	SOAP	JAVA	フリー
Open SOAP	SOAP	C	フリー

（http://mikilab.doshisha.ac.jp/dia/research/report/2002/0606/015/report20020606015.html）

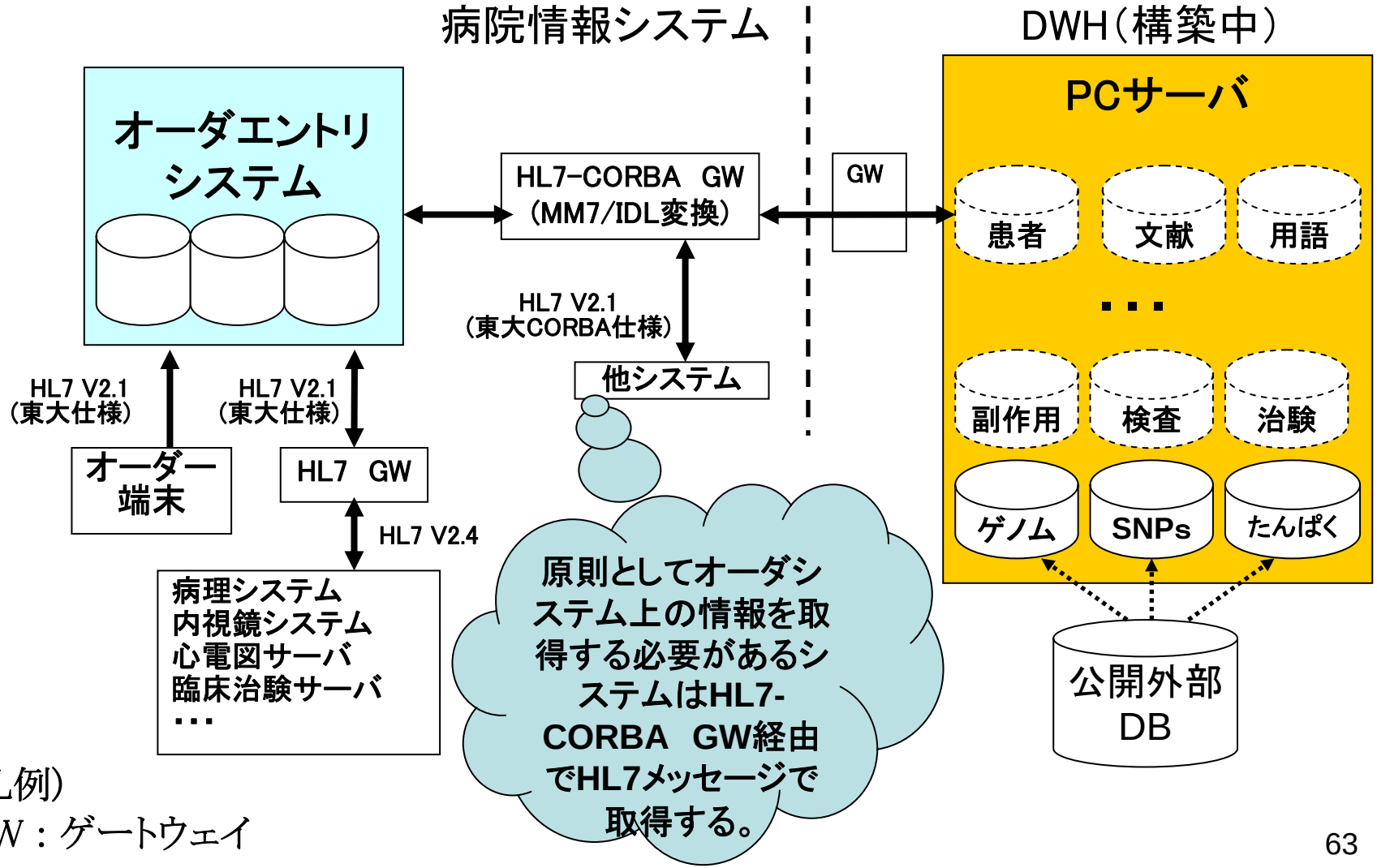
CORBAとEJB

- CORBA とEJB (EnterpriseJavaBeans)
  - EJBはコンポーネントモデルをサーバサイドで実現した仕様
  - 機能的にCORBAはEJBのスーパーセット
  - EJBアプリケーションはEJBコンテナの中で実行される
  - EJB 2.0では、CORBAのプロトコルIIOPや幾つかのCORBAサービスが必須



（http://www.atmarkit.co.jp/fjava/rensai/corba01/corba02.html）

オーダシステムからのデータ取得例



（凡例）
GW：ゲートウェイ

データ洗浄

- データ洗浄（データクレンジング）とはデータの情報の質を向上させるための一連の作業
- データ洗浄なくしては、よいDWHは絶対できない（ここにもっとお金を投資してもよい）
- 市販やフリーのデータ洗浄ツールを活用しよう

データは汚れている

- Dummy Values,
- Absence of Data,
- Multipurpose Fields,
- Cryptic Data,
- Contradicting Data,
- Inappropriate Use of Address Lines,
- Violation of Business Rules,
- Reused Primary Keys,
- Non-Unique Identifiers, and
- Data Integration Problems

Source: This article is on the Teradata University Network and the Teradata Student Network.

データ洗浄の手順

- Parsing （解析）
- Correcting （修正）
- Standardizing （標準化）
- Matching （マッチング）
- Consolidating （集約）

以下、簡単に説明

Source: This article is on the Teradata University Network and the Teradata Student Network.

Parsingの例

入力データ

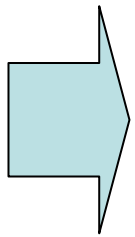
東大　太郎、　研究担当部長

XXX製薬(株)

赤門bldg.

本郷一丁目二番三号

文京、東京



解析済データ

姓:　東大

名:　太郎

役職:　研究担当部長

会社:　XXX製薬(株)

ビル:　赤門bldg.

番地:　本郷一丁目二番三号

市:　文京

県:　東京都

Correctingの例

解析済データ

姓: 東大

名: 太郎

役職: 研究担当部長

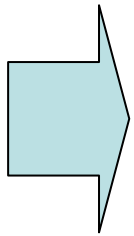
会社: XXX製薬(株)

ビル: 赤門bldg.

番地: 本郷一丁目二番三号

市: 文京

県: 東京都



修正済データ

姓: 東大

名: 太郎

役職: 研究担当部長

会社: XXX製薬(株)

ビル: 赤門bldg.

番地: 本郷一丁目二番三号

市: 文京区

県: 東京都

郵便番号: 113−8888

Standardizingの例

修正済データ

姓: 東大

名: 太郎

役職: 研究担当部長

会社: XXX製薬(株)

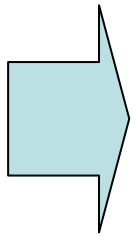
ビル: 赤門bldg.

番地: 本郷一丁目二番三号

市: 文京

県: 東京都

郵便番号:113−8888



標準化済データ

姓: 東大

名: 太郎

役職: 研究担当部長

会社: XXX製薬株式会社

ビル: 赤門ビル

番地: 本郷1−2−3

市: 文京区

県: 東京都

郵便番号:113−8888

データベース設計指針（1）

- 複雑なチューニングをしなくてもある程度のレスポンスが期待でき、DB管理が簡単なDBMSを選択するのが設計以前の大前提。
- トランザクション処理とデータウェアハウスとはデータベースの設計が異なる。
- トランザクション処理のためのデータベース構造は正規化され、データ更新の性能追及とデータの一貫性保持が目的。

データベース設計指針（2）

- データウェアハウスの設計は主に読み取り専用で問合せやレポート作成のために単純化されることが多い（スタースキーマ）。
- 実運用を想定したデータ量で設計をする。
- 論理データモデルを設計したら、各ベンダーにベンチマーキングを依頼するのもよい（実運用レベルの量のデータを使う）。

論理データモデルの作成

- 研究領域の基本調査
  - 研究範囲に沿ってシステムの基本的内容を調査する。
- 情報の分析
  - 研究領域について利用者が要求する情報を調査し分析する。情報構造を明らかにしDWHに格納するデータ要素を抽出。
- データモデル図の作成
  - 情報分析により得られたデータ要素を用いてデータモデルを組み立てる。
- 入力源となるデータの調査
  - データモデルのエンティティや属性に該当するデータを調査し分析する。

論理データモデルとは

- 研究領域でのデータの要件をまとめ図で表現
  - 研究領域: 研究対象となる世界でDBやハードウェアから独立
  - データの要件: 必要な情報を組み立てることができるデータ
  - 図で表現: データの構造を図で表現する（ex. ER図）
- 正規化されたデータ構造
  - 研究で使用されていた（る）データ項目を、「ひとつのデータは一ヶ所に」の考え方で整理したもの
- DWHの中核となるDBに格納されるデータ構造

業務ロジック設計

- SQLでの利用が基本
  - データのブラウジングとインテリジェントな問合せ
  - 標準的なレポート出力
  - データ分析とモデリング・ツール（データマイニングなど）
- ユーザがある程度自由に使えることが重要
  - 研究には前例のないことも多いので、事前にロジックを組み込めないことがある
  - アプリケーション開発は必要最低限に抑えたい

クライアントツール

- Webベース
  - Webブラウザを使用してサーバ上のツールを使う
- クライアントソフト
  - 特定のDBMS向けのツール（OLAPツールなど）
  - データマイニングソフトなど
- B/I（Business Intelligence）ツール
  - より幅広いユーザーから、それぞれの問題を手軽に解決できる柔軟なデータ分析環境を実現するためのプラットフォーム

運用と展開

- 一度成功したDWHは成長し続ける（一度失敗すると利用者は二度と使わないが・・・）
- 基本機能の拡張
  - データの追加
  - 問合せと画面の変更・追加
- システム機能の拡張
  - ハードウエア/ソフトウェア/ツール
- 新しい目標設定

まとめ

- DWHはevidence-basedな意思決定を支援するのが主目的
- TCOを考えると、どのようなプラットホーム（DBMS,ツール,など）を選択するかがカギ
- 質の高いデータをどのように素早く正確に収集・作成するのかが本質的に重要
- 臨床系のDWHは困難な課題が多くあり、今までの常識にはとらわれないことも必要

おまけ

DWHの70%が成功しない理由！

- 現在の緊急ニーズにだけ対応しDWH設計をする（将来のニーズを考慮しない）
- システム部門（開発側）が利用者の意見を取り入れずに、一緒に検討しないで開発する
- データ量やユーザ数の増加に耐えうるスケーラビリティを考えないでハード/ソフトを選定する
- 将来の技術を当てにしてDSSを構築する

etc.

（Meta Groupによる調査）

参考資料

- 図書
  - ウェスターマン、ウォルマートに学ぶデータ・ウェアハウジング、翔泳社
  - バーキン他、データウェアハウスの戦略と設計、日経BP社
- 規格・規準
  - 日本HL7協会
  - 財団法人医療情報システム開発センター（http://www.medis.or.jp/）
  - 臨床検査のホームページ （http://square.umin.ac.jp/clin-lab/）
  - http://www2.kiryu-jc.ac.jp/local/jmm/lec/H13info/3kai.html
  - DICOM規格(日本語草稿) (http://www.jfcr.or.jp/DICOM/dicom_draft-j.html)
  - http://www006.upp.so-net.ne.jp/ebisu/iryou.htm
  - 標準関係情報集（http://www2u.biglobe.ne.jp/~standard/）
  - 規格・データベース関連（http://www.aist.go.jp/renraku-kaigi/fukushi/standard.htm）
- ベンチマーキング
  - http://www.tpc.org/
  - http://www.gartner.co.jp/