



MAYO CLINIC

Center for Translational Science Activities

Collaborative Standards Within CTSA – Discovery Workflow



Brian Wilson
BioInformatic Core

Motivation

- Create a library of reusable software utilities which can be accessed & executed by a non programmer via a graphical or command line user interface.
- Support the execution of repetitive tasks which includes routine daily tasks and more complex processes which may be run less frequently (e.g. updating probe set annotation).
- Foster collaboration and inter exchange of analytics with other groups.
- The systems must be extensible including the application of custom scripts and programs
- Create a flexible system to meet the rapidly changing demands of the world of genomic research

What Is Workflow ?

Workflow at its simplest is the *movement* of information and/or tasks through a series of *synchronized*, time bounded, *reusable* components.

A workflow component can be many things:

- Database search
- Statistical function
- Bioinformatic algorithm
- Ordering a clinical test.....

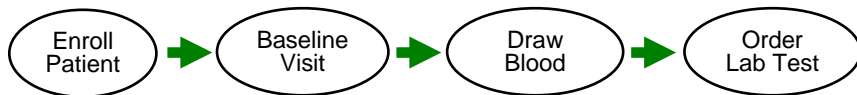
Types Of Workflow

Process Flows

Event driven - **Messages** flow between components - small data foot print

Examples:

- *Standard Operating Procedures*
- *Clinical Research Protocol Development*
- *Specimen Collection....*

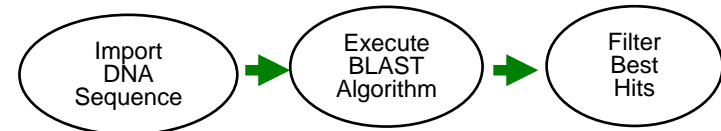


Analytical (Data) Flows

Large datasets – **Data** flows between components

Examples:

- *Normalizing Gene Expression Data*
- *Genomic Sequence Alignment*
- *Laboratory Test Quality Control Analysis*
- *Text Processing (UIMA)*



Often a process flow can execute a dataflow as a component in order to satisfy an event condition

Workflow Challenges

Building A Workflow

➤ Finding Components That Meet Your Needs

- Describe components by their function, input(s) and output(s)
- Map descriptions to controlled vocabularies & ontology's

➤ Linking Components

- Syntactic Interoperability



- Semantic Interoperability



➤ Do I need To Become a Programmer... ???

- Process parallelization – distributed computing
- Training

➤ Regulatory Compliance

- Workflows supporting clinical trials need to satisfy regulatory demands such as [CFR 21 Part 11](#) – difficult within a distributed environment

Workflow Challenges Cont....

Executing A Workflow

➤ Security

- *User has permission to execute workflow ?*
- *User permission to execute components within a workflow ?*
- *Access to data consumed or produced within a workflow*
- *Audit trail*

➤ Data Management

- *Accessing large datasets from multiple components*
- *Standardized – meta data described - generated data: reusable outside of the workflow environment*

➤ Orchestration

- *Load Balancing*
- *Job Management*

➤ Progress Tracking

- *Error handling – triaging*
- *Restarting long running work flows*
- *Providing real time status feed back to the user*

A Key Issue: Interoperability !

- **Standardizing how components are described** using ontology's and controlled vocabularies can help describe the function of each component in a workflow, what it produces and what it consumes.
 - *Enabling more effect searches*
 - *Facilitates 'intelligent' component linking*

- **Workflow distribution between institutions** is greatly enhanced through the use of a standard workflow design format.

- **Data Management:** Standards can help describe how data created by a workflow is stored and enable its reuse in secondary workflows.

- **Empowering researchers** by reducing the need for direct programmer involvement during workflow creation.

Workflow Standards & Formats

- The XML Process Definition Language (XPDL)

A format standardized by the Workflow Management Coalition (WfMC) to interchange Business Process definitions between different workflow products like modeling tools and workflow engines. XPDL defines a XML schema for specifying the declarative part of workflow. BPEL

- Business Process Execution Language (BPEL)

A business process modeling language that is executable. It is serialized in XML and aims to enable programming in the large. The concepts of programming in the large and programming in the small distinguish between two aspects of writing the type of long-running asynchronous processes that one typically sees in business processes. OASIS standard.

- Yet Another Workflow Language (YAWL)

Provides a very powerful, yet fundamentally simple language for process modelers to describe complex control flow relations between business processes. Consequently, YAWL enables businesses to own and manage very flexible and dynamic business processes

BioInformatic Core: A Brief History Of Analytical Workflows

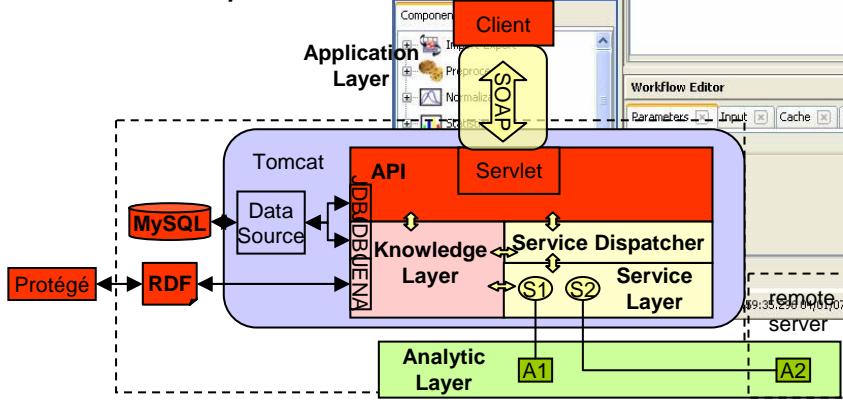
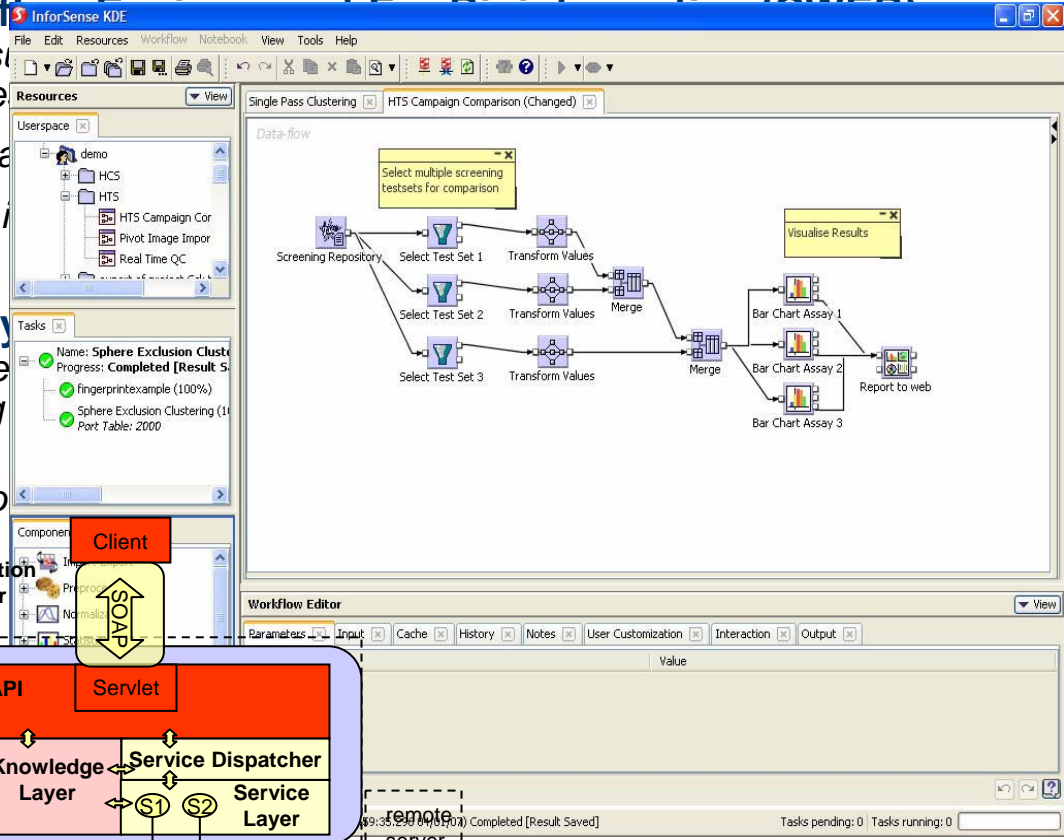
➤ **Perlscripts, Java, R-scripts, SAS.....**

➤ **Collaborative Workflow Editor (CWE)**
 A joint project with the... sponsored by the Minne...

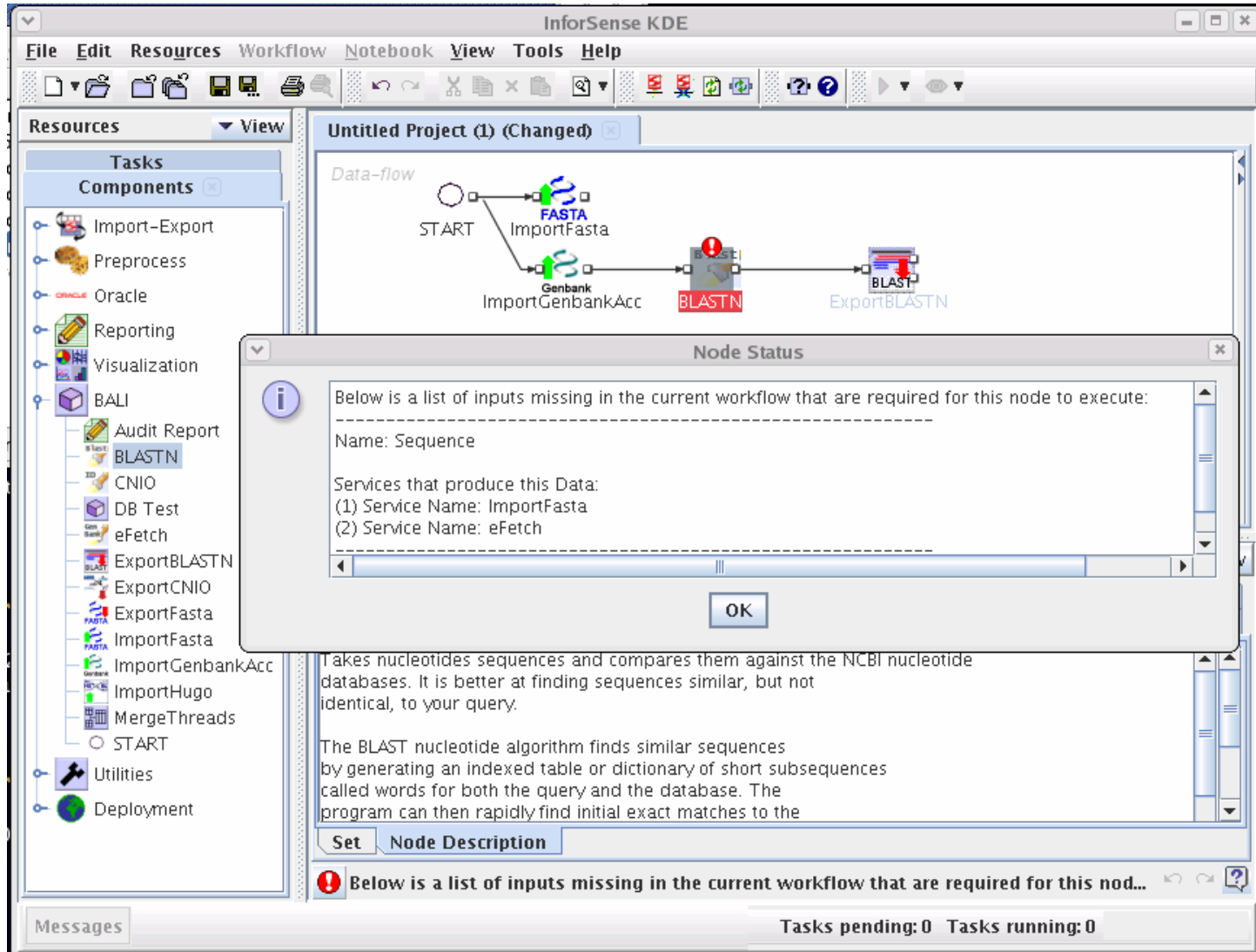
- Focused on the a...
- InforSense KDE i... environment

➤ **BioAnalytics Library**
 Built on the CWE exper... environment by defining... line user interface.

- Developed a Pro...



BALI Proof of Concept - InforSense



The screenshot displays the InforSense KDE interface. The main window shows a workflow diagram titled "Untitled Project (1) (Changed)" with a "Data-flow" view. The workflow consists of the following nodes: START, ImportFasta (FASTA), ImportGenbankAcc (Genbank), BLASTN (BLAST), and ExportBLASTN (BLAST). The BLASTN node is highlighted with a red error icon, indicating a missing input.

A "Node Status" dialog box is open, providing details for the BLASTN node:

Node Status

Below is a list of inputs missing in the current workflow that are required for this node to execute:

Name: Sequence

Services that produce this Data:

(1) Service Name: ImportFasta
 (2) Service Name: eFetch

OK

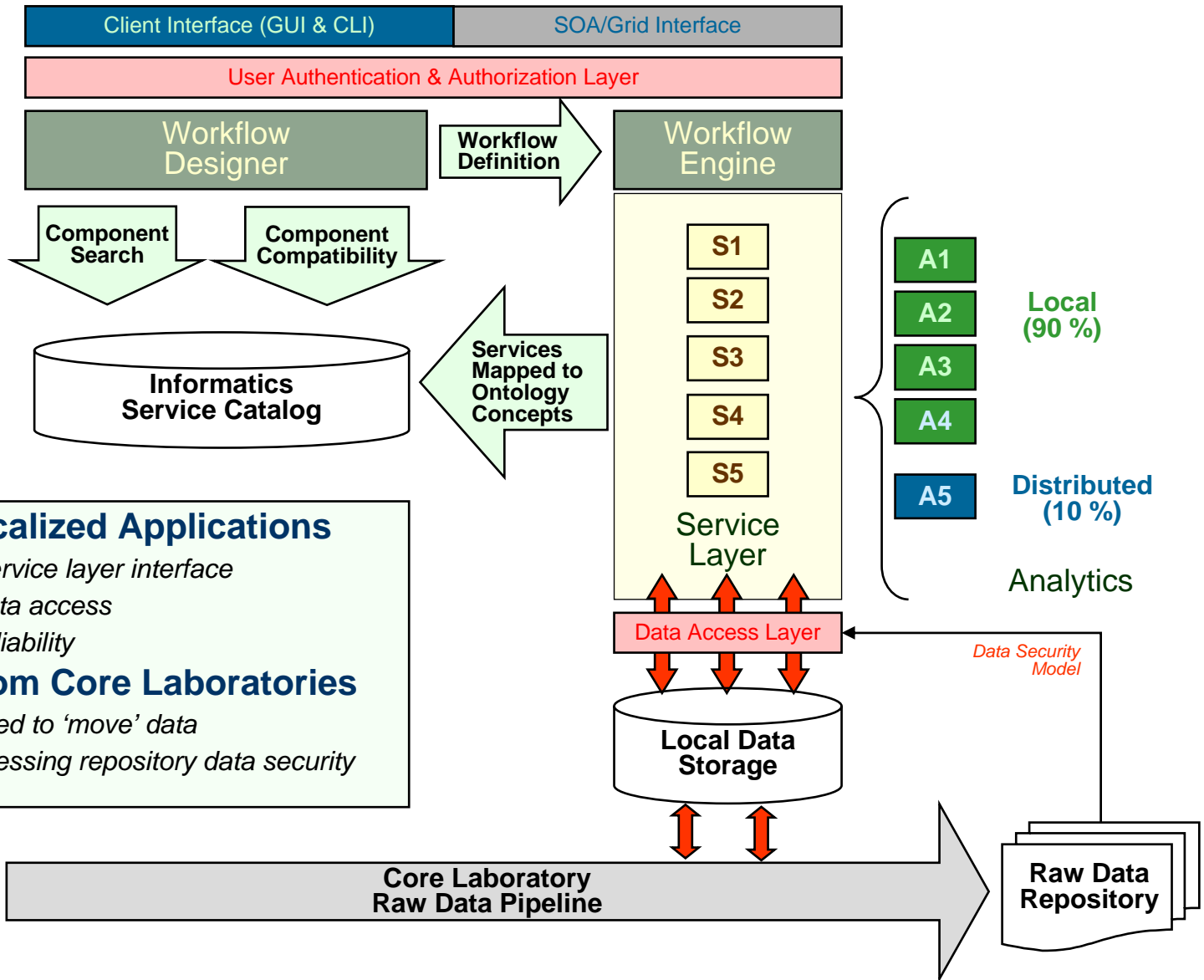
Below the dialog box, a "Node Description" tab is active, showing the following text:

Takes nucleotides sequences and compares them against the NCBI nucleotide databases. It is better at finding sequences similar, but not identical, to your query.

The BLAST nucleotide algorithm finds similar sequences by generating an indexed table or dictionary of short subsequences called words for both the query and the database. The program can then rapidly find initial exact matches to the

At the bottom of the interface, a status bar shows "Tasks pending: 0" and "Tasks running: 0".

Future Work

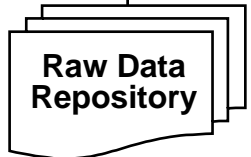


Focus On Localized Applications

- Simplifies service layer interface
- Improves data access
- Improved reliability

Data Feed From Core Laboratories

- Reduces need to 'move' data
- Exploits existing repository data security model



Acknowledgements

BMI – BioInformatic Core

Mat Wiepert, David Mead, Corey Carlson, Asif Hossain, David Rider, Patrick Duffy, Dana Carrington, & J.P Kocher.

Proteomics

Chris Mason & Rudi Chiarito

Biostatistics

Bruce Morlan, Jeanette Eckel-Passow, Diane Grill, Keith Anderson, Shannon McDonald & Beth Atkinson

Genotyping Core Facility

Julie Cunningham, Scott hammer & Sarah Anderson

University Of Minnesota

Ben Lynch & Birali Runesha

Inforsense

Raveen Sharma