Epidemiology Experiment and Simulation Management through Schema-Based Digital Libraries

Jonathan Leidig
Edward A. Fox
Madhav Marathe
Henning Mortveit

ECDL: 2nd DL.org Workshop, Sept. 9-10, 2010

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Network Dynamics and Simulation Science Laboratory

Digital Library Research Laboratory (DLRL) @ Virginia Tech

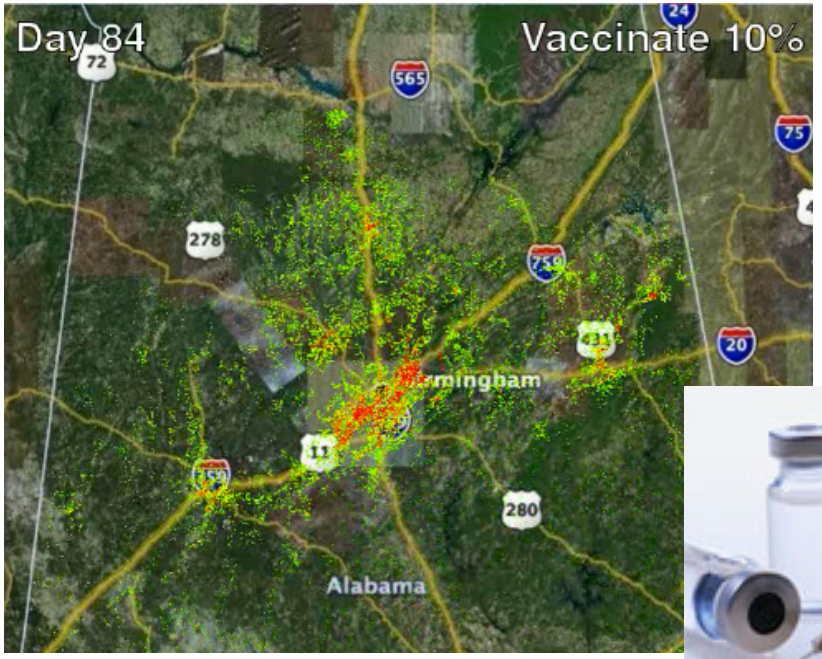
Virginia Tech

Overview

- Community: Computational Epidemiology
- Simulation Digital Library
- Content interoperability
- User interoperability
- System interoperability
- Future work

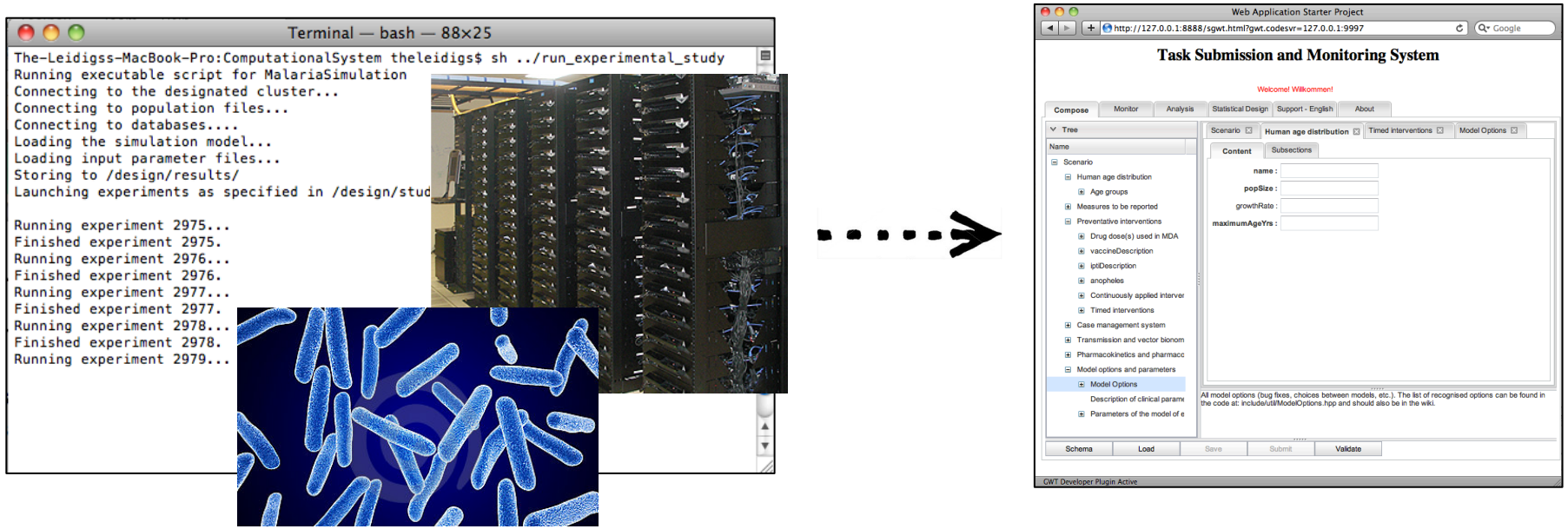
Domain: Computational Epidemiology

- Model diseases and populations
- Simulation of disease spread
- Test public policies, interventions, adaptive behavior, & economics



User Community Goals

- Expose models to diverse user groups
  - Abstract away need for computational expertise
- Automate generation of data management and user interfaces
  - Based on a schema for the model



Simulation Digital Library (SimDL) Goals

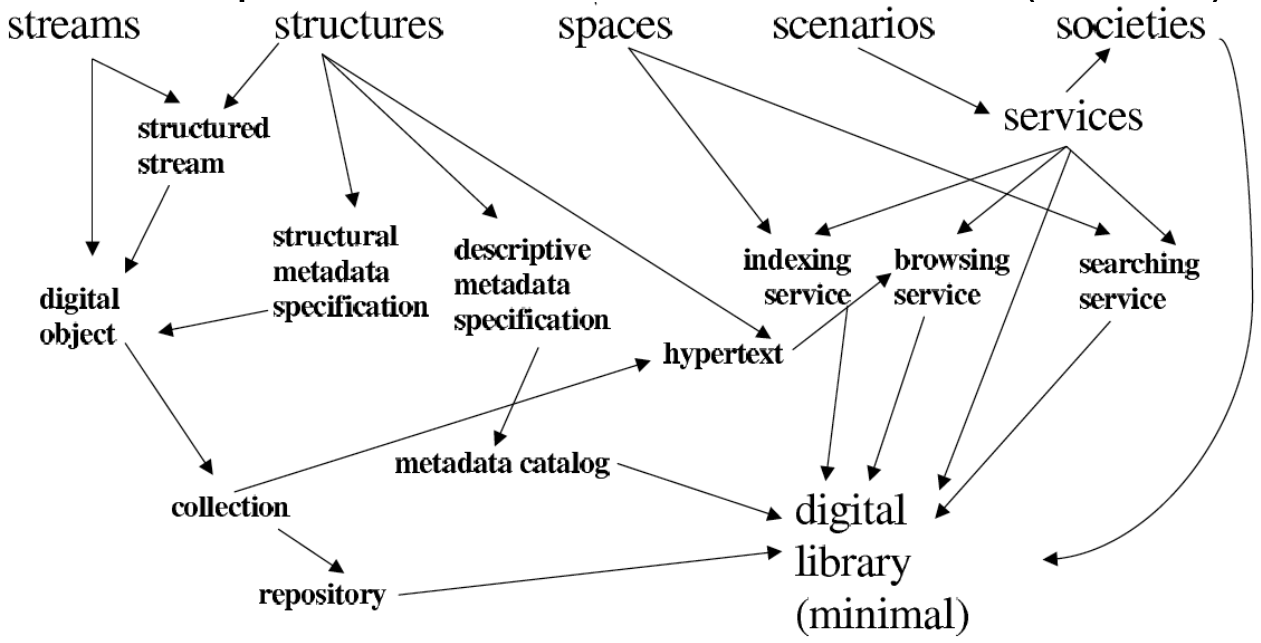
- Manage provenance of scientific content
- Formalize the components of a digital library
- Provide system interoperability:
  - user interface, data management, resources, and software
- Support user interoperability:
  - community collaboration and sharing
  - datasets, models, annotations, message boards

Challenges

- Lack of ontologies in simulation domains
- Heterogeneity between model and term definitions by modelers
- Disincentive to freely contribute models & software implementations
- Computational resources
- SimDL intellectual property of funding institutions

5S: Societies, Scenarios, Spaces, Structures, Streams

- Formalized semi-automated generation of specialized DLs
- Formal descriptions of Societies (users)
- Formal descriptions of Scenarios, activities, tasks (functionality)
- Formal descriptions of Structures and Streams (content)

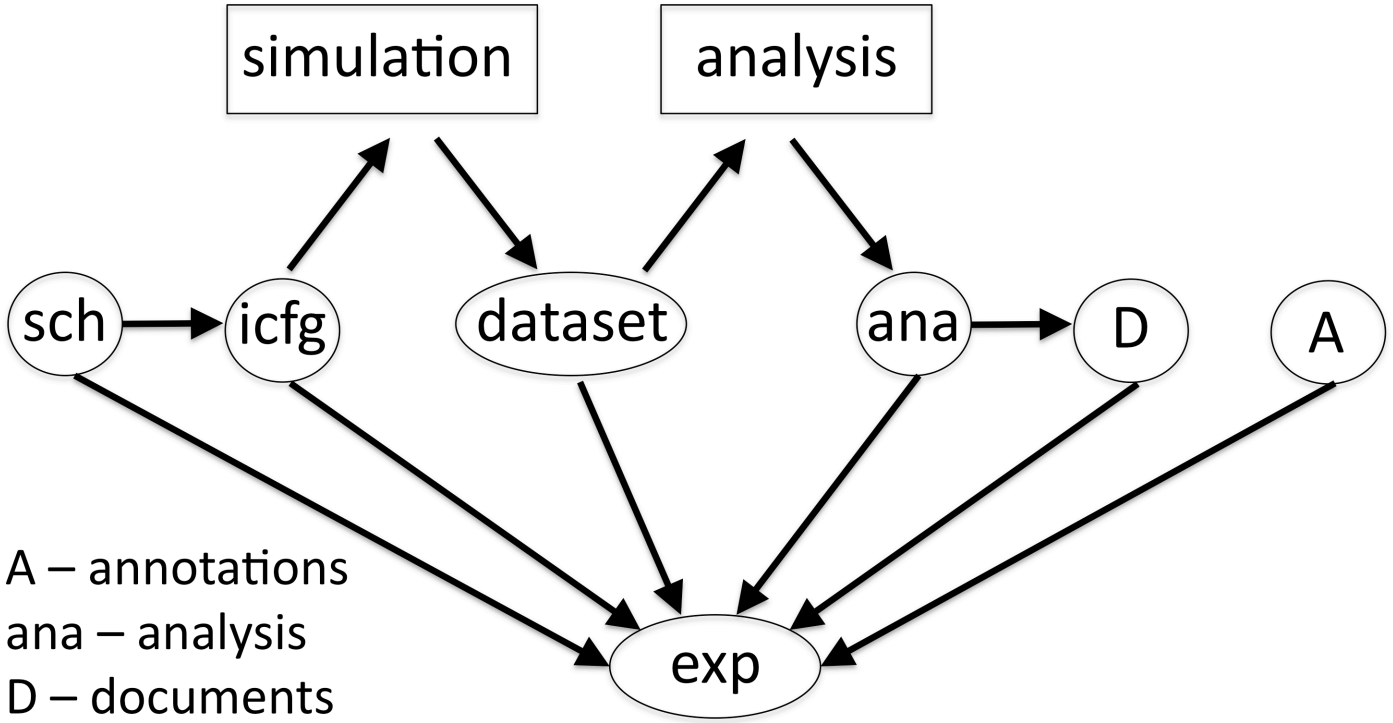


Simulation Datasets

Experimentation process:

- XSD **model schema**
- Study **input configuration**
- Input files
- **Result** Datasets
- Result summaries
- **Analysis** & plots
- Annotations
- Publications
- Wiki and collaboration posts
- Contributed datasets
- Notes repository

Simulation Content Types



A – annotations
ana – analysis
D – documents
exp – experiment
icfg – input configuration
sch – schema

Simulation Content Definitions

- Schema
  - *sch = (h, sm, S)*
- Input configuration
  - *icfg = (h, sm, ELE, ATT)*
- Sub-configuration
  - *sub-icfg = (h, sm, icfg, ELE, ATT)*
- Structured dataset
  - *dataset = (h, sm, S)*
- Analysis
  - *ana = (h, SCDO = DO ∪ SM, S, icfg)*
- Experiment
  - *exp = (h, SM, sch, icfg, dataset, ana, D, A)*

h - *h* exists in H, a set of unique handles (labels)

sm - *sm* is a stream

S - *S* is a structure that composes a digital object into a specific format

ELE - *ELE* is a set of XSD elements

ATT - *ATT* is a set of XSD attribute values for an element ELE_i

DO - *DO* is a set of digital objects

SM - *SM* is a set of streams

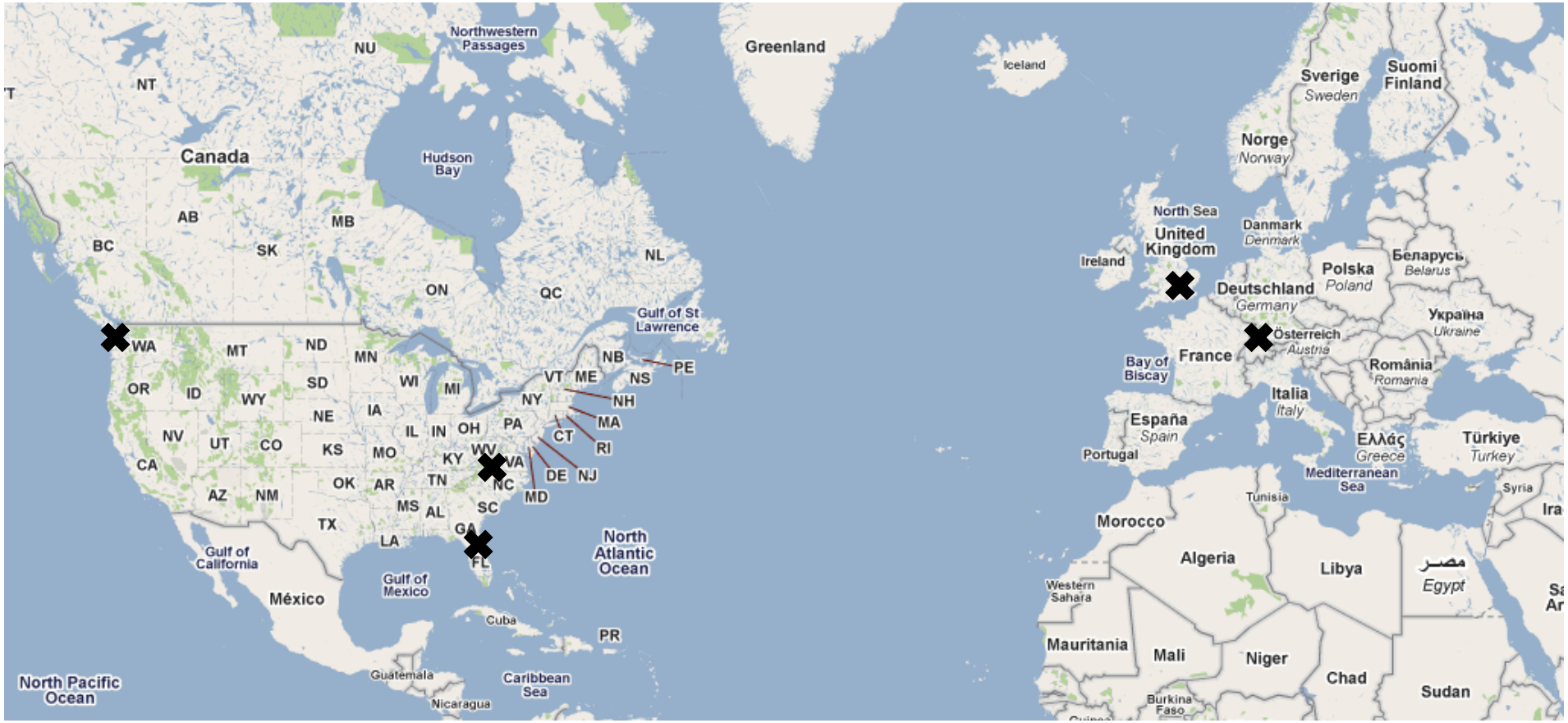
D - a set of additional documents

A - a set of annotations

Content Interoperability

Definitions describe:	Definitions support:
Content structure	semantic mapping between collections structured by models
Stage of content production in the simulation process	collaboration between users with different roles and stages
Provenance sequences	provenance investigations
Input & output structure for processes	automated chaining of processes

Malaria Institutional Collaboration Map



SimDL User Groups

User Roles	Services
Tool builder	Develop simulation model Develop simulation infrastructure Define and create DL
System & DL administrator	Manage content, users, processes
Related systems	Support simulations, analysis, & data mining
Study designer / Experimenter	Produce experiments and studies
Analyst	Produce analysis of studies
Annotator	Review and mark studies
Explorer	Review studies and determine public policies

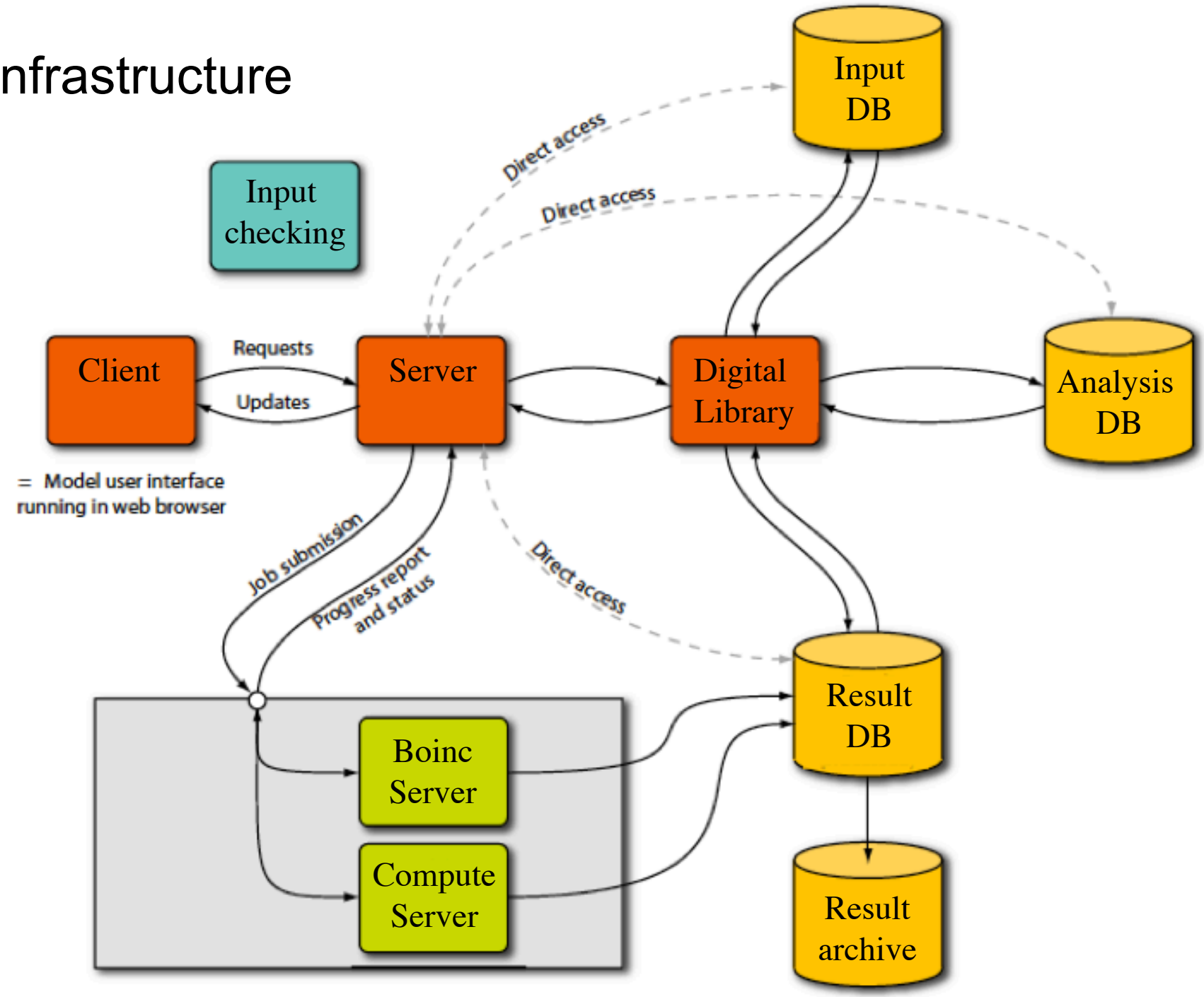
SimDL User Interoperability

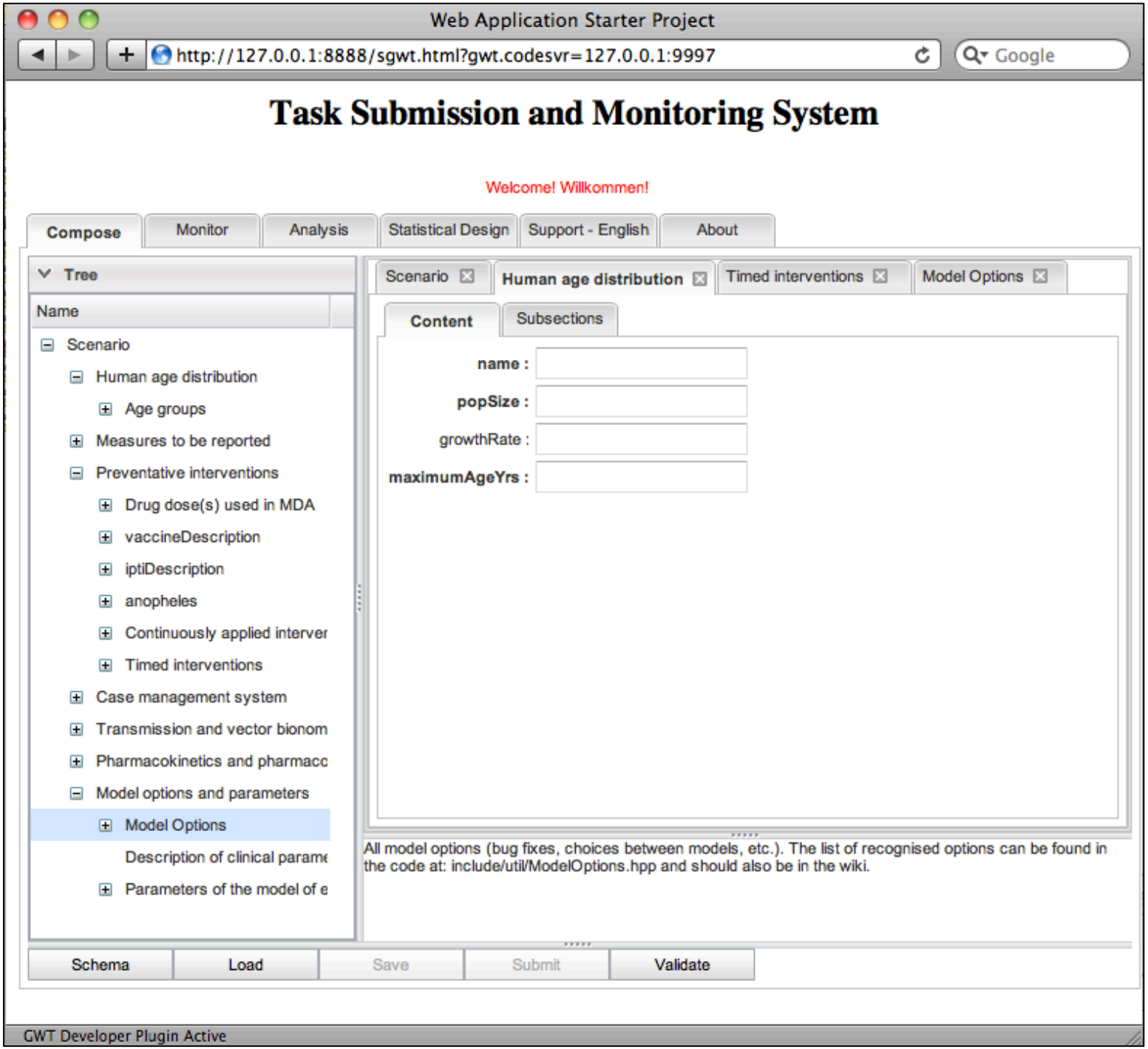
- Collaboration and cooperation
- M&S studies require detailed contributions between user groups
- DLs manage contributions and workflow processes
- Experts in study areas contribute high-quality content
  - Disease modeler – model and schema
  - Software engineer – model implementation
  - Local population expert provides population – regional dataset
  - Vector expert – mosquito and virus descriptions
  - Public health official – study design
  - Analyst – review and analysis of study

SimDL User Abstractions

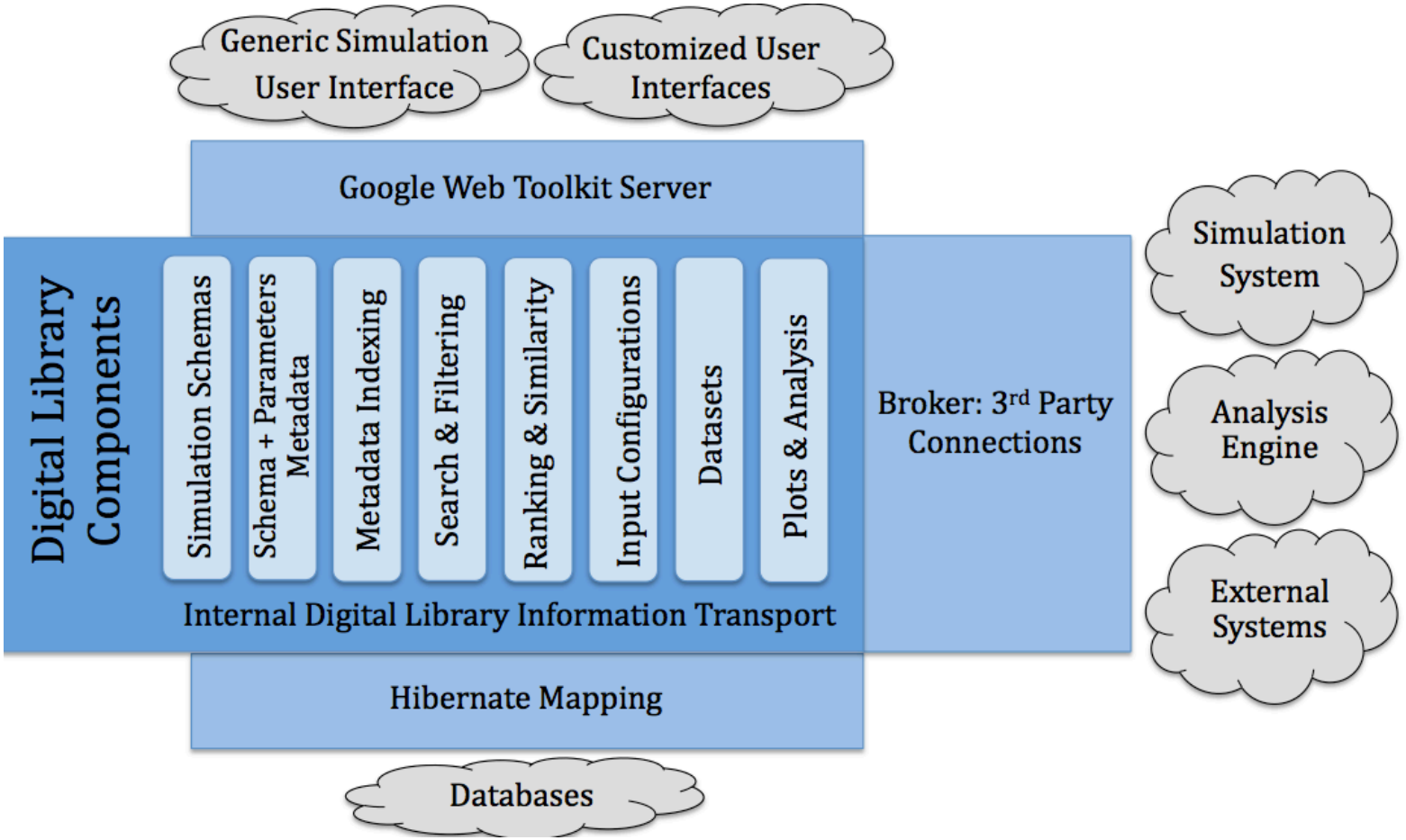
- Automated Coordination
  - Single interface to the system for all user groups
  - Experts work within a defined workflow to conduct studies
  - User tasks build upon previous stages of content
  - Users query content from multiple workflow stages
  - Studies build upon previous work

Infrastructure




Web Application Starter Project
http://127.0.0.1:8888/sgwt.html?gwt.codesvr=127.0.0.1:9997
Google
Task Submission and Monitoring System
Welcome! Willkommen!
Compose
Monitor
Analysis
Statistical Design
Support - English
About
Tree
Name
Scenario
Human age distribution
Age groups
Measures to be reported
Preventative interventions
Drug dose(s) used in MDA
vaccineDescription
iptiDescription
anopheles
Continuously applied interver
Timed interventions
Case management system
Transmission and vector bionom
Pharmacokinetics and pharmacc
Model options and parameters
Model Options
Description of clinical parame
Parameters of the model of e
Scenario
Human age distribution
Timed interventions
Model Options
Content
Subsections
name :
popSize :
growthRate :
maximumAgeYrs :
All model options (bug fixes, choices between models, etc.). The list of recognised options can be found in the code at: include/util/ModelOptions.hpp and should also be in the wiki.
Schema
Load
Save
Submit
Validate
GWT Developer Plugin Active

System Functionality



System Functionality Interoperability

- Model & software independent
- SimDL provides seamless connectivity between:
  - a generic user interface
  - computational backend
  - model simulation software
  - analysis software
  - input, result, analysis, annotation, and publication repositories
  - external system API
- Model schema (XSD file) provides:
  - UI parameters layout
  - domain specific database schema
  - simulation input requirements
  - contextual search

Infrastructure Status

User Roles	Services	Implementation
Tool builder	Define and create DL Develop simulation model Develop infrastructure	☑ One instance of SimDL ☑ Manages multiple models ■ Backend infrastructure
System & DL administrator	Manage content, users, processes	☑ Input configurations ☐ Backend content ☐ User management
Related systems	Simulations, analysis, & data mining	☑ Construct input files ☐ 3rd party connections
Study designer / experimenter	Design and run studies	☑ Construct experiments ☐ Factorial study designs
Analyst	Produce analysis of studies	☐ Automated analysis/model
Annotators	Review and mark studies	☐ Mark streams of content
Explorer	Review studies and determine public policies	☐ Provenance ☐ Query and access content

Summary

- Goal: provide a deployable DL to coordinate simulation efforts
- SimDL provides:
  - generic UI, data management, and connections
  - support for models with an XSD schema description
  - scientific process workflow management
  - interoperability between infrastructure components
  - content provenance
  - user collaboration on studies
  - definitions of content, users, and services

Future Work

- Plugins:
  - perform semantic queries across SimDL instances and models
  - communication, annotation, and recommendation
  - dataset and model visualization
  - user personalization
  - semantic federation of models with an ontology
- Planned studies:
  - User studies per user role
  - Multiple computational epidemiology lab reviews

Acknowledgments

This work has been partially supported by:

- Swiss Tropic and Public Health Institute: OpenMalaria subcontract
- National Institutes of Health: MIDAS project 2U01GM070694-7
- Department of Defense: DTRA CNIMS Grant HDTRA1-07-C-0113
- Department of Defense: DTRA R&D Grant HDTRA1-0901-0017
- National Science Foundation: HSD Grant SES-0729441
- National Science Foundation: PetaApps Grant OCI-0904844

Special thanks to:

- Drs. Fox, Marathe, and Mortviet
- NDSSL and DLRL members