G-SRS software for substance registration

TYLER PERYEA

NATIONAL CENTER FOR ADVANCING TRANSLATIONAL SCIENCES

NATIONAL INSTITUTES OF HEALTH

7 SEPTEMBER 2015

NCATS

NIH National Center for Advancing Translational Sciences

Outline

- Why we are here
- How we got here
- Where we were
- Where we are now
- Where we are going

Outline

- **Why we are here**
- How we got here
- Where we were
- Where we are now
- Where we are going

Why we are here:
NCATS Mission

- **N**ational
- **C**enter for
- **A**dvancing
- **T**ranslational
- **S**ciences

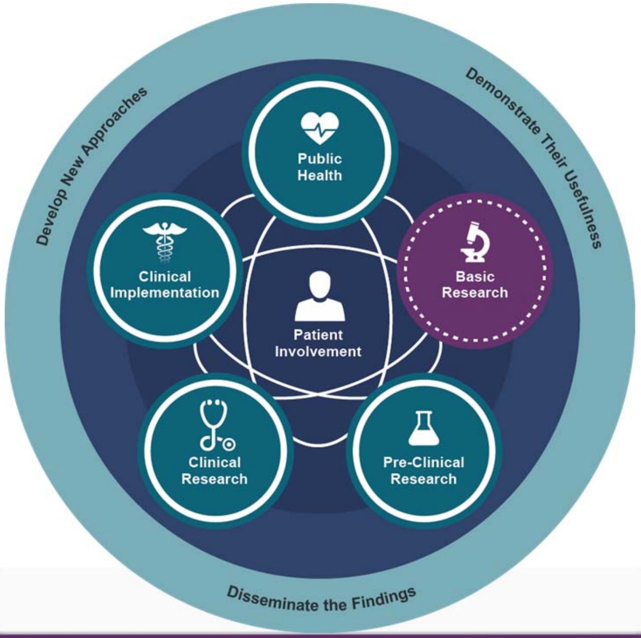
Translational science is a multidisciplinary form of science that bridges the recalcitrant gaps that sometimes exist between fundamental science and applied science, necessitating something in between **to translate knowledge into applications**.

Why we are here:
NCATS Mission

Translational science is a multidisciplinary form of science that bridges the recalcitrant gaps that sometimes exist between fundamental science and applied science, necessitating something in between **to translate knowledge into applications**.

Three D's:

1. Demonstrate
2. Develop
3. Distribute



Why we are here: NCATS Interest

- **NCATS need to define substances**
  - Informatics Discovery / Exploration
  - Drug Repurposing
  - Regulatory inquiries
- **Need for:**
  - Good identifiers
  - Curated data
  - Common data elements
  - Common data strategy

Why we are here:
Substance Problem

- **Fundamental Science**
  - Groundwork set by International Standard (ISO IDMP)
  - Great Subject Matter Experts with detailed knowledge of substances
  - Supporting software libraries are available
- **Application Need**
  - Global identifier
  - Common strategy
  - Common understanding
  - Common data system

Outline

- Why we are here
- **How we got here**
  - **Familiarization with the Standard**
  - Familiarization with the Data
  - Familiarization with the Procedures
- Where we were
- Where we are now
- Where we are going

Familiarization with the Standard

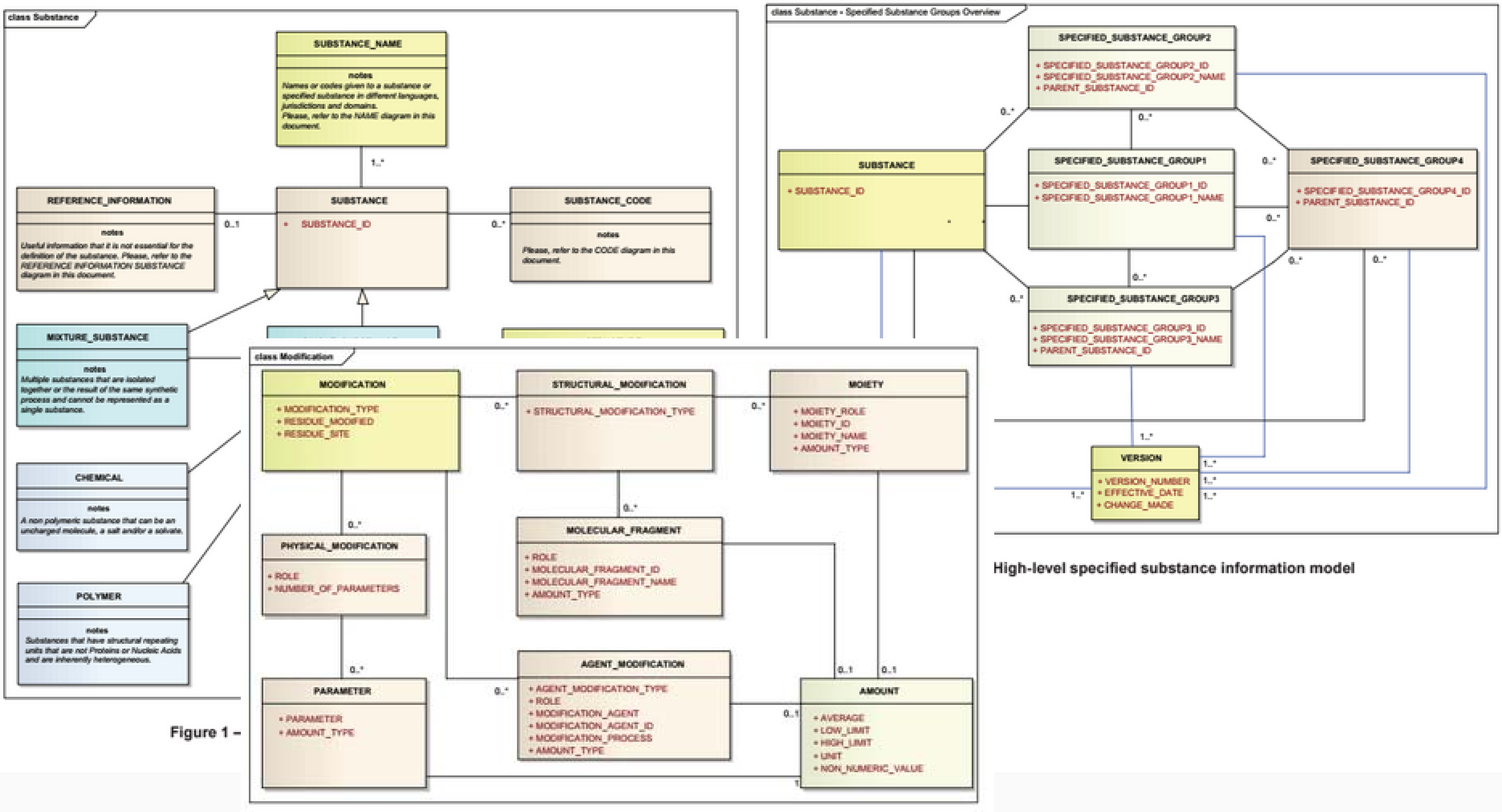

class Substance
SUBSTANCE_NAME
notes
Names or codes given to a substance or specified substance in different languages, jurisdictions and domains.
Please, refer to the NAME diagram in this document.
1..*
REFERENCE_INFORMATION
notes
Useful information that it is not essential for the definition of the substance. Please, refer to the REFERENCE INFORMATION SUBSTANCE diagram in this document.
0..1
SUBSTANCE
+ SUBSTANCE_ID
0..*
SUBSTANCE_CODE
notes
Please, refer to the CODE diagram in this document.
MIXTURE_SUBSTANCE
notes
Multiple substances that are isolated together or the result of the same synthetic process and cannot be represented as a single substance.
CHEMICAL
notes
A non polymeric substance that can be an uncharged molecule, a salt and/or a solvate.
POLYMER
notes
Substances that have structural repeating units that are not Proteins or Nucleic Acids and are inherently heterogeneous.
class Substance - Specified Substance Groups Overview
SPECIFIED_SUBSTANCE_GROUP2
+ SPECIFIED_SUBSTANCE_GROUP2_ID
+ SPECIFIED_SUBSTANCE_GROUP2_NAME
+ PARENT_SUBSTANCE_ID
0..*
0..*
SUBSTANCE
+ SUBSTANCE_ID
SPECIFIED_SUBSTANCE_GROUP1
+ SPECIFIED_SUBSTANCE_GROUP1_ID
+ SPECIFIED_SUBSTANCE_GROUP1_NAME
0..*
SPECIFIED_SUBSTANCE_GROUP4
+ SPECIFIED_SUBSTANCE_GROUP4_ID
+ PARENT_SUBSTANCE_ID
0..*
0..*
0..*
0..*
0..*
SPECIFIED_SUBSTANCE_GROUP3
+ SPECIFIED_SUBSTANCE_GROUP3_ID
+ SPECIFIED_SUBSTANCE_GROUP3_NAME
+ PARENT_SUBSTANCE_ID
1..*
VERSION
+ VERSION_NUMBER
+ EFFECTIVE_DATE
+ CHANGE_MADE
1..*
1..*
1..*
1..*
class Modification
MODIFICATION
+ MODIFICATION_TYPE
+ RESIDUE_MODIFIED
+ RESIDUE_SITE
0..*
STRUCTURAL_MODIFICATION
+ STRUCTURAL_MODIFICATION_TYPE
0..*
MOIETY
+ MOIETY_ROLE
+ MOIETY_ID
+ MOIETY_NAME
+ AMOUNT_TYPE
0..*
0..*
PHYSICAL_MODIFICATION
+ ROLE
+ NUMBER_OF_PARAMETERS
MOLECULAR_FRAGMENT
+ ROLE
+ MOLECULAR_FRAGMENT_ID
+ MOLECULAR_FRAGMENT_NAME
+ AMOUNT_TYPE
0..*
PARAMETER
+ PARAMETER
+ AMOUNT_TYPE
0..*
AGENT_MODIFICATION
+ AGENT_MODIFICATION_TYPE
+ ROLE
+ MODIFICATION_AGENT
+ MODIFICATION_AGENT_ID
+ MODIFICATION_PROCESS
+ AMOUNT_TYPE
0..1
0..1
0..1
AMOUNT
+ AVERAGE
+ LOW_LIMIT
+ HIGH_LIMIT
+ UNIT
+ NON_NUMERIC_VALUE
1

Figure 1 –

High-level specified substance information model

Figure 6 — Information model for modifications

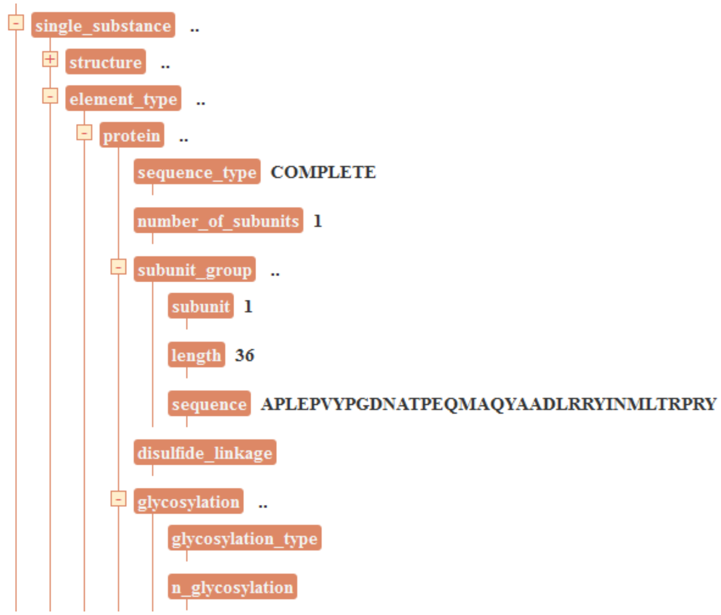
Familiarization with the Data

FDA SRS XML

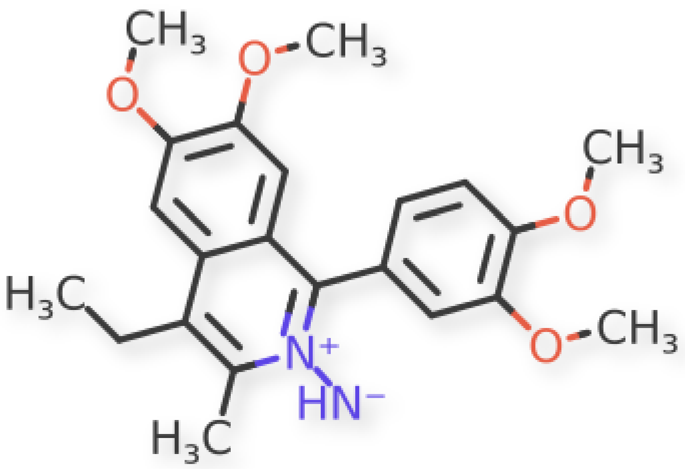
FDA SRS Chemical Structures

FDA SRS AUX Data

External Data



```
- single_substance ..
  + structure ..
  - element_type ..
    - protein ..
        sequence_type COMPLETE
        number_of_subunits 1
      - subunit_group ..
          subunit 1
          length 36
          sequence APLEPVYPGDNATPEQMAQYAADLRRYINMLTRPRY
        disulfide_linkage
      - glycosylation ..
          glycosylation_type
          n_glycosylation
```

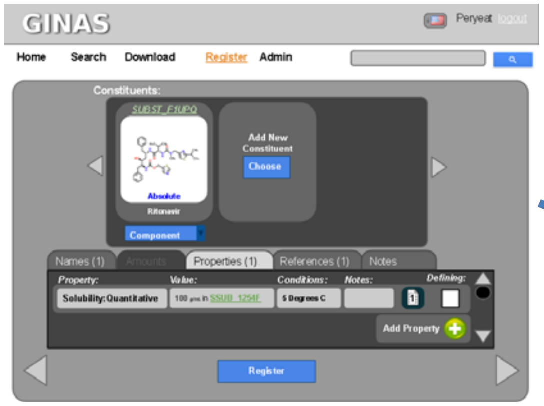


Familiarize with procedures

- Ginas working group, mockups for how-to at user-experience level
  - Feedback on each mockup proposal for every substance class
  - Implementation to functional version
  - Feedback on functional version

Familiarize with the procedures

Mockups



GINAS

Peryeat logout

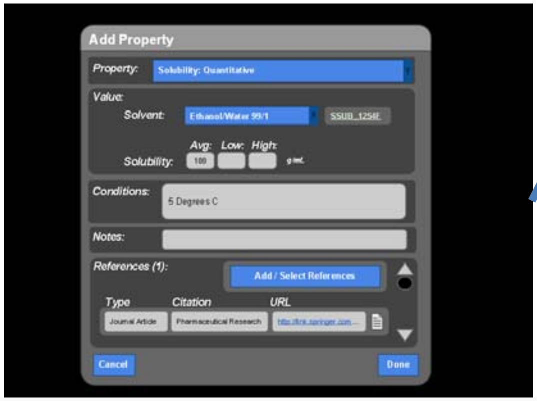
Constituents:

SUBST_F1UPO

Absolute

Ritonavir

Property:	Value:	Conditions:	Notes:	Defining:
Solubility: Quantitative	100 g/mL in SSUB_1254F	5 Degrees C		□



Add Property

Property: Solubility: Quantitative

Value:

Solvent: Ethanol/Water 99/1 SSUB_1254F

	Avg:	Low:	High:	
Solubility:	100			g/mL

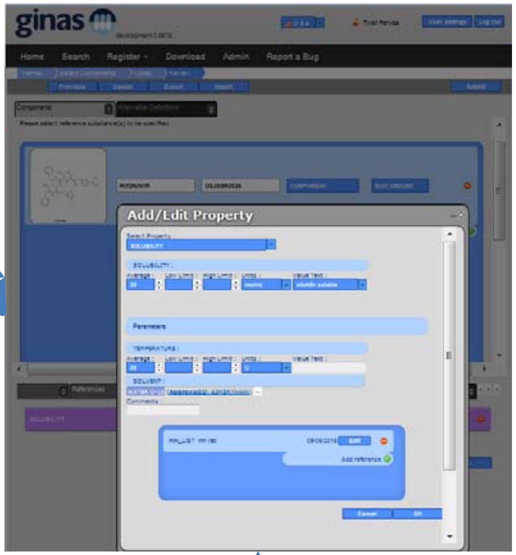
Conditions: 5 Degrees C

Notes:

References (1):

Type	Citation	URL
Journal Article	Pharmaceutical Research	http://link.springer.com...

Live Front End



Please select reference substance(s) to be specified

Add/Edit Property

Select Property: SOLUBILITY

SOLUBILITY:

Average: Low Limit: High Limit: Units: Value Text:

Parameters

TEMPERATURE:

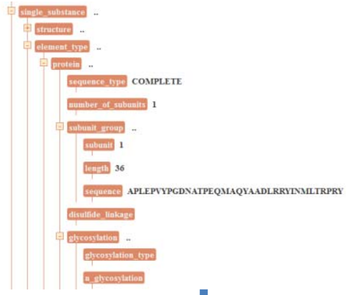
Average: Low Limit: High Limit: Units: C Value Text:

SOLVENT:

Comments:

Add reference

Seed data



```
single_substance ..
  structure ..
  element_type ..
    protein ..
      sequence_type COMPLETE
      number_of_subunits 1
      subunit_group ..
        subunit 1
        length 36
        sequence APLEPVYPGDNATPEQMAQYAADLRRYINMLTRPRY
      disulfide_linkage
      glycosylation ..
        glycosylation_type
        n_glycosylation
```

Transformation

Standardization

Validation

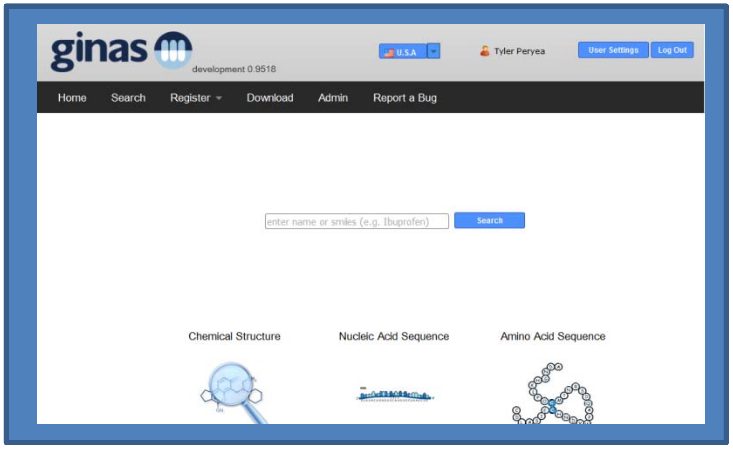
Data Model and API

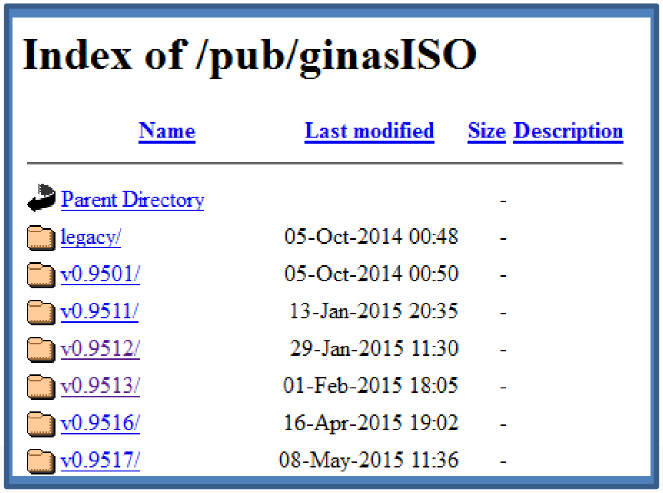
Putting it all together

- REST API
- Distributable Messages and Data dumps
- User’s guide
- Conformance to standard
- User Acceptance Testing
- Rollout

Where we were

- **February 2015**
  - Version 0.9
    - feature complete
      - Registration for all substance classes, up to group 1 specified substance
      - Alternative definitions
      - Approval process
      - Basic search available for textual, structural, sequence-based data
    - Self-contained and distributable
    - Programmatically Accessible
  - Launched on Health Canada site





Index of /pub/ginasISO

Name	Last modified	Size	Description
Parent Directory		-	
legacy/	05-Oct-2014 00:48	-	
v0.9501/	05-Oct-2014 00:50	-	
v0.9511/	13-Jan-2015 20:35	-	
v0.9512/	29-Jan-2015 11:30	-	
v0.9513/	01-Feb-2015 18:05	-	
v0.9516/	16-Apr-2015 19:02	-	
v0.9517/	08-May-2015 11:36	-	

Where we were

What we learned

The Good

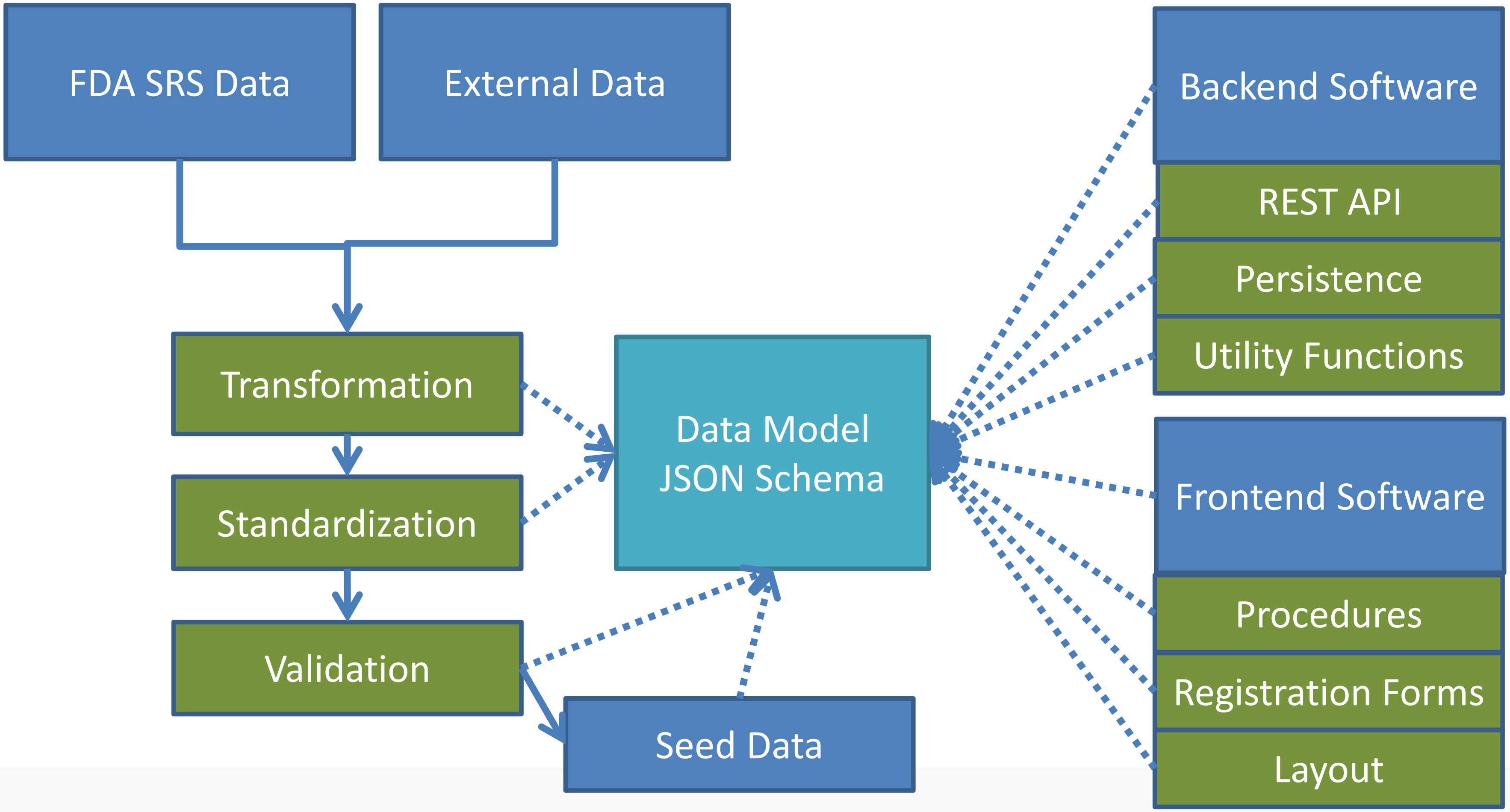
- Lots of features
- Lots of details
- Lots of discussion
- Physical instantiation of data
- Utility functions for registration

The Bad

- Lots of features
- Lots of details
- Lots of discussion
- Complex installation
- Difficult to customize, beyond substances
- Lack of mobile support
- Persistence layer not as flexible / accessible
- **Browsing / discoverability lacking**

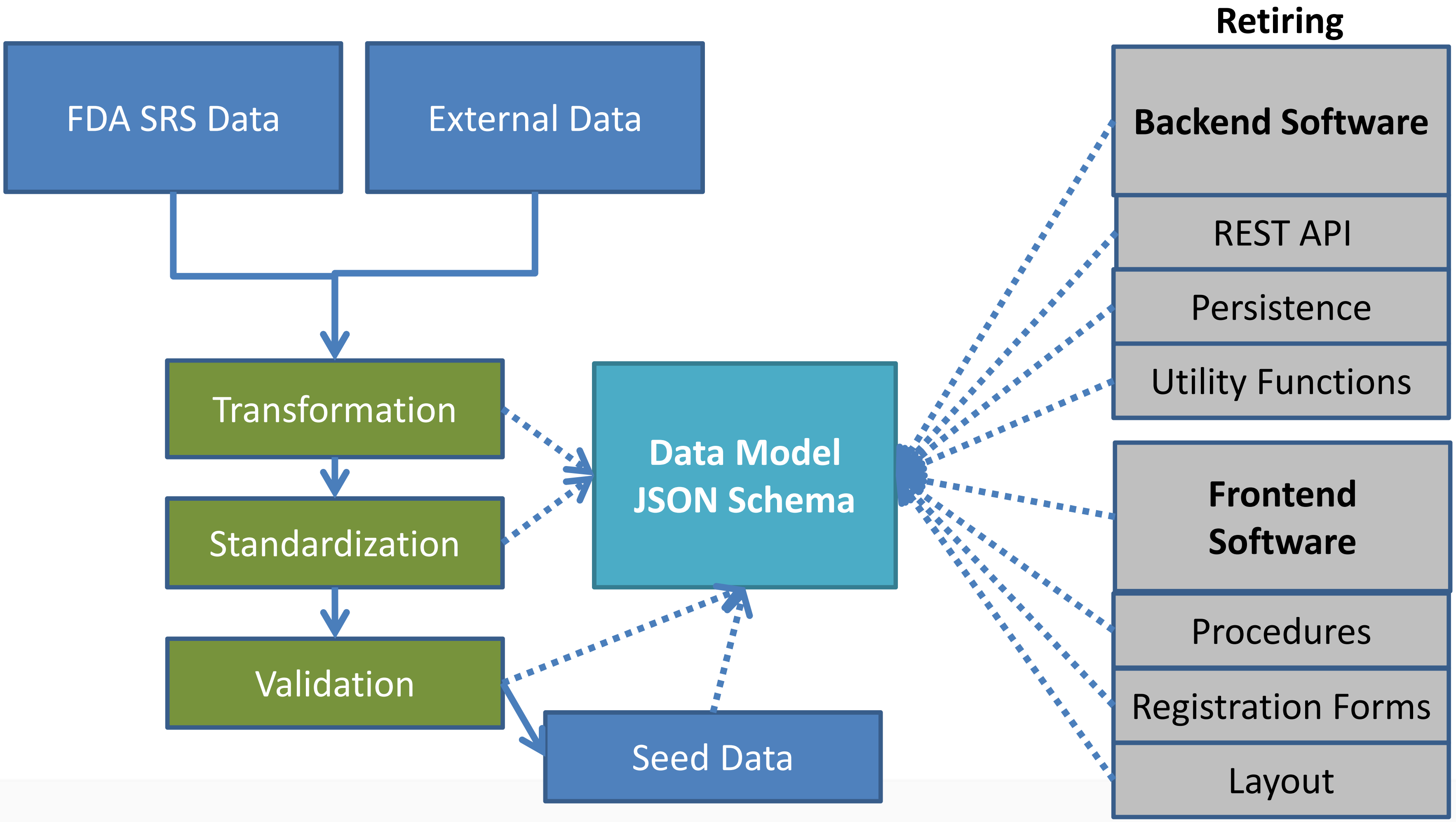
Where we were

What we learned



Where we were

What we learned

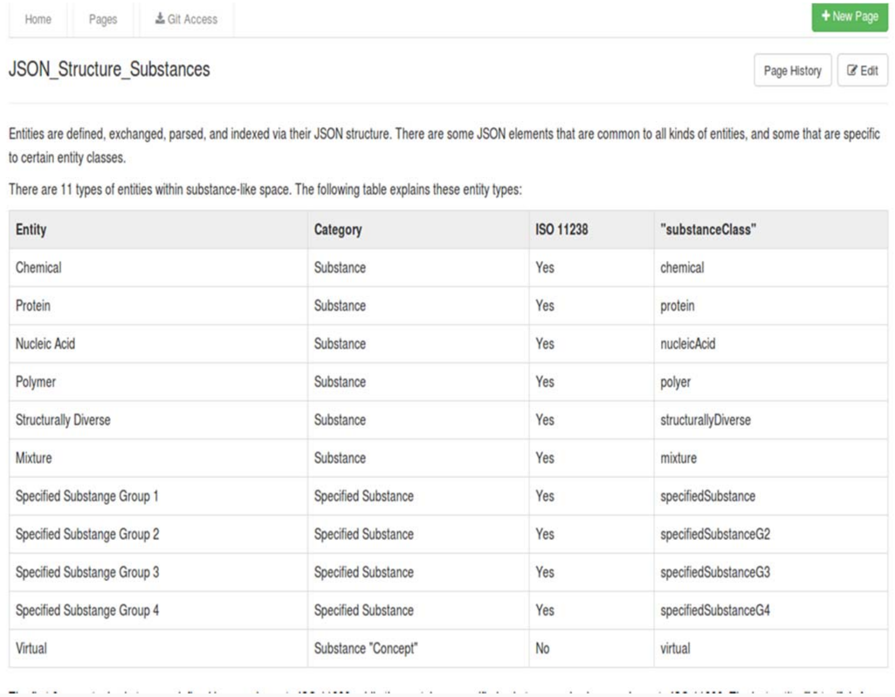


Where we were

What we learned

Data Model JSON Schema

- **JSON Schema Overview**
- JSON Schema versioning
- Schema Review



JSON_Structure_Substances

Entities are defined, exchanged, parsed, and indexed via their JSON structure. There are some JSON elements that are common to all kinds of entities, and some that are specific to certain entity classes.

There are 11 types of entities within substance-like space. The following table explains these entity types:

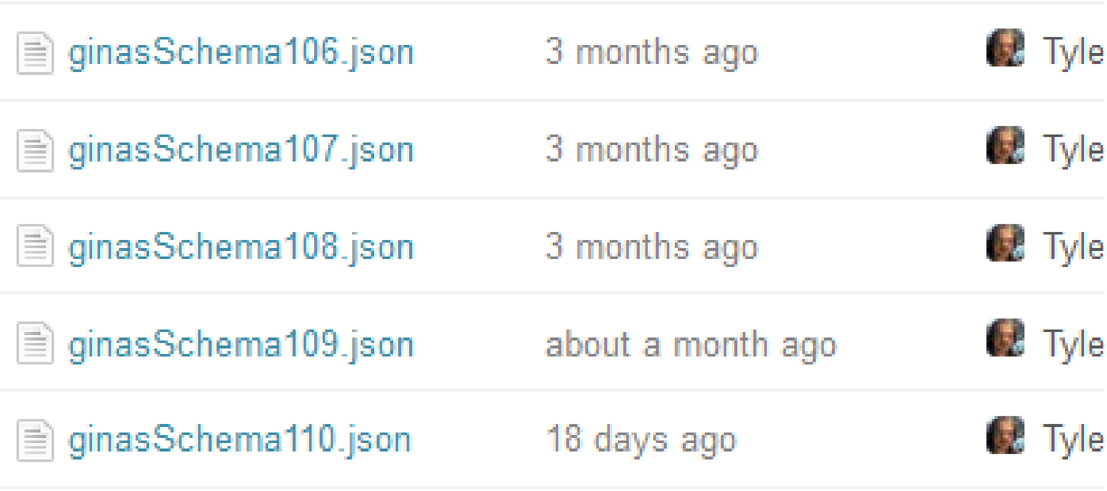
Entity	Category	ISO 11238	"substanceClass"
Chemical	Substance	Yes	chemical
Protein	Substance	Yes	protein
Nucleic Acid	Substance	Yes	nucleicAcid
Polymer	Substance	Yes	polyer
Structurally Diverse	Substance	Yes	structurallyDiverse
Mixture	Substance	Yes	mixture
Specified Substange Group 1	Specified Substance	Yes	specifiedSubstance
Specified Substange Group 2	Specified Substance	Yes	specifiedSubstanceG2
Specified Substange Group 3	Specified Substance	Yes	specifiedSubstanceG3
Specified Substange Group 4	Specified Substance	Yes	specifiedSubstanceG4
Virtual	Substance "Concept"	No	virtual

Where we were

What we learned

Data Model JSON Schema

- JSON Schema Overview
- **JSON Schema versioning**
- Schema Review



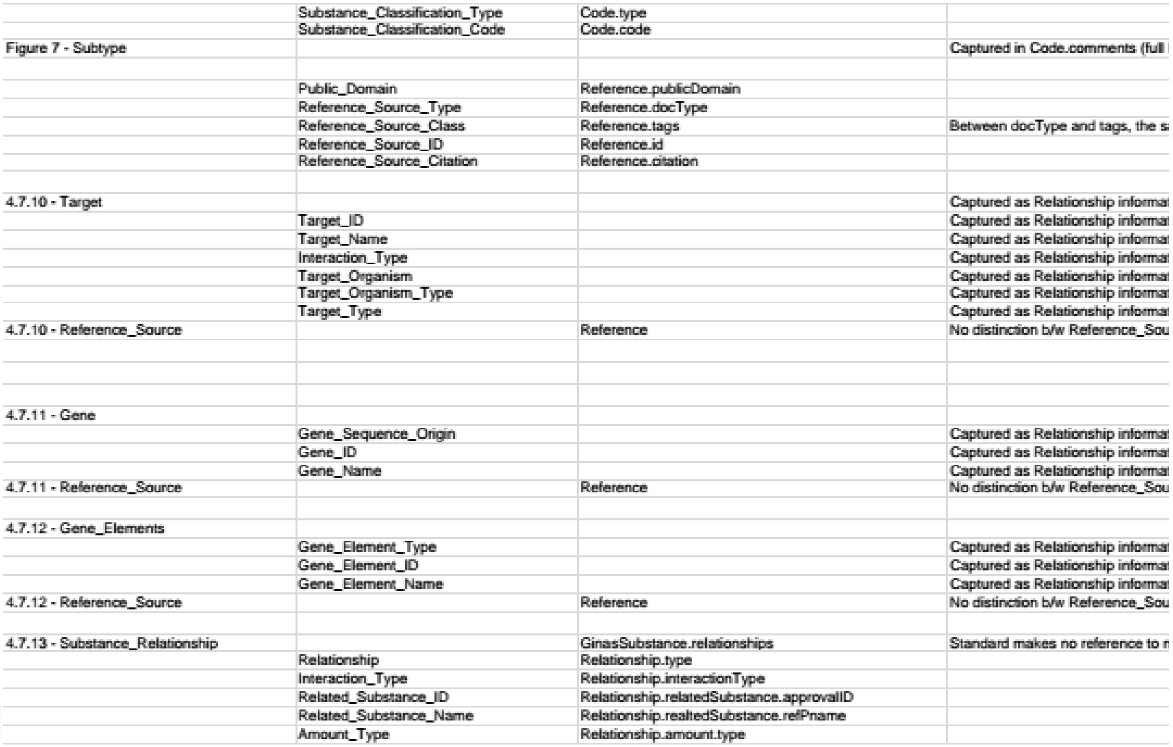
ginasSchema106.json	3 months ago	Tyle
ginasSchema107.json	3 months ago	Tyle
ginasSchema108.json	3 months ago	Tyle
ginasSchema109.json	about a month ago	Tyle
ginasSchema110.json	18 days ago	Tyle

Where we were

What we learned

Data Model JSON Schema

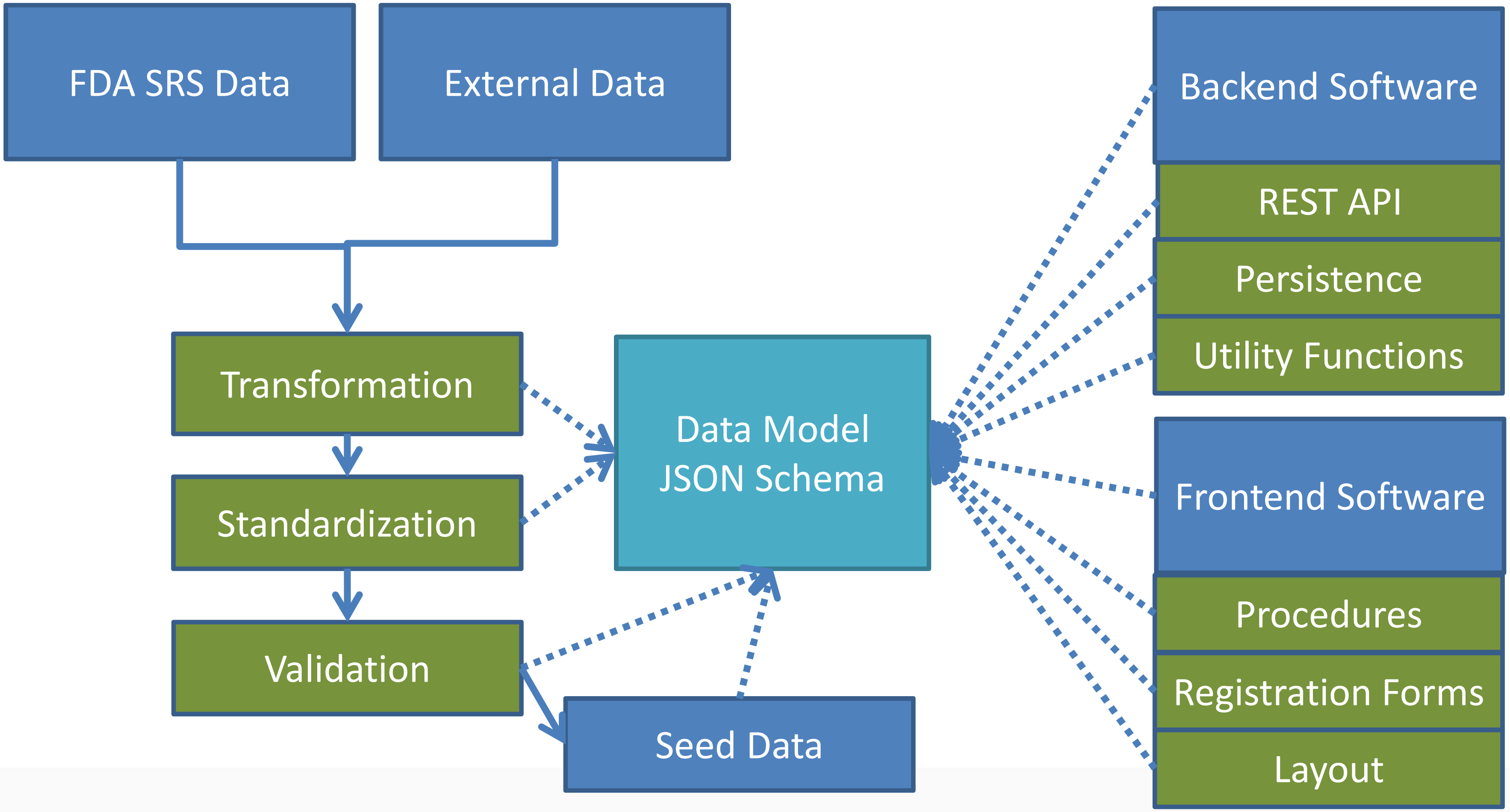
- JSON Schema Overview
- JSON Schema versioning
- **Schema Review**



	Substance_Classification_Type	Code.type	
	Substance_Classification_Code	Code.code	
Figure 7 - Subtype			Captured in Code.comments (full
	Public_Domain	Reference.publicDomain	
	Reference_Source_Type	Reference.docType	
	Reference_Source_Class	Reference.tags	Between docType and tags, the s
	Reference_Source_ID	Reference.id	
	Reference_Source_Citation	Reference.citation	
4.7.10 - Target			Captured as Relationship informat
	Target_ID		Captured as Relationship informat
	Target_Name		Captured as Relationship informat
	Interaction_Type		Captured as Relationship informat
	Target_Organism		Captured as Relationship informat
	Target_Organism_Type		Captured as Relationship informat
	Target_Type		Captured as Relationship informat
4.7.10 - Reference_Source		Reference	No distinction b/w Reference_Sou
4.7.11 - Gene			
	Gene_Sequence_Origin		Captured as Relationship informat
	Gene_ID		Captured as Relationship informat
	Gene_Name		Captured as Relationship informat
4.7.11 - Reference_Source		Reference	No distinction b/w Reference_Sou
4.7.12 - Gene_Elements			
	Gene_Element_Type		Captured as Relationship informat
	Gene_Element_ID		Captured as Relationship informat
	Gene_Element_Name		Captured as Relationship informat
4.7.12 - Reference_Source		Reference	No distinction b/w Reference_Sou
4.7.13 - Substance_Relationship		GinasSubstance.relationships	Standard makes no reference to r
	Relationship	Relationship.type	
	Interaction_Type	Relationship.interactionType	
	Related_Substance_ID	Relationship.relatedSubstance.approvalID	
	Related_Substance_Name	Relationship.realtedSubstance.refPname	
	Amount_Type	Relationship.amount.type	

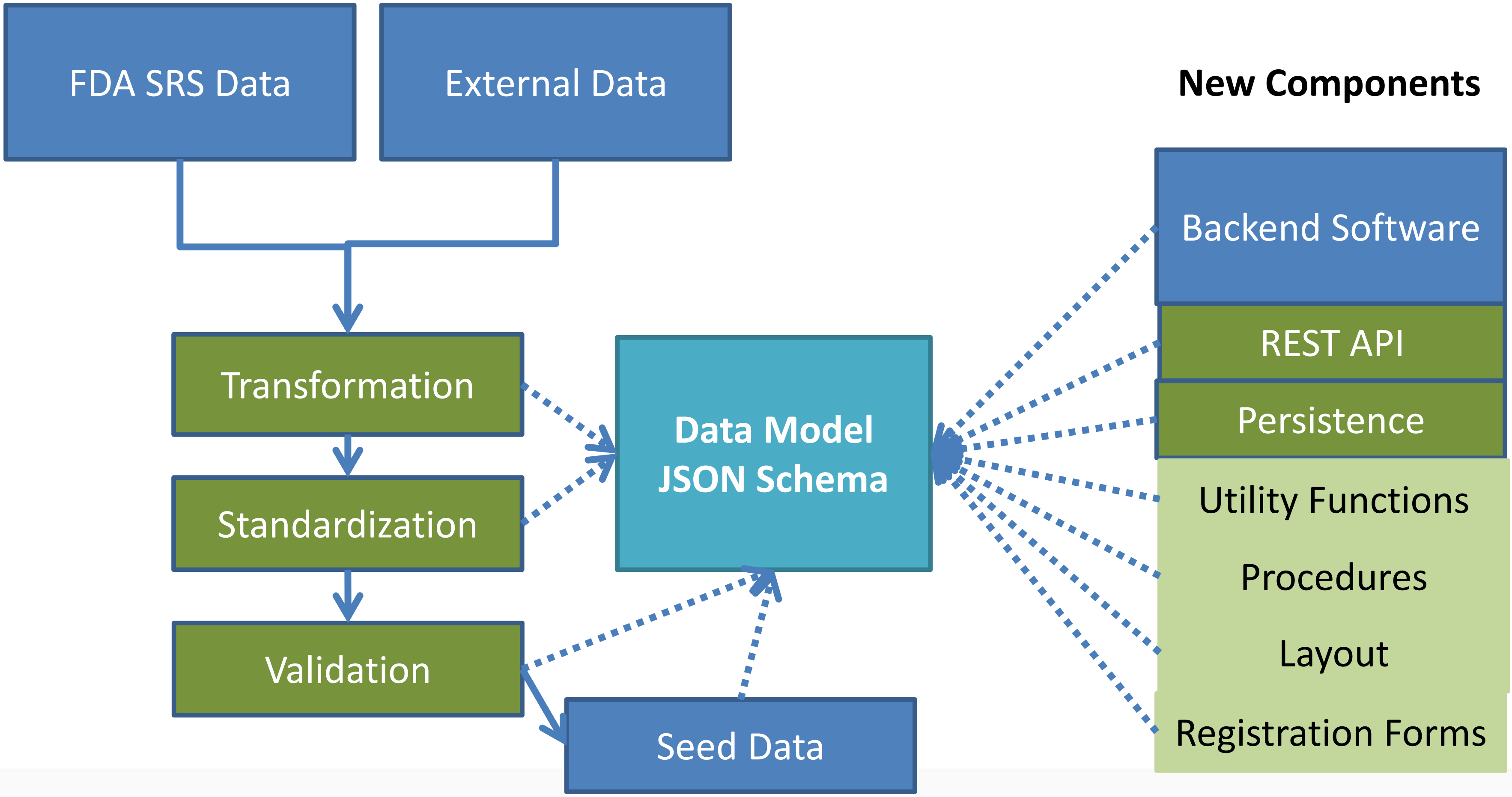
Where we were

What we learned



Where we were

What we learned

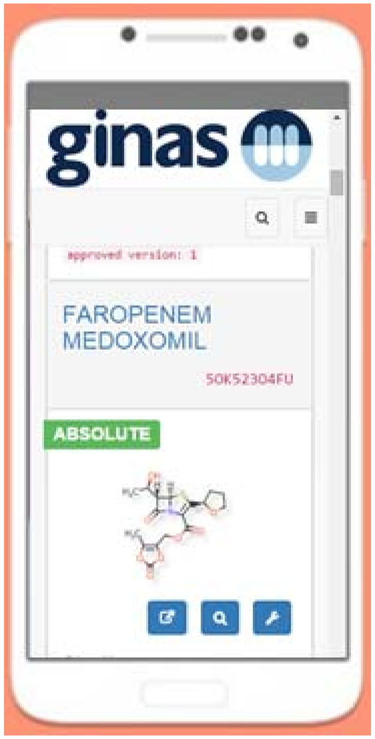


Outline

- Why we are here
- How we got here
- Where we were
- **Where we are now**
- Where we are going

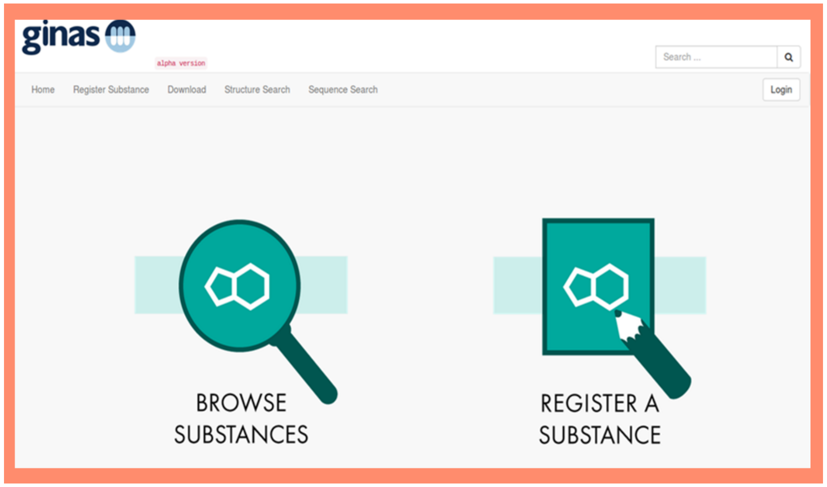
What we are now

- Frontend / Backend rewrite
  - ➢ Discoverability first
  - ➢ Mobility first
  - ➢ Focus on browsing data



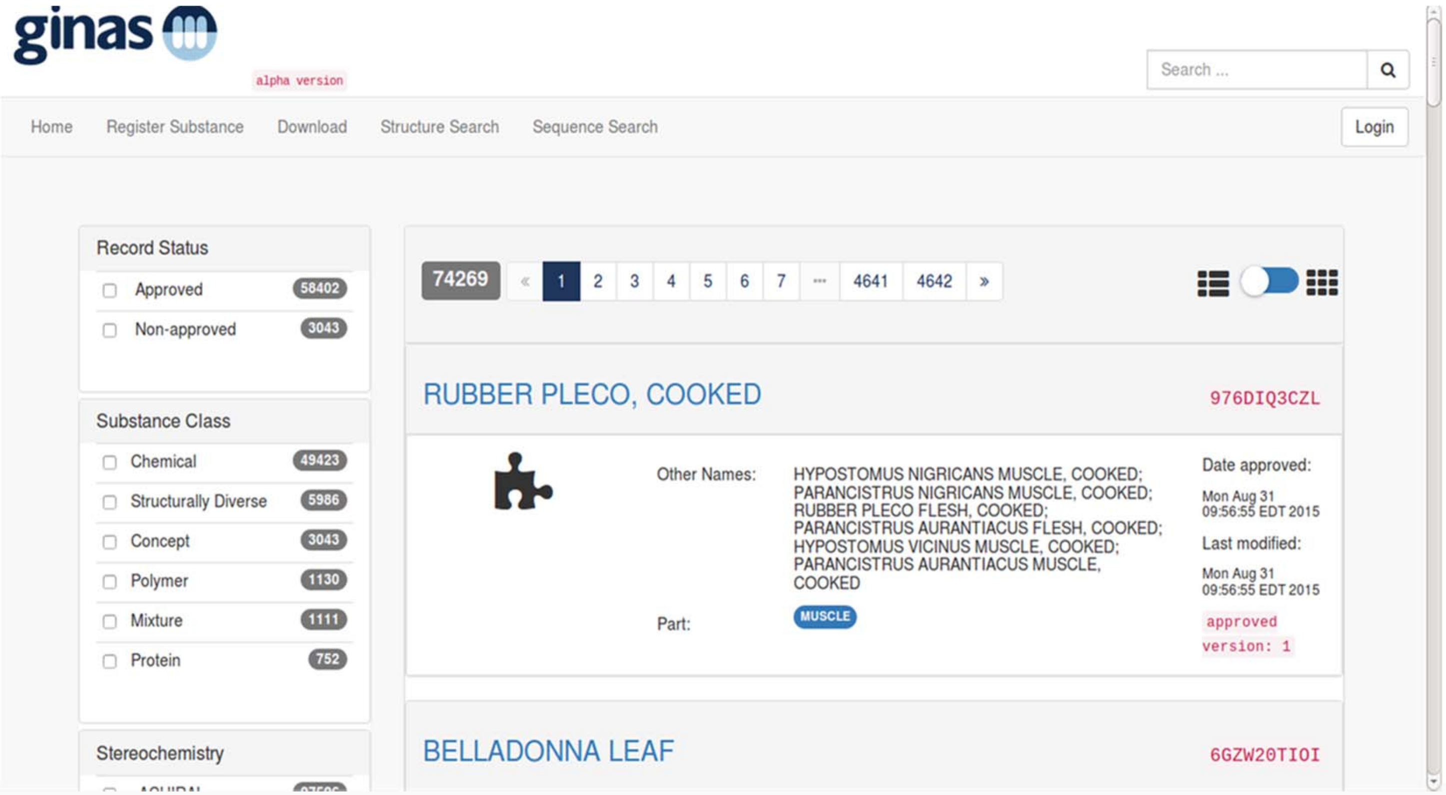
FAROPENEM MEDOXOMIL

50K52304FU



Where we are now

Browse all substances in seed data



Record Status

□ Approved	58402
□ Non-approved	3043

Substance Class

□ Chemical	49423
□ Structurally Diverse	5986
□ Concept	3043
□ Polymer	1130
□ Mixture	1111
□ Protein	752

Stereochemistry

RUBBER PLECO, COOKED 976DIQ3CZL

Other Names: HYPOSTOMUS NIGRICANS MUSCLE, COOKED; PARANCISTRUS NIGRICANS MUSCLE, COOKED; RUBBER PLECO FLESH, COOKED; PARANCISTRUS AURANTIACUS FLESH, COOKED; HYPOSTOMUS VICINUS MUSCLE, COOKED; PARANCISTRUS AURANTIACUS MUSCLE, COOKED

Part: MUSCLE

Date approved: Mon Aug 31 09:56:55 EDT 2015

Last modified: Mon Aug 31 09:56:55 EDT 2015

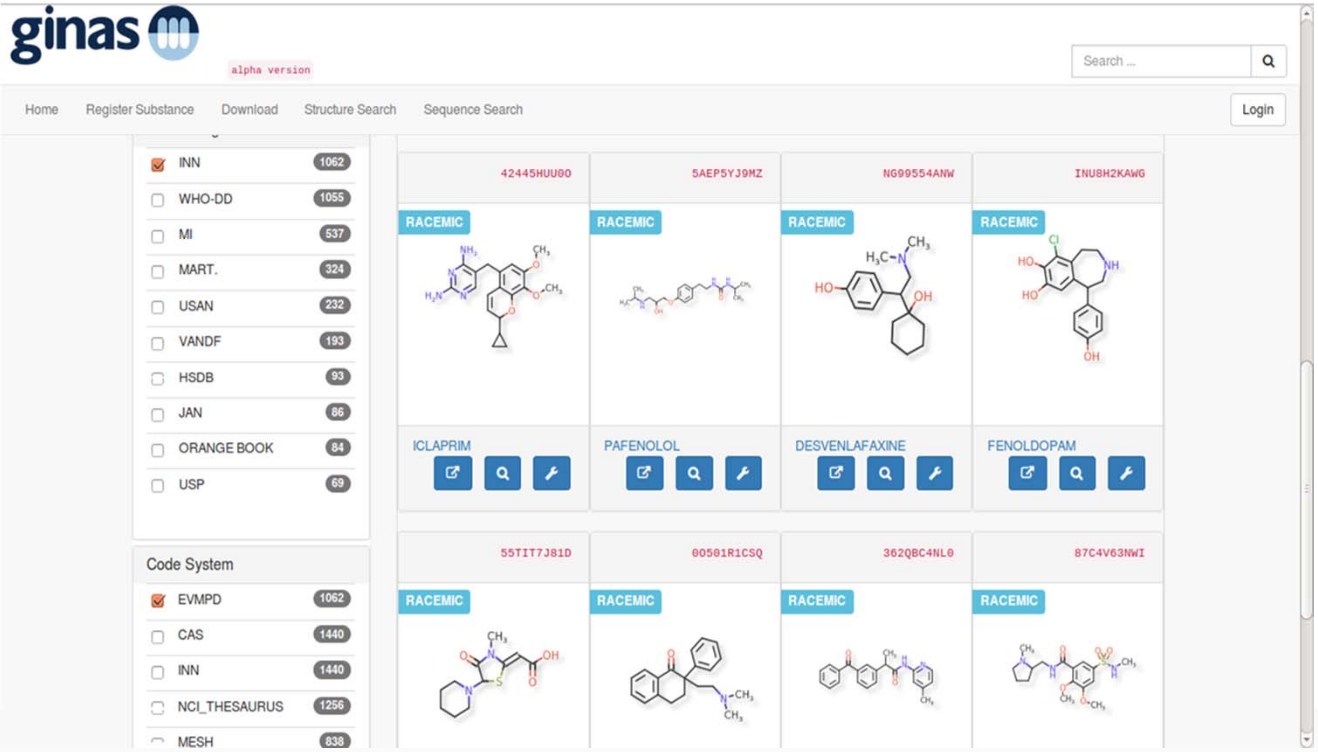
approved

version: 1

BELLADONNA LEAF 6GZW20TIOI

Where we are now

Quick filtering to entities of interest *("racemic INN-named chemical substances with EVMPD codes")*



☑	INN	1062
□	WHO-DD	1055
□	MI	537
□	MART.	324
□	USAN	232
□	VANDF	193
□	HSDB	93
□	JAN	86
□	ORANGE BOOK	84
□	USP	69

Code System

☑	EVMPD	1062
□	CAS	1440
□	INN	1440
□	NCI_THESAURUS	1256
□	MESH	838

42445HUU0O RACEMIC ICLAPRIM

5AEP5YJ9MZ RACEMIC PAFENOLOL

NG99554ANW RACEMIC DESVENLAFAXINE

INU8H2KAWG RACEMIC FENOLDOPAM

55TIT7J81D RACEMIC

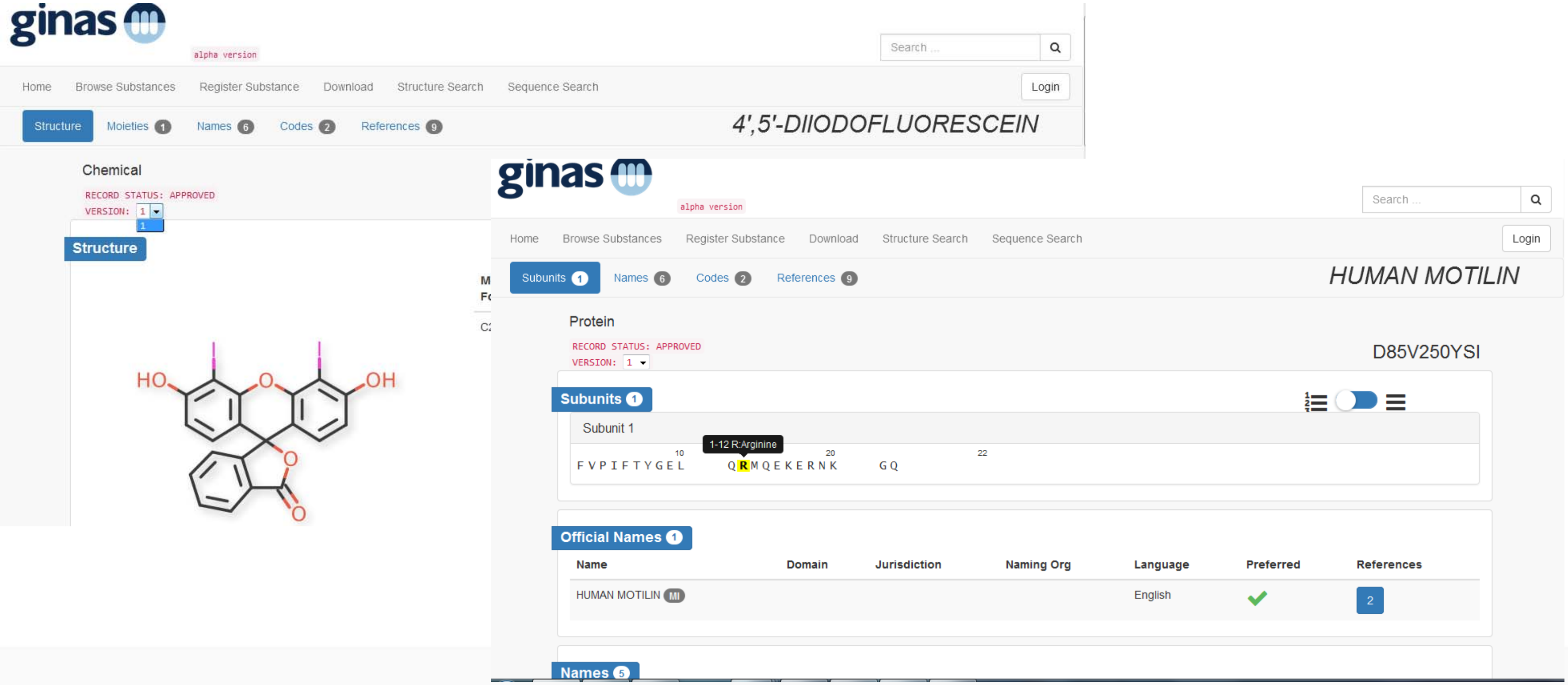
0O501R1CSQ RACEMIC

362QBC4NL0 RACEMIC

87C4V63NWI RACEMIC

Where we are now

Full substance views for substance class



4',5'-DIIODOFLUORESCEIN

Chemical

RECORD STATUS: APPROVED
VERSION: 1

HUMAN MOTILIN

Protein

RECORD STATUS: APPROVED
VERSION: 1

D85V250YSI

Subunit 1

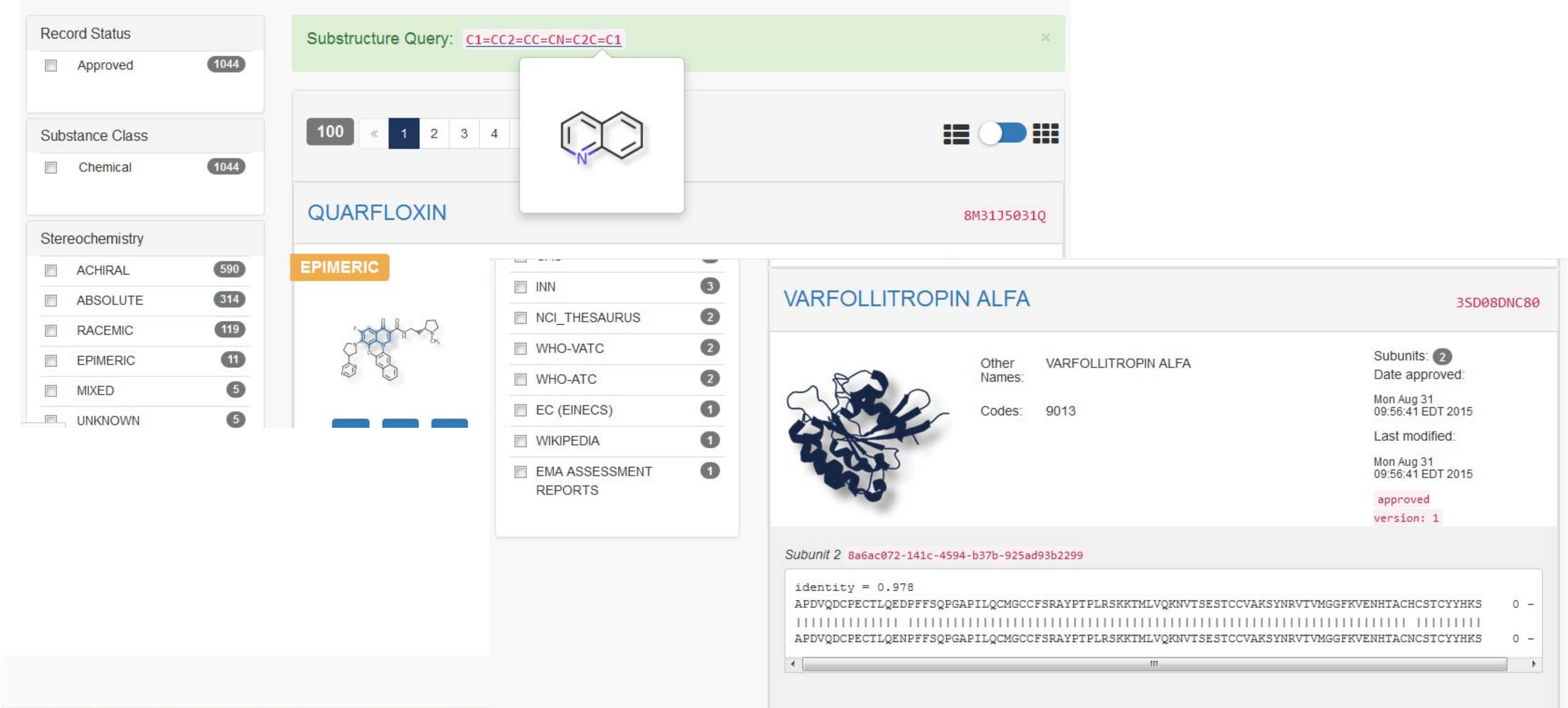
1-12 R:Arginine

```
                10                     20            22
F V P I F T Y G E L    Q R M Q E K E R N K    G Q
```

Name	Domain	Jurisdiction	Naming Org	Language	Preferred	References
HUMAN MOTILIN MI				English	✓	2

Where we are now

Advanced searching



Record Status

□ Approved	1044

Substance Class

□ Chemical	1044

Stereochemistry

□ ACHIRAL	590
□ ABSOLUTE	314
□ RACEMIC	119
□ EPIMERIC	11
□ MIXED	5
□ UNKNOWN	5

Substructure Query: C1=CC2=CC=CN=C2C=C1

QUARFLOXIN 8M31J5031Q

EPIMERIC

□ INN	3
□ NCI_THESAURUS	2
□ WHO-VATC	2
□ WHO-ATC	2
□ EC (EINECS)	1
□ WIKIPEDIA	1
□ EMA ASSESSMENT REPORTS	1

VARFOLLITROPIN ALFA 3SD08DNC80

Other Names: VARFOLLITROPIN ALFA

Codes: 9013

Subunits: 2

Date approved:

Mon Aug 31 09:56:41 EDT 2015

Last modified:

Mon Aug 31 09:56:41 EDT 2015

approved

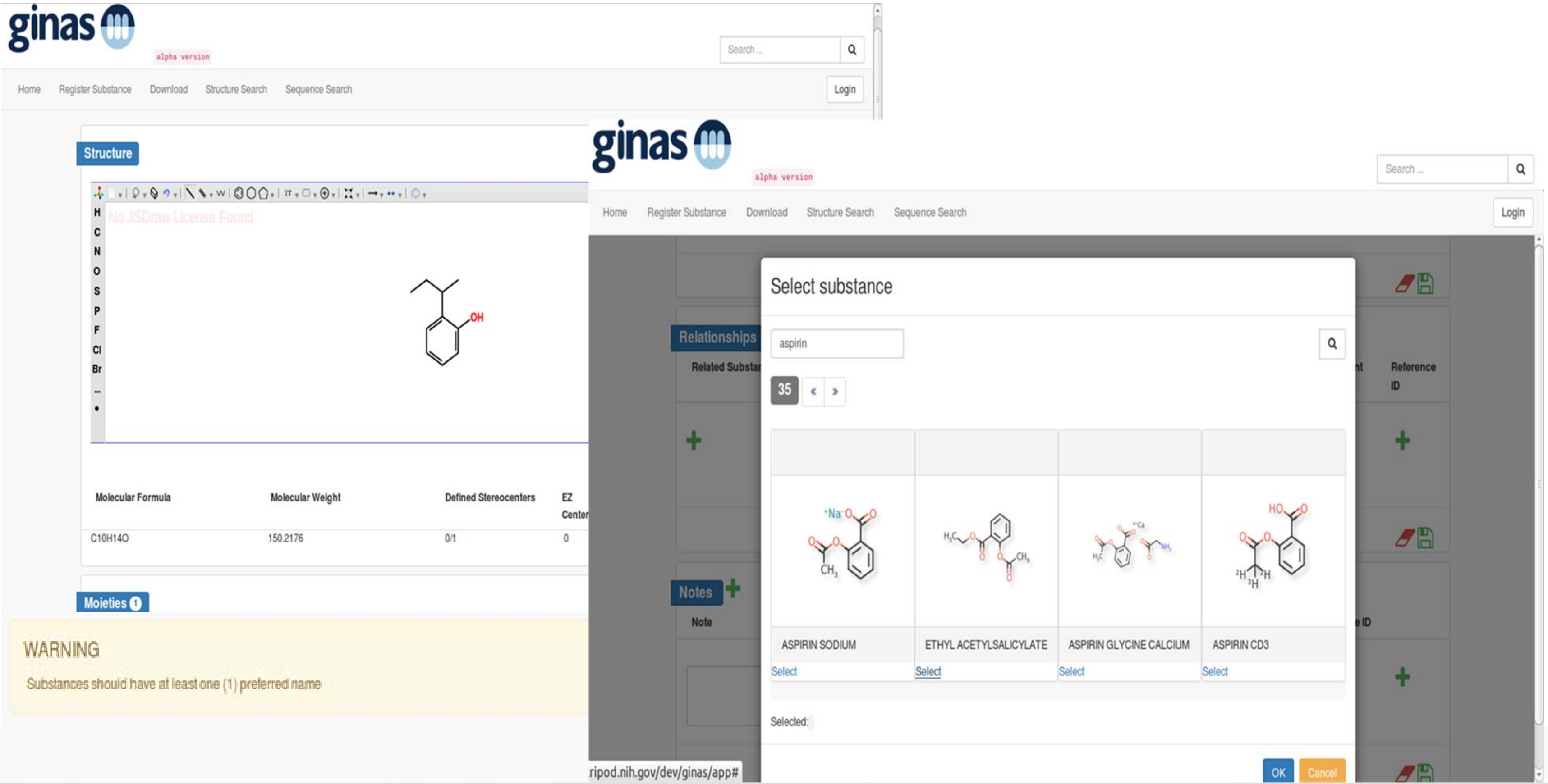
version: 1

Subunit 2 8a6ac072-141c-4594-b37b-925ad93b2299

```
identity = 0.978
APDVQDCPECTLQEDPFFSQPGAPILQCMGCCFSRAYPTPLRSKKTMLVQKNVTSESTCCVAKSYNRVTVMGGFKVENHTACHCSTCYYHKS    0 -
|||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||||
APDVQDCPECTLQENPFFSQPGAPILQCMGCCFSRAYPTPLRSKKTMLVQKNVTSESTCCVAKSYNRVTVMGGFKVENHTACNCSTCYYHKS    0 -
```

Where we are now

Registration and Validation



Molecular Formula	Molecular Weight	Defined Stereocenters	EZ Center
C10H14O	150.2176	0/1	0

WARNING

Substances should have at least one (1) preferred name

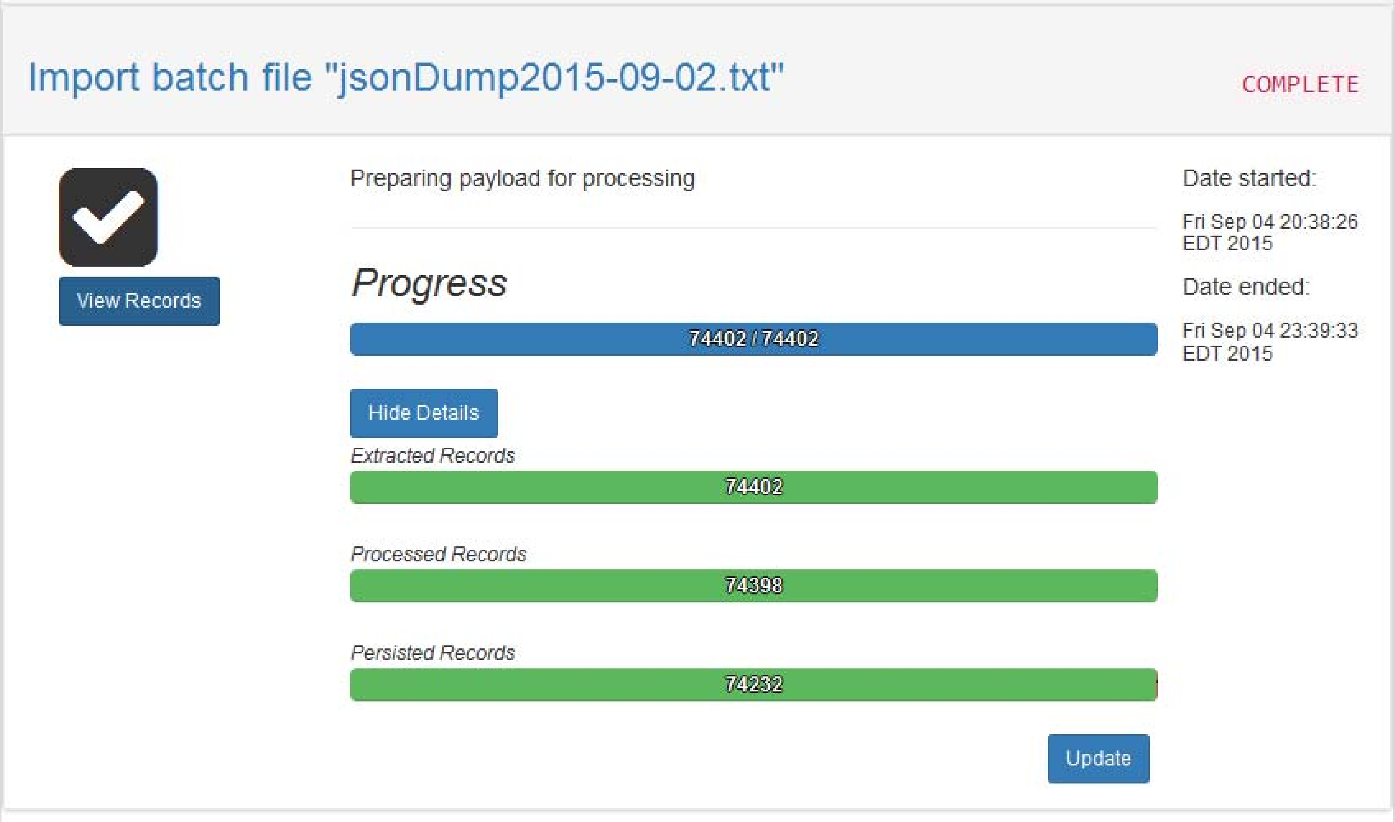
aspirin

Selected:

ripod.nih.gov/dev/ginas/app#

Where we are now

Batch import of records



Import batch file "jsonDump2015-09-02.txt"

COMPLETE

Preparing payload for processing

Progress

74402 / 74402

Extracted Records

74402

Processed Records

74398

Persisted Records

74232

Date started:

Fri Sep 04 20:38:26 EDT 2015

Date ended:

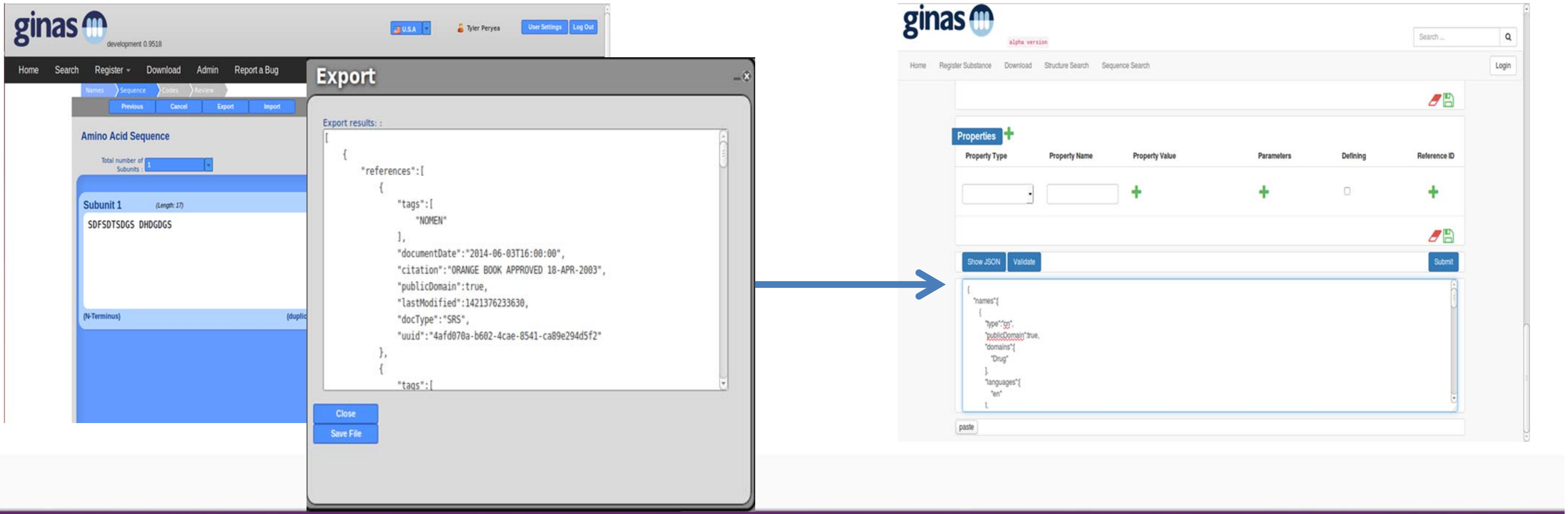
Fri Sep 04 23:39:33 EDT 2015

Outline

- Why we are here
- How we got here
- Where we were
- Where we are now
- **Where we are going**

Where we are going

- **Complete migration of Registration forms to new pages**
  - ➢ Nucleic Acids
  - ➢ Polymers
  - ➢ Proteins



SDFSDTSDGS DHDGDGS

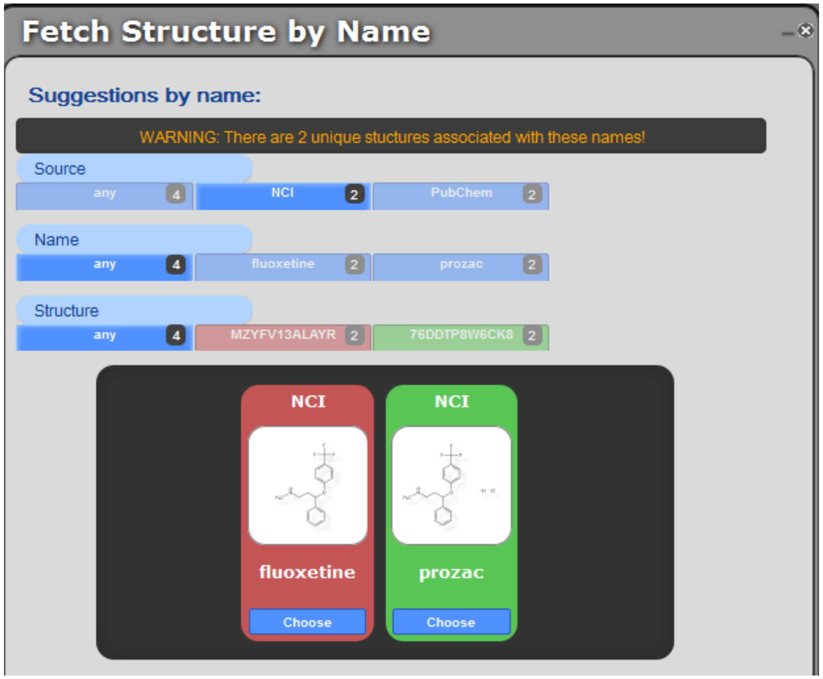
Property Type	Property Name	Property Value	Parameters	Defining	Reference ID
		+	+	□	+

```
[
  {
    "references":[
      {
        "tags":[
          "NOMEN"
        ],
        "documentDate":"2014-06-03T16:00:00",
        "citation":"ORANGE BOOK APPROVED 18-APR-2003",
        "publicDomain":true,
        "lastModified":1421376233630,
        "docType":"SRS",
        "uuid":"4afd070a-b602-4cae-8541-ca89e294d5f2"
      },
      {
        "tags":[
```

```
{
  "names"{
   {
     "type":"cn",
     "publicDomain":true,
     "domains"{
       "Drug"
     ],
     "languages"{
       "en"
     ],
```

Where we are going

- **Complete migration of Registration forms to new pages**
  - Nucleic Acids
  - Polymers
  - Proteins
- **Complete migration of utility functions**
  - Name-to-definition resolver
  - Name analyzers (INN protein format, etc)



Fetch Structure by Name

Suggestions by name:

WARNING: There are 2 unique stuctures associated with these names!

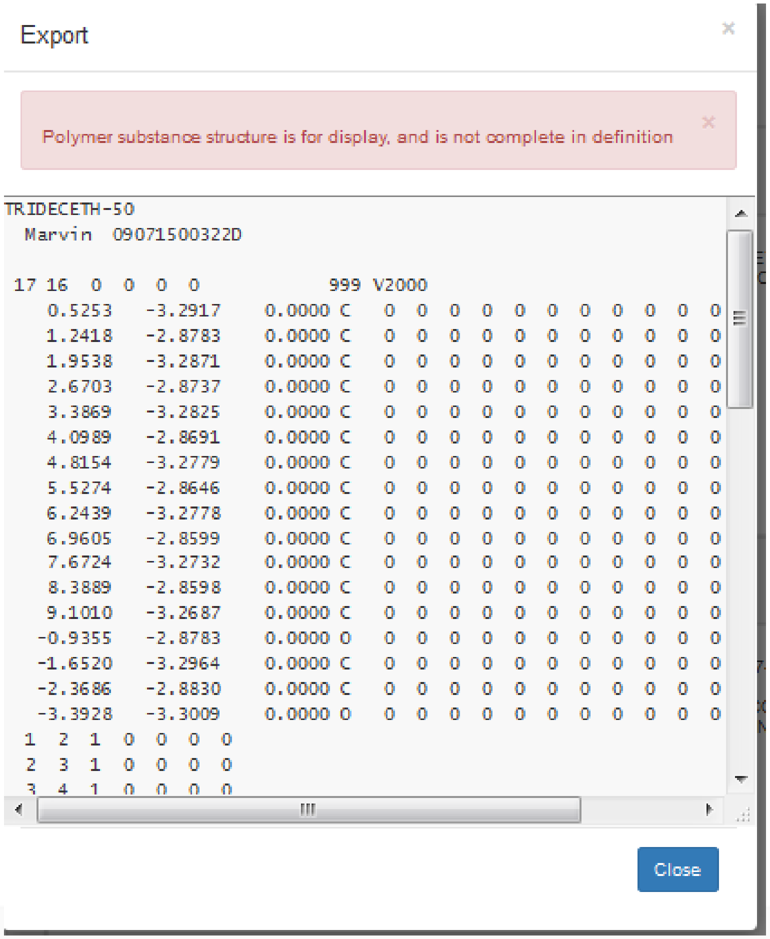
Source: any 4 | NCI 2 | PubChem 2

Name: any 4 | fluoxetine 2 | prozac 2

Structure: any 4 | MZYFV13ALAYR 2 | 76DDTP8W6CK8 2

Where we are going

- **Complete migration of Registration forms to new pages**
  - Nucleic Acids
  - Polymers
  - Proteins
- **Complete migration of utility functions**
  - Name-to-definition resolver
  - Name analyzers (INN protein format, etc)
- **Simple Export to variety of formats**
  - **JSON**
  - **SDF**
  - **Excel**



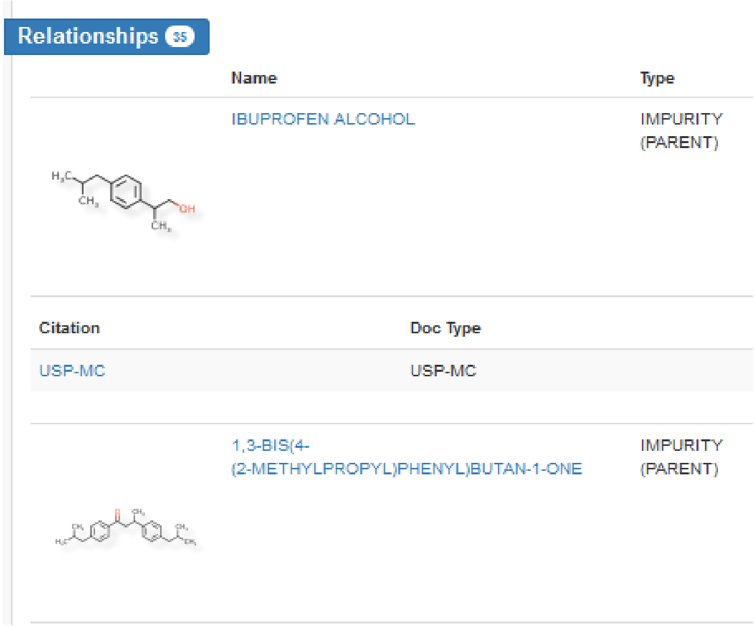
Polymer substance structure is for display, and is not complete in definition ×

```
TRIDECETH-50
  Marvin  09071500322D

 17 16  0  0  0  0            999 V2000
    0.5253   -3.2917    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    1.2418   -2.8783    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    1.9538   -3.2871    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    2.6703   -2.8737    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    3.3869   -3.2825    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    4.0989   -2.8691    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    4.8154   -3.2779    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    5.5274   -2.8646    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    6.2439   -3.2778    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    6.9605   -2.8599    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    7.6724   -3.2732    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    8.3889   -2.8598    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
    9.1010   -3.2687    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
   -0.9355   -2.8783    0.0000 O   0  0  0  0  0  0  0  0  0  0  0
   -1.6520   -3.2964    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
   -2.3686   -2.8830    0.0000 C   0  0  0  0  0  0  0  0  0  0  0
   -3.3928   -3.3009    0.0000 O   0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
```

Where we are going

- **Complete migration of Registration forms to new pages**
  - Nucleic Acids
  - Polymers
  - Proteins
- **Complete migration of utility functions**
  - Name-to-definition resolver
  - Name analyzers (INN protein format, etc)
- **Simple Export to variety of formats**
  - **JSON**
  - **SDF**
  - **Excel**
- **Integrate and validate other data and entities into seed set**
  - Metabolism
  - Monograph information
  - Seed examples for products



	Name	Type
	IBUPROFEN ALCOHOL	IMPURITY (PARENT)

Citation	Doc Type
USP-MC	USP-MC

	Name	Type
	1,3-BIS(4-(2-METHYLPROPYL)PHENYL)BUTAN-1-ONE	IMPURITY (PARENT)

Where we are going

- Demonstration
  - Later today
  - Tomorrow during lunch
- Software Details / Architecture
  - Tim Sheils will give more information tomorrow

ginas

Acknowledgements

U.S. Food and Drug Administration

Yulia Borodina
Larry Callahan
Ta-Jen Chen
Ramez Ghazzaoui
Elaine Johansen
Archana Newatia
Vada Perkins
Mary-Ann Slack
Frank Switzer
Alex Welsch

U.S. Pharmacopeial Convention

Fouad Atouf
Tina Morris
Andrej Wilk

Federal Institute for Drugs and Medical Devices (Germany)

Thomas Balzer

SwissMedic

Philipp Weyermann

European Directorate Quality of Medicines

Christopher Jarvis

Dow Corning

Katherine Ulman

NIH National Center for Advancing Translational Sciences

NIH/NCATS

Chris Austin
Ajit Jadhav
Dac-Trung Nguyen
Tyler Peryea
Tim Sheils
Tongan Zhao

Medicines Evaluation Board (Netherlands)

Herman Diederik
Marcel Hoefnagel
Joris Kampmeijer
Burt Kroes
Ciska Matai

Health Canada

Vikesh Srivastava

Royal Botanic Gardens, Kew (UK)

Bob Allkin
Elizabeth Dauncey

European Medicines Agency

Paolo Alcini
Sabine Brosch
Telonis Pangiotis
Ilaria Del Seppia

Uppsala Montintoring Centre / WHO

Malin Jakobsson
Martin Strömberg
Malin Fladvad