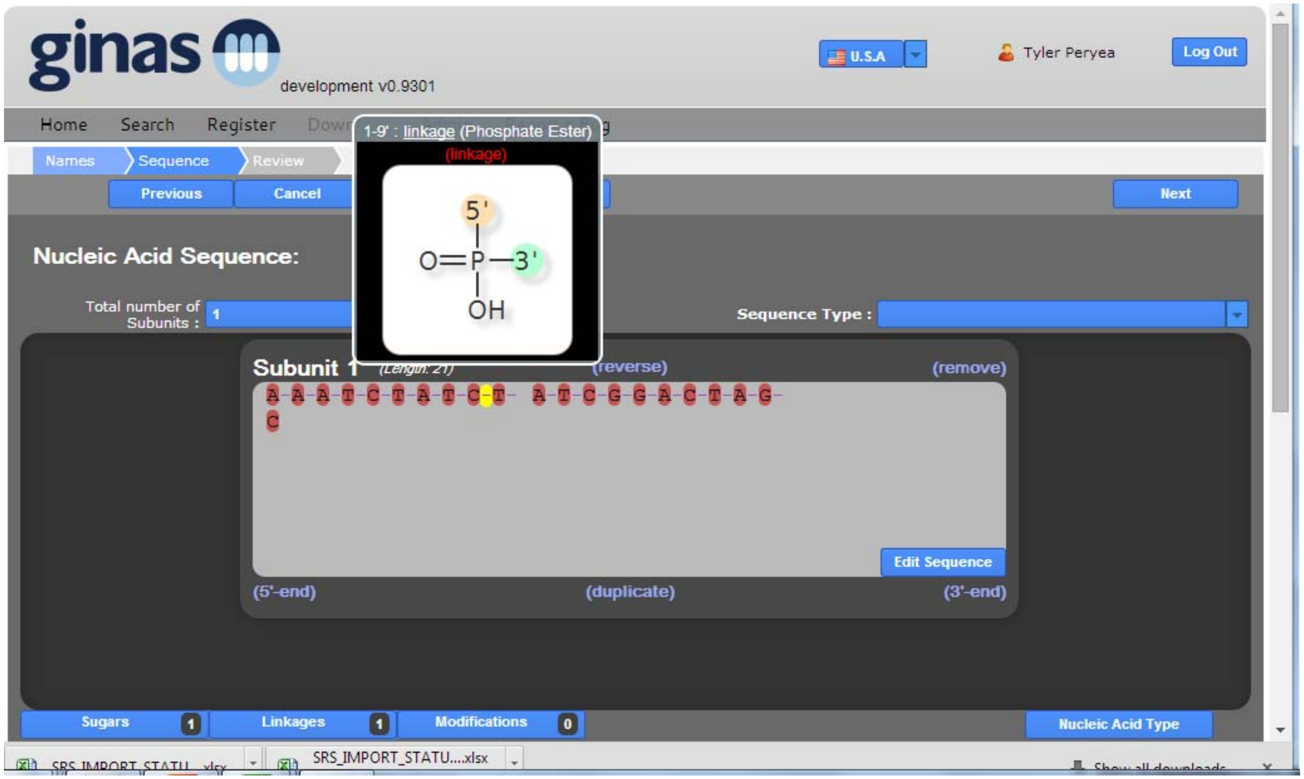
ginas Software Status



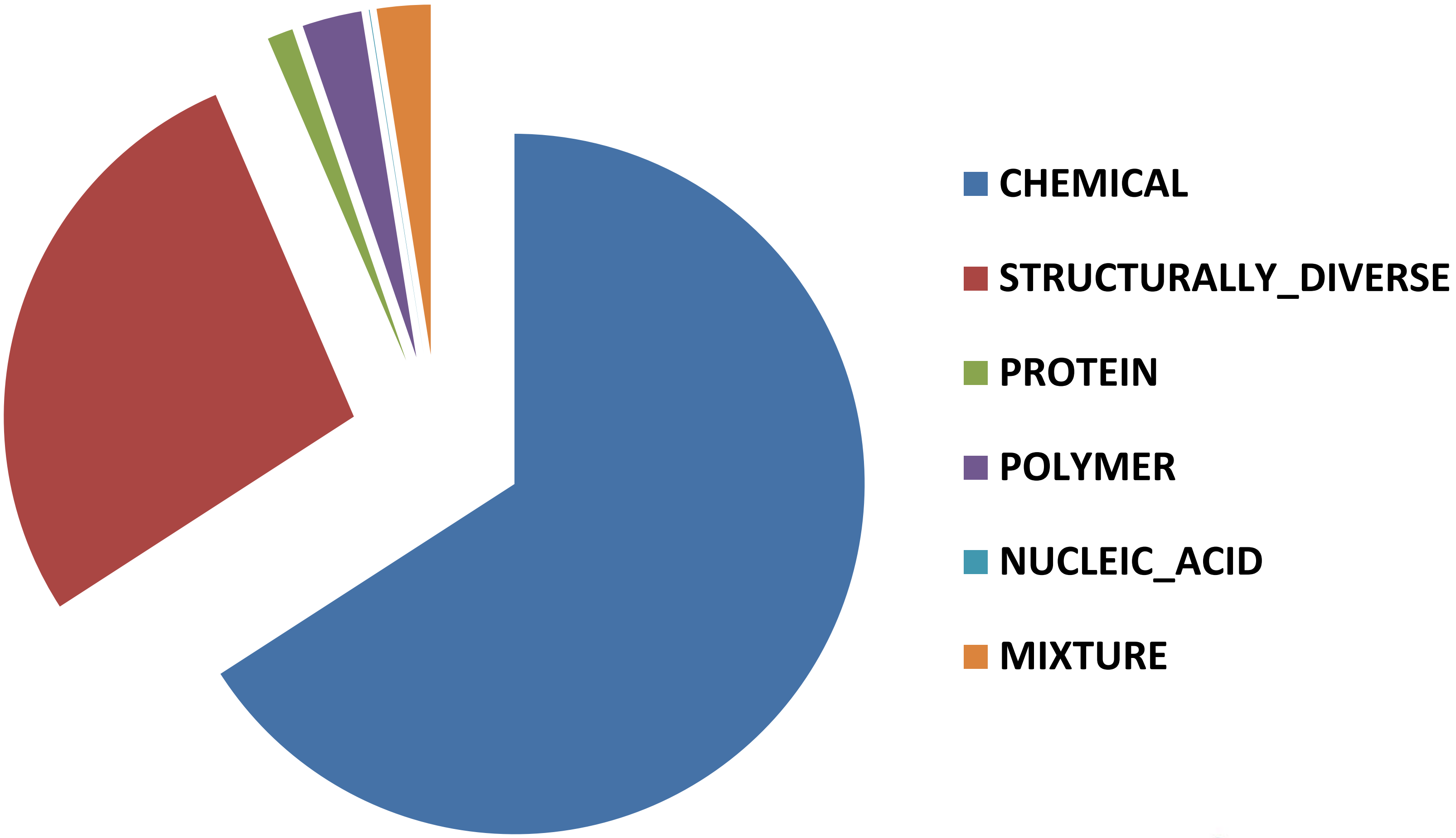
Dac-Trung Nguyen, Tyler Peryea

NCATS

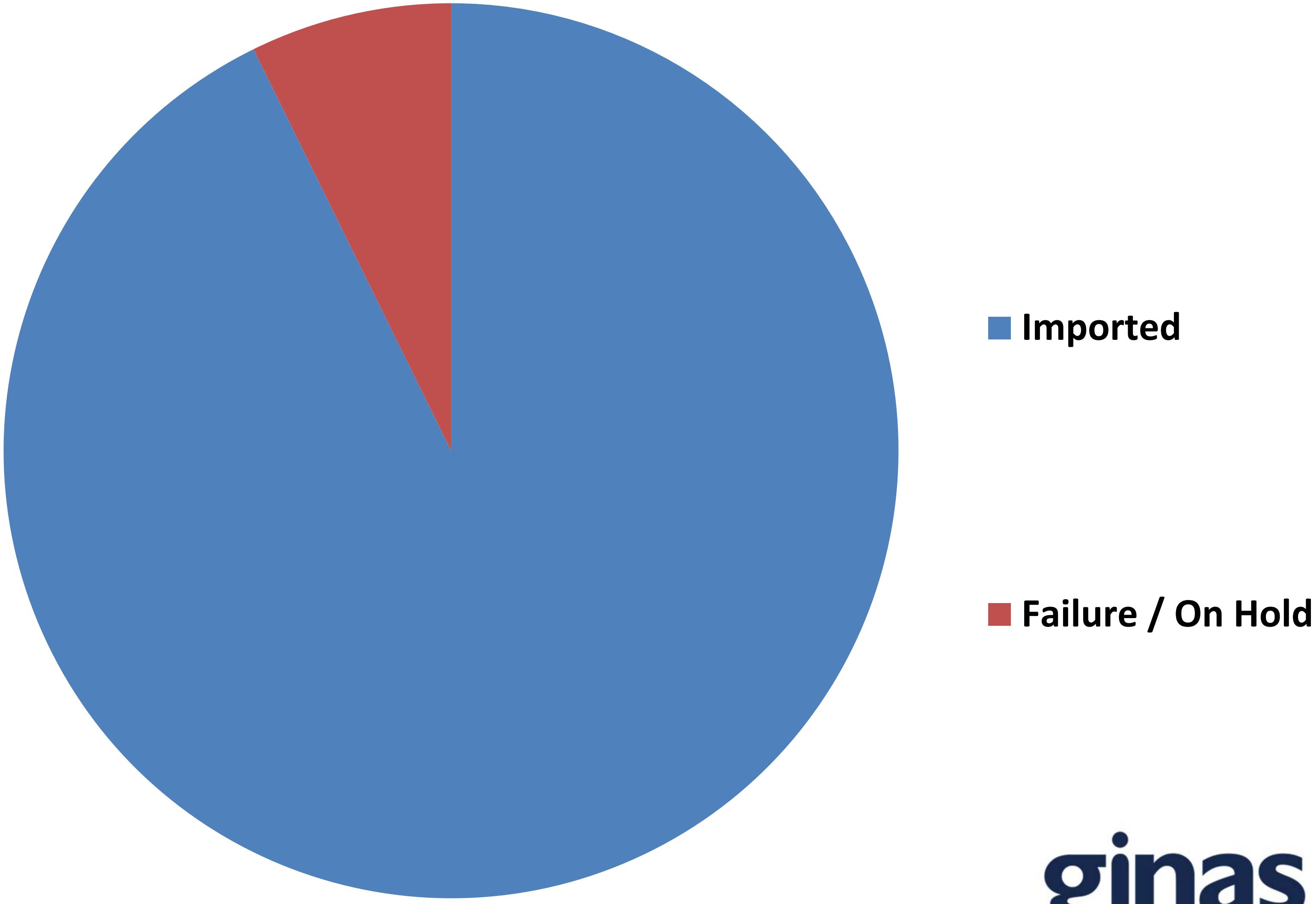
Software Status

- Registration Coverage
  - Web interface and data import path
- Global-ness
- Distributable-ness
- Usability
- Security and User management
- Data Quality and Data Management
- Data Exchange
- “Open Source” status
- Links

SRS Distribution of Substance classes



SRS Transformation Count



ginas

Software Status : Registration

	Conceptual Case Study	Implemented	Live Case Study	Large Scale Import Study
Protein	YES	YES	YES	In process
Chemical	YES	YES	YES	In process
Nucleic Acid	YES	YES	YES	In process
Mixture	YES	YES	YES	In process
Structurally Diverse	YES	YES	YES	In process
Polymer	YES	YES	YES	In process
G1SS	YES	YES	YES	NO
G2SS	NO	NO	NO	NO
G3SS	NO	NO	NO	NO
G4SS	NO	NO	NO	NO

ginas

SRS Import Tests

- Developing adapter from SRS format to ginas format, which is run periodically
- Some things fail due to the adapter, some things fail due to the data

June 6th Status

Status	Count	Meaning
UNEVALUATED	326	Not yet attempted or fundamental I/O problem
LOADED	317	Basic table info assembled from SRS tables, but no XML/structural parsing possible
PARSED	807	Description parsed correctly, but no adapter yet, or in unexpected/invalid state
ADAPTED	3436	Light-version of GINAS adaptation / validation
PROCESSED	1766	Heavy-version of GINAS validation / formatting
SUBMITTED	56150	Successfully entered into a GINAS instance
TOTAL	**62802**	89%

CLASS	Total	Submitted	Preliminary Percentage
CHEMICAL	40937	37487	92%
STRUCTURALLY_DIVERSE	17189	17051	99%
PROTEIN	744	508	68%
POLYMER	1708	1469	86%
NUCLEIC_ACID	22	0	0%
MIXTURE	1542	1103	72%
Total	62142	57618	93%

Software Status : **Global**-ness

- Full Unicode support for entry/searching across many different languages
- Translations for *most* INN names now present for 6 languages:
  - Russian
  - Spanish
  - English
  - Chinese
  - Arabic
  - French

Software Status : **Global**-ness

- To do:
  - Regional translations for software text
  - Regional translations for controlled vocabulary
  - Regional preferences for naming display

Software Status : **Distributable**-ness

- Embedded instance now compiled:
  - Self-contained H2 Database
  - Self-contained Java Web Servlet
  - Bootable image for CD/USB stick
    - Live-boot GNU/Linux (slax) with preset configuration
    - Portable virtual machine
  - Pre-compiled with preliminary sample data

Software Status : **Distributable**-ness

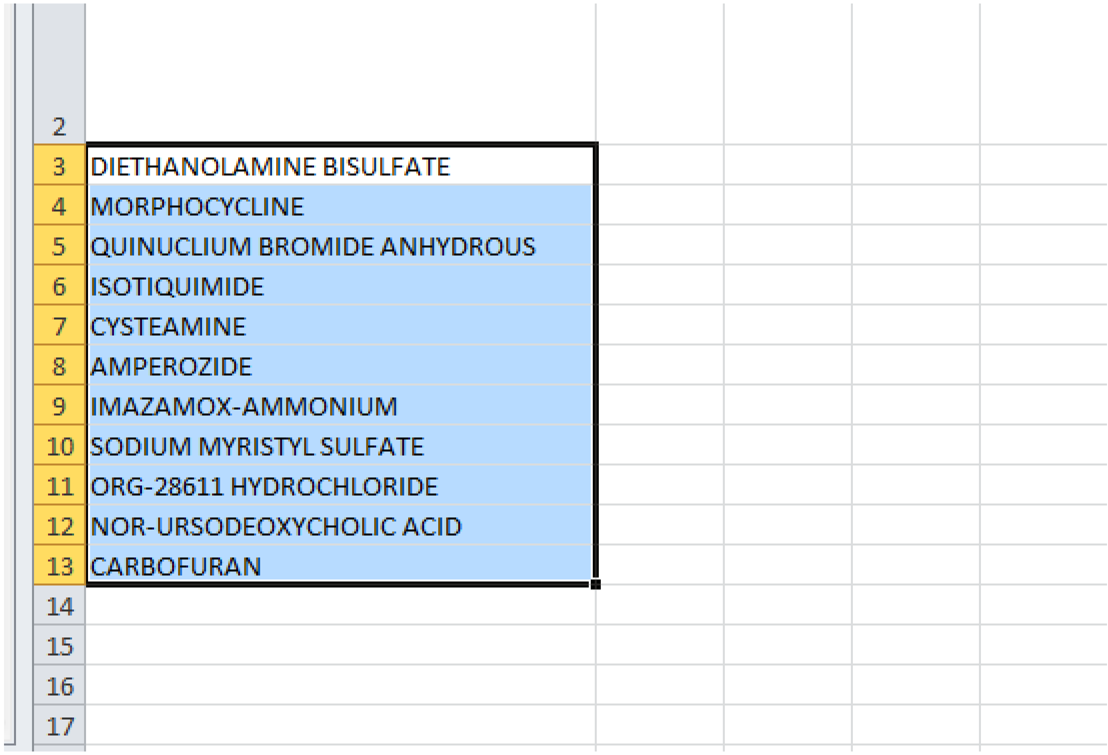
- To do:
  - Detailed documentation of specific setup
  - Optimize initialization
  - Optimize memory footprint
  - Procedure for updates
    - For data
    - For software

Software Status : **Use-ability**

- Many convenience tools for searching and registration:
  - Copy/paste chemical structure browser plugin
  - Image-to-structure browser plugin
  - “Draw-ahead” substructure searching
  - Real-time validation and feedback in many areas
  - REST API excel plugin for quickly resolving names/structures to ginas ID
  - Sequence homology searching (proteins and nucleic acids)

Software Status : **Use-ability**

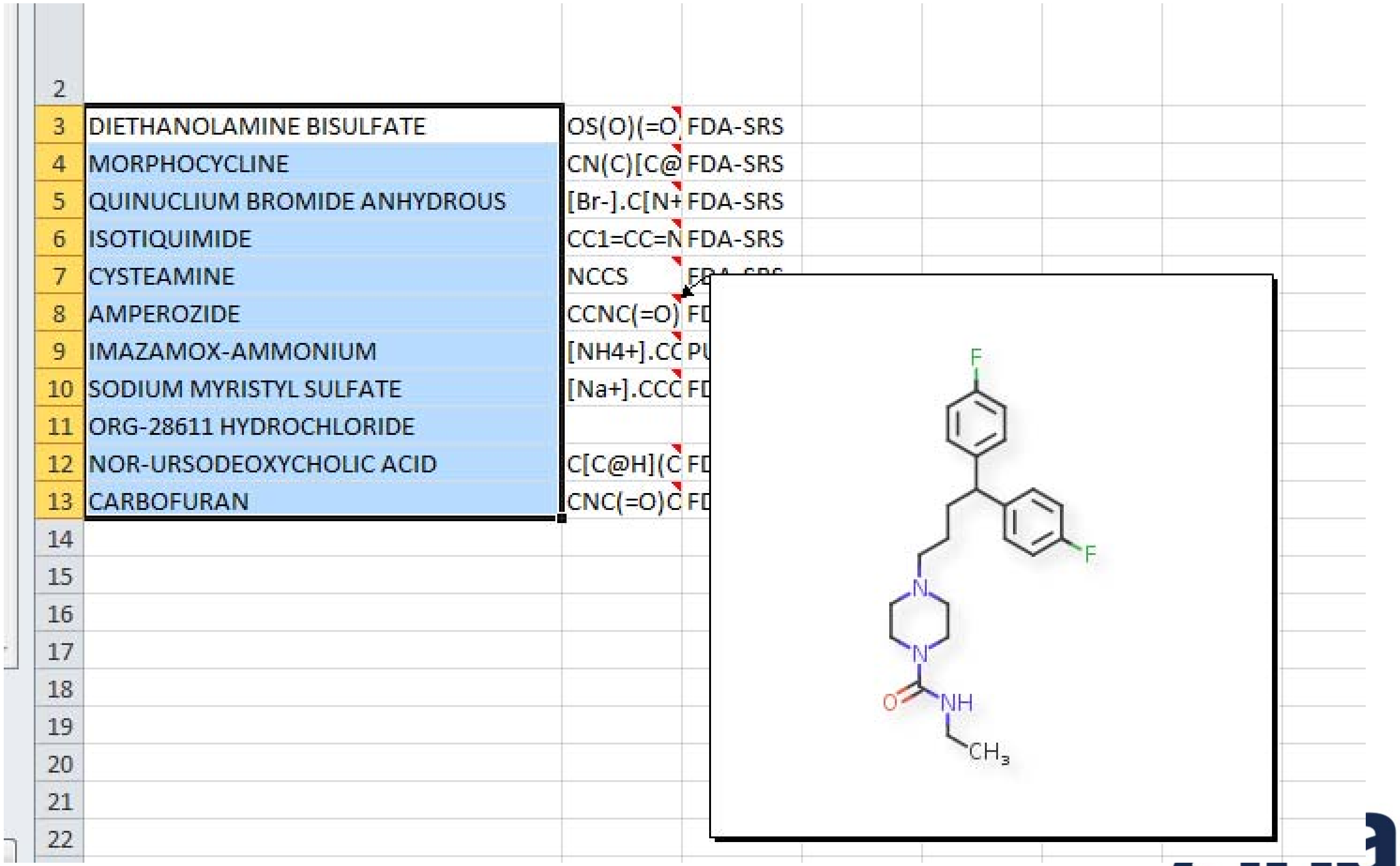
- Excel example



2	
3	DIETHANOLAMINE BISULFATE
4	MORPHOCYCLINE
5	QUINUCLIUM BROMIDE ANHYDROUS
6	ISOTIQUIMIDE
7	CYSTEAMINE
8	AMPEROZIDE
9	IMAZAMOX-AMMONIUM
10	SODIUM MYRISTYL SULFATE
11	ORG-28611 HYDROCHLORIDE
12	NOR-URSODEOXYCHOLIC ACID
13	CARBOFURAN
14	
15	
16	
17	

Software Status : **Use-ability**

- Excel example



2			
3	DIETHANOLAMINE BISULFATE	OS(O)(=O	FDA-SRS
4	MORPHOCYCLINE	CN(C)[C@	FDA-SRS
5	QUINUCLIUM BROMIDE ANHYDROUS	[Br-].C[N+	FDA-SRS
6	ISOTIQUIMIDE	CC1=CC=N	FDA-SRS
7	CYSTEAMINE	NCCS	FDA-SRS
8	AMPEROZIDE	CCNC(=O)	FD
9	IMAZAMOX-AMMONIUM	[NH4+].CC	PU
10	SODIUM MYRISTYL SULFATE	[Na+].CCC	FD
11	ORG-28611 HYDROCHLORIDE		
12	NOR-URSODEOXYCHOLIC ACID	C[C@H](C	FD
13	CARBOFURAN	CNC(=O)C	FD
14			
15			
16			
17			
18			
19			
20			
21			
22			

Software Status : **Use-ability**

- To Do:
  - Improve registration step-through wizards
    - (particularly in polymers, structurally diverse)
  - Improve record display for browse-ability rather than strict data elements
  - Improve searching/browsing and filtering capabilities
  - Implement export procedures to commonly used formats (sdf, excel, etc)

Software Status :
Security and User management

- Users allowed with various roles
- Roles control what that user is allowed to do
  - Register
  - Approve
  - View
  - Update
- Done now as static users, inherent to embedded system

Software Status :
Security and User management

- To Do:
  - All public releases have only public data, so deeper security model has been triaged
  - Authentication and private/public key encryption for all information sent
  - Database-level security implementation for embedded and production system

Software Status :
Data Quality and Data Management

- REST API in place to query and return full substance object (in JSON format)
- Fairly static format for objects
- Timestamps, owners, and references for every piece of information
- Very simple duplication detection for chemicals

Software Status :
Data Quality and Data Management

- To Do:
  - Allow for easy backend SQL querying inherent to model
  - Model for each substance class must be more solidified in a few areas (cardinality issues, etc)
  - Controlled Vocabularies preliminary, and must be re-evaluated, and pointed to external authorities where available (e.g. Kew Gardens)
  - Proper “fuzzy” duplication detection for all substance classes
  - REST API exposure of change log and versioning

Software Status : **Data Exchange**

- Object structure can be readily exported / imported into different ginas instances
- To Do:
  - Common exchange mechanism (semi-automatic imports into other systems)
  - Merging / conflict resolution reporting on exchange

ginas

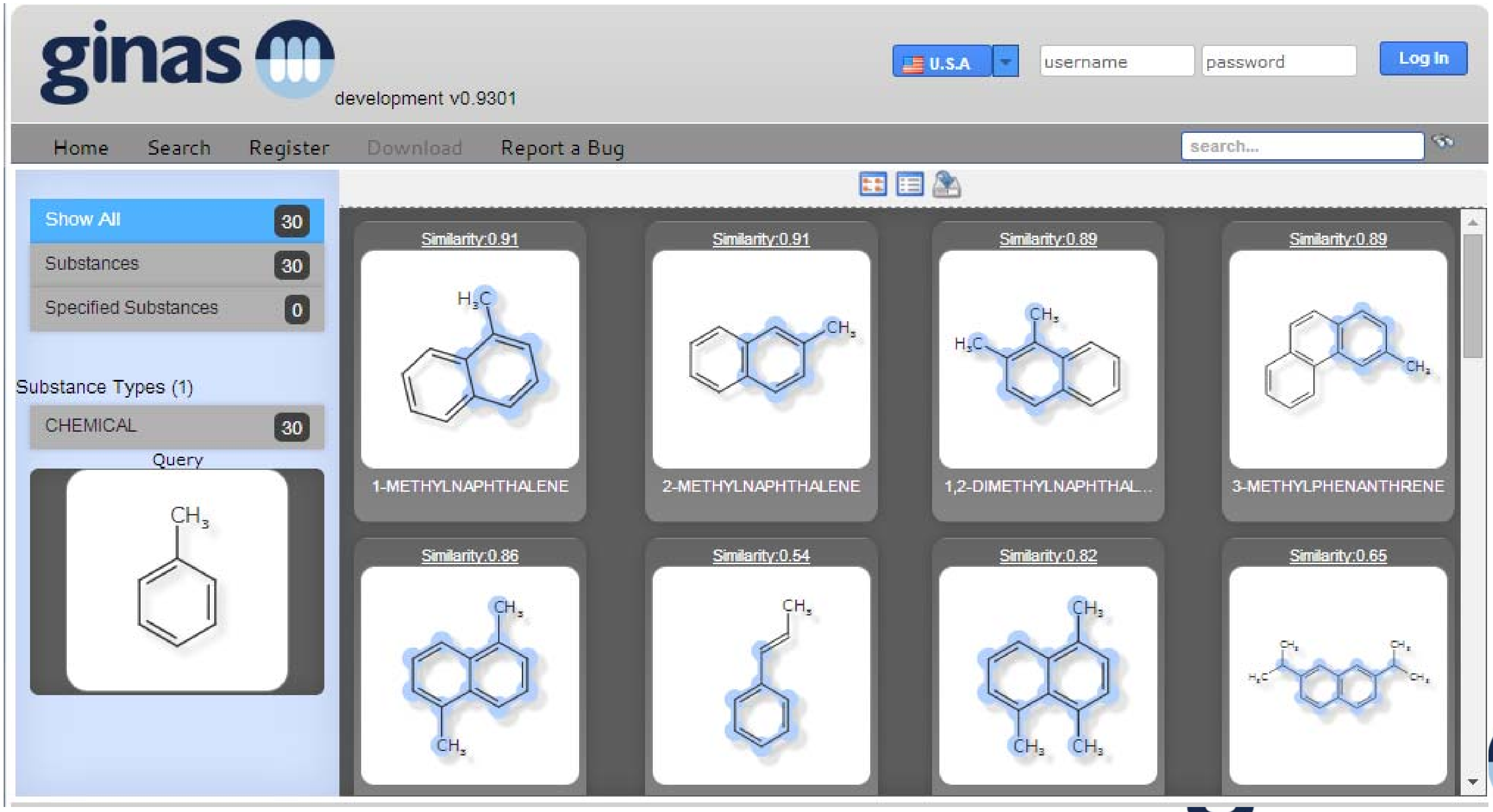
Software Status : **"Open source" Status**

- Mostly open source software, with a few licensed, distributable commercial packages
- Code available on private NCATS GitHub account
  - Limited number of seats
- NCATS-specific git repository will be available soon for all who request access
- REST API documentation available on github wiki
- To Do:
  - Transition to public completely public git repository
  - Publish API specifications for other developers

ginas

Links:

- http://ginas.hc.ircan-rican.org/ginas/
- http://tripod.nih.gov/ginas/


ginas
development v0.9301
U.S.A
username
password
Log In
Home
Search
Register
Download
Report a Bug
search...
Show All 30
Substances 30
Specified Substances 0
Substance Types (1)
CHEMICAL 30
Query
CH_3
Similarity:0.91
1-METHYLNAPHTHALENE
Similarity:0.91
2-METHYLNAPHTHALENE
Similarity:0.89
1,2-DIMETHYLNAPHTHAL...
Similarity:0.89
3-METHYLPHENANTHRENE
Similarity:0.86
Similarity:0.54
Similarity:0.82
Similarity:0.65