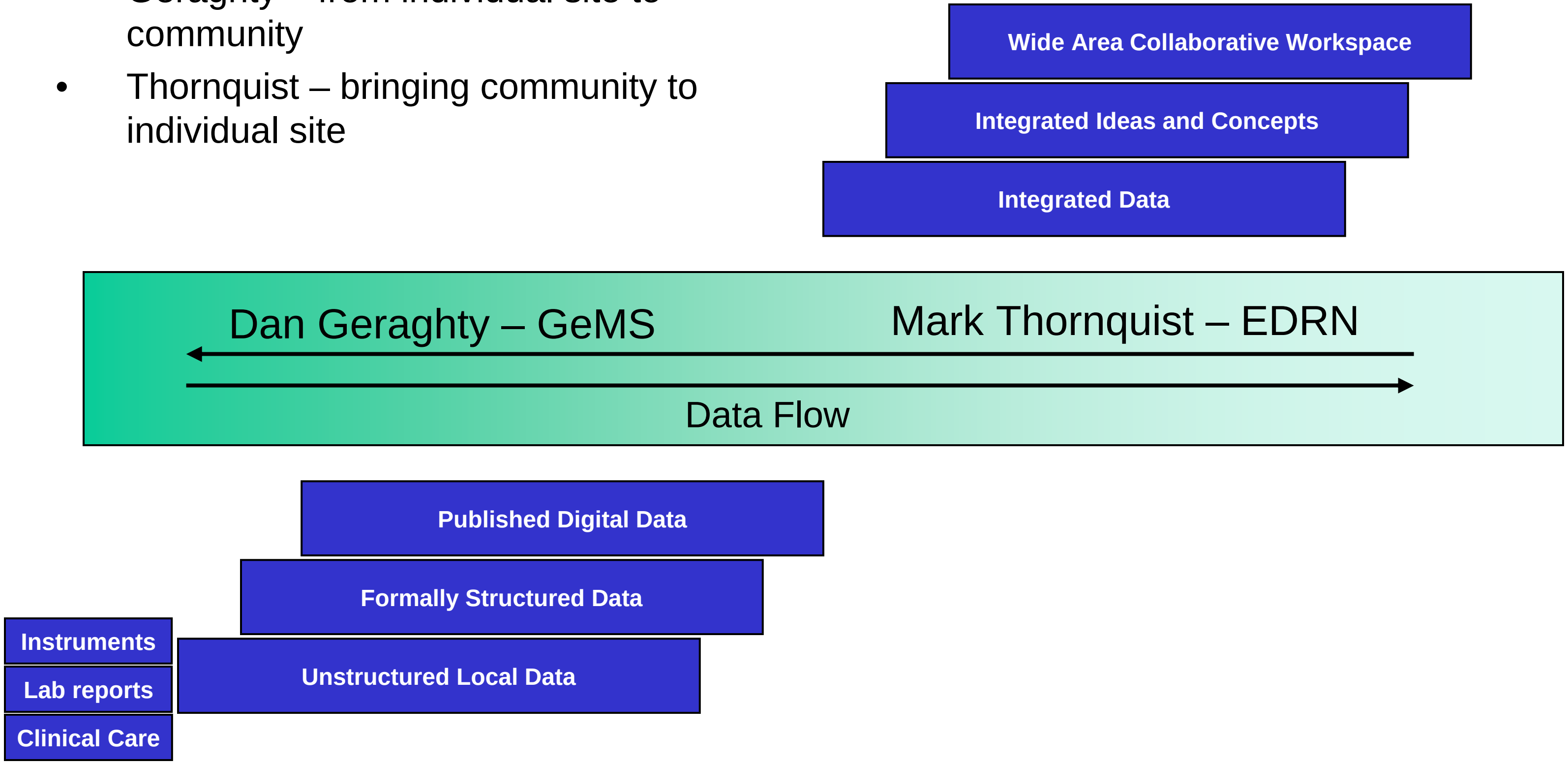
caBIG Architecture Kickoff Meeting Presentation
Fred Hutchinson Cancer Research Center

Mark Thornquist, Derek Walker, Heather Kincaid, Rahul Joshi, Dan Geraghty, Robert Robbins.

Data Sharing Continuum

- Geraghty – from individual site to community
- Thornquist – bringing community to individual site



Development Principles

- Roadmap Driven: all pieces align with a reference architecture / roadmap
- Flexibility in inputs and outputs: allows variety of data types and meta data classifications to co-exist within the same system
- Scalable Design: retain system performance under increasing system load
- Wide Ranging: retain consistency with other information technology initiatives
- Technology Agnostic: allow for variety of technologies to exchange data
- Open source: allow interested parties to adopt, modify and improve the current state

Different Approaches for Different Circumstances

Geraghty – GeMS

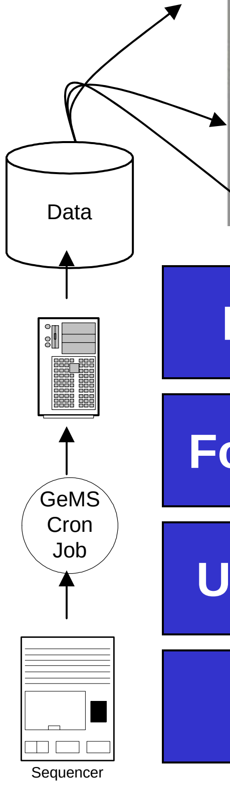
- Integration through usage
- Provide useful, needed tools – resulting in *de facto* common data

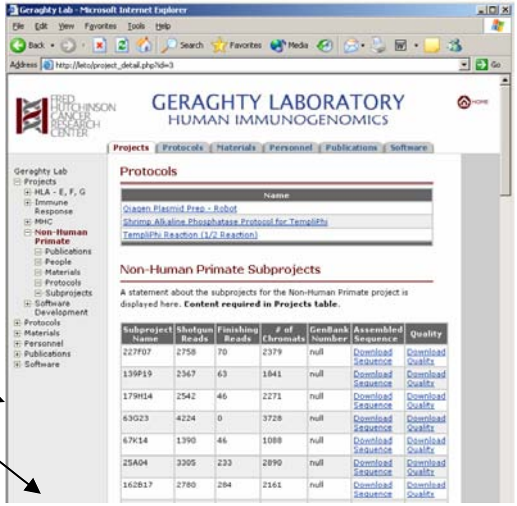
Thornquist – EDRN

- Integration through middleware
- Map existing databases to common data elements

From Data Generation to Data Publication

- Nightly Data pick up by system
- Unstructured and unrelated data sent to GeMS server for processing
- Data related to associated parameters
- Subset of data made available to the Geraghty website




Geraghty Lab - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://leto/project_detail.php?id=3
Go
FRED HUTCHINSON CANCER RESEARCH CENTER
HOME
Projects | Protocols | Materials | Personnel | Publications | Software
Geraghty Lab
Projects
HLA - E, F, G
Immune Response
MHC
Non-Human Primate
Publications
People
Materials
Protocols
Subprojects
Software Development
Protocols
Materials
Personnel
Publications
Software

GERAGHTY LABORATORY
HUMAN IMMUNOGENOMICS

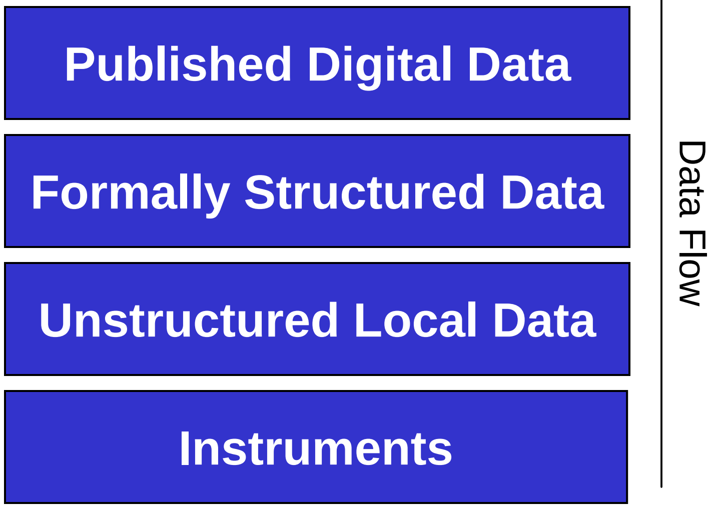
Protocols

Name
Qiagen Plasmid Prep - Robot
Shrimp Alkaline Phosphatase Protocol for TempliPhi
TempliPhi Reaction (1/2 Reaction)

Non-Human Primate Subprojects

A statement about the subprojects for the Non-Human Primate project is displayed here. **Content required in Projects table.**

Subproject Name	Shotgun Reads	Finishing Reads	# of Chromats	GenBank Number	Assembled Sequence	Quality
227F07	2758	70	2379	null	Download Sequence	Download Quality
139P19	2367	63	1841	null	Download Sequence	Download Quality
179H14	2542	46	2271	null	Download Sequence	Download Quality
63G23	4224	0	3728	null	Download Sequence	Download Quality
67K14	1390	46	1088	null	Download Sequence	Download Quality
25A04	3305	233	2890	null	Download Sequence	Download Quality
162B17	2780	284	2161	null	Download Sequence	Download Quality



Published Digital Data

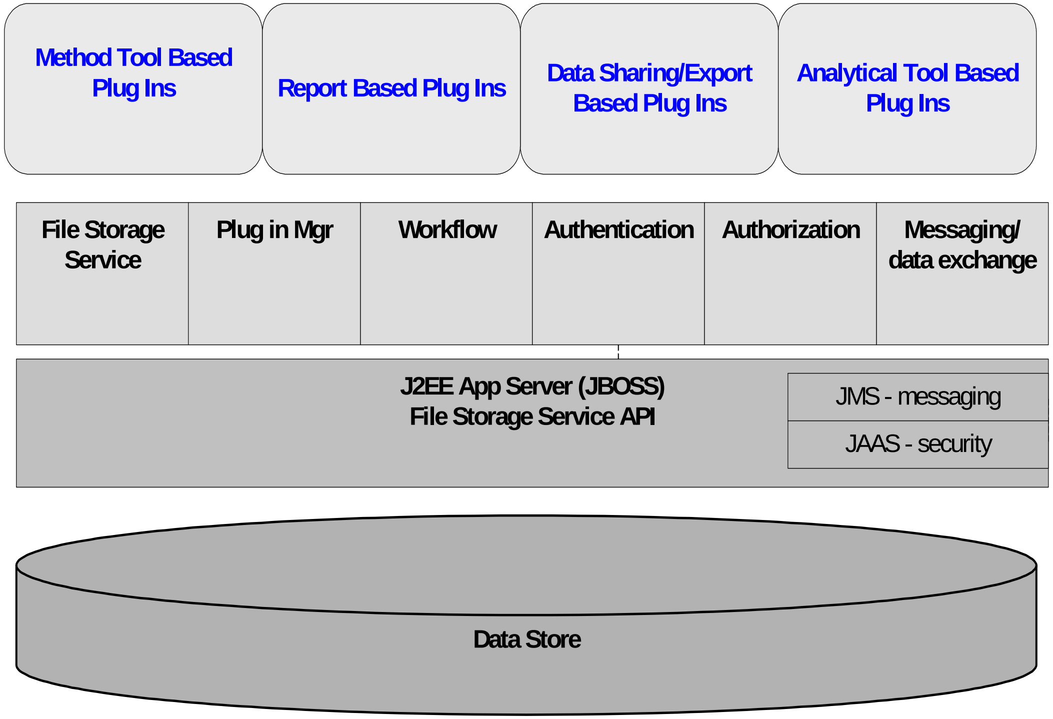
Formally Structured Data

Unstructured Local Data

Instruments

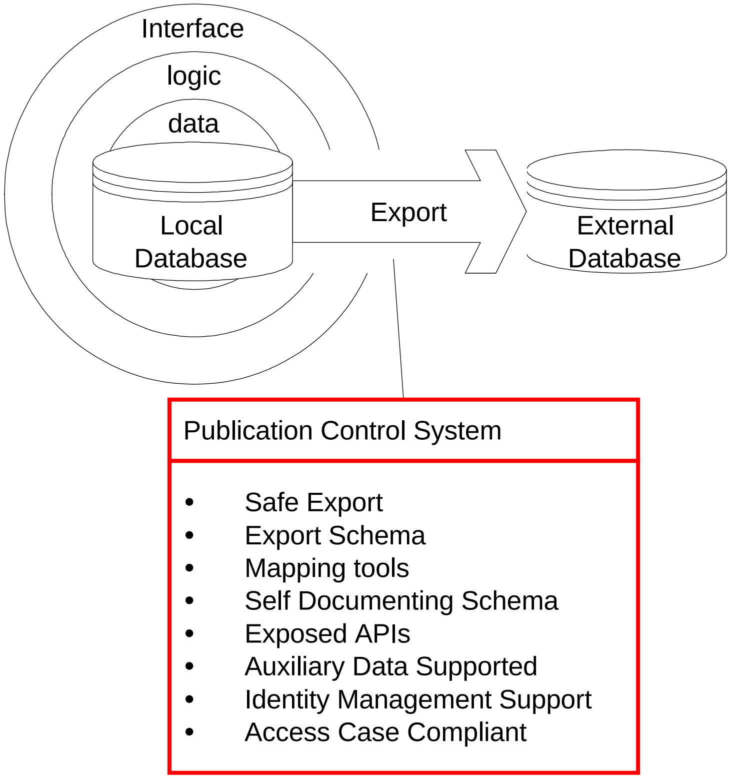
GeMS Architecture

- The data store is accessed through a file storage service API that acts as a DAO (Data Access Object) Layer.
- Core services is made available above J2EE application server. These services are used by the plugins to carry out their functions.
  - *File Storage Service– manages file system*
  - *Authentication – identify validation*
  - *Authorization – users level of access*
  - *Messaging – local workflow processes andcollaboration with remot sites*
  - *Plugin Manager – manages the resigration of plugin components*
  - *Workflow – manages the workflow agents, their states, and the associated triggers*
- Plugins represent the functional components that use the core services.



Generalizing the Data over a grid

- Next phase to build data sharing mechanism based on development of generic publication control system (export server)
- Test publication control and data sharing across disciplines with the Thornquist's EDRN/ERNE development efforts



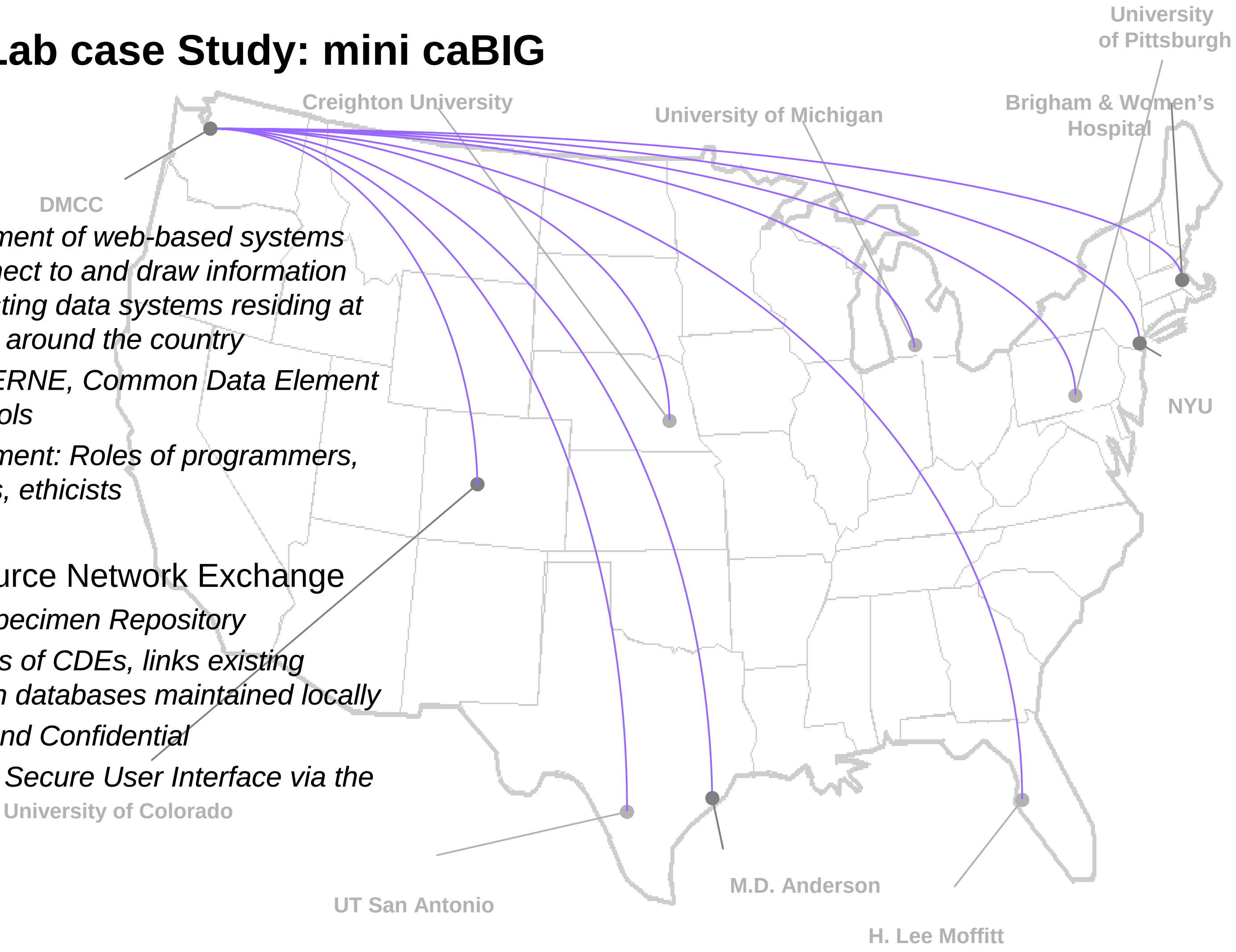
Publication Control System

- Safe Export
- Export Schema
- Mapping tools
- Self Documenting Schema
- Exposed APIs
- Auxiliary Data Supported
- Identity Management Support
- Access Case Compliant

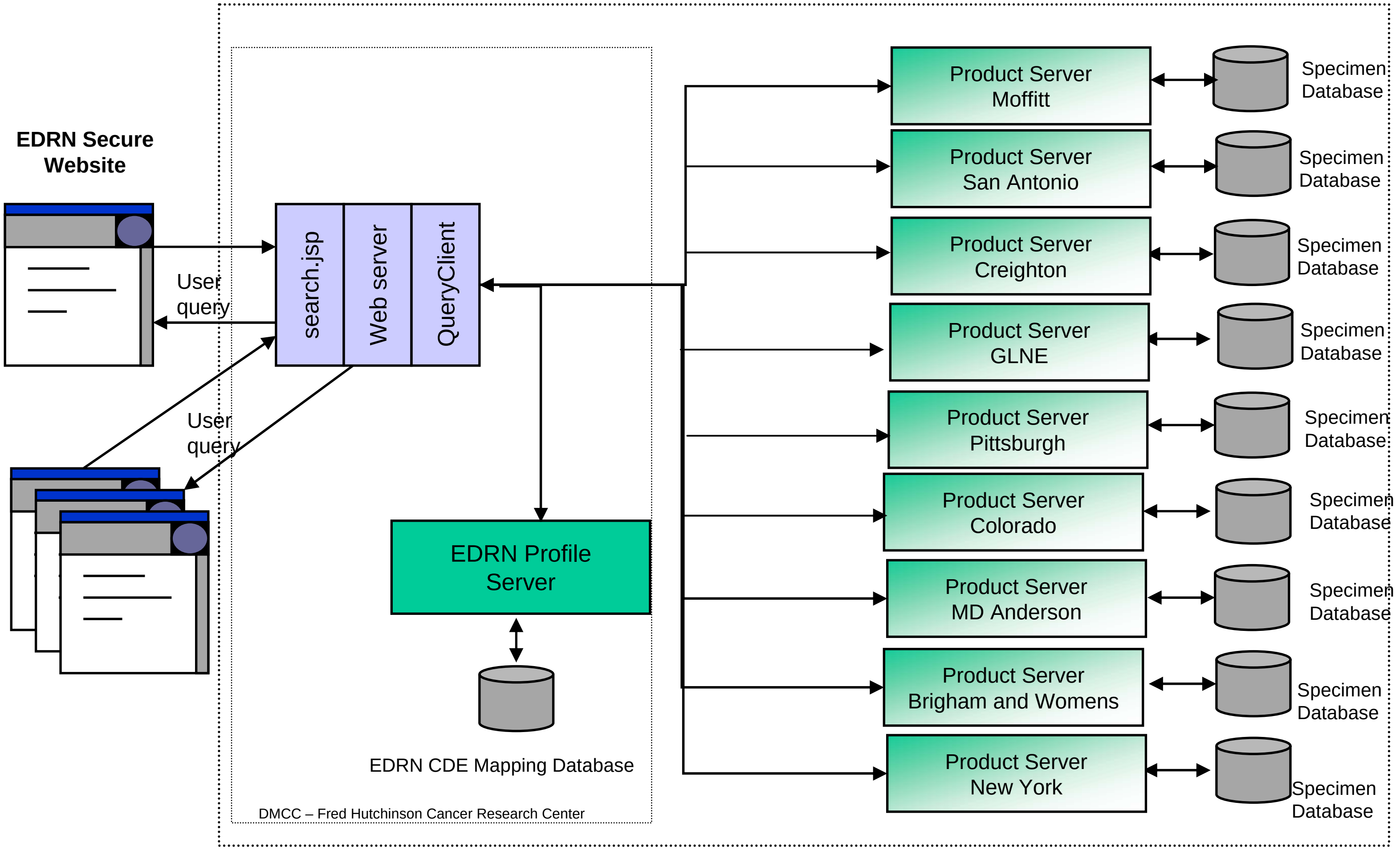
Thornquist Lab case Study: mini caBIG

- Structure
  - *Development of web-based systems that connect to and draw information from existing data systems residing at locations around the country*
  - *Pieces: ERNE, Common Data Element (CDE) tools*
  - *Development: Roles of programmers, scientists, ethicists*

- EDRN Resource Network Exchange
  - *Virtual Specimen Repository*
  - *By means of CDEs, links existing specimen databases maintained locally*
  - *Secure and Confidential*
  - *Dynamic Secure User Interface via the Internet*



Software Component Deployment

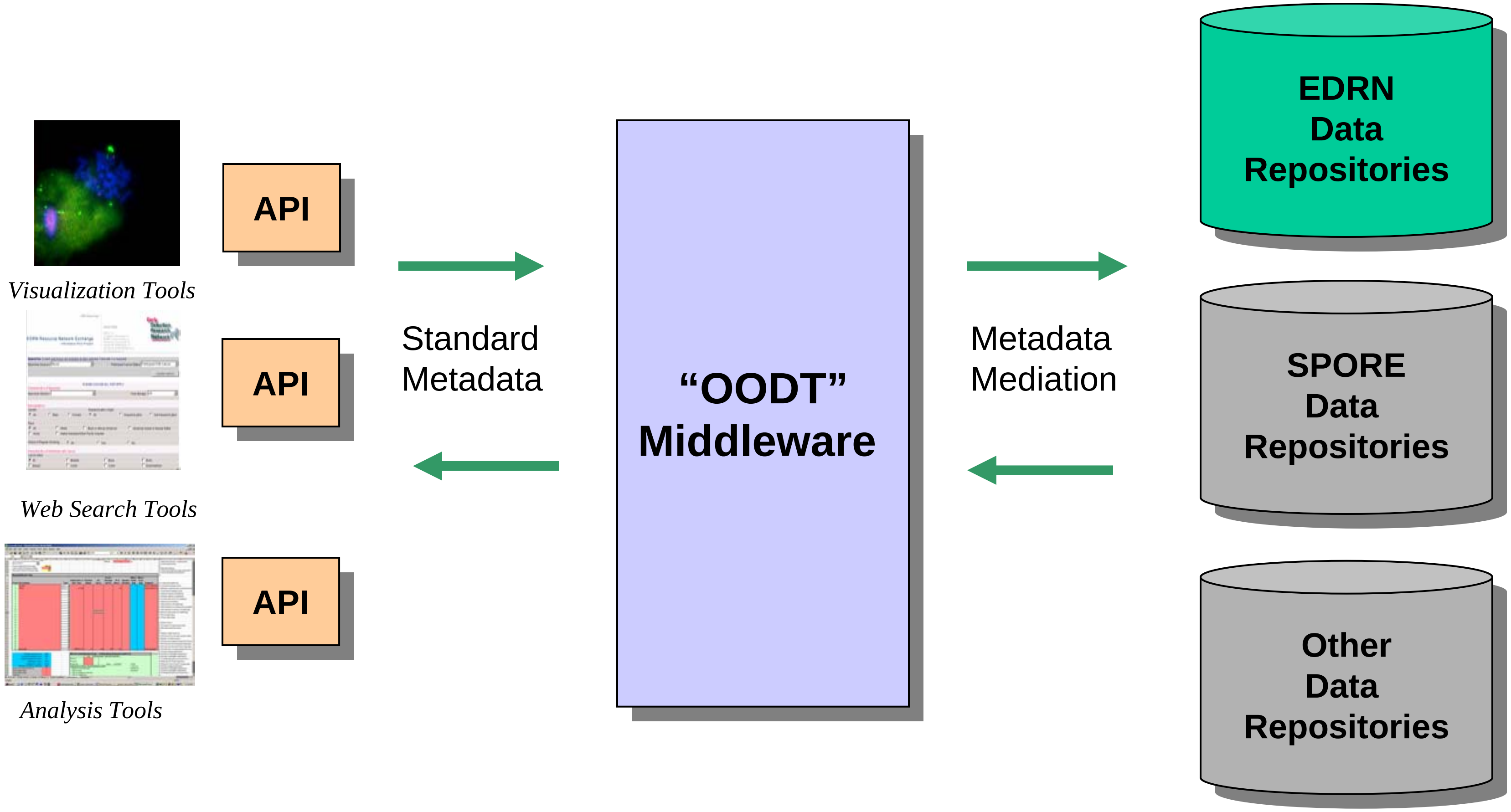


EDRN Bioinformatics Architecture

1. **API's exposed** for Bioformatics tools and applications

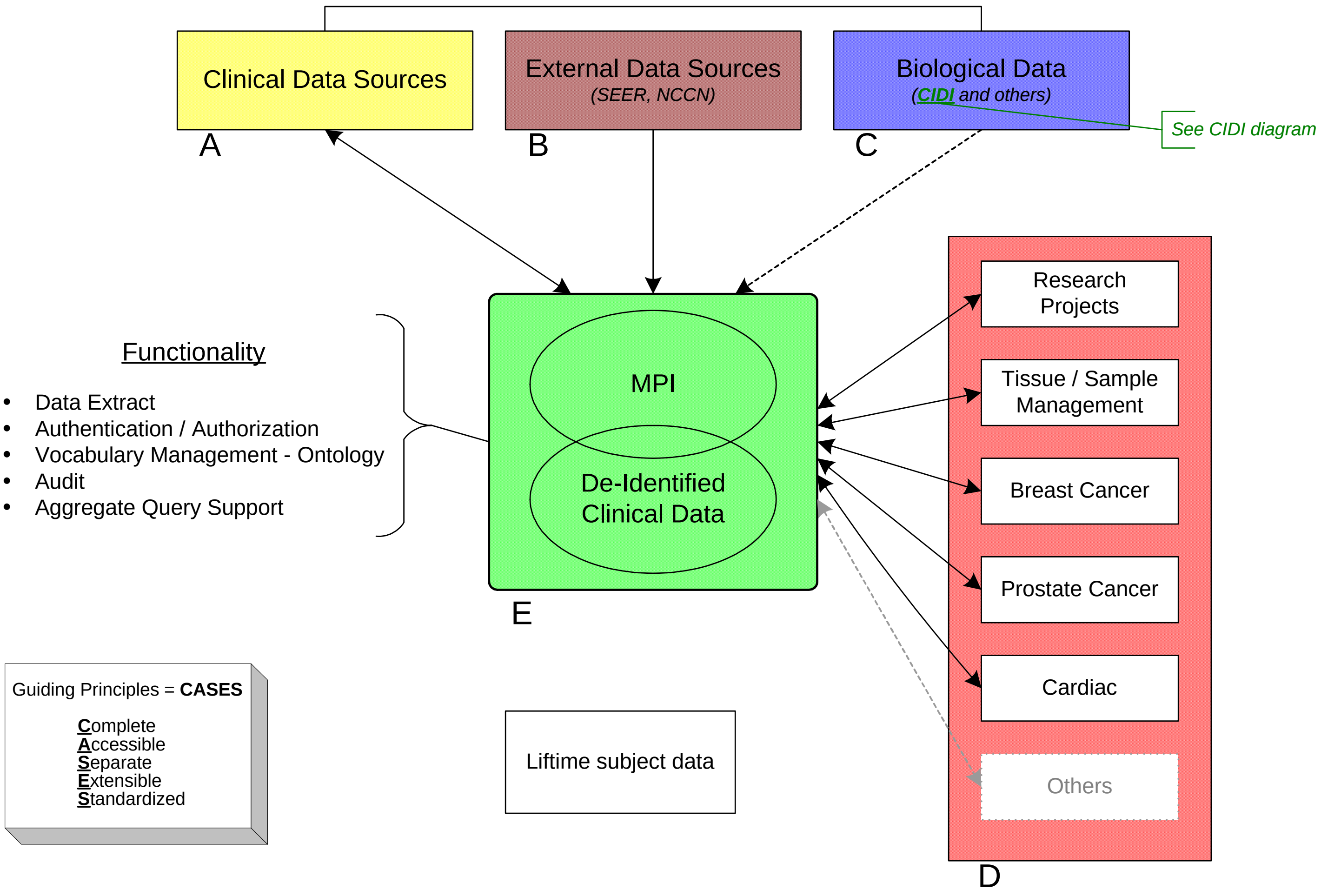
2. **Middleware** creates the informatics infrastructure connecting systems and data

3. **Repositories** for storing and retrieving many data types data



Reference Architecture – Conceptual Design

High Level View



Functionality

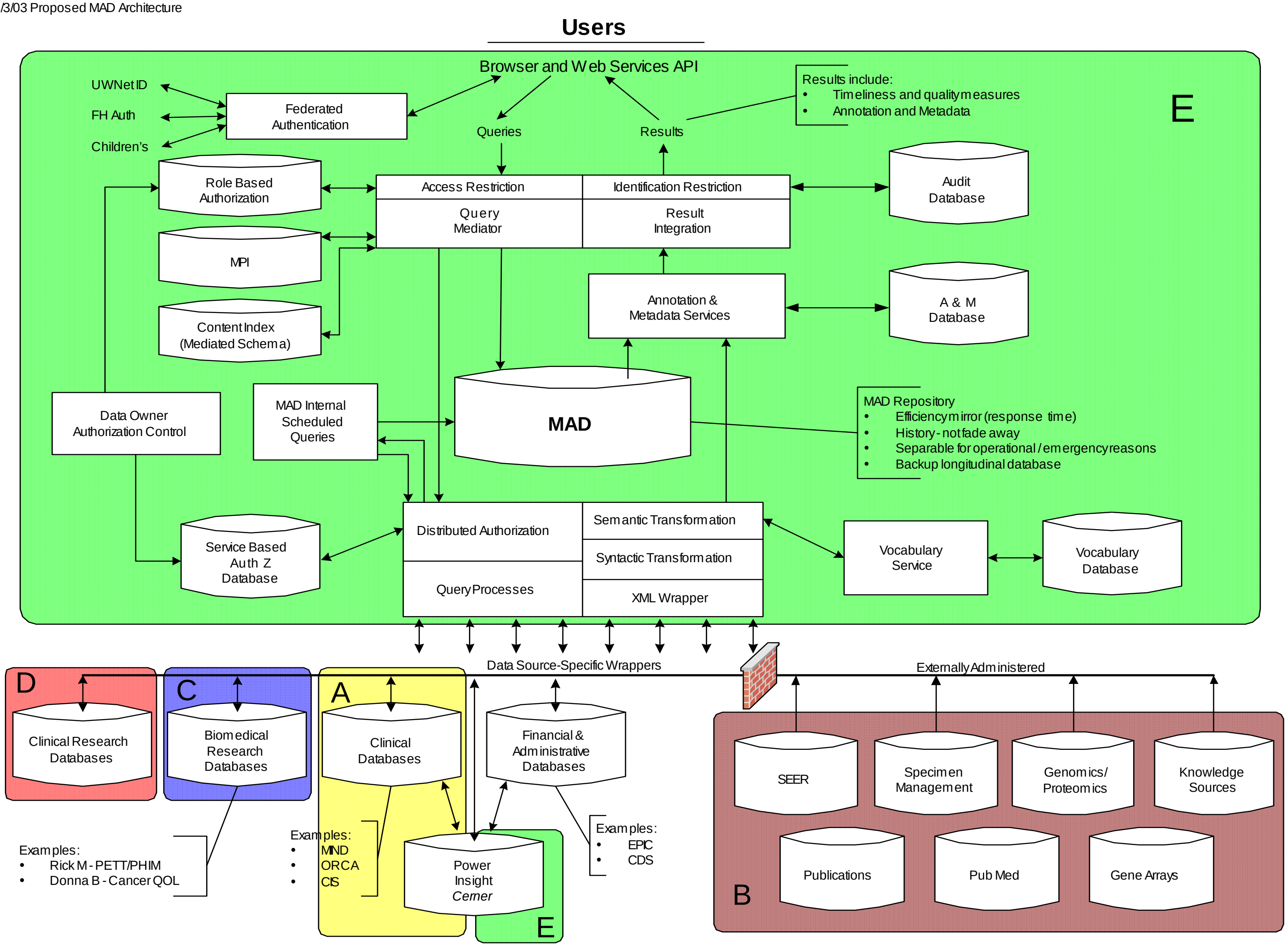
- Data Extract
- Authentication / Authorization
- Vocabulary Management - Ontology
- Audit
- Aggregate Query Support

Guiding Principles = **CASES**

Complete
Accessible
Separate
Extensible
Standardized

Reference Architecture – Detailed View

4/3/03 Proposed MAD Architecture



Users

Browser and Web Services API

UWNet ID
FH Auth
Children's

Federated Authentication

Queries
Results

Results include:
- Timeliness and quality measures
- Annotation and Metadata

E

Role Based Authorization

MPI

Content Index (Mediated Schema)

Access Restriction
Identification Restriction
Query Mediator
Result Integration

Audit Database

Annotation & Metadata Services

A & M Database

Data Owner Authorization Control

MAD Internal Scheduled Queries

MAD

MAD Repository
- Efficiency mirror (response time)
- History - not fade away
- Separable for operational / emergency reasons
- Backup longitudinal database

Service Based Auth Z Database

Distributed Authorization
Query Processes
Semantic Transformation
Syntactic Transformation
XML Wrapper

Vocabulary Service

Vocabulary Database

Data Source-Specific Wrappers

Externally Administered

D
Clinical Research Databases

C
Biomedical Research Databases

A
Clinical Databases

Financial & Administrative Databases

Power Insight
Cerner

E

Examples:
- Rick M - PETT/PHIM
- Donna B - Cancer QOL

Examples:
- MIND
- ORCA
- CIS

Examples:
- EPIC
- CDS

SEER
Specimen Management
Genomics/ Proteomics
Knowledge Sources
Publications
Pub Med
Gene Arrays

B

Summary

- Support the establishing and maintenance of common architecture
  - *Fostering alignment with a common vision in software design with an eye to collaboration*
  - *Development of tools that can interoperate between institutions/research initiatives*
  - *Understand the need to build and share these tools in a systematic way*

- Experience and Lessons Learned
  - *Managing and integrating systems from a variety of sources*
  - *Data publishing in real time as it becomes available*
  - *Challenges in supporting a variety of hardware and software systems*

- Flexibility is Essential
  - *Existing variability in data sets/systems/vocabularies/implementations that must be assembled in a grid environment*
  - *Depending on degree of expertise and budget available to the individual researcher*
  - *Based on the evolving nature of data elements in discovery oriented research*
  - *Based on the evolving nature of technology (connectivity, software platforms, hardware platforms)*