Database Core

- Database completion
  - a common, secure database established in Europe for all relevant scientific information in GenomEUtwin
- First ten months
  - a database structure established

Database Core Personnel

Prof. **Jan-Eric Litton**, Dept. of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Dr. **Kari Kuulasmaa**, Director of MONICA Data Centre -organising and overseeing communications, Finland

Prof. **Nancy Pedersen**, Swedish Twin Registry -quality of the data

Zygimantas Cepaitis Eng, Systems analyst -design of data systems, database management and data quality control

Kauko Heikkilä, Phil. Lic, Finnish Twin Cohort Study, -twin database manager

Dr. **Juha Muilu**, NPHI, -integration of genotype-phenotype databases; software issues

Lars Hvidberg, Danish Twin Registry, -twin database manager

Lars Bäckström, Uppsala University, SNP database in Uppsala

Jaason Haapakoski, NPHI - NPHI sample database issues, Finland

Ann Björklund, Karolinska Institutet, Stockholm, Sweden - core database manager

Jenny Carlsson, Karolinska Institutet, Stockholm, Sweden - Swedish twin registry database manager

Axel Skytte, Karolinska Institutet**,** Stockholm, Sweden

Rodolfo Cotichini, Istituto Superiore di

Ingunn Brandt Norwegian Institute of Public Health, Norway

Anne K.Leinonen, The Finnish Genome Center, Helsinki, Finland

Fagnani Corrado, Istituto Superiore di Sanità, Rome, Italy

Emad Qweitin, St Thomas's Hospital, London , England

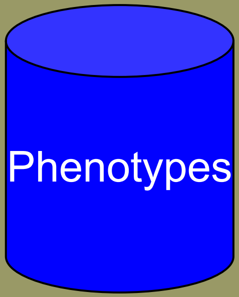
Leiden, Holland

Database Core Harmonization

- Actions taken
  - Data Format and Variable Standard for GenomEUtwin's Phenotype Database Prototype

    Version: 3.2

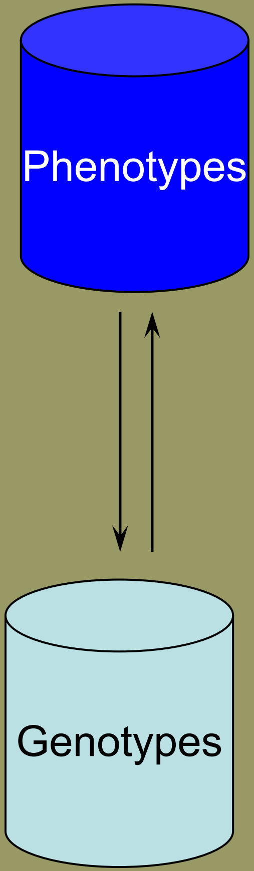
Each center contributed 100 twins



Database Core Harmonization

- **EUid number (EUIDNUM)** 752000021210
  - The EUid number consists of four parts:
    - Country code 3 digits – ISO 3166
    - Randomized number 7 digits
    - Identification number 1 digit
    - Check sum 1 digit

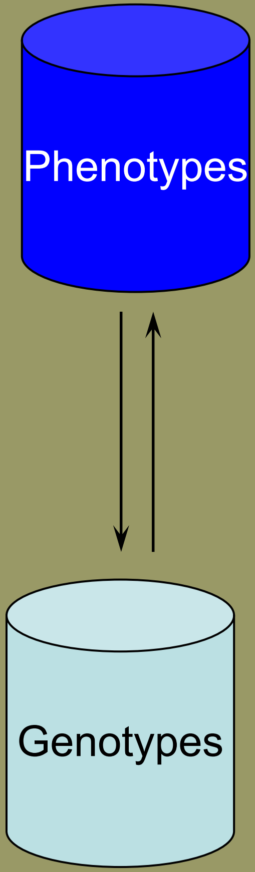
Database Core



- A Data warehouse extracts data from data sources across an entire enterprise and Acts as a centralized repository of information.

- A Data mart is a “small” warehouse designed to support a specific activity

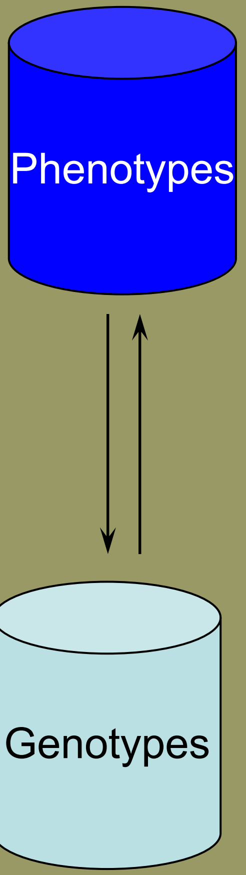
SQL



SQL = Structured Query Language.
is used to communicate with a database.
According to ANSI (American National Standards Institute), it is the standard language for relational database management systems.

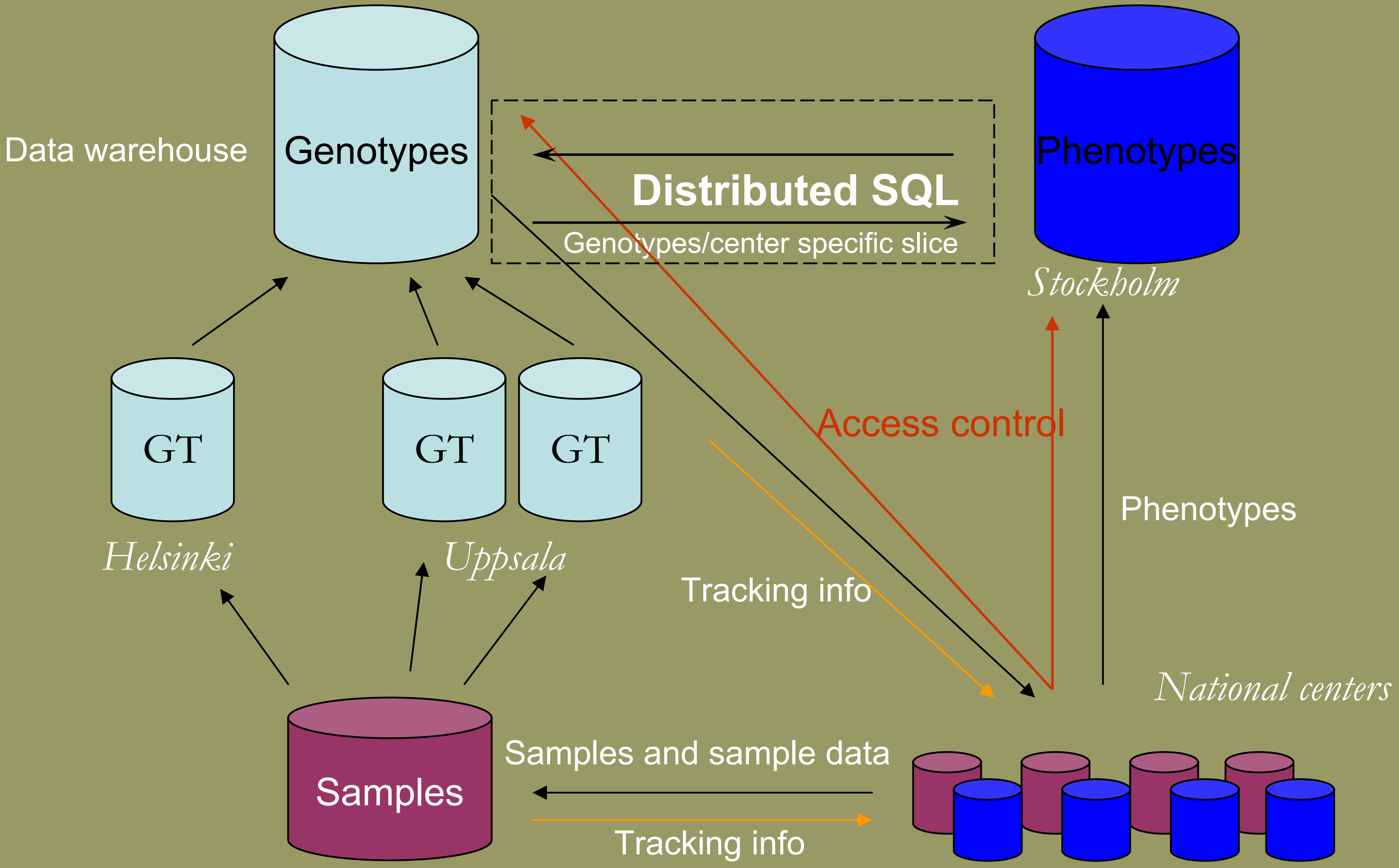
```
select "column1"  [,"column2",etc]
from "tablename"  [where "condition"];
[] = optional
```

Distributed SQL



- •Synchronous Direct Access to remote database
  - –DB links
- •Location Transparency
  - –Tables residing in the remote databases look local
- •Data integrity maintained using Two-phase commit
- •Distributed SQL
  - –Supports DML and Query
  - –Intelligently optimizes execution plans

Database Core



Steering&SAB Rotterdam
October 5-7, 2003

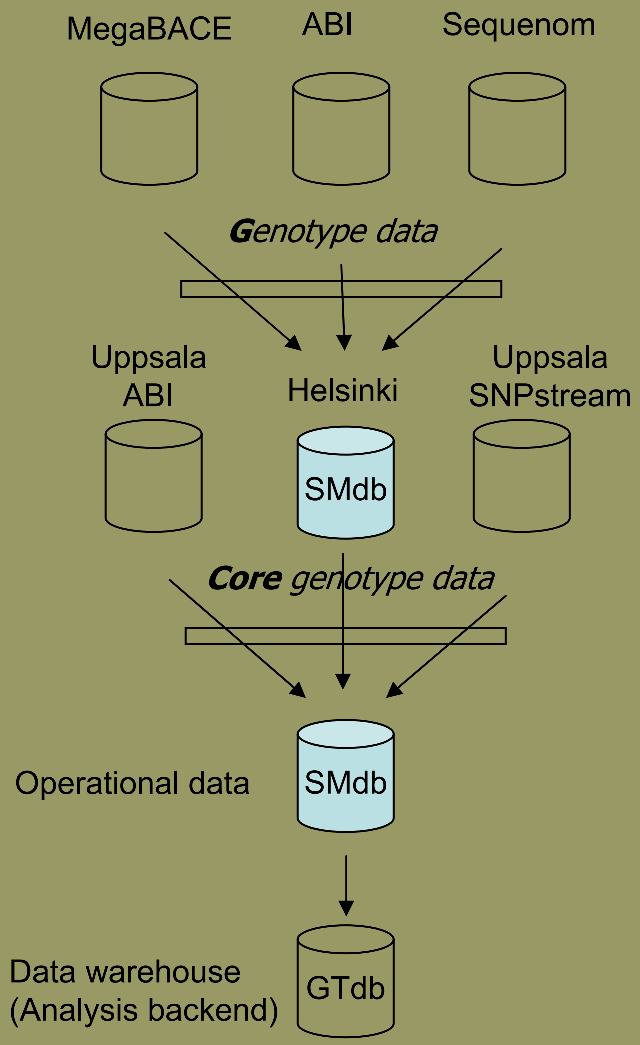
GENOMEUTWIN

Development of Genotype Database

”Open source” project

Genotyping Core ↔ Database Core

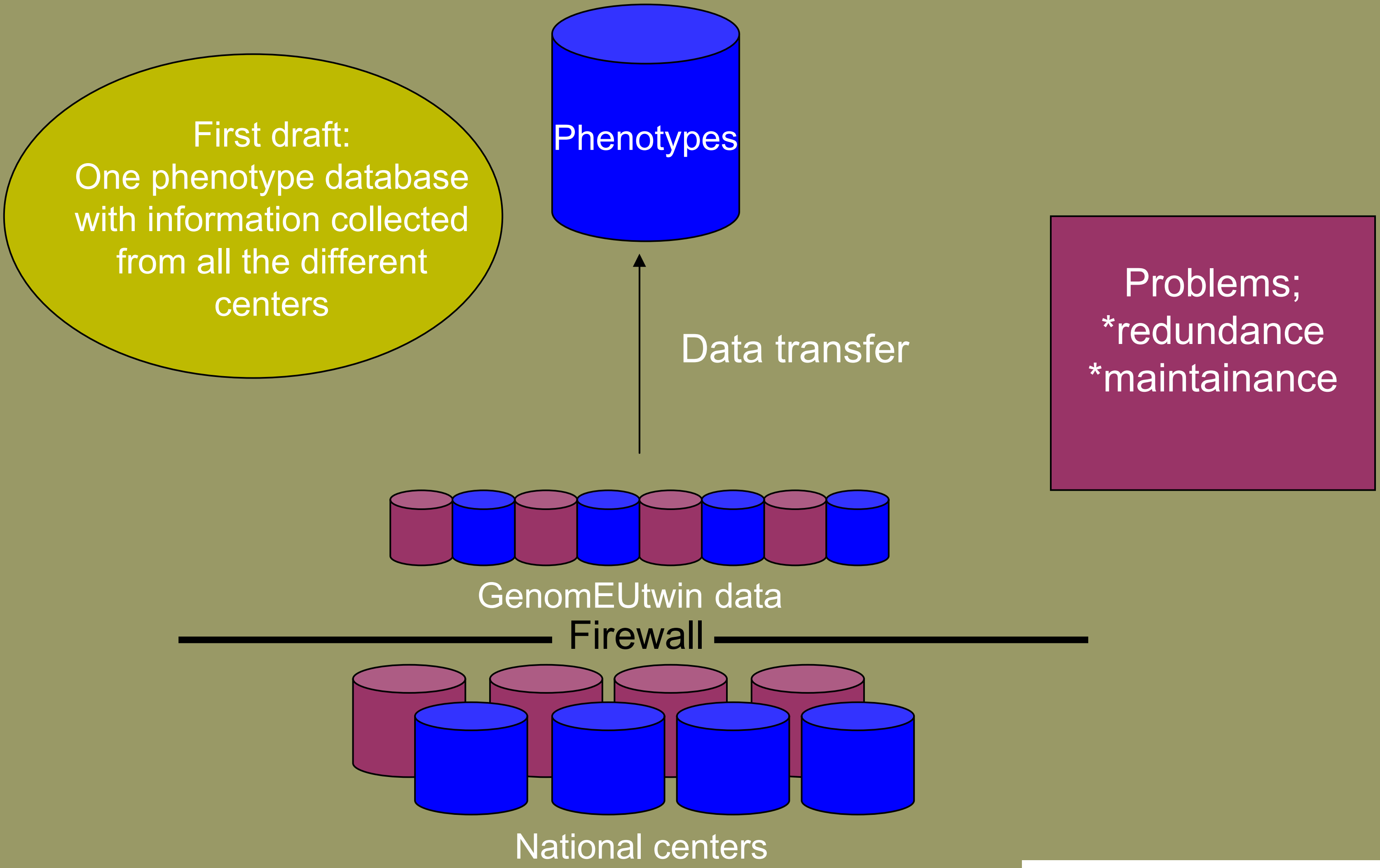
Management of data produced by Genotyping Core: Submission database in Helsinki



- Repository for data submissions
- Submission log
  - Source, date, type of submission operation, name of operator..
- Full history of data
  - Nothing is deleted
  - Full operation log
- Submission operations
  - Insertion (new data)
  - Update (modifies existing data)
    - = Deletion + Insertion
  - Deletion
    - Data are marked as deleted
  - Export
    - Data is sent further

Juha Muilu, FGC

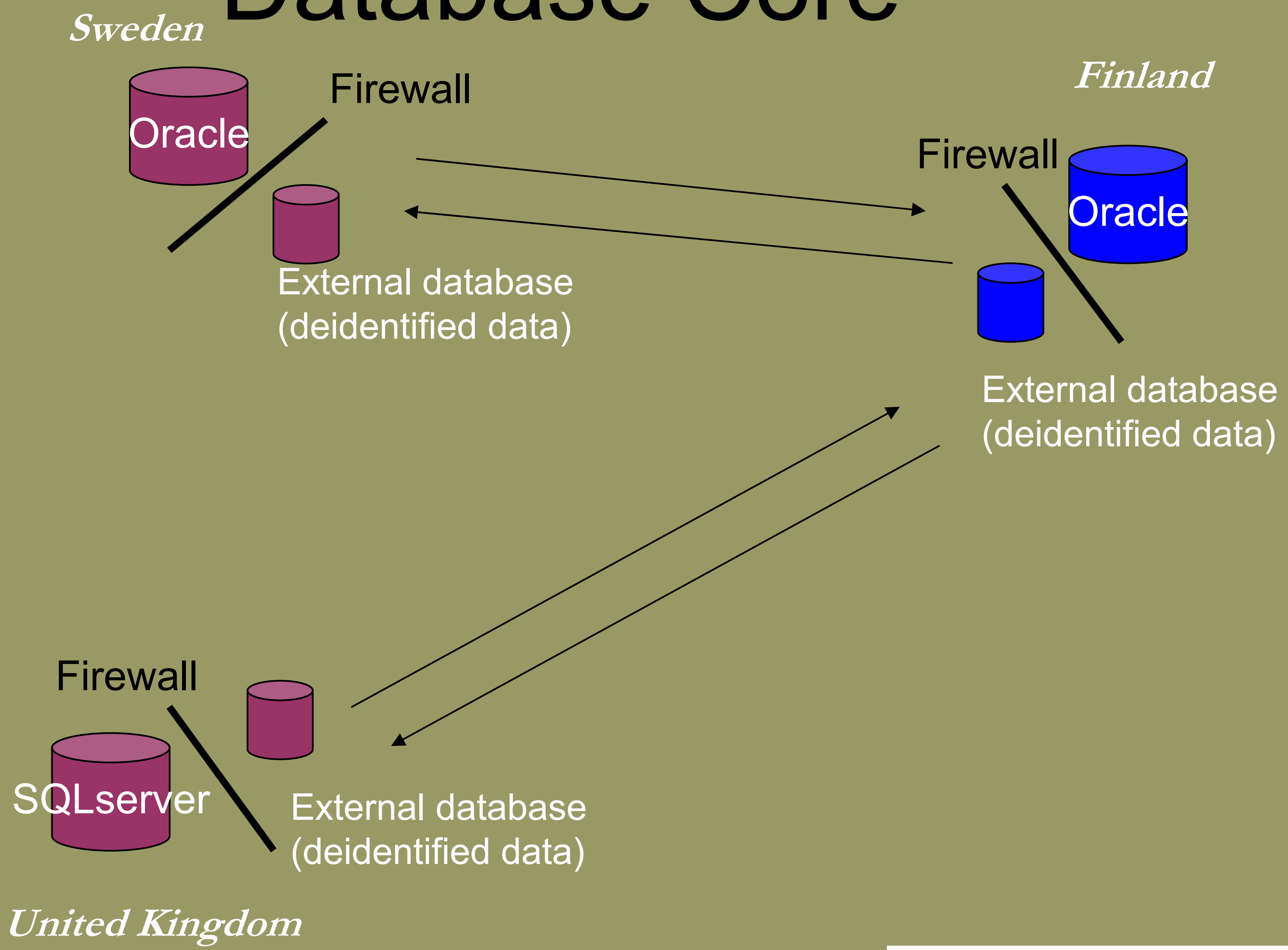
Database Core



First draft:
One phenotype database with information collected from all the different centers

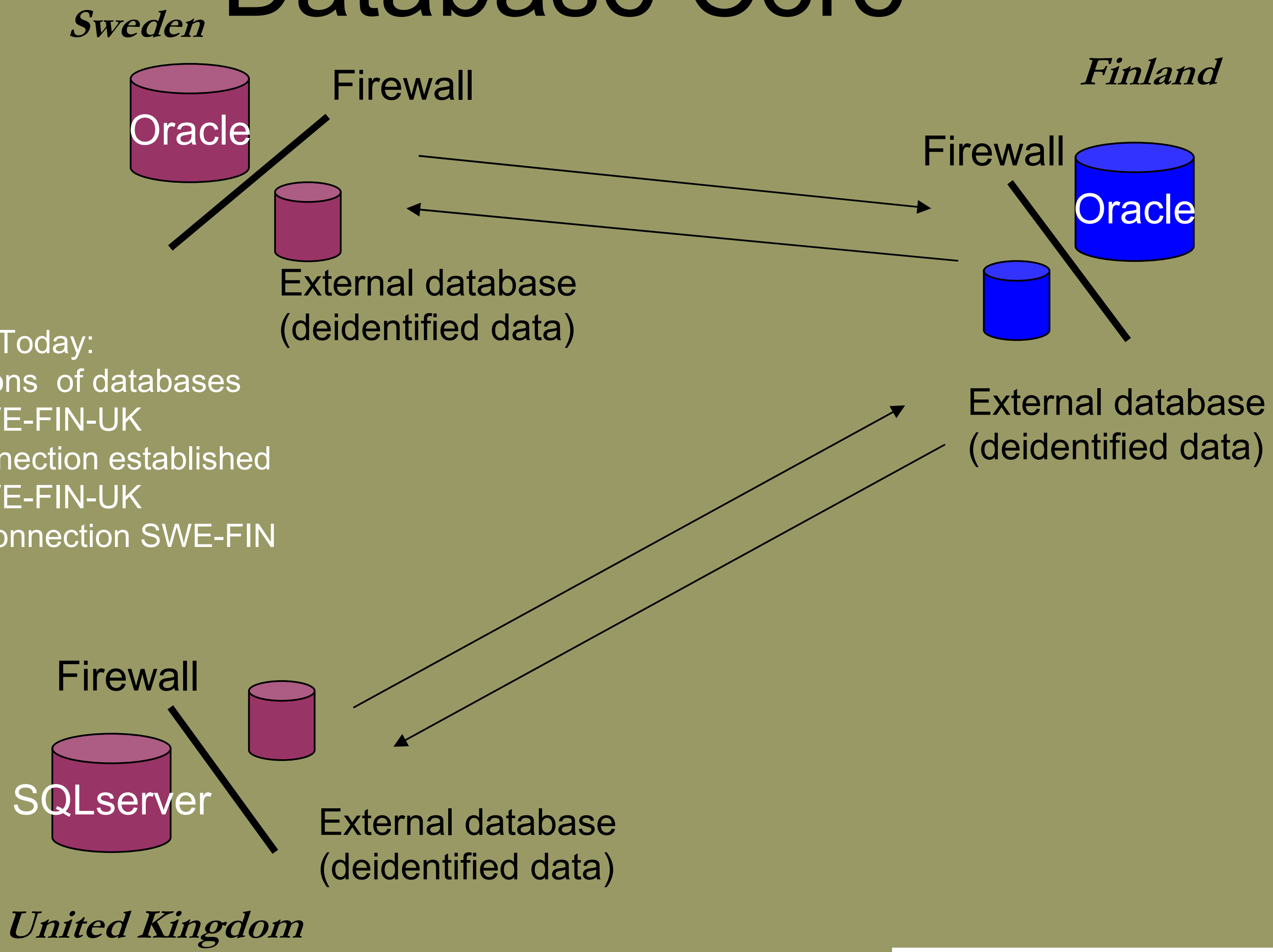
Problems;
*redundance
*maintainance

Database Core



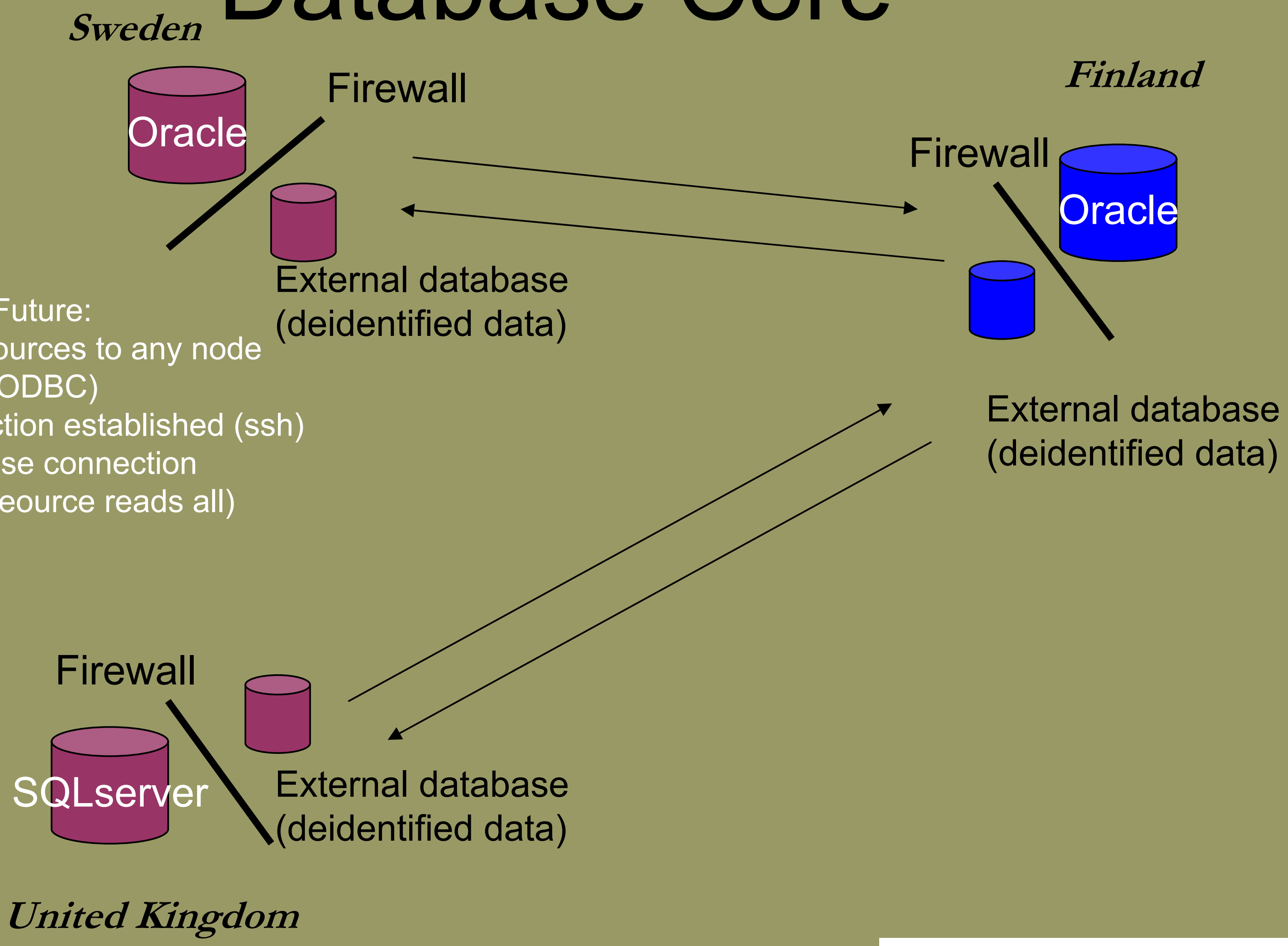
GENOMEUTWIN

Database Core



Today:
*Replications of databases SWE-FIN-UK
*Secure connection established SWE-FIN-UK
*Database connection SWE-FIN

Database Core



Sweden

Oracle

Firewall

External database
(deidentified data)

Finland

Firewall

Oracle

External database
(deidentified data)

Firewall

SQLserver

External database
(deidentified data)

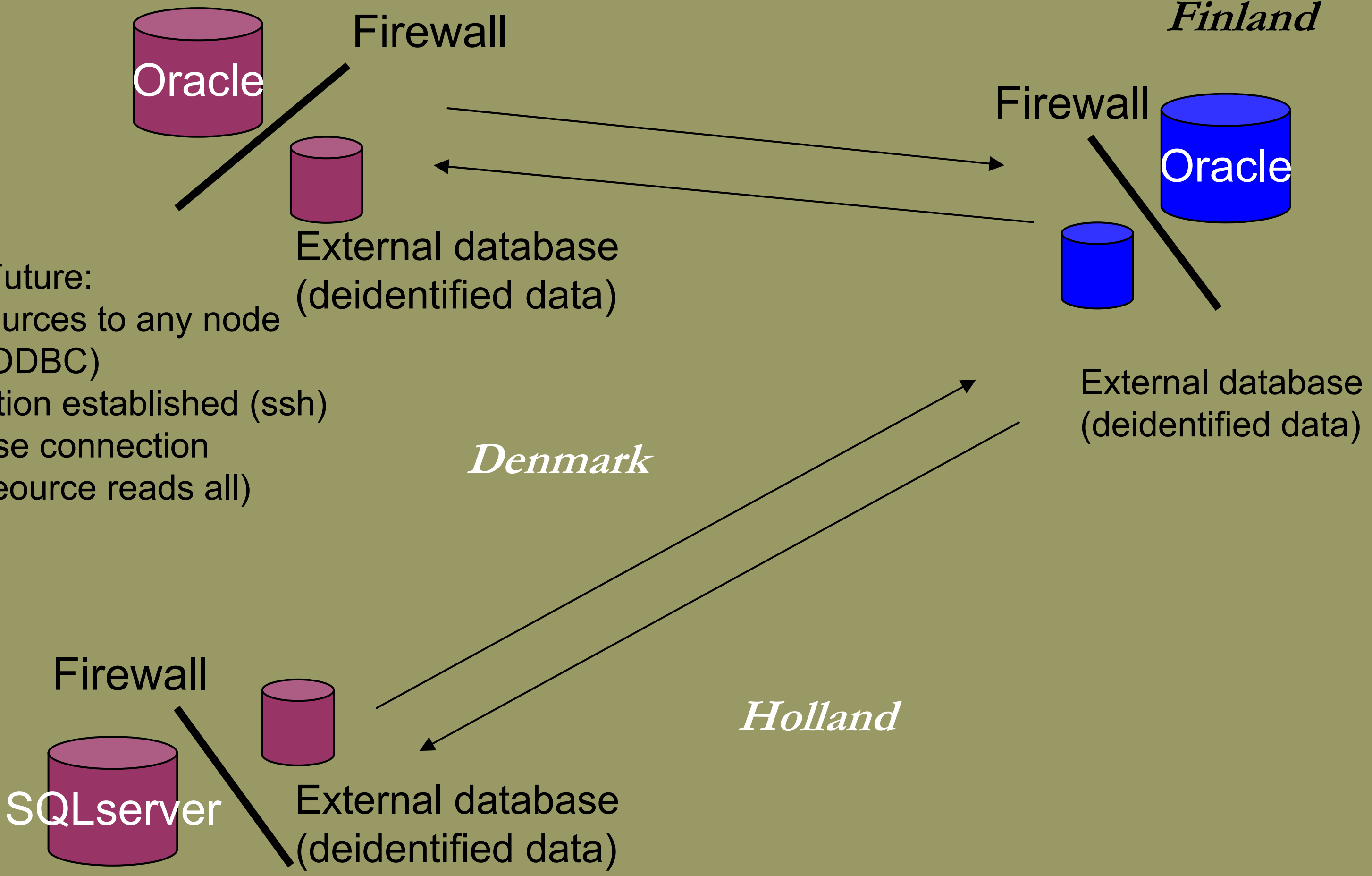
United Kingdom

Future:
*adding db-sources to any node
(ODBC)
*Secure connection established (ssh)
*Database connection
(one dataseource reads all)

Database Core

Norway

Sweden



Oracle

Firewall

External database
(deidentified data)

Finland

Firewall

Oracle

External database
(deidentified data)

Firewall

SQLserver

External database
(deidentified data)

Future:
*adding db-sources to any node
(ODBC)
*Secure connection established (ssh)
*Database connection
(one dataseource reads all)

Denmark

Holland

Italy

United Kingdom

Australia

GENOMEUTWIN

Database Core Result

- a database structure has been established
    - a common format and variable standard for phenotypes has been launched
    - a distributed SQL model between Stockholm - Helsinki -London has been demonstrated

Next... a Federated Database

- A federated database for GenomEUtwin data
    - Data remain in the original separate sources
    - All operational data sources accessible with a single query
    - Query optimization of all data sources

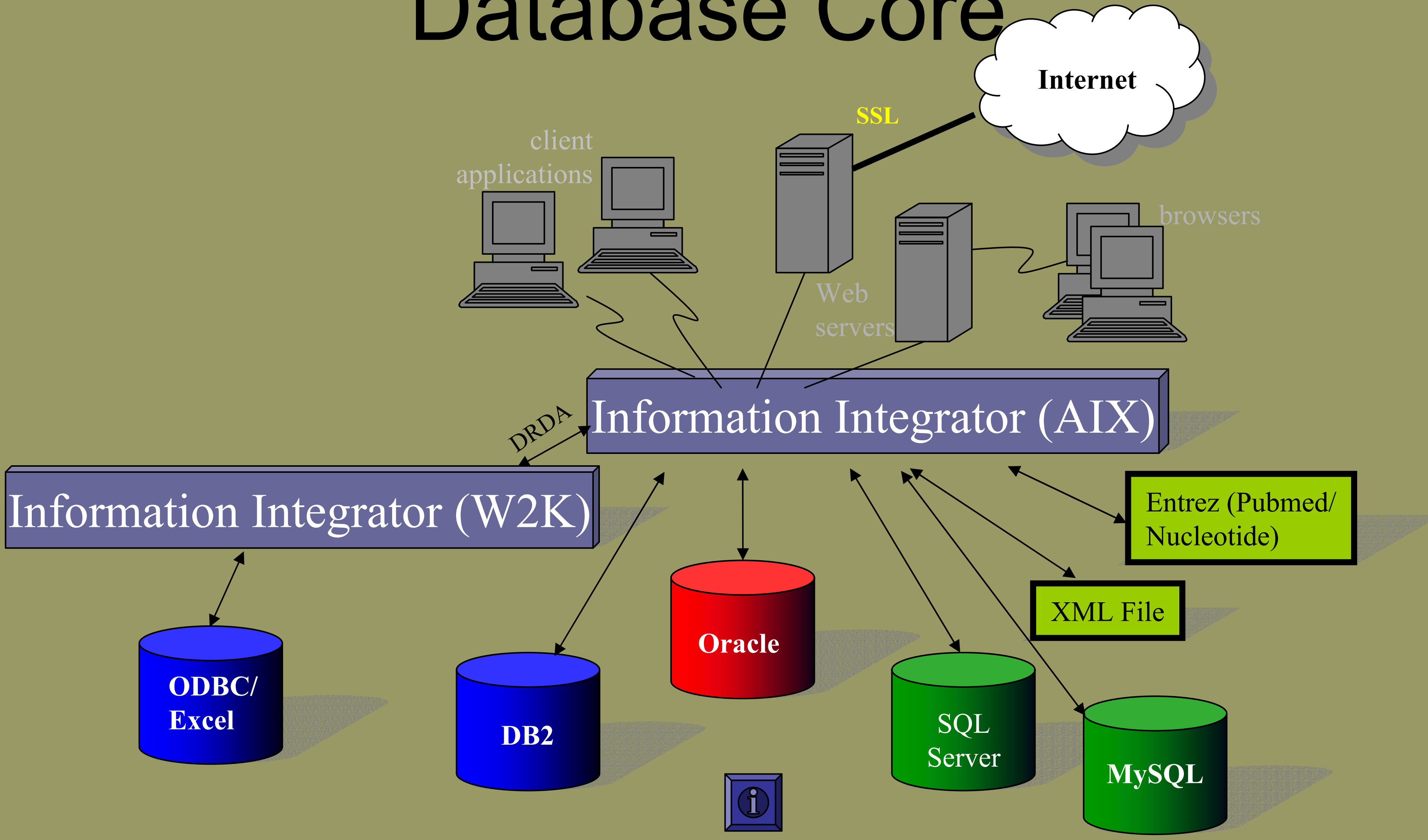
- Proof of concept together with IBM, using the middleware Information Integrator (Discovery Link)

- Information Integrator provides the researcher with a view of their data as one “virtual relational database”
    - This can be for relational and non-relational data

Distributed SQL vs Federated Database

- Distributed SQL model
  - + cheap, easy to use for database administrators
  - - outside firewall
  - - no web portal
- Federated database using IBM's Information Integrator
  - + all kind of data (incl. flatfile, XML, SAS, Internet db)
  - + inside firewall
  - + web portal
  - - cost
  - - needs a server with IBM Information Integrator/site

Database Core

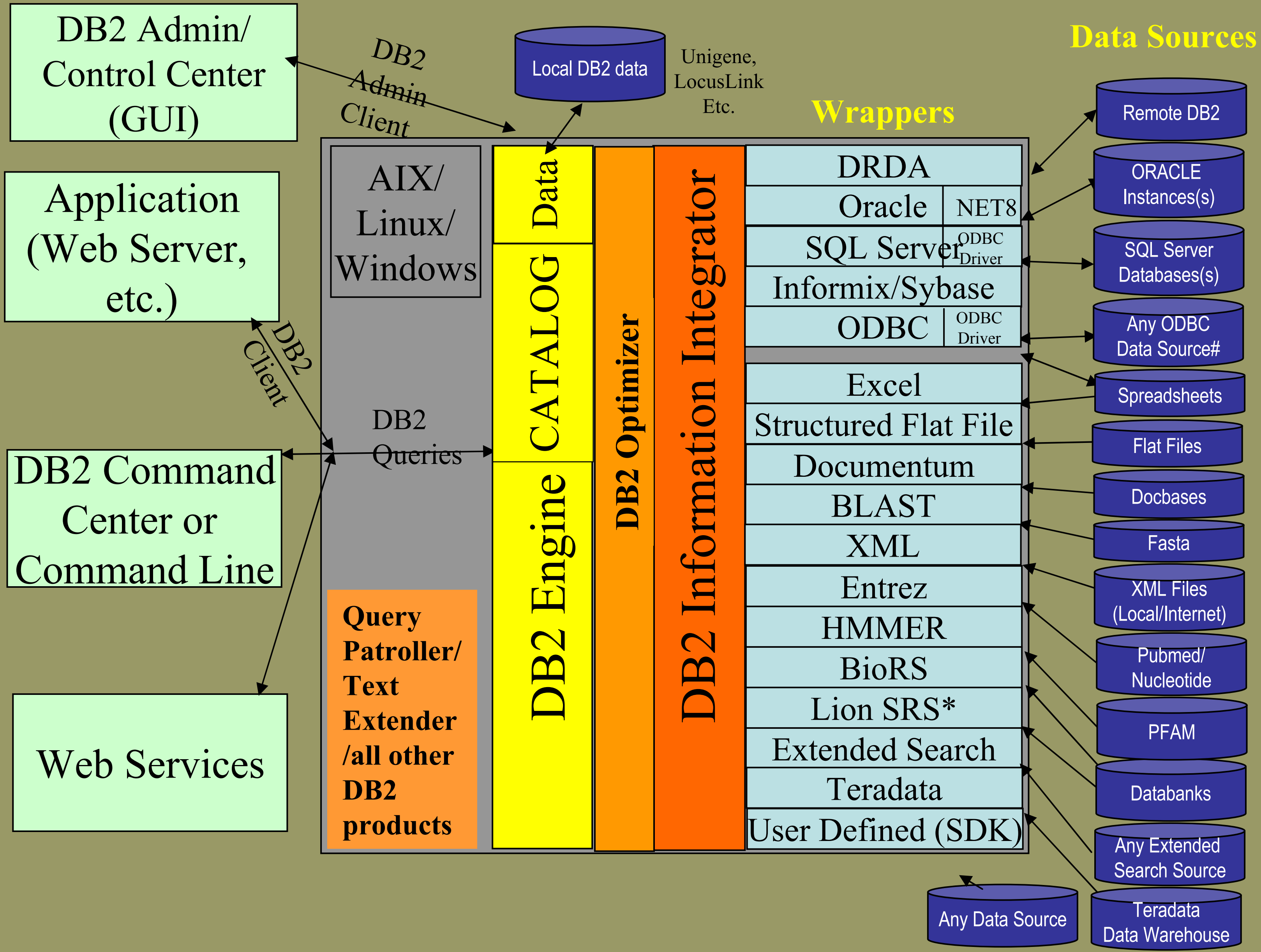


Steering&SAB Rotterdam
October 5-7, 2003

GENOMEUTWIN

Information Integrator

- Standardization of research data access
- Supports common relational and non relational data sources (including life sciences data such as BLAST, XML, etc.)
- Adaptable, robust and extensible ("wrappers") foundation for knowledge discovery
- Makes all data sources "SQL aware"
- Smarter, more efficient than "gateways"



DB2 Admin/ Control Center (GUI)

DB2 Admin Client

Local DB2 data

Unigene, LocusLink Etc.

Wrappers

Data Sources

Application (Web Server, etc.)

DB2 Client

DB2 Command Center or Command Line

Web Services

AIX/ Linux/ Windows

DB2 Queries

Query Patroller/ Text Extender /all other DB2 products

DB2 Engine CATALOG Data

DB2 Optimizer

DB2 Information Integrator

DRDA

Oracle NET8

SQL Server ODBC Driver

Informix/Sybase

ODBC ODBC Driver

Excel

Structured Flat File

Documentum

BLAST

XML

Entrez

HMMER

BioRS

Lion SRS*

Extended Search

Teradata

User Defined (SDK)

Remote DB2

ORACLE Instances(s)

SQL Server Databases(s)

Any ODBC Data Source#

Spreadsheets

Flat Files

Docbases

Fasta

XML Files (Local/Internet)

Pubmed/ Nucleotide

PFAM

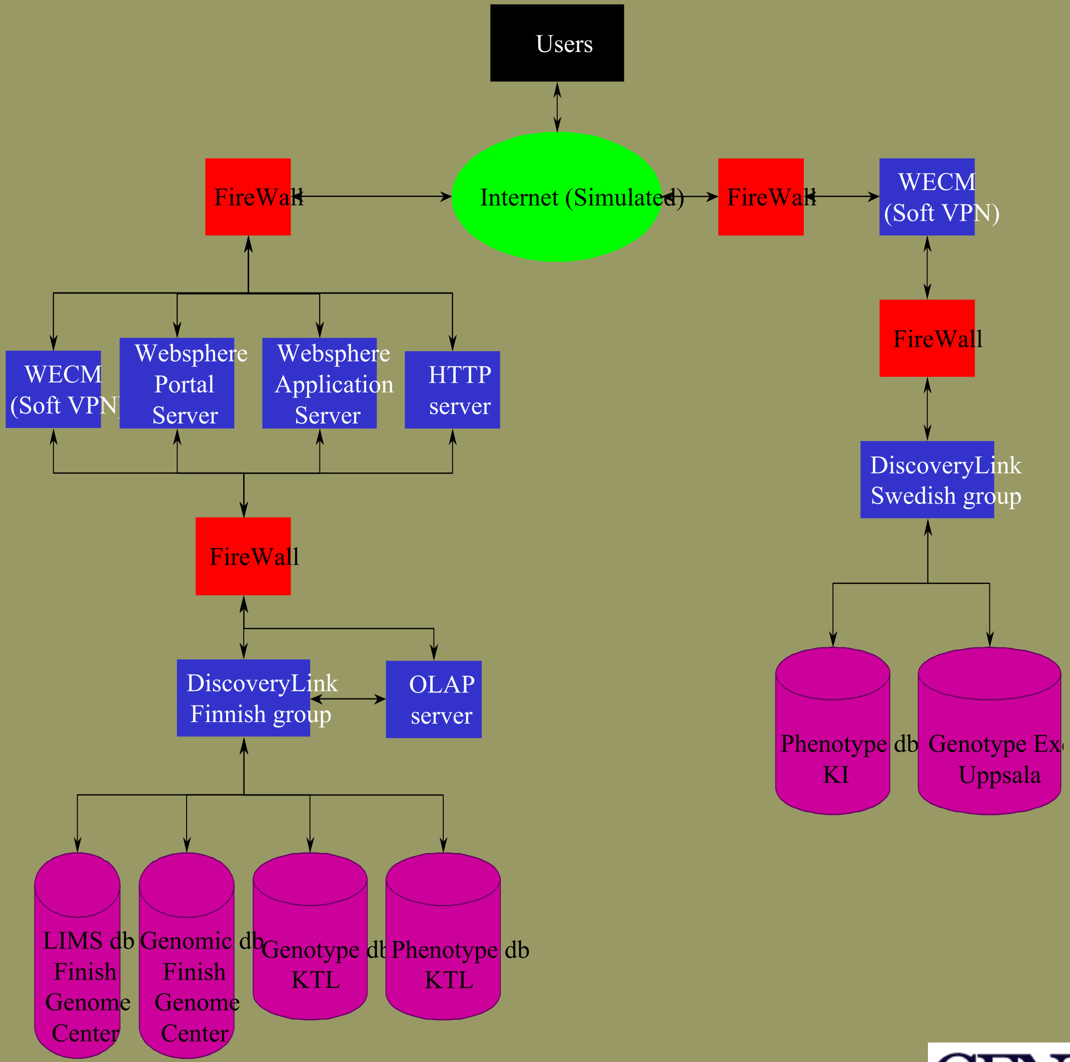
Databanks

Any Extended Search Source

Any Data Source

Teradata Data Warehouse

Database Core - Nice



Users

FireWall

Internet (Simulated)

FireWall

WECM (Soft VPN)

WECM (Soft VPN)

Websphere Portal Server

Websphere Application Server

HTTP server

FireWall

FireWall

DiscoveryLink Swedish group

DiscoveryLink Finnish group

OLAP server

Phenotype db KI

Genotype Exp Uppsala

LIMS db Finish Genome Center

Genomic db Finish Genome Center

Genotype db KTL

Phenotype db KTL

Steering&SAB Rotterdam
October 5-7, 2003

GENOMEUTWIN

Database Core

Questions ?