

Cyber Dumpster Diving – creating new software systems for less

**Ian Gorton,
R&D Manager,
Data Intensive Scientific Computing,
Computational Sciences and Math Division
Pacific Northwest National Lab**



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Pacific Northwest National Lab

- ▶ Department of Energy Science Lab
 - Fundamental sciences
 - National security
- ▶ 4500+ people
- ▶ Business volume of over \$1b per annum
- ▶ Large scale experimental facilities, e.g.
 - Environmental Molecular Sciences Lab (EMSL)
 - 161 Tflop supercomputer



Proudly Operated by Battelle Since 1965

DISC@PNNL

► Data Intensive Scientific Computing

- User platforms
- Data management
- Tool integration
- Workflows
- Provenance

► Applications in e.g.

- Bioinformatics
- Climate modeling
- Carbon sequestration
- Subsurface modeling



High Performance Computing



Scientific User Environments



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

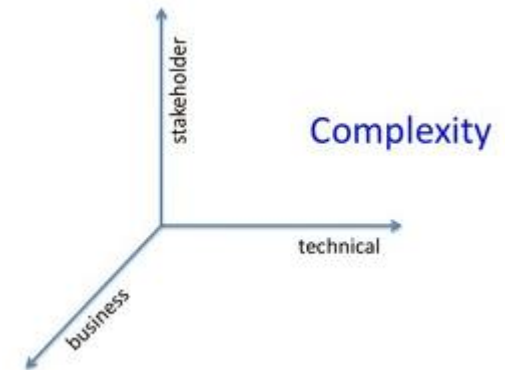
The middle is a hard place ...

► Requirements

- Need to understand science domain
- Need to understand HPC
- Difficult to define, constant refinement, negotiations, communications
- “The hardest single part of building a software system is deciding precisely what to build.”

► Design

- Conflicting quality requirements
- Complex, heterogeneous technologies
- Large data
- Proliferation of tools, variable quality

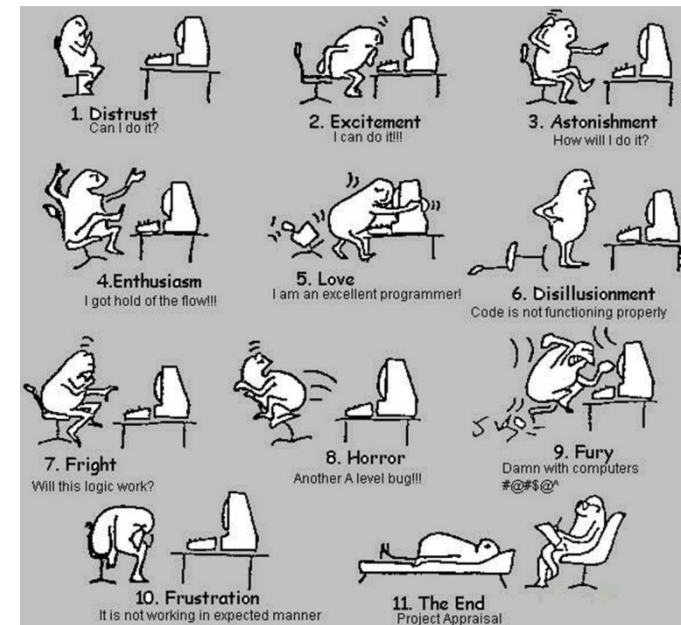


Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Project Funding Profiles

- ▶ Typically fixed amounts
 - What can we build with X dollars?
 - Fixed amounts per year, 1-3 year lifecycle
- ▶ Limited funding
 - From .25 to 10 team size per year
 - 1-2 people per year most common
- ▶ High expectations
 - Scientists think 'software is easy'
 - it's just coding, right?



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

**The most radical possible solution
for constructing software is not to
construct it at all.**

Fred Brooks: No Silver Bullet: Essence and Accidents of Software Engineering



Pacific Northwest
NATIONAL LABORATORY

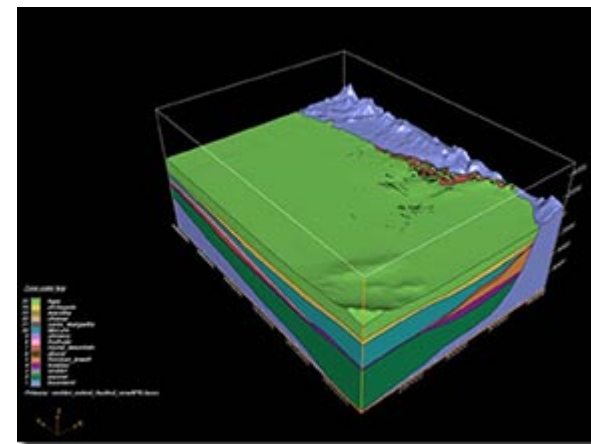
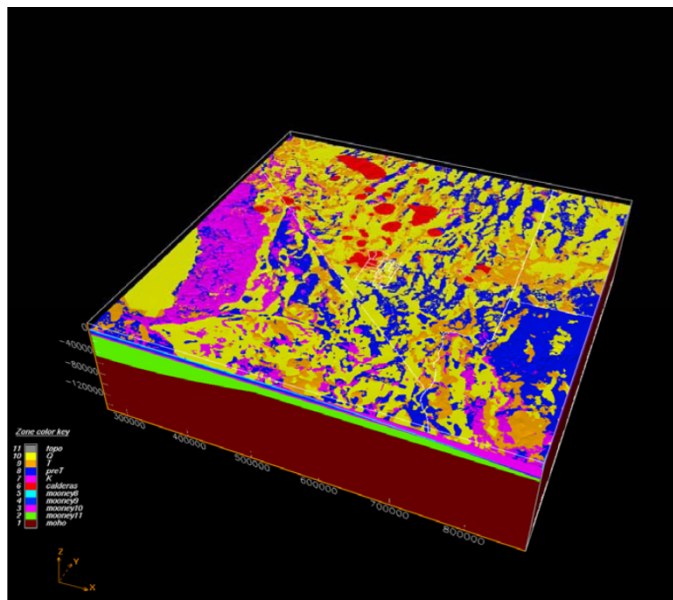
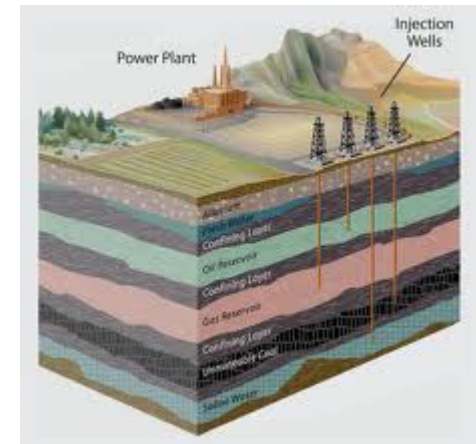
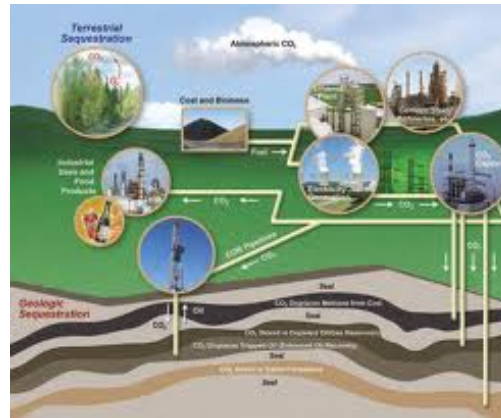
Proudly Operated by Battelle Since 1965



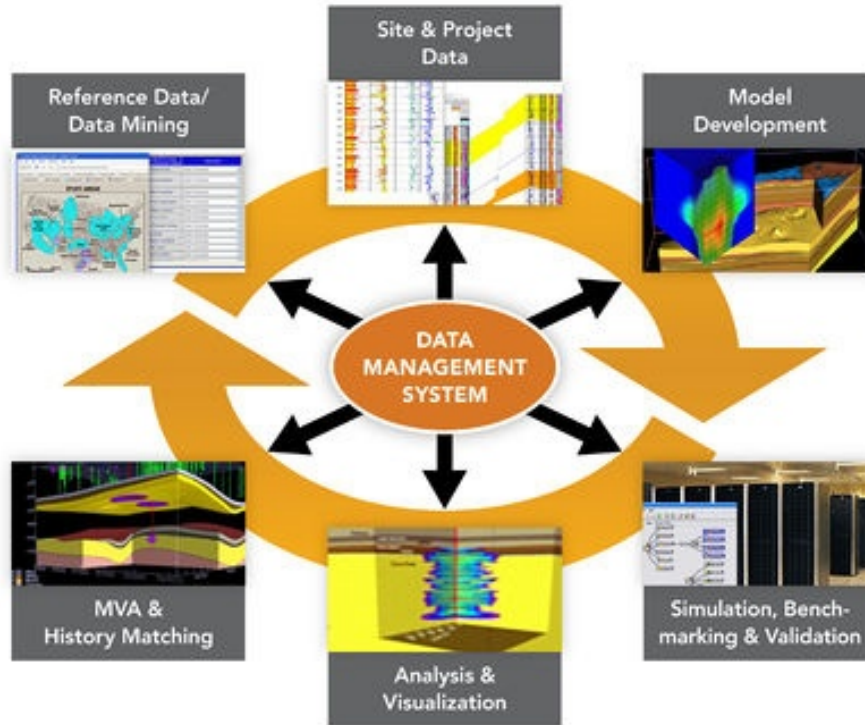
Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Carbon Sequestration (Storage)

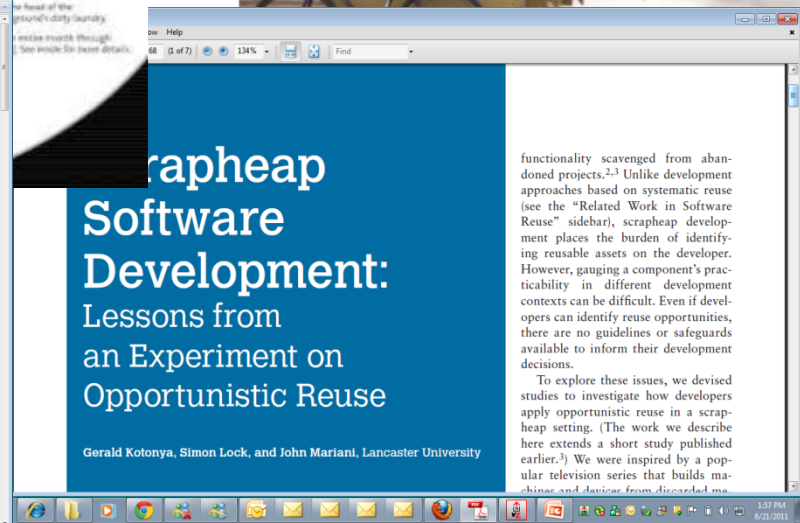
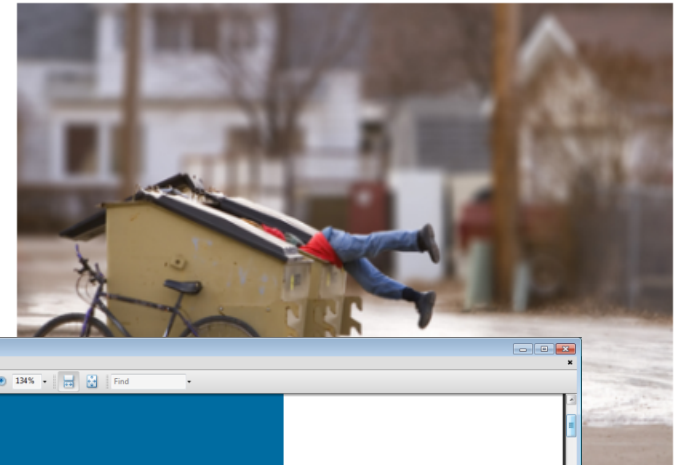
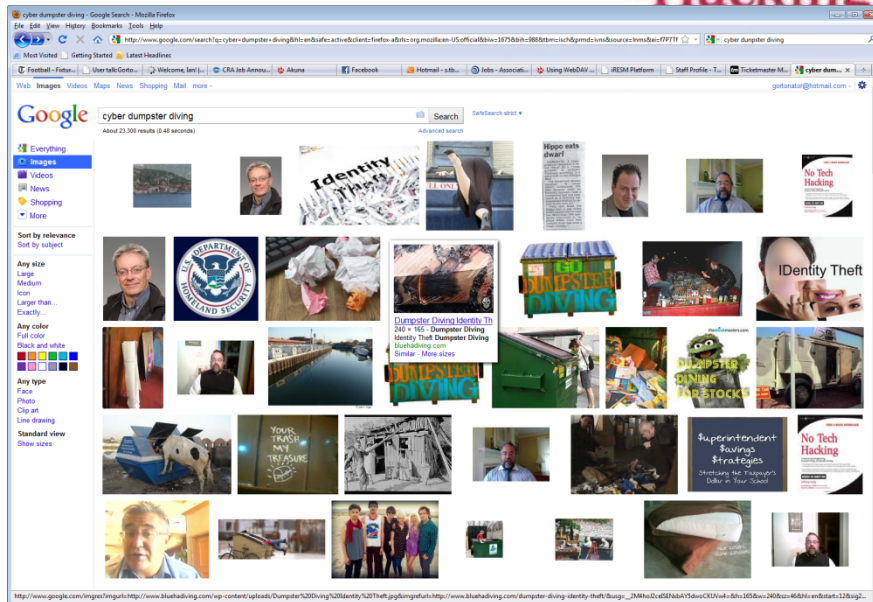


Geological Sequestration Software Suite (GS3)



- ▶ Large-scale, complex data
 - Experimental
 - HPC Simulation inputs/outputs
 - Multiple realizations for uncertainty quantification
- ▶ Long-lived projects
 - Modeling
 - Analysis
 - Monitoring (100+ years)

A powerful, usually legal, source of information that isn't seriously defended because of social taboos.



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

‘Write-as-little-code-as-possible’ Reuse

► Approach:

- Leverage open source frameworks and tools
- Extend to support science applications
- Generalize to support multiple science domains

► Requires:

- Careful technology selection
- Creative design
- Robust architectures



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Velo – Knowledge Management for Modeling and Simulation



Pacific Northwest
NATIONAL LABORATORY

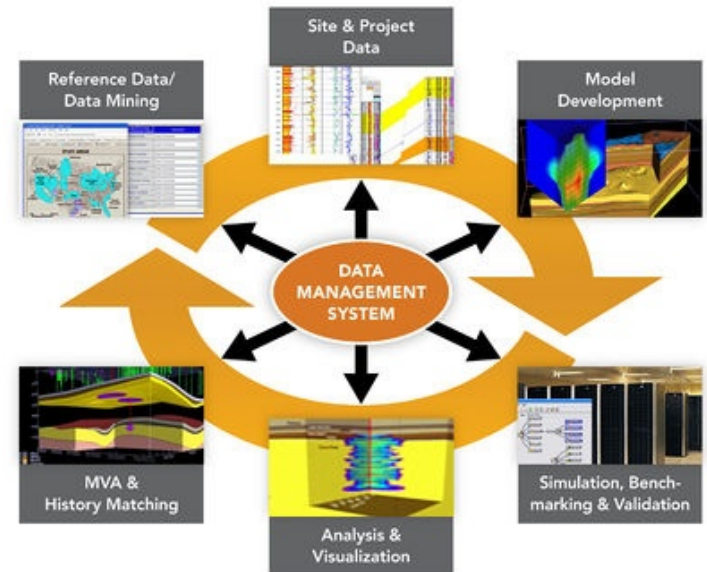
Proudly Operated by Battelle Since 1965

Supporting Carbon Sequestration Modeling

► Requirements

- Collaboration
- Sharing data
- Metadata management
- User-driven customization
- Extensibility
- Model and data versioning
- Provenance and user annotation
- Robust, scalable

► Small project, team ~1.75 people, 3 years



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Cyber Dumpster Diving Process ;)

- ▶ Open source
- ▶ Candidate technology assessments:
 - Quality of docs
 - Release schedule
 - Community scope
 - APIs
 - Code/architecture
 - Install and workout, simple tests



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Feature-Reuse Matrix

Feature	Solution	Notes	Reuse
Collaboration	Mediawiki	Core wiki features support this	100%
Sharing data	Mediawiki Alfresco	Requires integration of MW and Alfresco	60%
Metadata management	Mediawiki Alfresco	Requires customization of MW and Alfresco basic features	80%
User-driven customization	Mediawiki	Core wiki features support this	100%
Extensibility	Mediawiki Alfresco	APIs support extension, but requires design of exact integration mechanisms	20%
Model versioning	Mediawiki Alfresco	Minor extensions for MW/Alfresco capabilities	75%
Provenance	Mediawiki	Some for free in MW, but advanced features need developing	20%
Role-based Security	Halo ACL	Mediawiki extension	100%



STORY

GS3 Examples - Semantic Capabilities - Metadata Extraction

► Metadata:

- Generic information e.g. file size, owner, preview/thumbnails
- Specific to the file type, e.g. keywords, geographic location

► Metadata is searchable

► Extensible architecture for custom data types ingest pipelines, e.g.

- Simulation outputs
- Spreadsheets
- Input files

The screenshot displays the 'Velo Demonstration' web application. The top navigation bar includes 'Home', 'Browse', 'Tools', 'Misc. Tools', and 'Account Links'. The main content area shows a file browser view for the path '/refdata/Illinois Basin/Zhou et al 2009.pdf'. The file is a PDF, and its first page is previewed. The preview text includes the title 'Modeling Basin- and Plume-Scale Processes of CO2 Storage for Full-Scale Deployment' and an abstract describing the simulation. To the right of the preview, a table titled 'Top Category/Keyword Matches' lists various geological and geophysical terms with their respective counts.

Top Category/Keyword Matches	
Rock Name (32)	Sandstone (17) Granite (8) Shale (7)
Geologic Formation Name (156)	Mt. Simon (119) Eau Claire (34) St. Peter (3)
Rock Property (188)	Permeability (93) Salinity (29) Porosity (22) Compressibility (19) Depths (13) Thick (12)
Data Identifiers (75)	Core (61) Wells (9) Geophysical (3) Seismic (2)
Geologic Sequestration Unit (29)	Caprock (25) Reservoir (4)

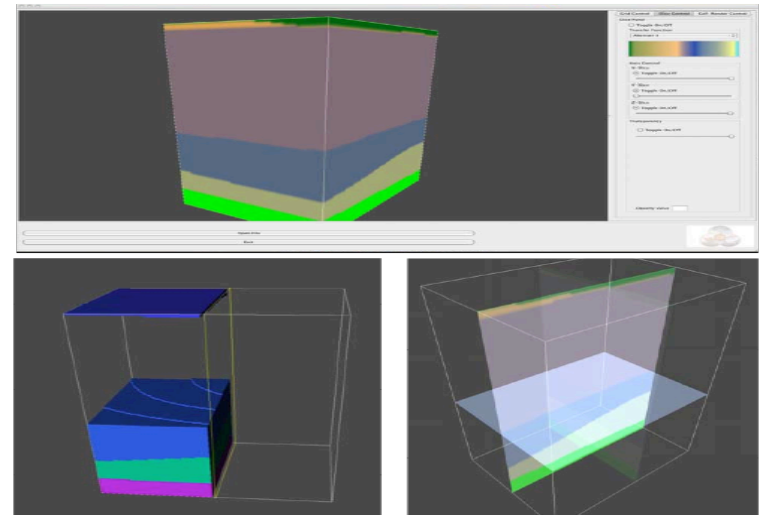
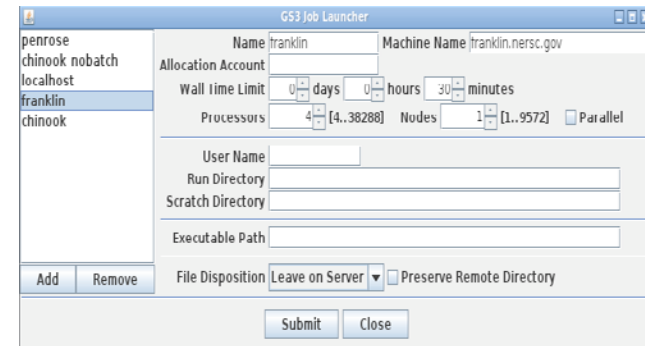


Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

GS3 Examples - Tool Integration

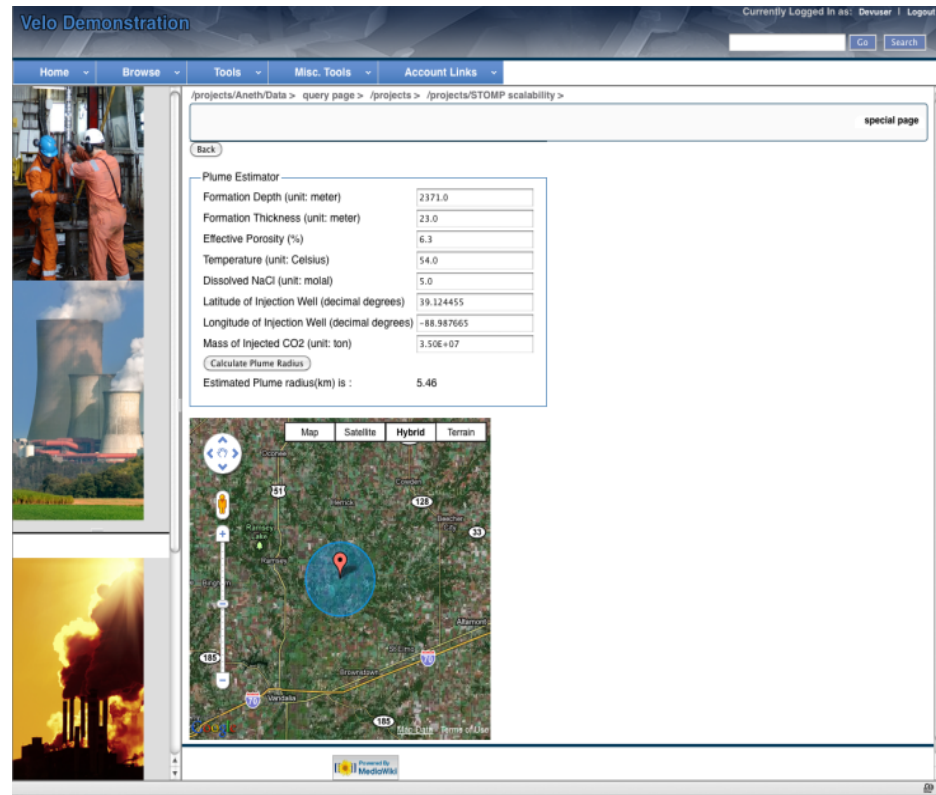
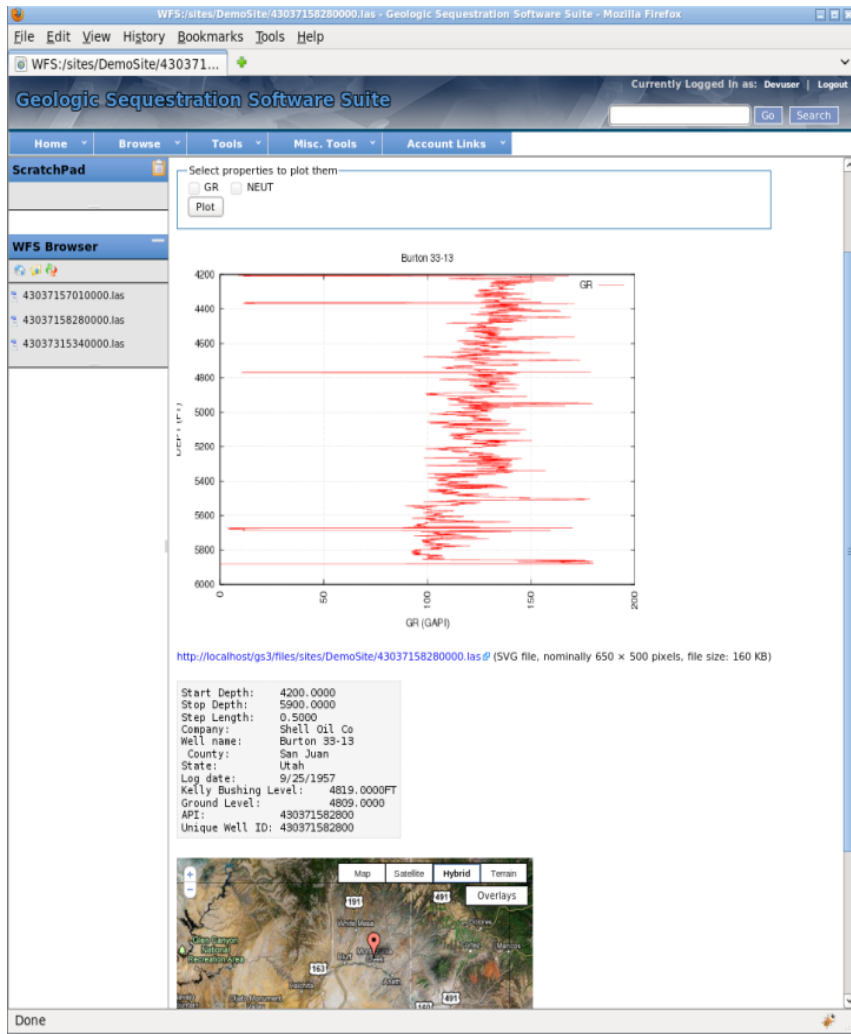
- ▶ Mediawiki plugins
- ▶ 'Black box' tools
- ▶ External 3rd party tools



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

GS3 Examples – Tool Plugins



GS3 Examples – Black box Tool Plugins

GS3 Model Attributes: General Site Description

Location Geography Reservoir Category **Target Formations** Seal Formations

Number of Target Formations within Reservoir: 1

Target Formation #1

Target Formation Name: Mt Simon Geologic Age: Cambrian

Target Formation Rock Types (check all that apply):

☒ Sandstone ☐ Limestone ☐ Dolomite ☐ Shale ☐ Coal Seam ☐ Basalt

Other (Specify):

Depositional Environment (check all that apply):

Continental: ☐ Alluvial ☐ Aeolian ☐ Fluvial ☐ Lacustrine

Transitional: ☐ Deltaic ☐ Tidal ☐ Lagoonal ☐ Beach

Marine: ☒ Shallow Water ☐ Deep Water ☐ Reef

Others: ☐ Evaporite ☐ Glacial

Sequestration Trapping Mechanisms (check all that apply):

☒ Dissolution and Diffusion ☐ Physical Containment ☐ Mineralization ☐ Residual Saturation

Other (Specify):

Target Reservoir Depth and Thickness:

Top Depth: Min: Max: Mean: 6705 ft

Bottom Depth: Min: Max: Mean: 9241 ft

Thickness: Min: Max: Mean: 2536 ft

Estimated Fracture Gradient: 0.8 psi/ft

Estimated Fracture Opening Pressure: 5200 psi

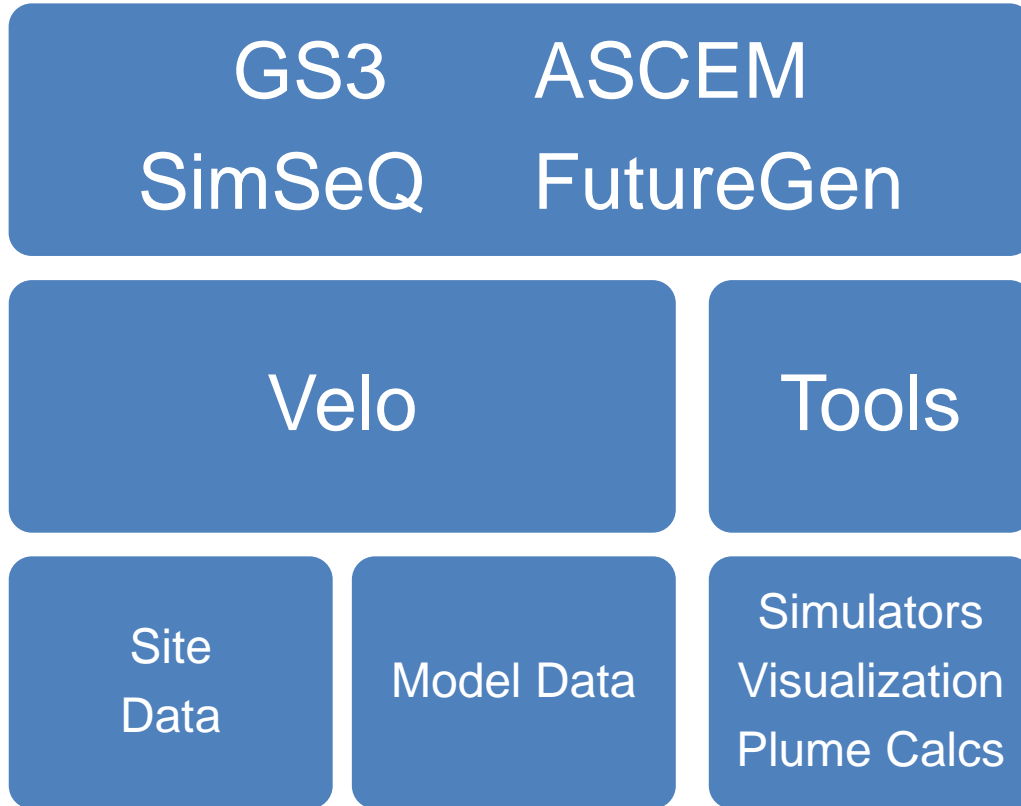
OK Cancel

What Happened?

- ▶ Iterative development process
 - Design, build and demo, repeat
- ▶ Interest from user community was strong
 - Power of mock-ups and prototypes
- ▶ New funding obtained
- ▶ Initial sites deployed
- ▶ And along the way ...



Velo - Flexible, Rigorous Scientific Knowledge Management

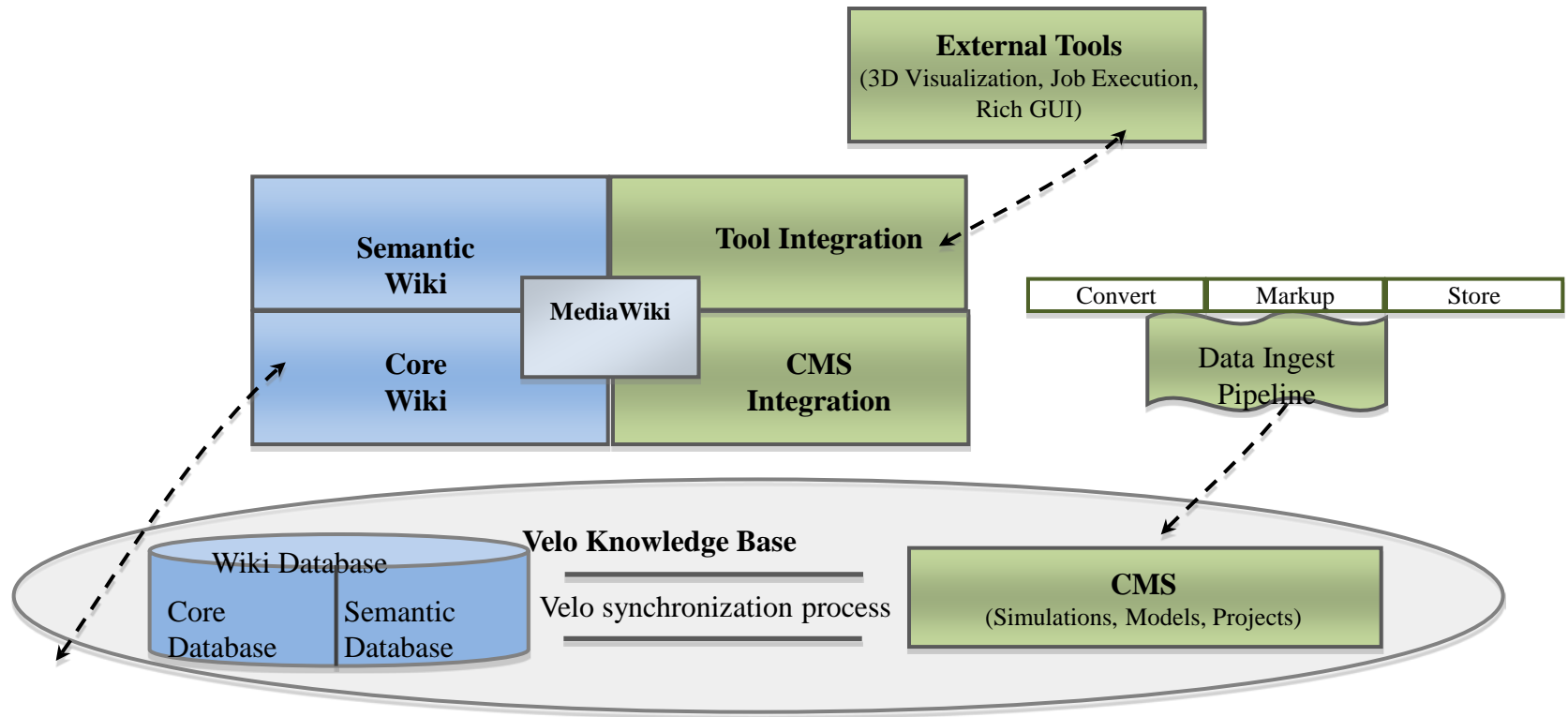


User customizable 'skins'
Web-based
Extensible

Raw data and metadata storage
Versioning
Provenance
Tool registry
Many deployment options

Extensible data types
Extensible tool repository
Programming interfaces

Velo Architecture



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Some reflections

- ▶ Science is a complex domain
 - Requirements, funding models
 - Diversity of software/data
 - Users who are pushing the boundaries
- ▶ Scientists don't (in general) understand complexity of software systems
 - Architectures, integration, testing
 - Different to implementing a set of equations
- ▶ Through deliberate, creative reuse and a strong focus on architecture, we've:
 - Built generically useful technologies at low cost)
 - They work ;)



Questions?

