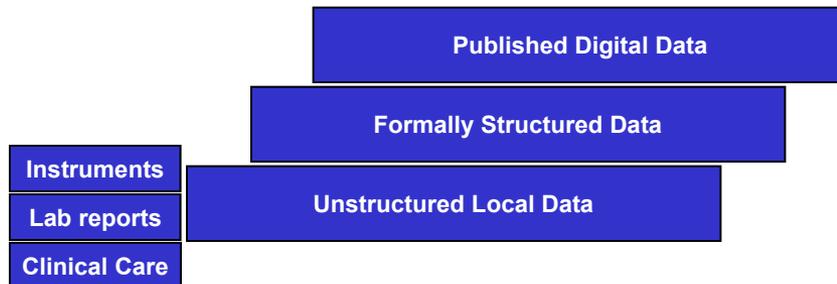
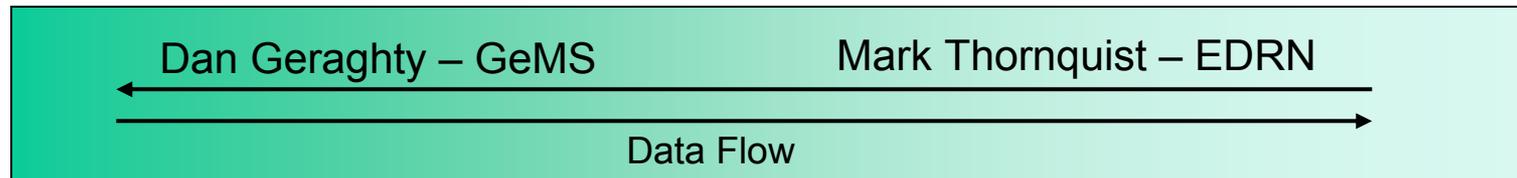


caBIG CDE/V Kickoff Meeting Presentation Fred Hutchinson Cancer Research Center

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

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

Thornquist – EDRN

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

Geraghty – GeMS

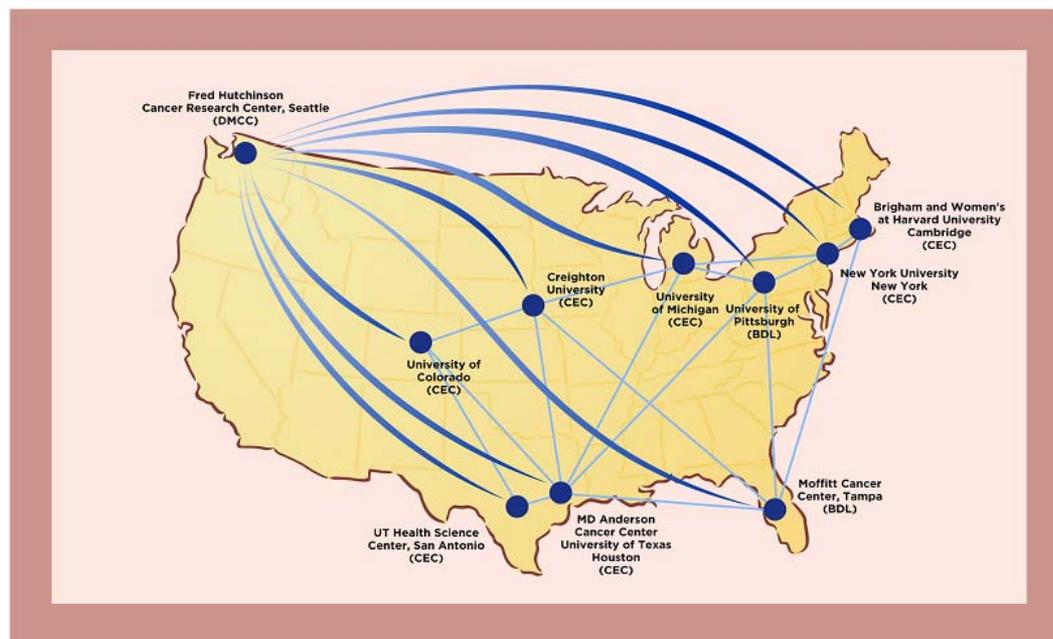
- Integration through usage
- Provide useful, needed tools – resulting in *de facto* common data elements

Thornquist Lab: Early Detection Research Network (EDRN)

- 5-Year collaboration supported by NCI
- Goal: Identify, evaluate, and validate promising biomarkers to support the early detection of cancer
- Comprised of:
 - *18 Biomarker Laboratories*
 - *9 Clinical and Epidemiology Centers*
 - *3 Biomarker Validation Laboratories*
 - *Data Management and Coordinating Center*
- Informatics Approach:
 - *Cross-disciplinary team of biomedical and computer science researchers*
 - *Common Data Elements to standardize data definitions for databases and forms*
 - *Informatics infrastructure that allows for capture and exchange of information across EDRN centers*
 - *Leverage JPL/NASA's experience and software in developing IT infrastructures to support planetary science*
 - *Use existing EDRN databases without requiring changes*
 - *Develop a common IRB protocol template*
 - *Common portals to access data (secure, validation, public, etc) as a single entry point*

EDRN Resource Network Exchange (ERNE)

- Virtual Specimen Repository
- Informatics infrastructure created for EDRN
- Existing sites specimen databases maintained locally
- Uses EDRN Common Data Elements (CDEs)
- Maps institutions local data definitions to EDRN CDEs
- Secure and Confidential
- Secure Dynamic Portal



EDRN Informatics Tools

- EDRN Secure Website - CDE Tools
 - CDE Repository
 - Form Tools
 - Mapping Tools
- EDRN Resource Network Exchange (ERNE)
 - An infrastructure for sharing data resources across EDRN
 - Supports real time (on demand) distribution of data to users
 - First release - Specimen sharing tool
 - EDRN CDE Mapping Tool
- Validation Infrastructure (VSIMS)
 - Provide common infrastructure across validation studies
 - Online Forms - Data Driven from CDE Repository

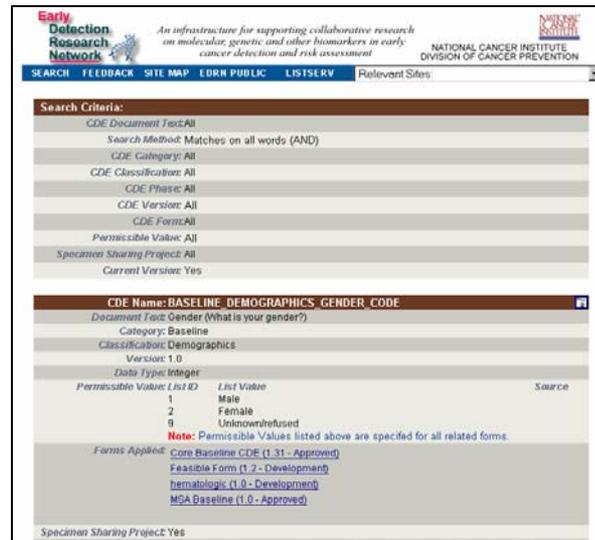


The image displays three screenshots of EDRN informatics tools:

- VSIMS (Validation Infrastructure for Supporting Microsatellite Analysis):** A web interface for a study titled "Detection of Bladder Cancer by Microsatellite Analysis (MSA) of Urinary Sediment: Multicohortational Study". It includes a navigation menu (Submit Data, Confirm Eligibility, Specimens, Reports, Issue Tracking, Study Info), a "Study Update" section with a timeline of events (e.g., "Timeline updated November 25, 2003"), and a "Message Board" with an upcoming two-day investigator meeting in Baltimore, Maryland.
- EDRN Resource Network Exchange (ERNE):** A search interface for "EDRN Resource Network Exchange - Informatics Pilot Project". It features a search bar, filters for "Specimen Source" (Blood) and "Participant Cancer Status" (Participant With Cancer), and a "Characteristics of Specimen" section with dropdowns for "Specimen Storage" and "Final Storage". It also includes a "Demographics" section with radio buttons for Gender, Hispanic/Latino Origin, Race, and History of Regular Smoking.
- EDRN Informatics Tools:** A search criteria page for "CDE Name: BASELINE, DEMOGRAPHICS, GENDER_CODE". It lists various search filters such as "CDE Document Type", "CDE Category", "CDE Classification", "CDE Phase", "CDE Version", "CDE Form", "Permissible Value", "Specimen Sharing Project", and "Current Version". It also includes a "CDE Name: BASELINE, DEMOGRAPHICS, GENDER_CODE" section with a dropdown for "Document Type: Gender (What is your gender?)" and a table for "Diels Type: Integer" with permissible values and list values.

Legacy Data and Mapping

- Semantic Architecture
 - Many institutions have existing specimen repositories with locally defined data models
 - EDRN Common Data Elements (CDEs)
 - ISO/IEC 11179
- Data Model Mapping
 - Communicating EDRN CDEs
 - EDRN CDE Mapping Tool
 - EDRN CDE Repository



Early Detection Research Network
An infrastructure for supporting collaborative research on molecular genetic and other biomarkers in early cancer detection and risk assessment
NATIONAL CANCER INSTITUTE
DIVISION OF CANCER PREVENTION

SEARCH | FEEDBACK | SITE MAP | EDRN PUBLIC | LISTSERV | Relevant Sites

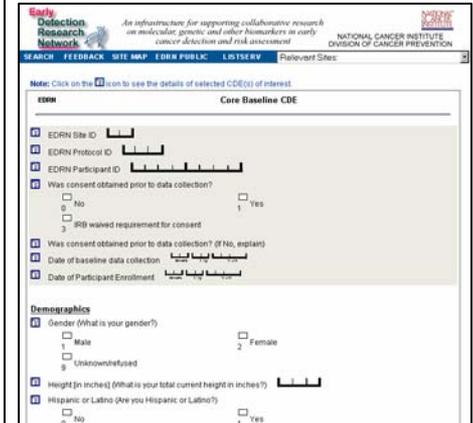
Search Criteria:

CDE Document Text: All
Search Method: Matches on all words (AND)
CDE Category: All
CDE Classification: All
CDE Phase: All
CDE Version: All
CDE Form: All
Permissible Value: All
Specimen Sharing Project: All
Current Version: Yes

CDE Name: BASELINE_DEMOGRAPHICS_GENDER_CODE

Document Text: Gender (What is your gender?)
Category: Baseline
Classification: Demographics
Version: 1.0
Data Type: Integer
Permissible Value List ID List Value Source
1 Male
2 Female
9 Unknown/refused
Note: Permissible Values listed above are specified for all related forms.
Forms Applied: Core Baseline CDE (1.31 - Approved)
Feasible Form (1.2 - Development)
hematologic (1.0 - Development)
MSA Baseline (1.0 - Approved)

Specimen Sharing Project: Yes



Early Detection Research Network
An infrastructure for supporting collaborative research on molecular genetic and other biomarkers in early cancer detection and risk assessment
NATIONAL CANCER INSTITUTE
DIVISION OF CANCER PREVENTION

SEARCH | FEEDBACK | SITE MAP | EDRN PUBLIC | LISTSERV | Relevant Sites

Note: Click on the  icon to see the details of selected CDE(s) of interest.

EDRN

Core Baseline CDE

EDRN Site ID
 EDRN Protocol ID
 EDRN Participant ID
 Was consent obtained prior to data collection?
 No
 IRB waived requirement for consent
 Was consent obtained prior to data collection? (if No, explain)
 Date of baseline data collection
 Date of Participant Enrollment

Demographics

Gender (What is your gender?)
 Male
 Female
 Unknown/refused
 Height (in inches) (What is your total current height in inches?)
 Hispanic or Latino (Are you Hispanic or Latino?)
 No
 Yes

Note: Click on the  icon to see the details of selected CDE(s) of interest.

Map ID: 1059 Last Modified Date: 2002/07/08

Site and Protocol Information

58 The Biomarker Knowledge System Informatics Pilot Project Supplement To The Biomarker Development Laboratory at Moffitt (Bedlam) 89 H. Lee Moffitt Cancer Center (BDL)

[Protocol Database and Site Technology Survey](#)

Mapping Results		Comments
	EDRN CDE 	EDRN2_Sput_Epi
Table Name	BASELINE_DEMOGRAPHICS_GENDER_CODE	M_Sput_Subject
CDE Name	BASELINE_DEMOGRAPHICS_GENDER_CODE	SEX
CDE Version	1.0	
CDE Data Type	Integer	Character
CDE Document Text	Gender (What is your gender?)	Gender
Permissible Value	1 Male 2 Female 9 Unknown/refused	F M U
CDE Comments		Case M.Cdem_SEX F -> E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 2 M -> E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 1 Else E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 9 EndCase Else E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 9 EndCase
Mapping Solution		
Mapping Comments		This is being done in Product Server
Mapping Status	Completed	Completed

EDRN CDE LU	EDRN2_Sput_Epi
CDE Category: Baseline	
CDE Classification: Demographics	
Table Name: BASELINE_DEMOGRAPHICS_GENDER_CODE	M_Sput_Subject
CDE Name: BASELINE_DEMOGRAPHICS_GENDER_CODE	SEX
CDE Version: 1.0	
CDE Data Type: Integer	Character
CDE Document Text: Gender (What is your gender?)	Gender
Permissible Value: 1 Male 2 Female 9 Unknown/refused	F M U
CDE Comments:	
Mapping Solution Proposed:	Case M.Cdem_SEX F -> E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 2 M -> E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 1 Else E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 9 EndCase Else E.BASELINE_DEMOGRAPHICS-GENDER_CODE = 9 EndCase
Others:	
Mapping Comments:	This is being done in Product Server

Shared needs of the local (small) genetics data generating labs.

- **Laboratory organization and data flow.**
 - *Solid informatics infrastructure essential for data retrieval (i.e., a lab notebook).*
 - *Efficient data tracking improves data quality and lowers costs.*
- **Collaboration Potential. Ability to easily share data in a secure manner.**
 - *Labs at different localities collaborating on a project.*
 - *Acquiring genomic data developed in another lab (e.g. for genotyping).*
 - *Pooling data among labs to increase sample size.*
 - *Pooling genetic data from common samples (e.g. building haplotypes).*
 - *Sharing data for standardization (e.g. STRs).*

To address these issues we are building a Genetics Management Software suite.

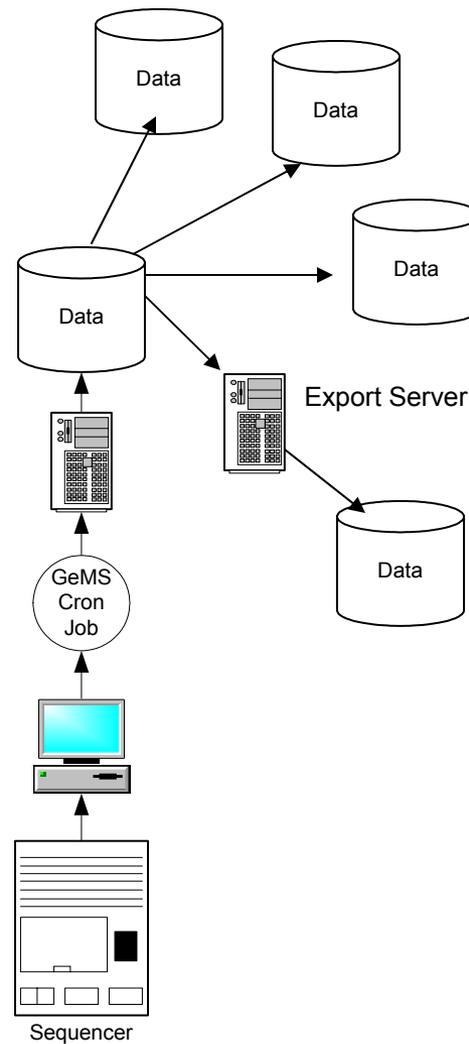
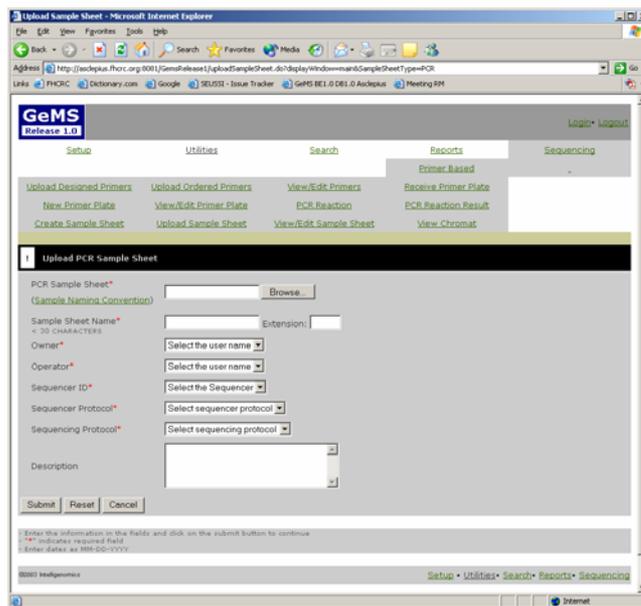
Geraghty Lab: GeMS Approach

- Wide area data integration is seen as stack of activities
- Focus on bringing full power of high throughput DNA sequencing instruments into hands of small (R01-funded) laboratory



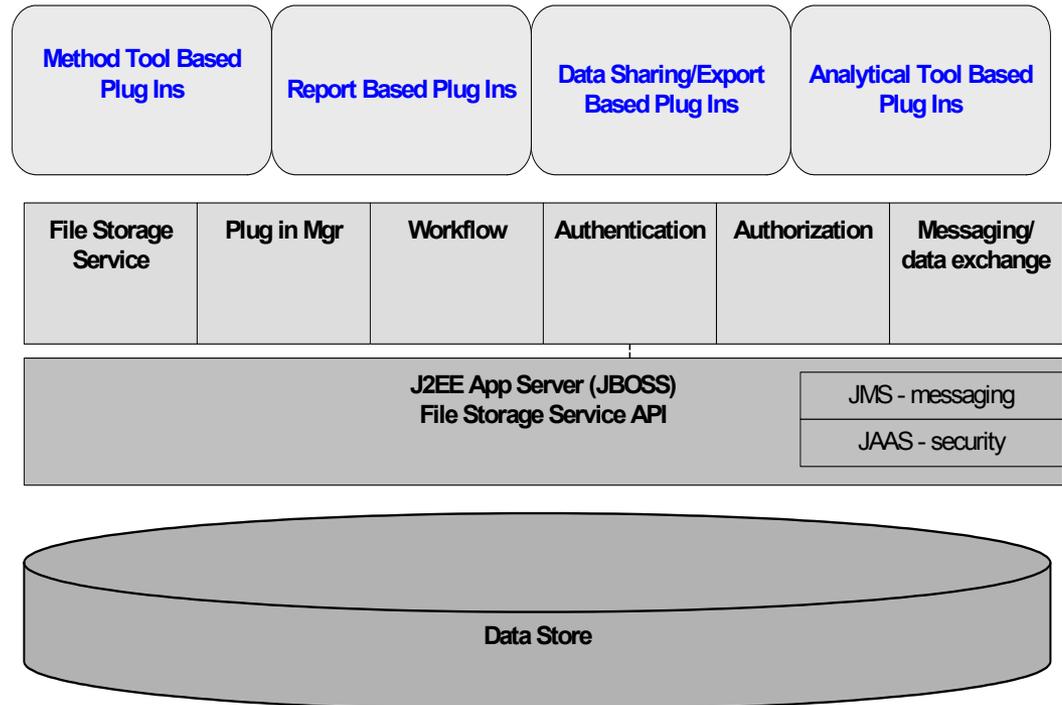
GeMS System Overview

- Data generated from sequencer
- Converted to standardized text formats
- Populated into published schema which relates variables to one another



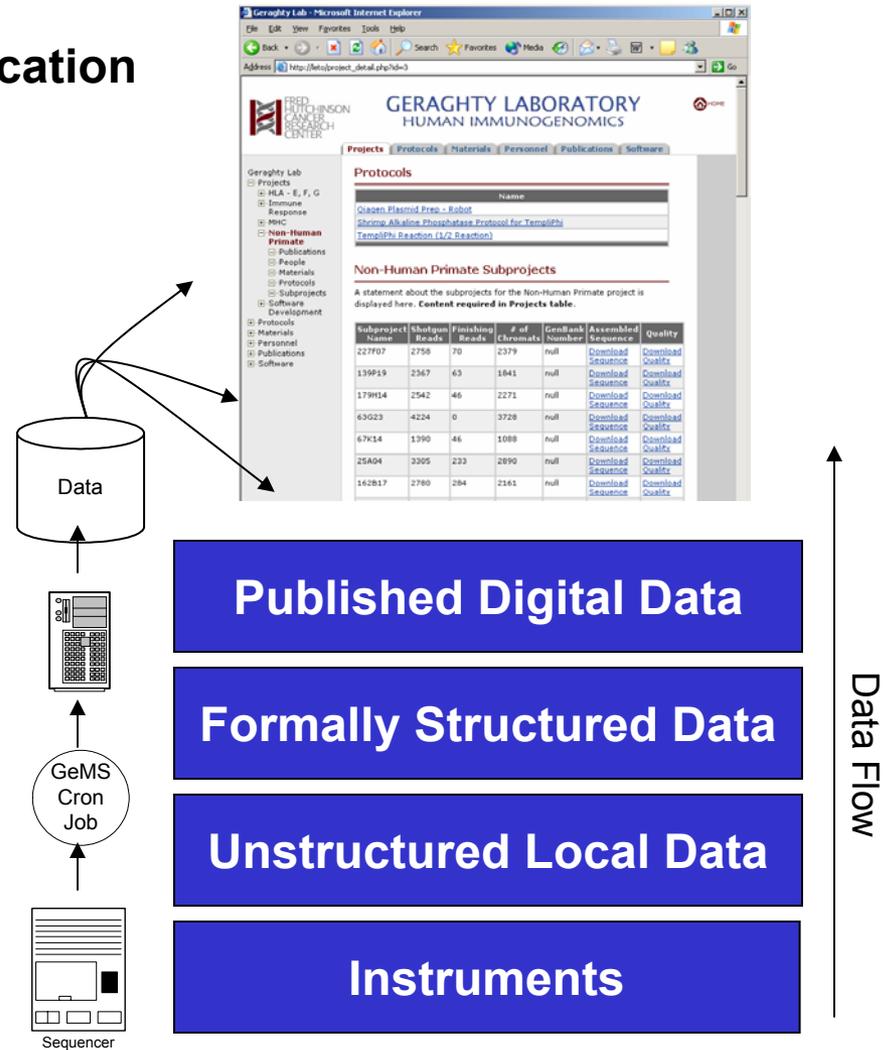
GeMS Architecture

- 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 and collaboration 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.

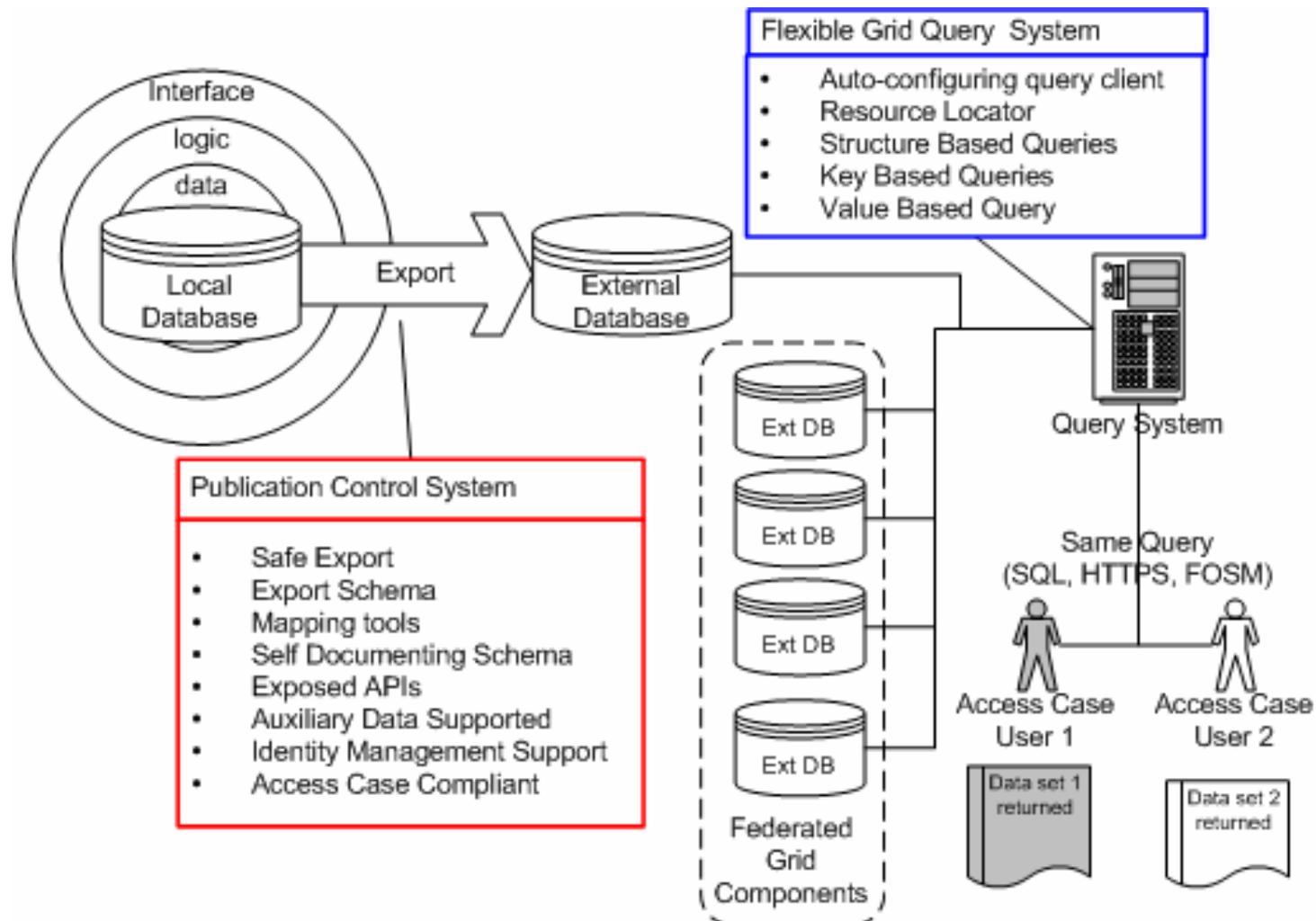


From Data Generation to Data Publication

- 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



Proposed Collaboration/Contribution to caBIG



Summary

- Support establishment and maintenance of Common Data Elements by
 - *Fostering the generation of CDE/Vs from the data gathering instruments (GeMS)*
 - *Developing tools to interpret and integrate legacy data (EDRN)*
 - *Understanding the need to build and share mapping tools (GeMS - EDRN)*
- Experience and Lessons Learned
 - *Managing and integrating data sets from a variety of sources*
 - *How to share data effectively across data grids*
 - *Data publishing in real time as it becomes available*
- Flexibility is Essential
 - *Consider the variability in data sets that must be assembled in a grid environment*
 - *Depends on the perspective of the study itself, or point of view of the researcher*
 - *Support for evolving data elements and classifications in discovery-oriented research*
 - *Supporting the scientist by delivering tools that add value as a mechanism for delivering established CDEs.*