Lessons Learned in Building the CompTox Chemicals Dashboard

Engineering a more sustainable web-based chemical database

Christopher Grulke[1]
Antony Williams[1]
Amar Singh[1]
Jeremy Dunne[1]
Ann Richard[1]
Jeff Edwards[1]

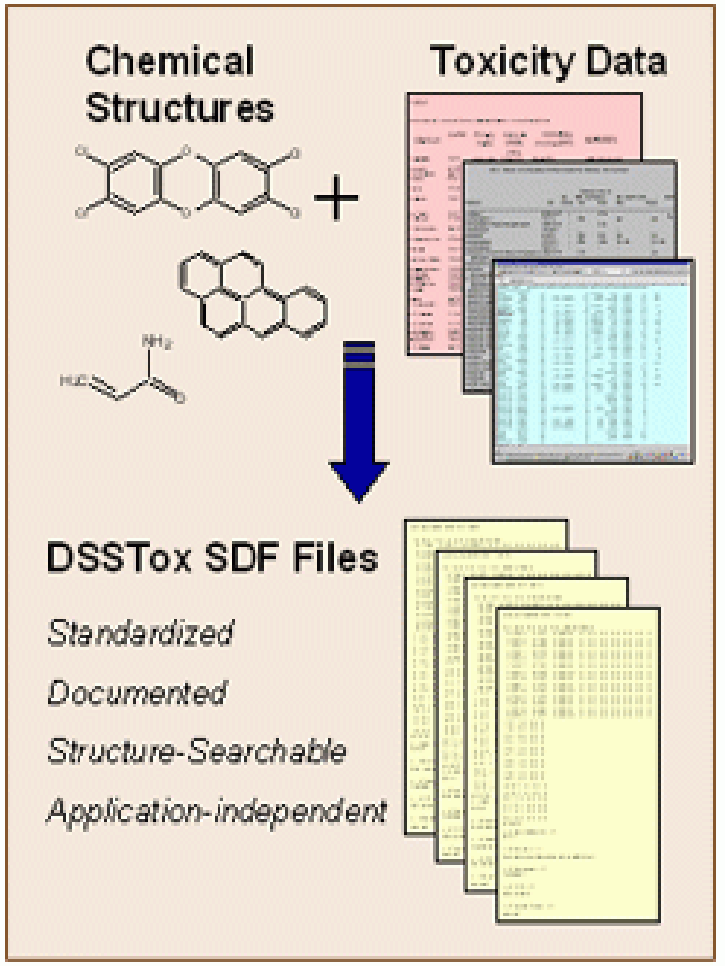
American Chemical Society Meeting, Fall 2019
28 August 2019, San Diego, FL

1. National Center of Computational Toxicology, U.S. EPA
(soon to be the Center for Computational Toxicology and Exposure)

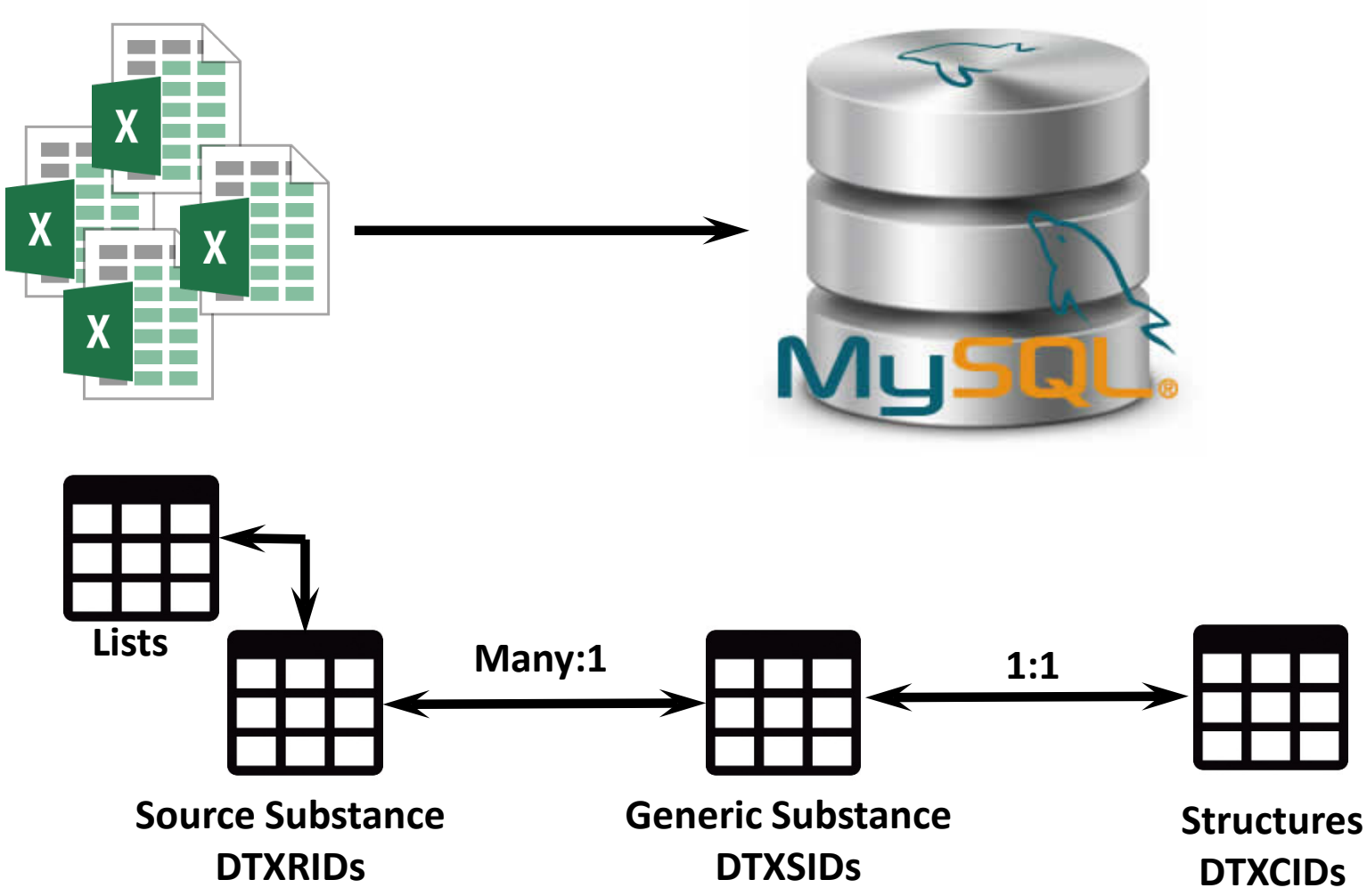
DSSTox Ancient History

Goal: Linking chemical structures to data enabling SAR

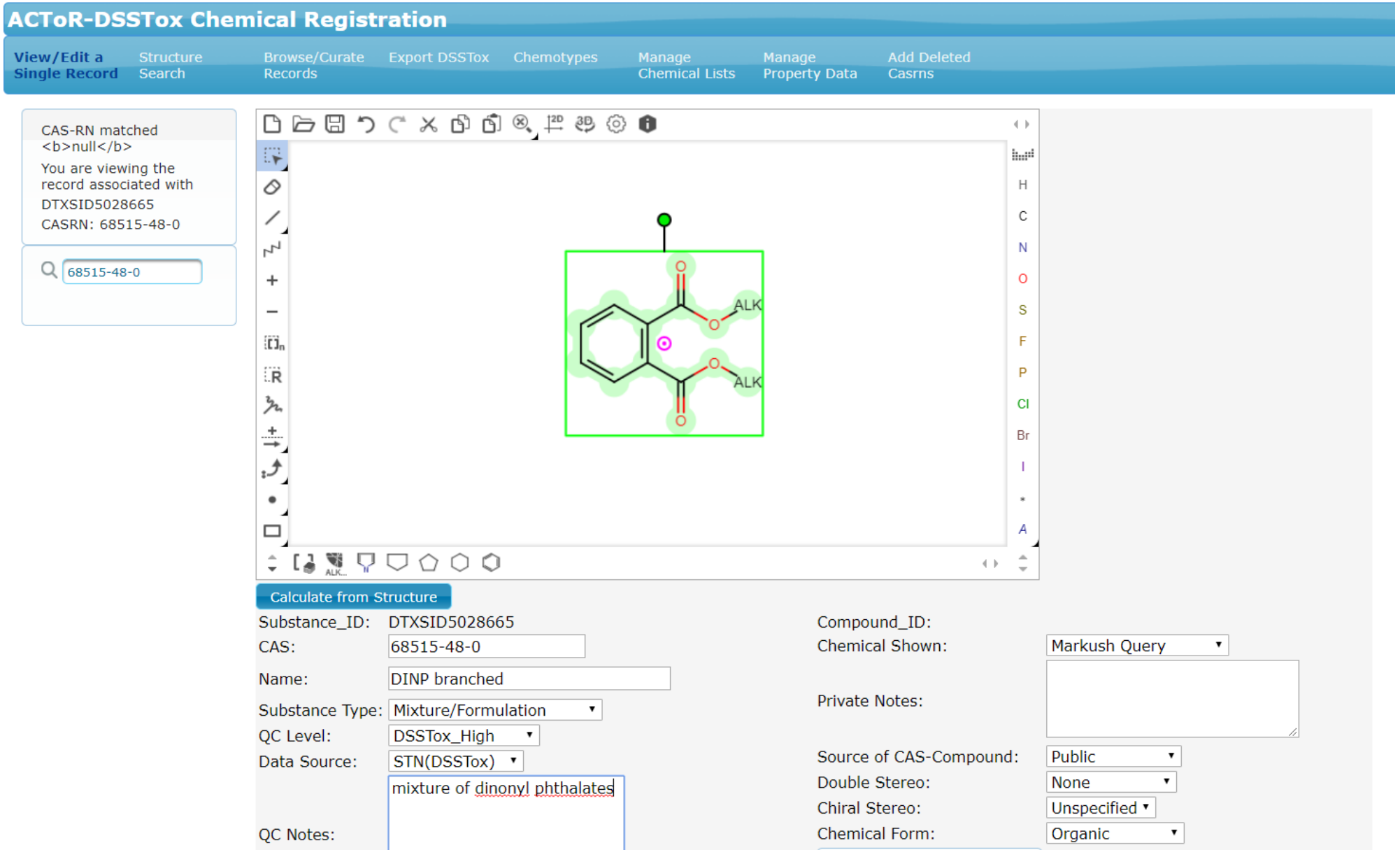
- First release of data files in 2004
- Focused on high impact sets of data
  - Carcinogenic Potency Database
  - Drinking water disinfection by-products
  - EPA's Integrated Risk Information System
  - FDA's Maximum Daily Dose dataset
  - EPA's Fat Head Minnow Toxicity dataset
  - etc...
- Managed all chemical registration for ToxCast and Tox21 chemicals
- By 2014, roughly 20K manually curated substance records



Generalized DSSTox Storage Architecture



Chemical Registration



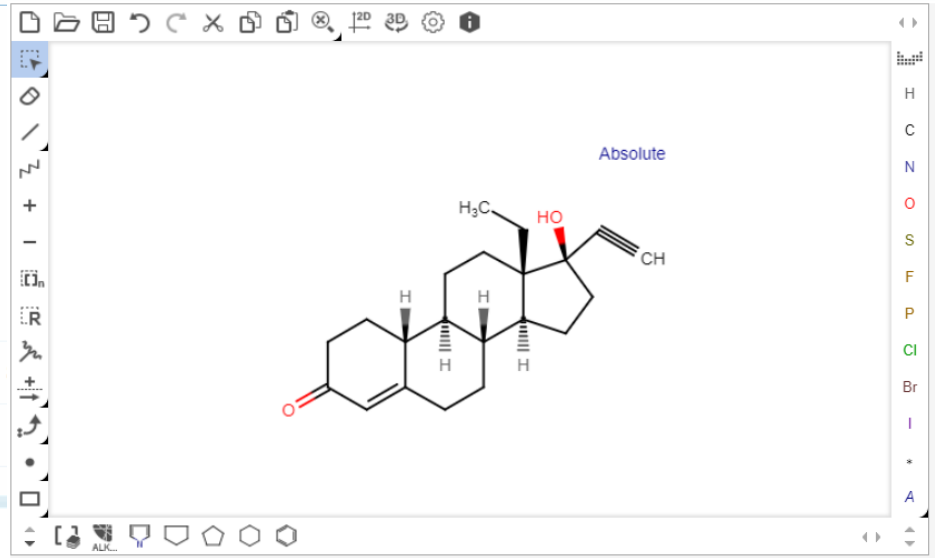
Chemical List Registration

Hits

	ssCAS-RN	ssName	Hit Desc	Hit Substance_ID	Hit Casrn	Hit Name
●		Norgestrel	Structure matched **SMILES**	DTXSID10859541	NOCAS_859541	13-Ethyl-17-ethynyl-17-hydroxy-1,2,6,7,8,9,10,11,12,13,14,15,16,1 tetradecahydro-3H-cyclopenta[a]phenanthren-3-one (non-preferred name)
●		Norgestrel	Mapped Identifier matched **NAME1**	DTXSID3036496	797-63-7	Levonorgestrel
●		Norgestrel	Mapped Identifier matched **NAME1**	DTXSID3047477	6533-00-2	dl-Norgestrel

Map hit | Cancel

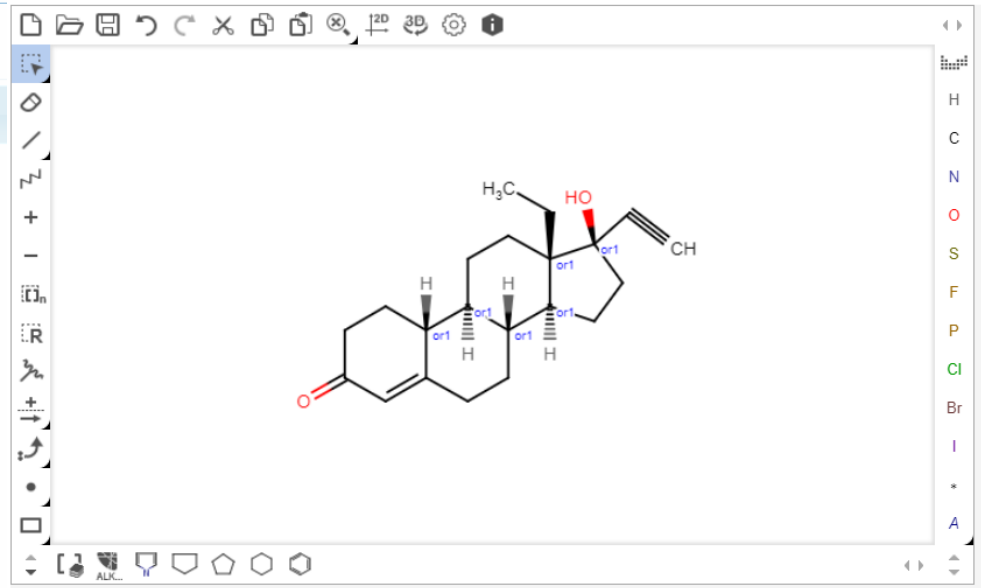
SMILES
Valid Synonym matched **NAME2**
Preferred Name matched **NAME1**
Valid Synonym matched other record: **NAME1**
Unique Synonym matched other record: **NAME2**
Unique Synonym matched other re
Mapped Identifier matched **NAME1**
Mapped Identifier matched **NAME2**
Name2Structure
Export All



Calculate from Structure

Substance_ID: DTXSID3036496
CAS: 797-63-7
Name: Levonorgestrel
Substance Type: Single Compound
QC Level: DSSTox_High
Data Source: STN(DSSTox)

Compound_ID:
Chemical Shown:
Private Notes:
Source of CAS-Compound:



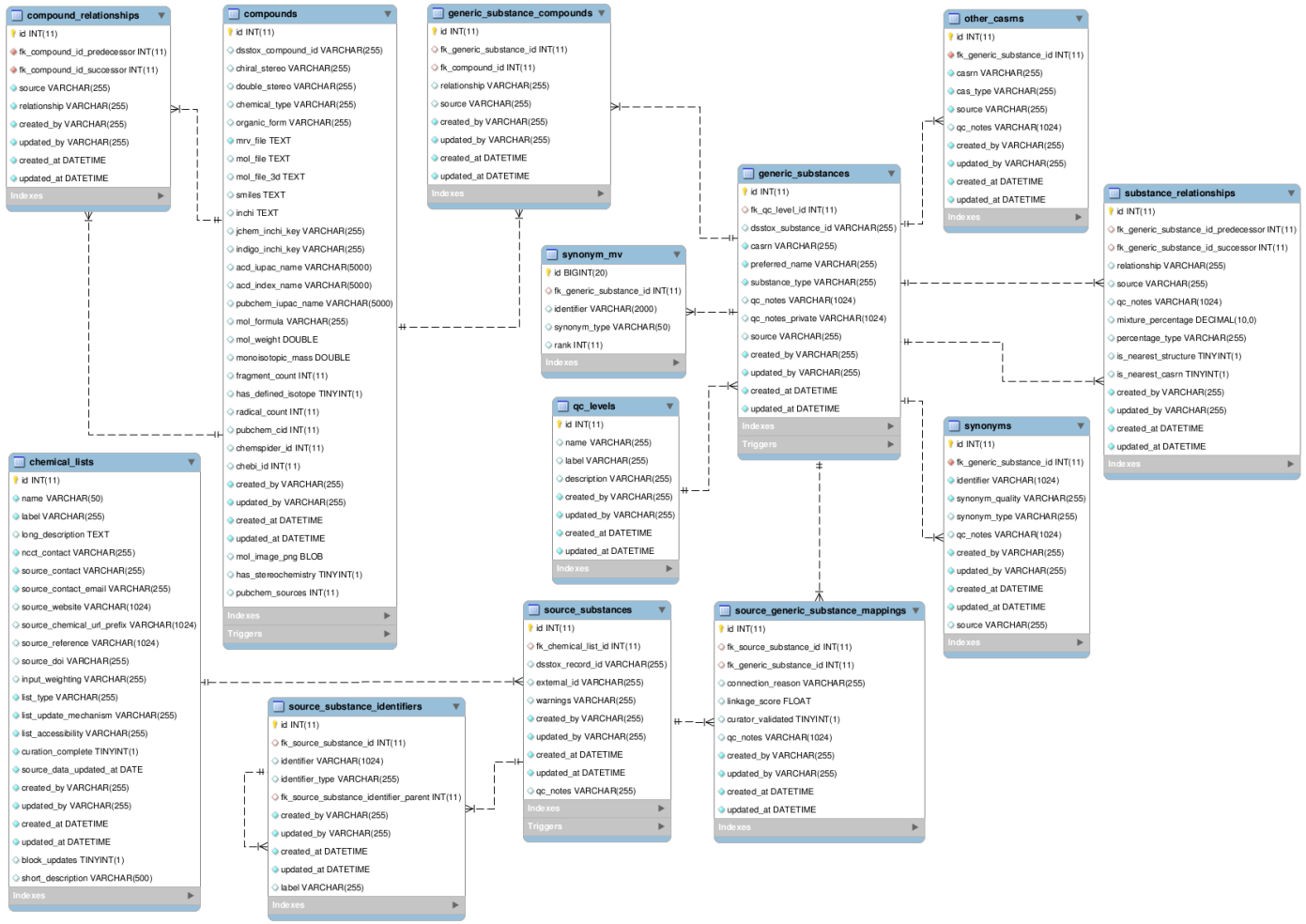
Calculate from Structure

Substance_ID: DTXSID3047477
CAS: 6533-00-2
Name: dl-Norgestrel
Substance Type: Mixture of Stereoisomers
QC Level: DSSTox_High

Compound_ID:
Chemical Shown:
Private Notes:

DSSTox Chemistry Schema

19 Tables



ChemTrack Bottles

ChemTrack | Inventory | Partnerships | Ordering | Data | Single Chemical Search | Hi Chris !

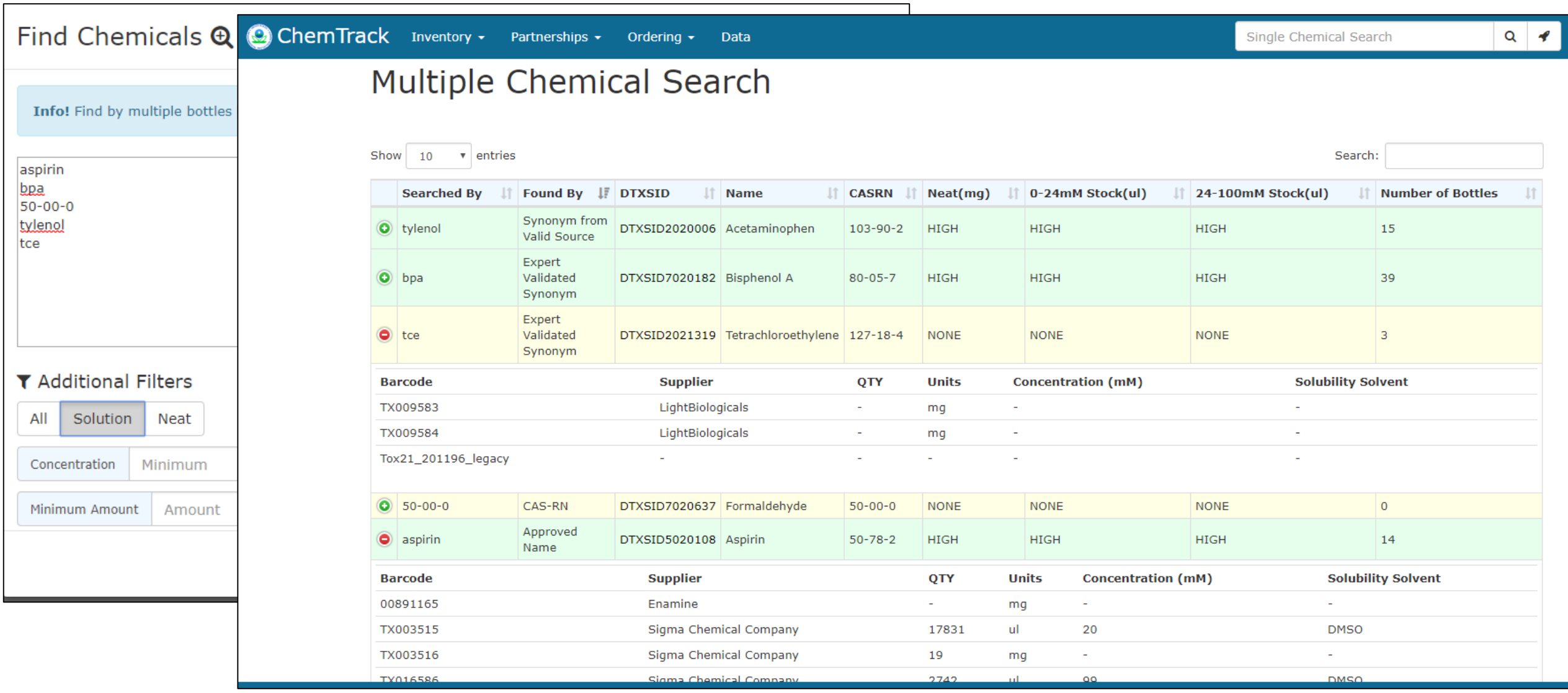
Uploaded MOSAIC Files 25 | Mapped & Available Bottles 29850 | Unmapped Bottles 449 | External Bottles 98 | Add New MOSAIC | Create External Bottle

Mapped & Available Bottles

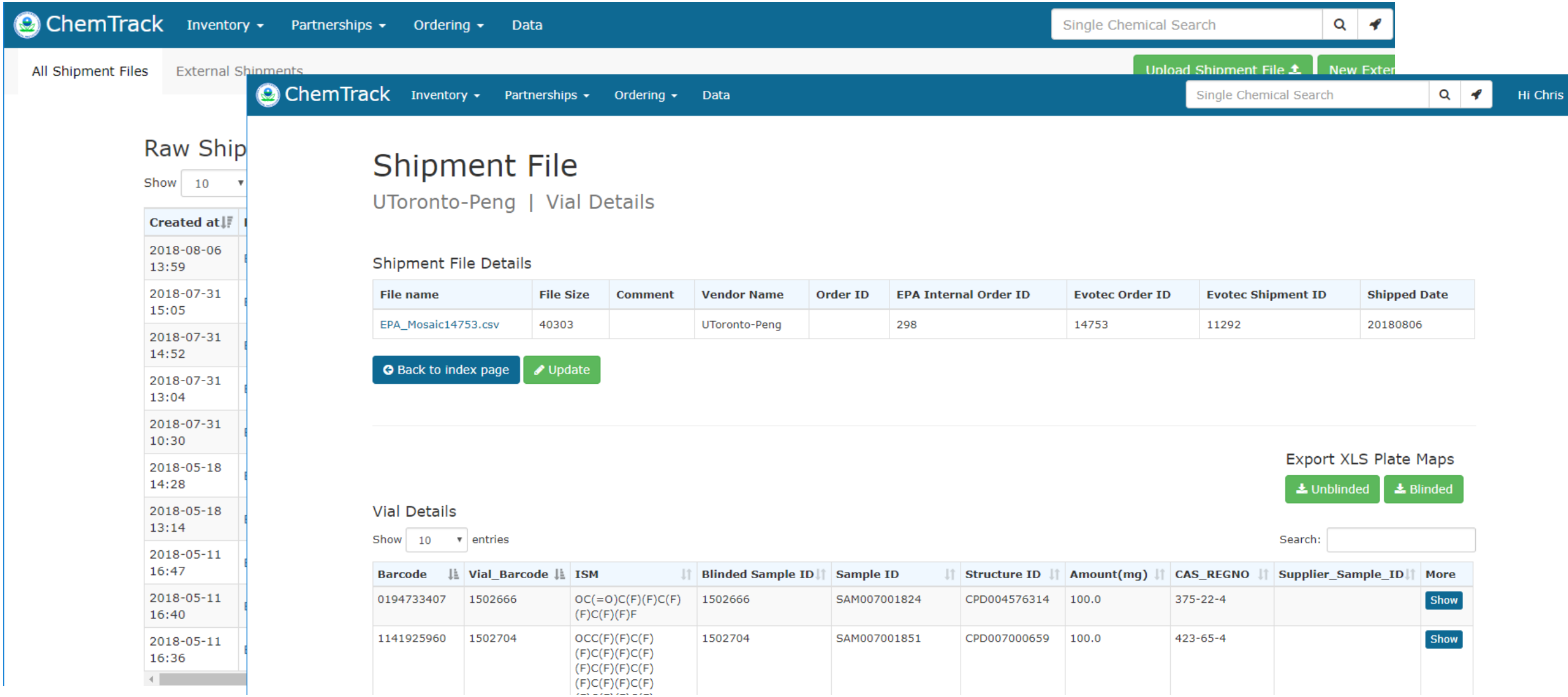
Show 10 entries | Search:

Barcode Type	Barcode	COA Summary	Compound Name	CAS	QTY Available	Units	Vendor	SAM	CPD	Can Plate?	Comment	
INVALID_SUPTX0013222_INVALID	ALID_SUPTX0013222_INVALID	2660	Aluminium phthalocyanine chloride	14154-42-8	1000	mg	Sigma Chemical Company	SAM004888816	CPD003650672	Yes		Edit
INVALID_SUPTX0013294_INVALID	ALID_SUPTX0013294_INVALID	2735	Nickle(III) carbonate basic hydrate	12607-70-4	250000	mg	Sigma Chemical Company	SAM004888887	CPD003650704	Yes		Edit
EPA_Vial_Source_Storage	BF00079587	18768	Pentabromophenol	608719	178	mg	Sigma Chemical Company	SAM006061824	CPD001224527	Yes		Edit
EPA_Vial_Source_Storage	BF00079581	18774	Phthalic anhydride	85-44-9	193	mg	Sigma Chemical Company	SAM006061820	CPD001252223	Yes		Edit
EPA_Vial_Source_Storage	BF00079583	18772	TRIPOLI	7631-86-9	196	mg	Sigma Chemical Company	SAM006061823	CPD001252283	Yes		Edit
EPA_Vial_Source_Storage	BF00079582	18773	4,4'-Methylenebis(2-chloroaniline)	101144	188	mg	Sigma Chemical Company	SAM006061819	CPD001307314	Yes		Edit
EPA_Vial_Source_Storage	BF00079585	18770	4-sec-Butyl-2,6-di-tert-butylphenol	17540-75-9	188	mg	Sigma Chemical Company	SAM006061825	CPD004560495	Yes		Edit
EPA_Vial_Source_Storage	BF00079584	18771	Creosote	8001-58-9	178	mg	Sigma Chemical Company	SAM006061821	CPD004757028	Yes		Edit
EPA_Vial_Source_Storage	BF00079586	18769	tert-Amyl methyl ether	994-05-8	194	mg	Sigma Chemical Company	SAM006061822	CPD004757029	Yes		Edit
EPA_Vial_Source_Storage	BF00079580	18775	3,3'-	91-94-	188	mg	Sigma Chemical	SAM006061826	CPD004757030	Yes		Edit

ChemTrack Search

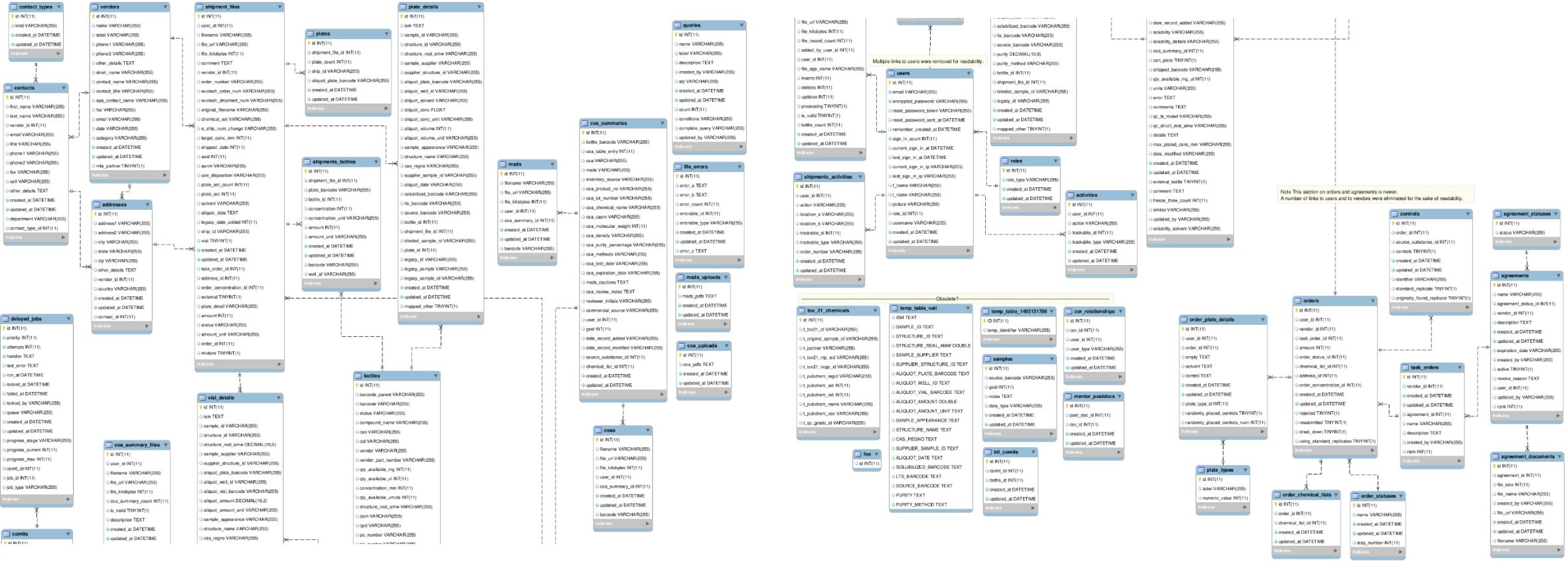


ChemTrack Shipments

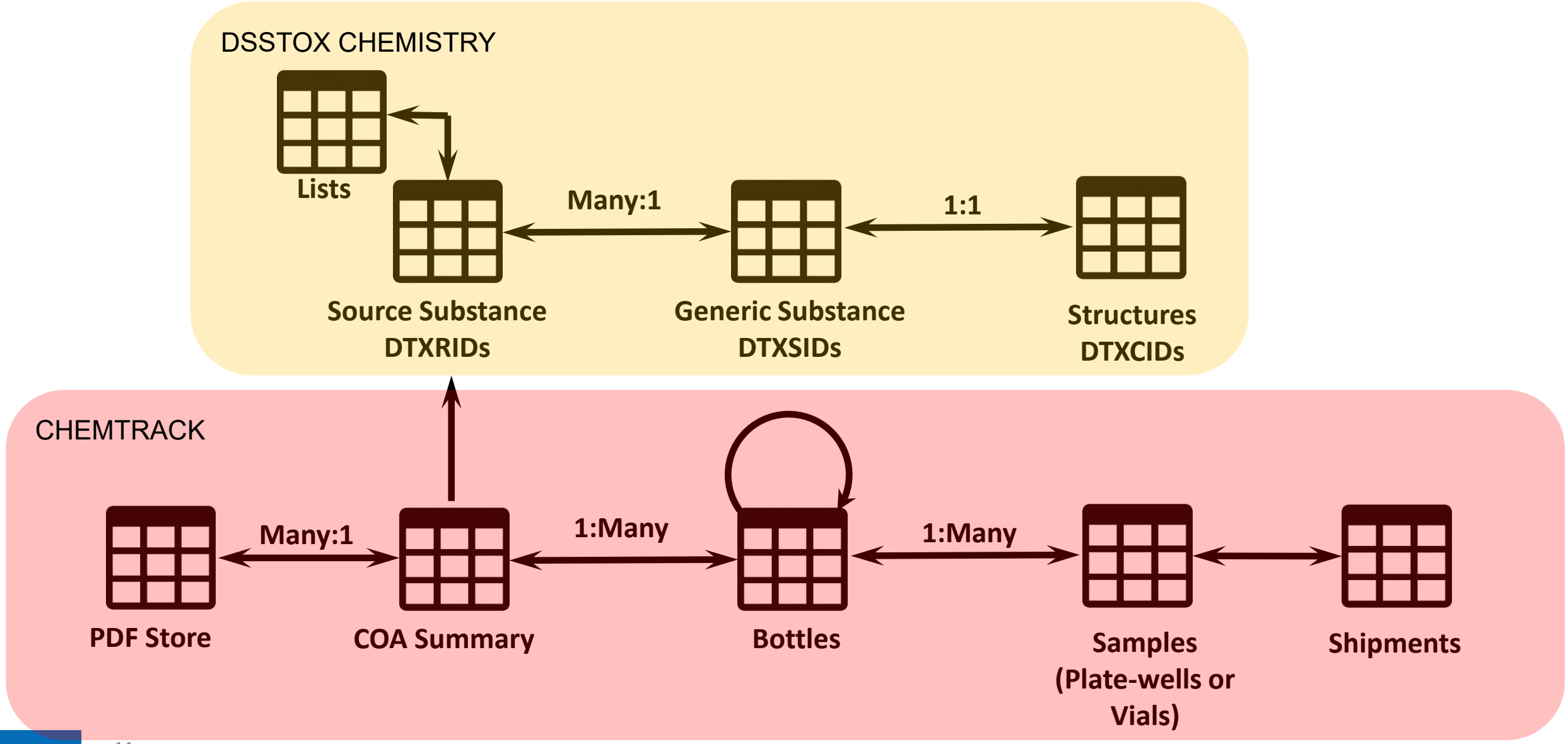


ChemTrack Schema

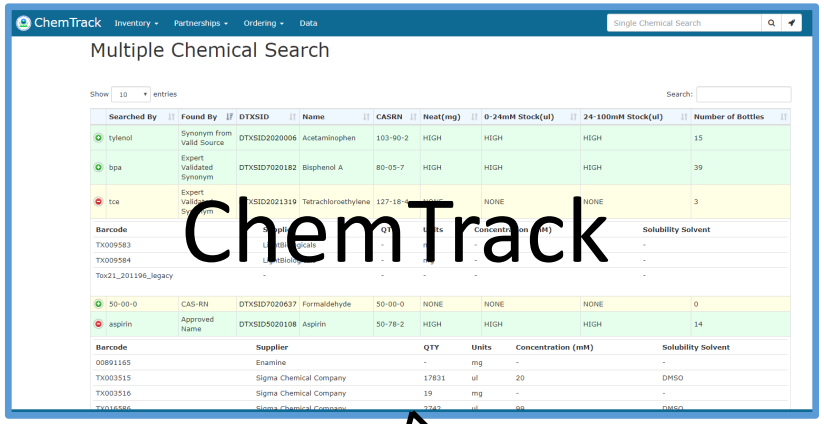
57 Tables

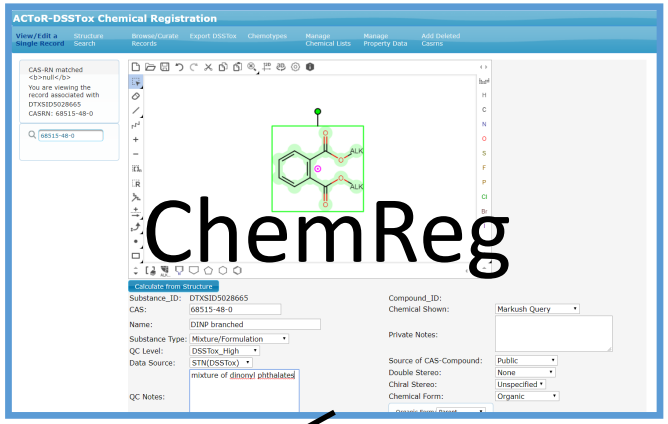


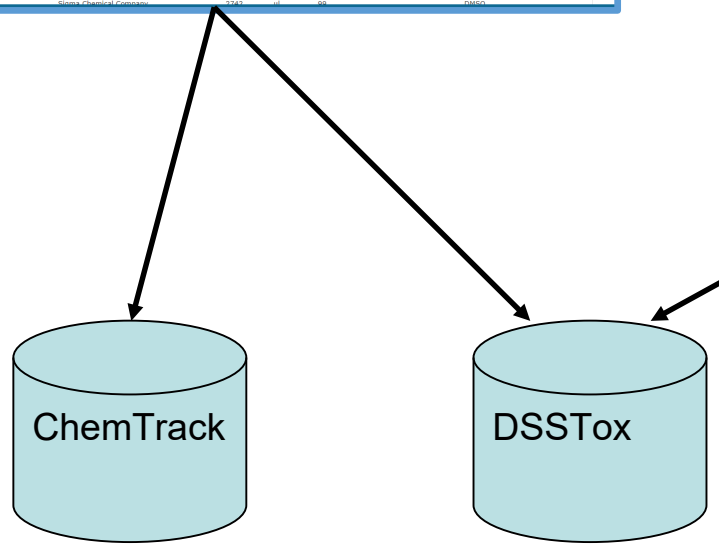
Generalized Data Model



Sharing the DSSTox Database



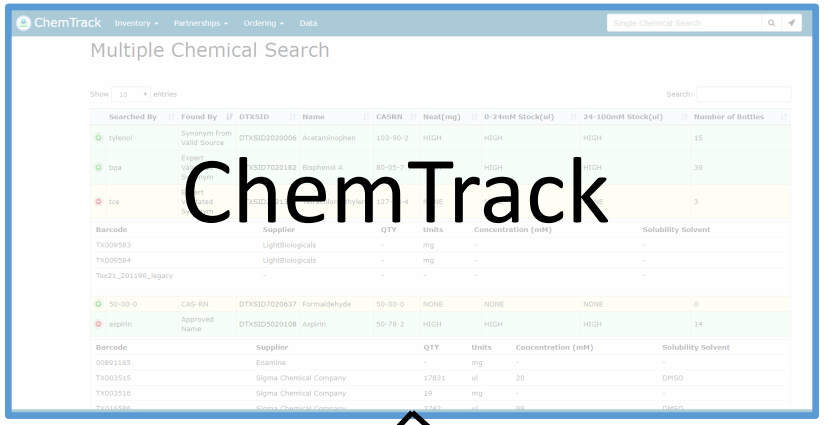




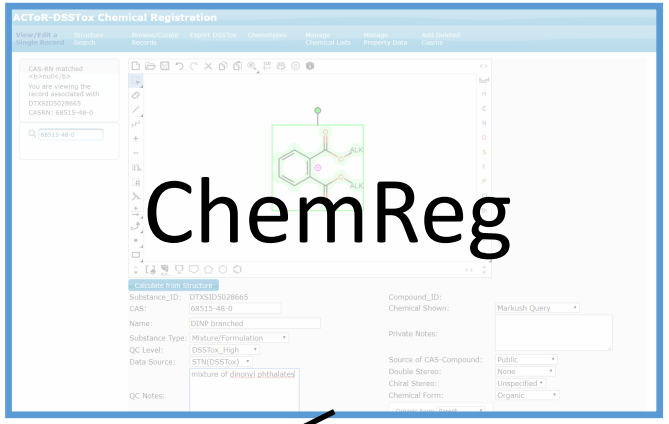
“ChemSpiderMan” comes to NCCT (2015)

- DSSTox: ~700K substance records (25K manually curated)

Sharing the DSSTox Database



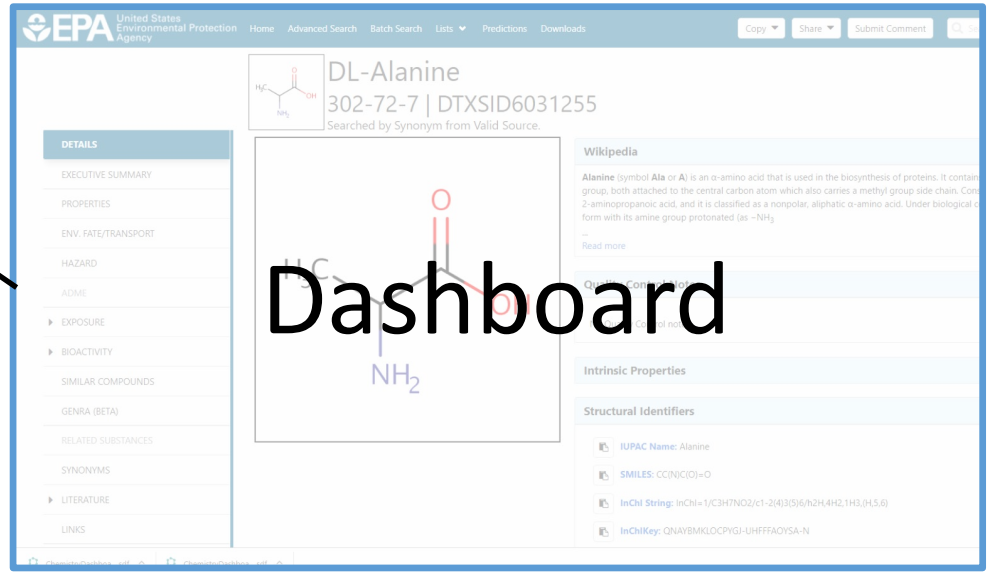
ChemTrack



ChemReg

ChemTrack

DSSTox



Dashboard

The Dashboard Taskmaster

I Need
QSAR Predictions

I Need
ToxCast Results

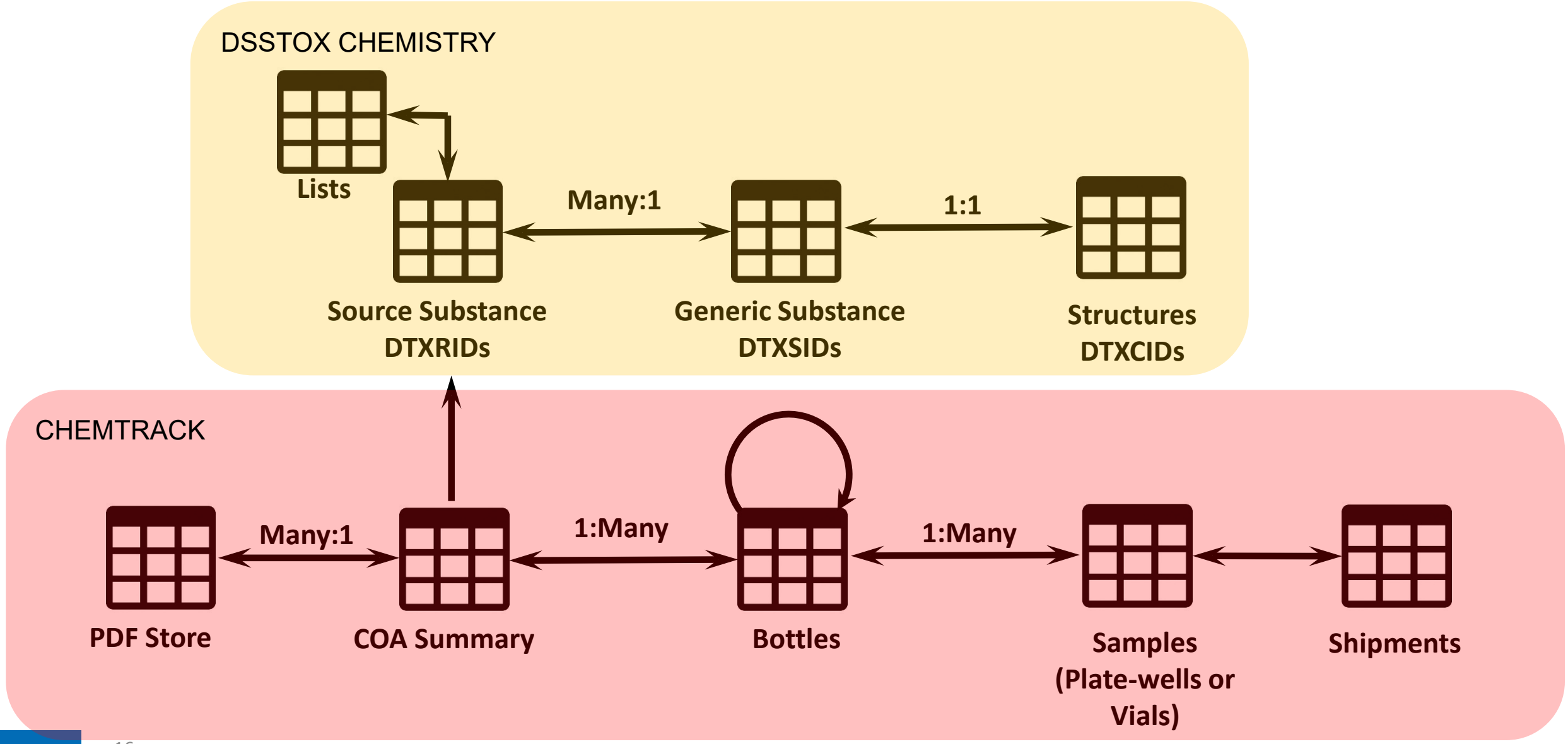
I Need
Analytical QC Data

I Need
PhysChem Data

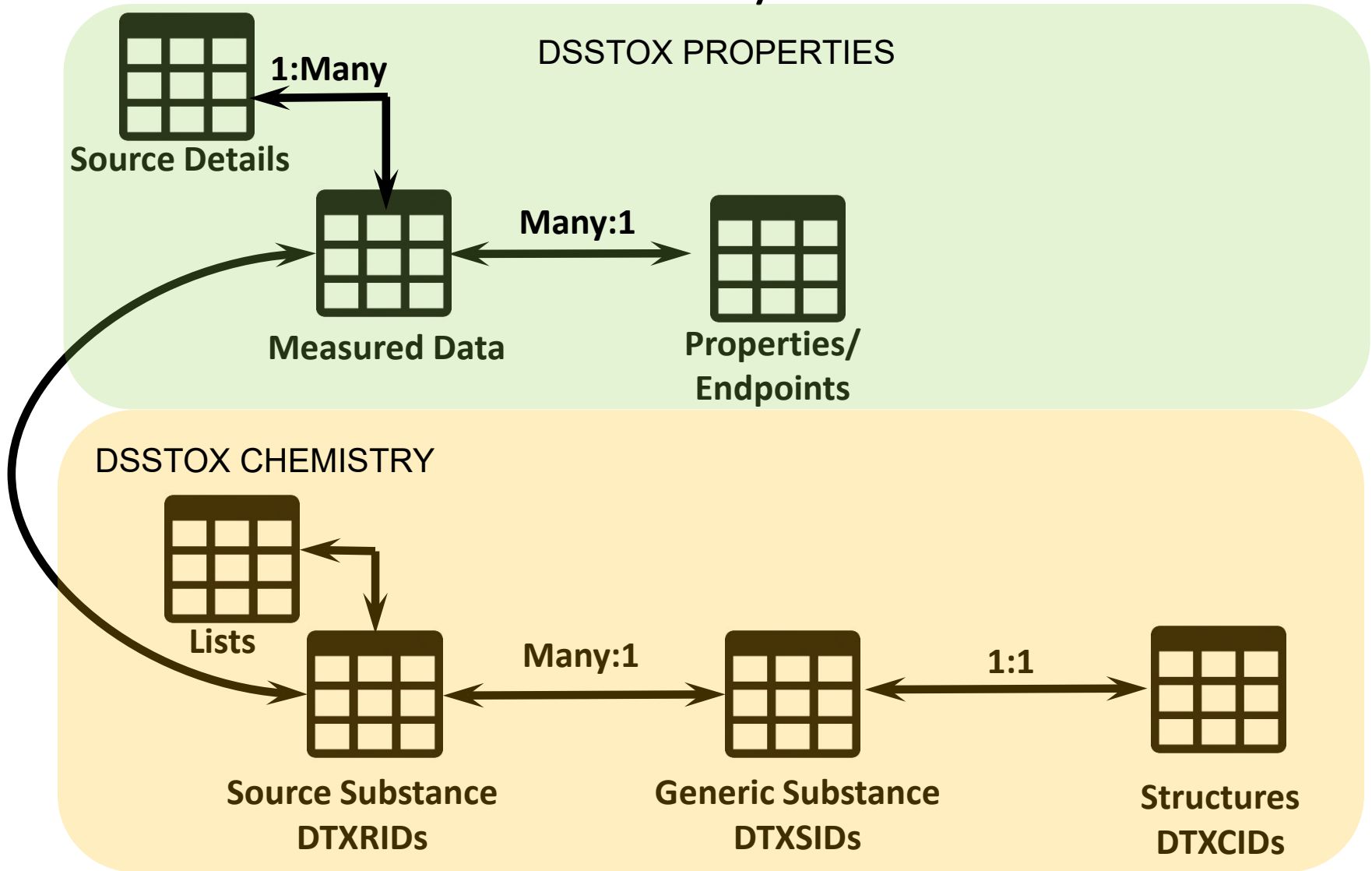
I Need
In Vivo Toxicity
Data

I Need
Consumer Product
Information

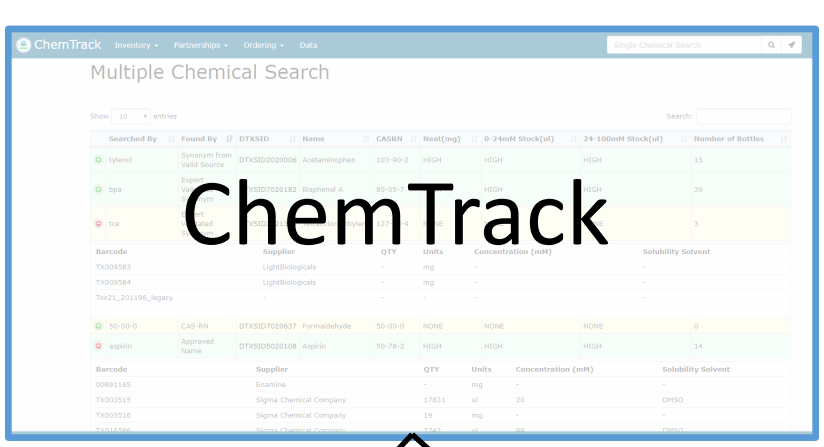
I Need PhysChem Data…

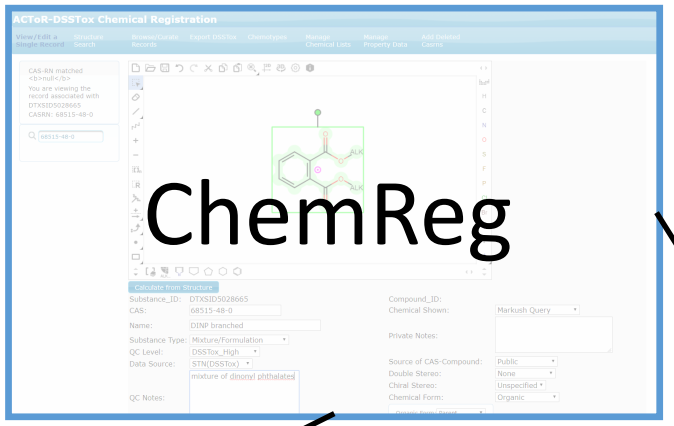


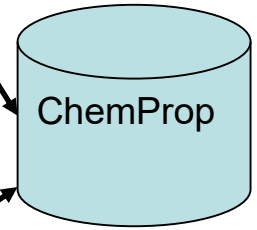
I Need PhysChem Data…

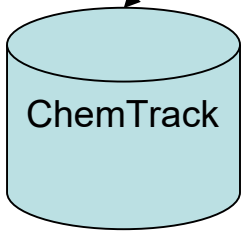


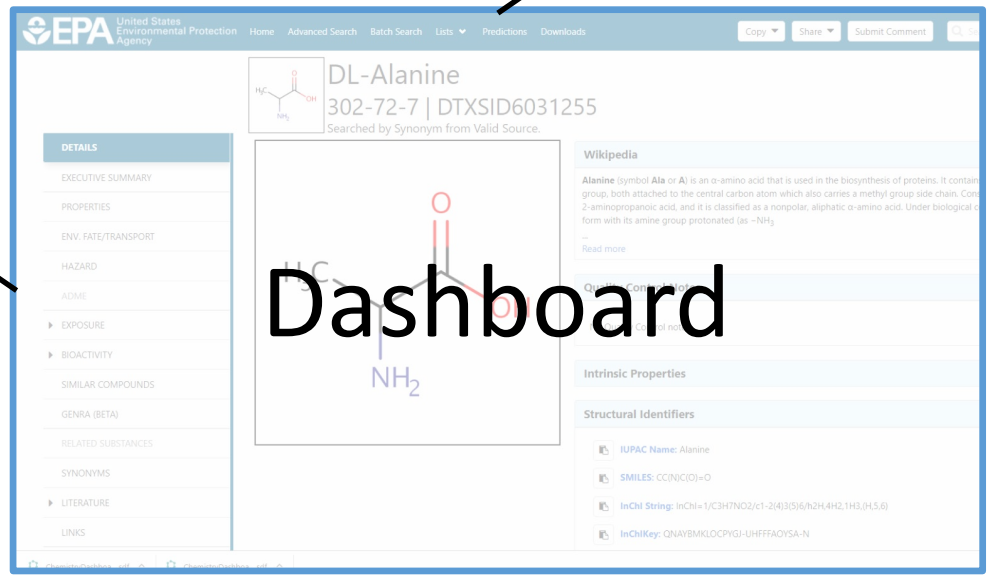
Sharing the DSSTox Database



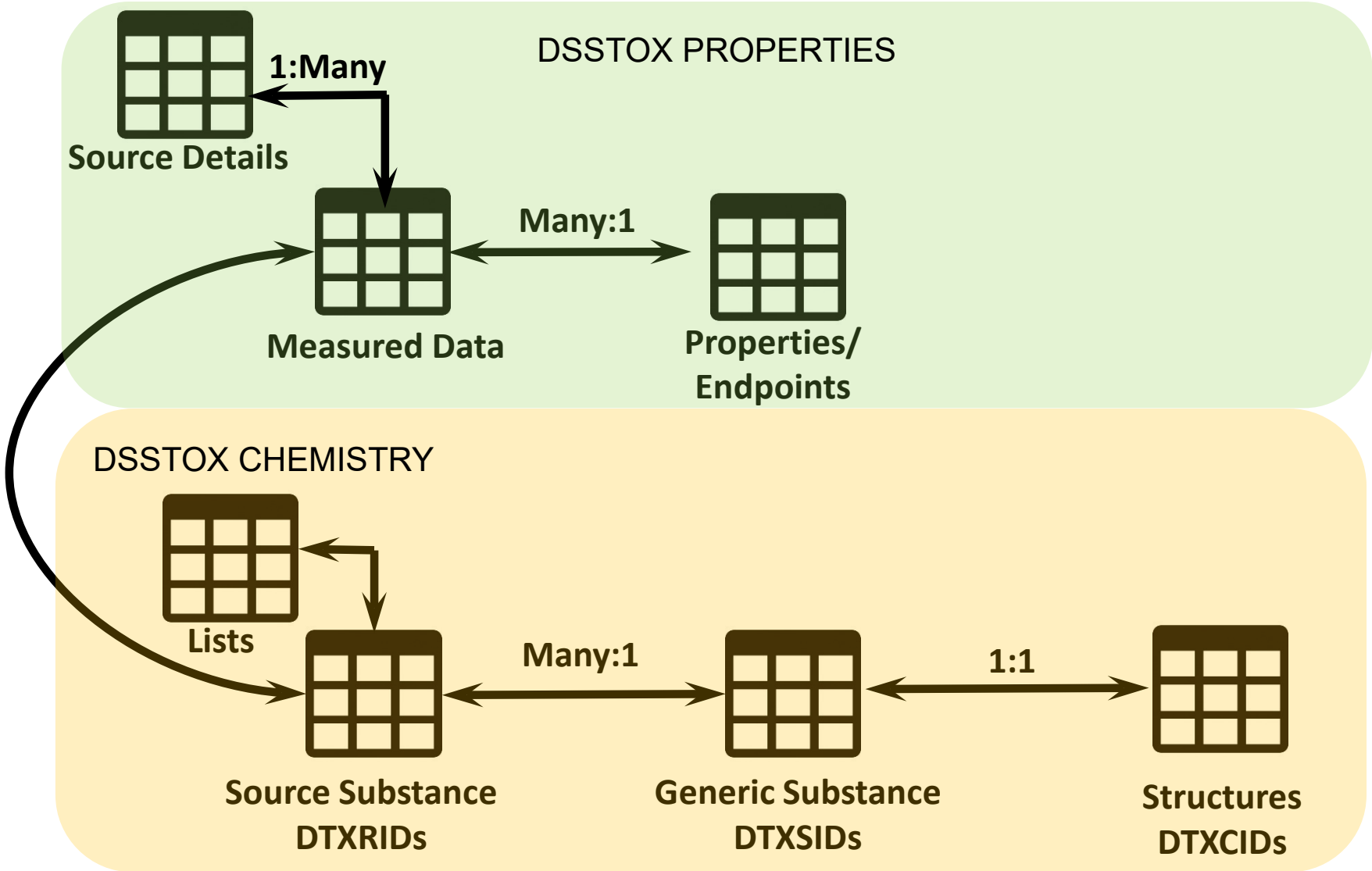




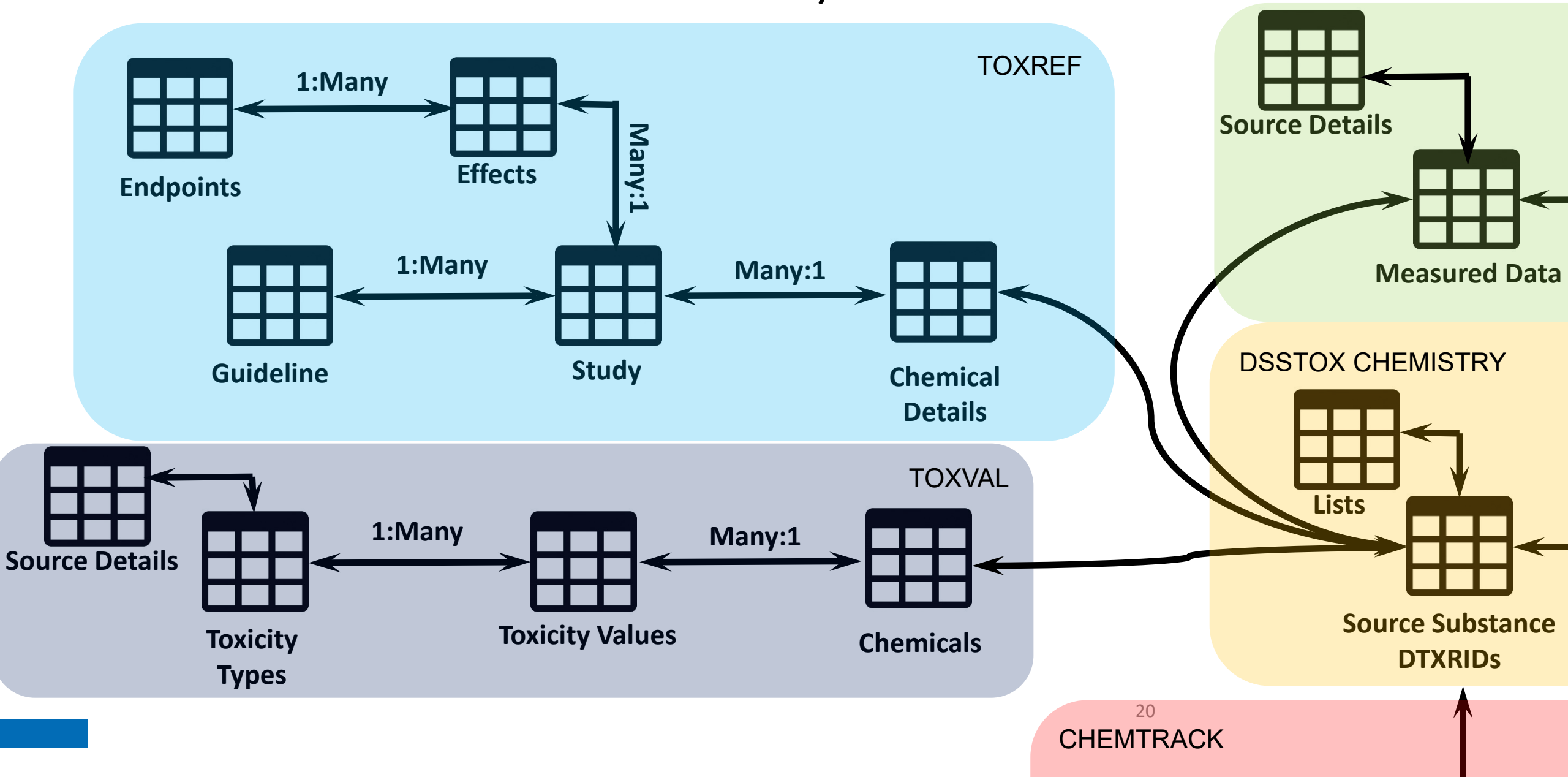




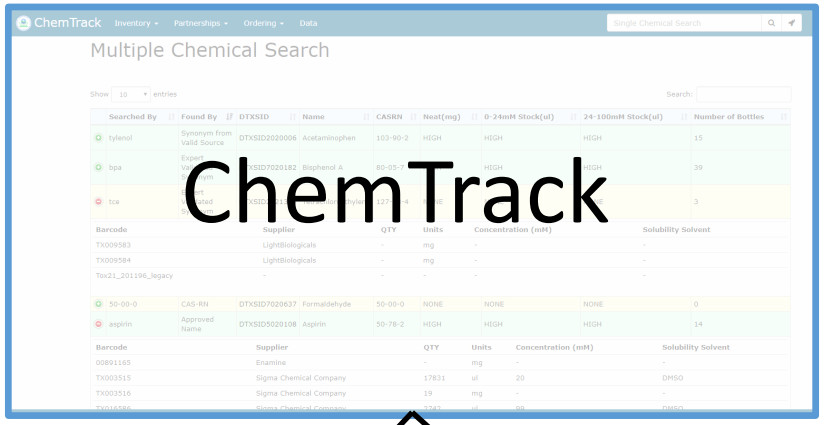
I Need *In Vivo* Toxicity Data…



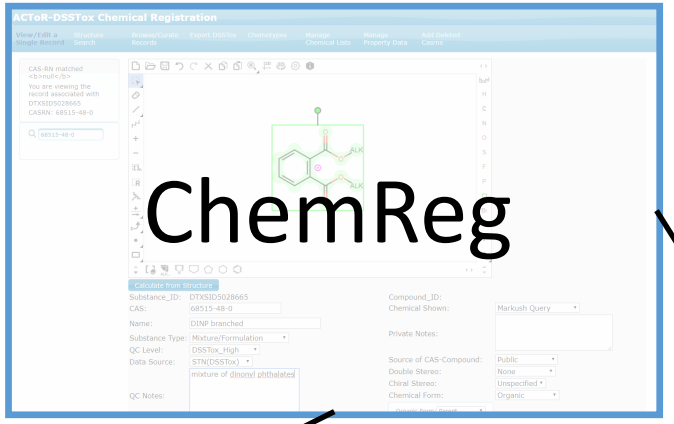
I Need *In Vivo* Toxicity Data…



Sharing the DSSTox Database



ChemTrack



ChemReg

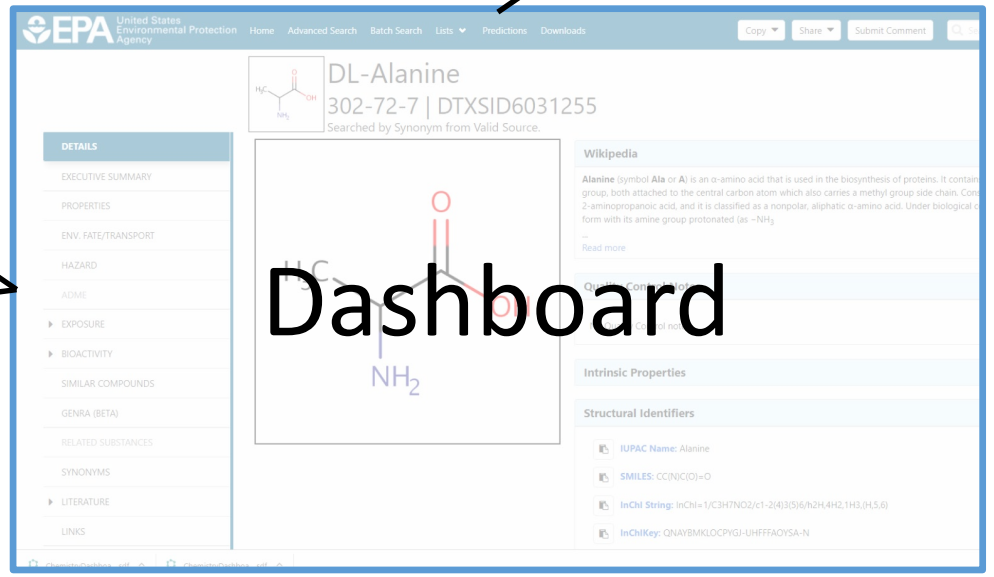
ChemProp

ChemTrack

DSSTox

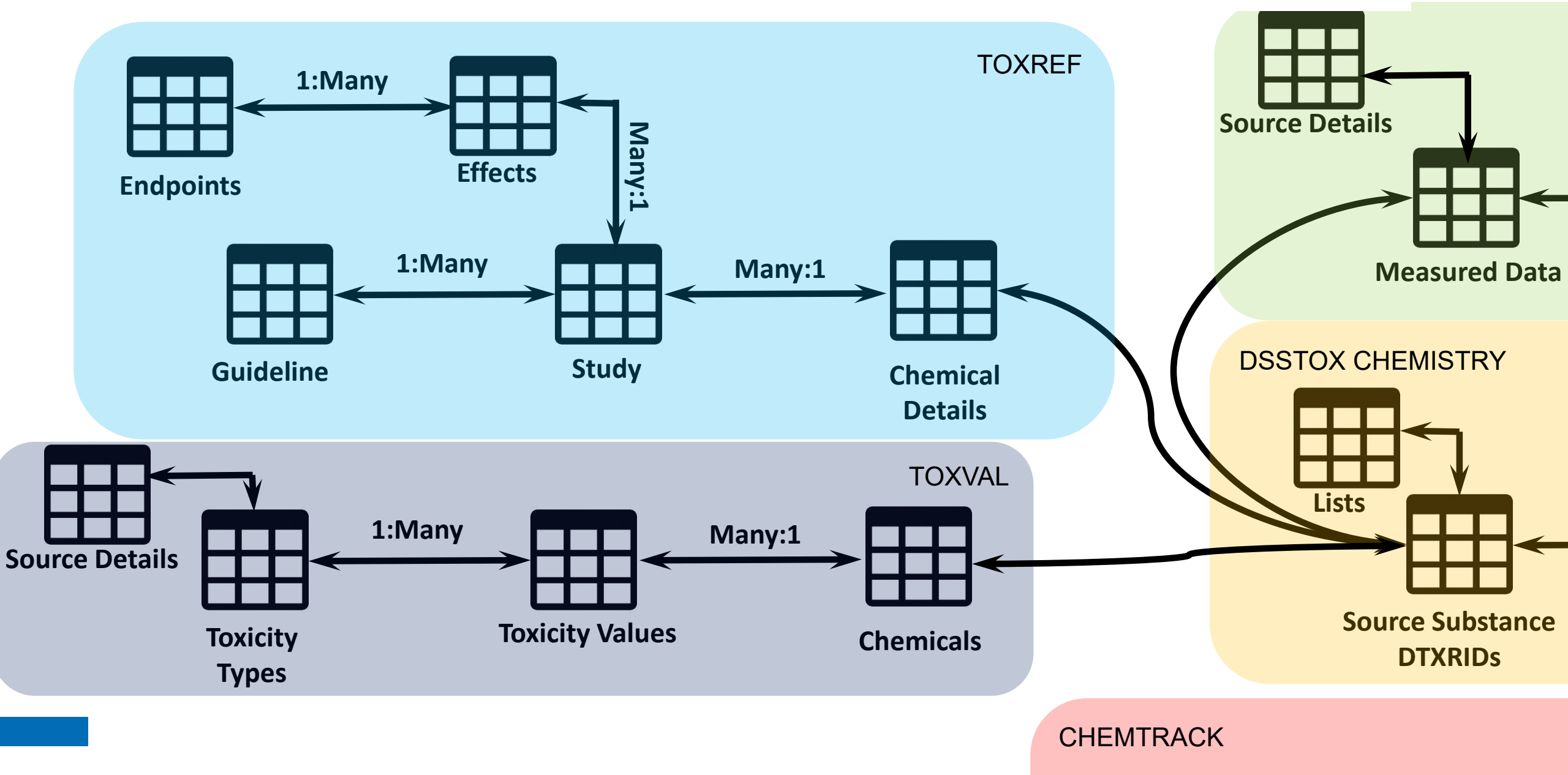
ToxVal Load Scripts

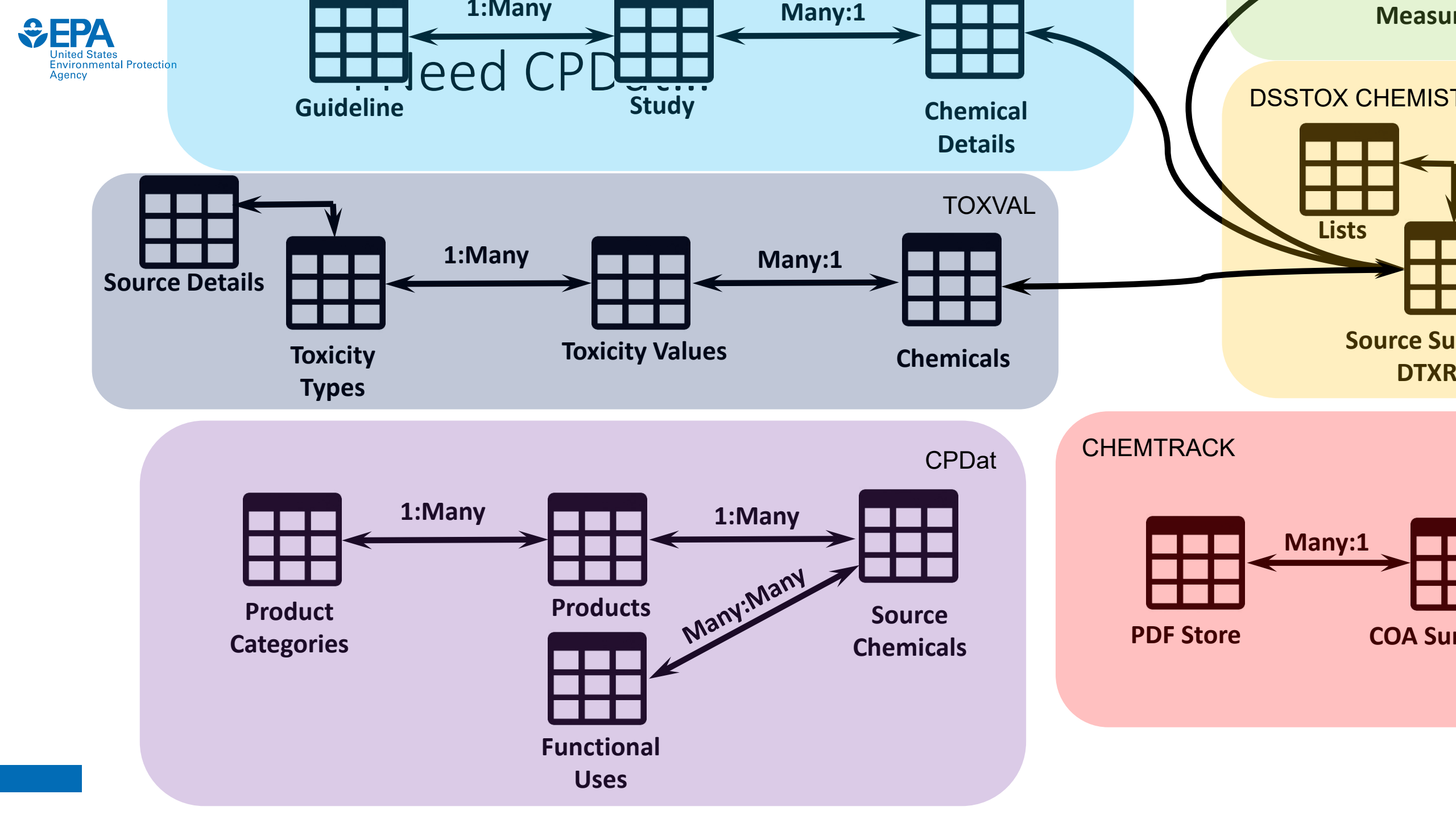
ToxVal



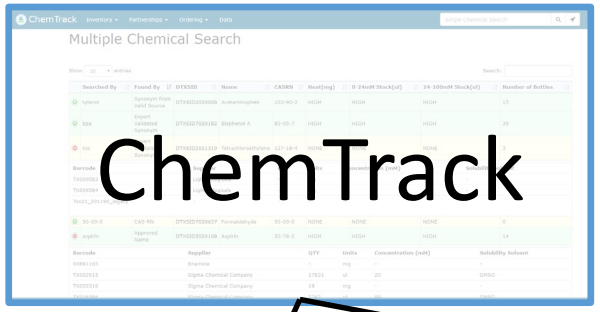
Dashboard

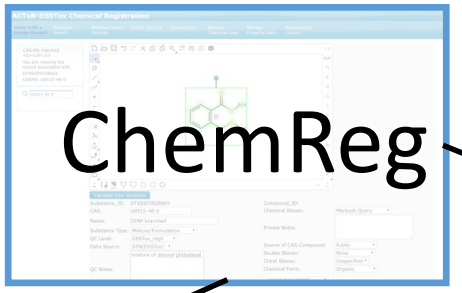
I Need CPDat…


TOXREF
1:Many
Endpoints
Effects
Many:1
1:Many
Guideline
Study
Many:1
Chemical Details
Source Details
Measured Data
DSSTOX CHEMISTRY
Lists
Source Substance DTXRIDs
TOXVAL
Source Details
1:Many
Toxicity Types
Toxicity Values
Many:1
Chemicals
CHEMTRACK


EPA
United States
Environmental Protection
Agency
1:Many
Many:1
Guideline
Study
Chemical
Details
DSSTOX CHEMIST
Lists
Source Su
DTXR
TOXVAL
Source Details
Toxicity
Types
1:Many
Toxicity Values
Many:1
Chemicals
CPDat
1:Many
1:Many
Product
Categories
Products
Source
Chemicals
Many:Many
Functional
Uses
CHEMTRACK
Many:1
PDF Store
COA Su

EPA United States Environmental Protection Agency





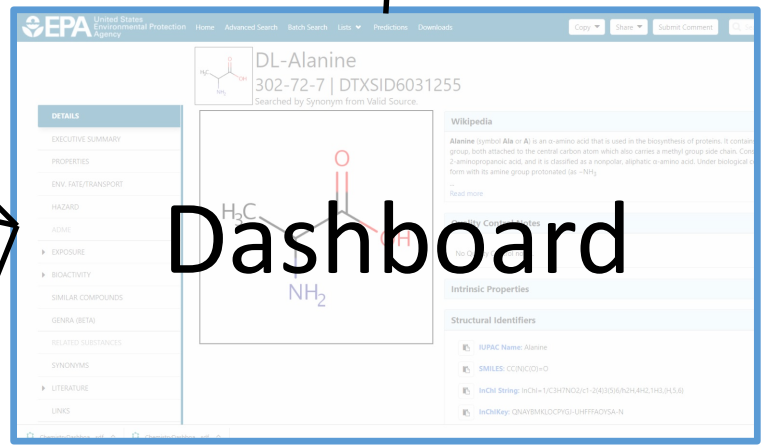
ChemProp

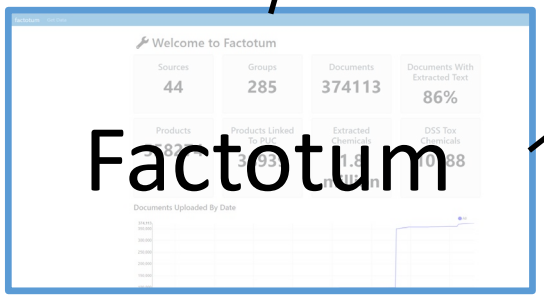
DSSTox

ChemTrack

ToxVal Load Scripts

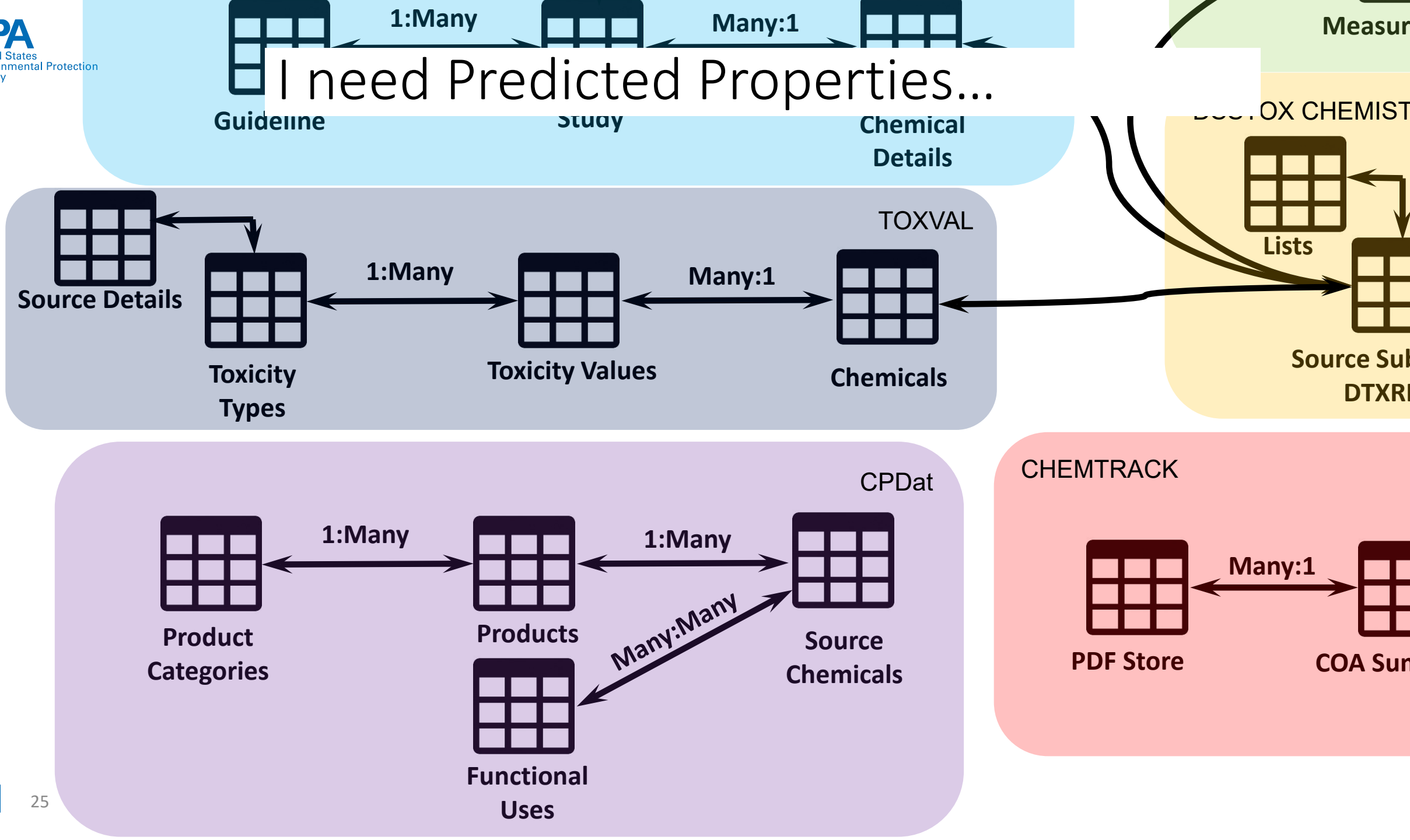
ToxVal



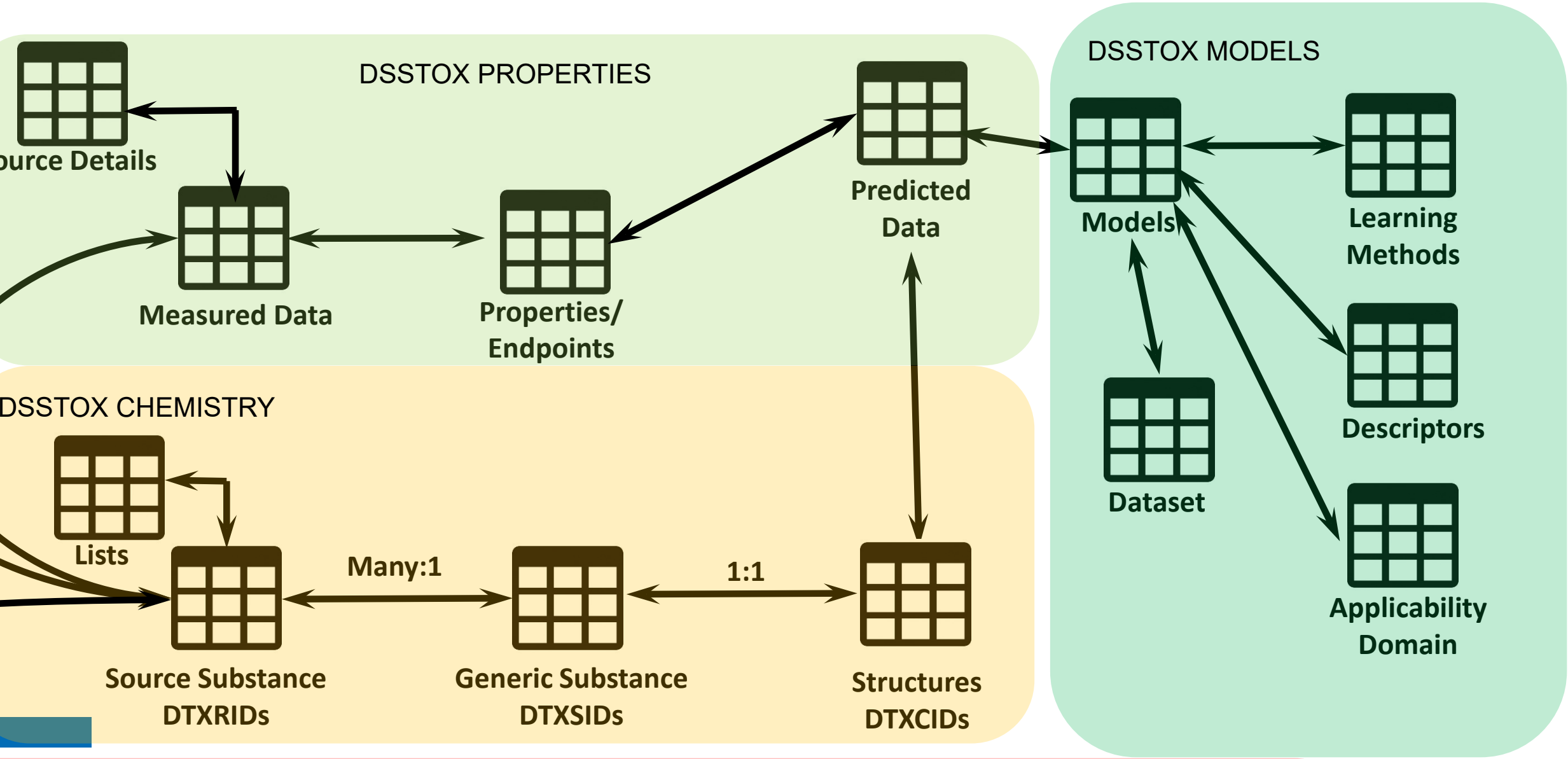


CPDat

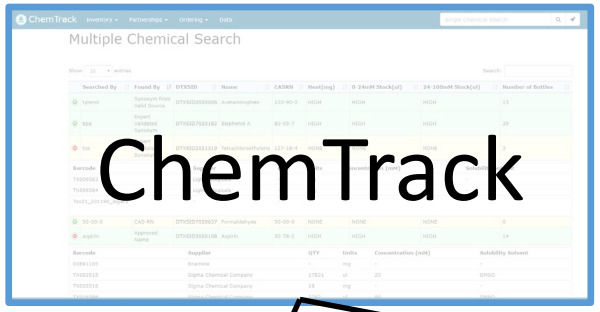
I need Predicted Properties…

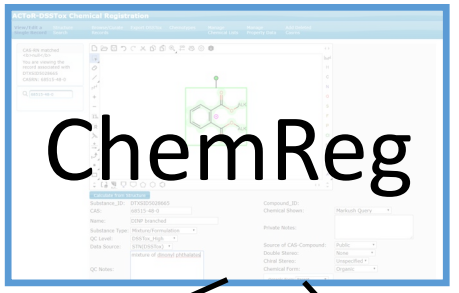


I need Predicted Properties…



EPA United States Environmental Protection Agency





Modeling Scripts

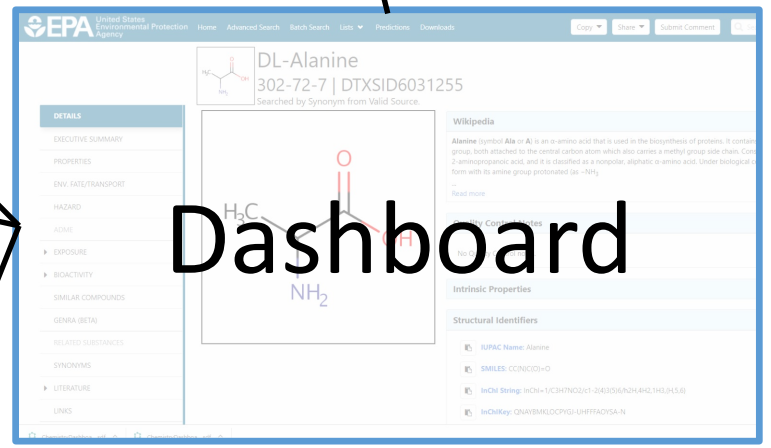
DSSTox

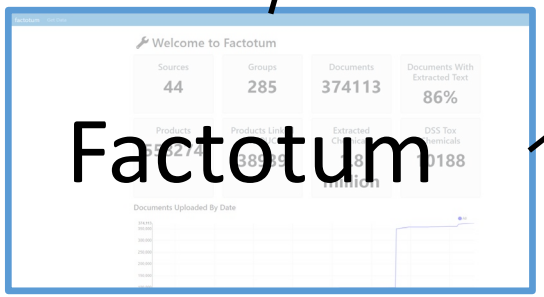
ChemTrack

ChemProp

ToxVal Load Scripts

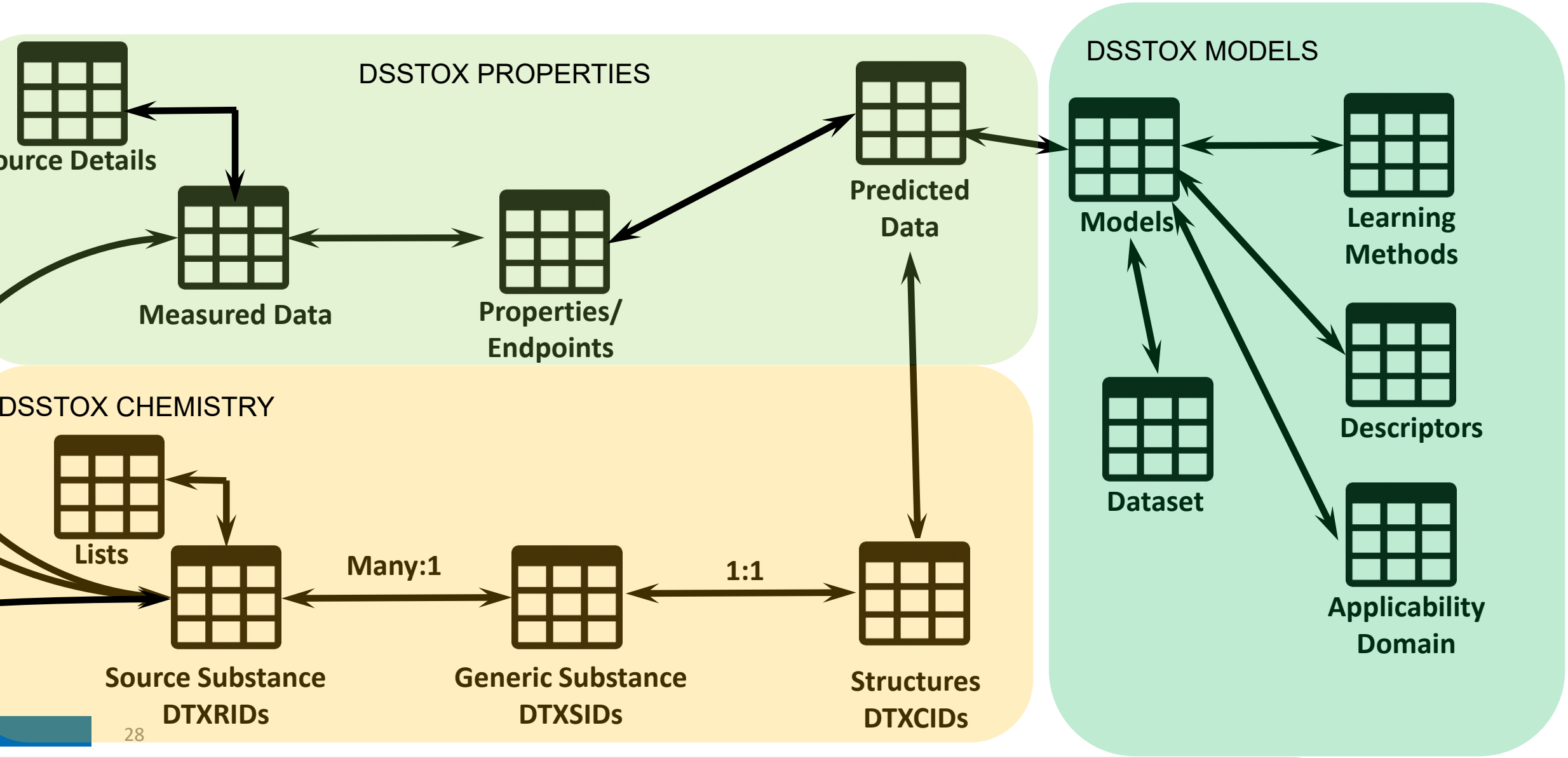
ToxVal



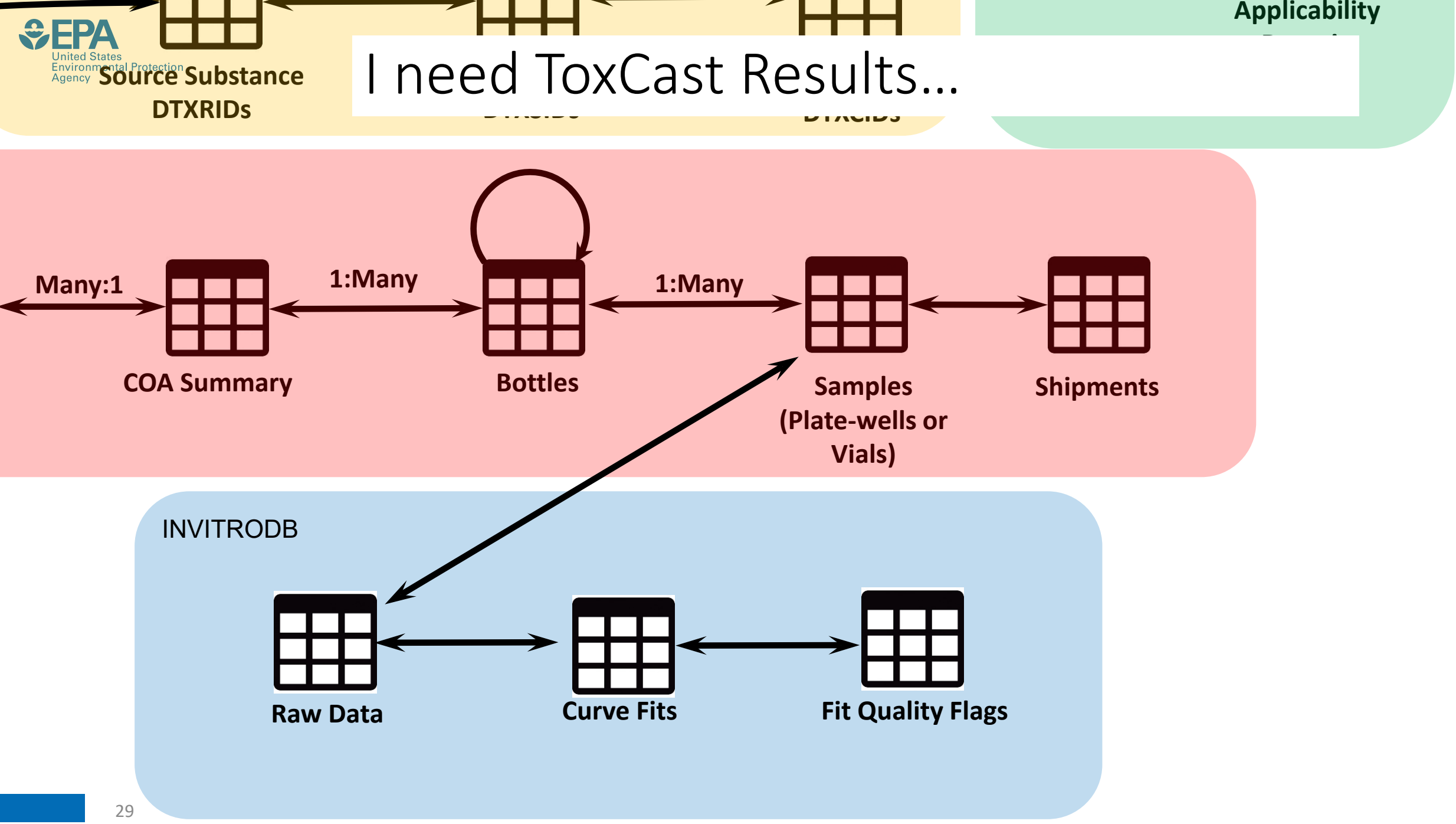


CPDat

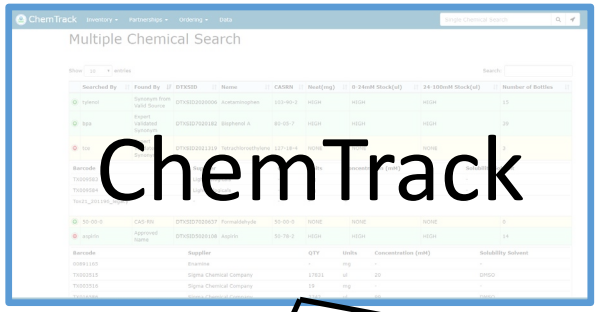
I need ToxCast Results…

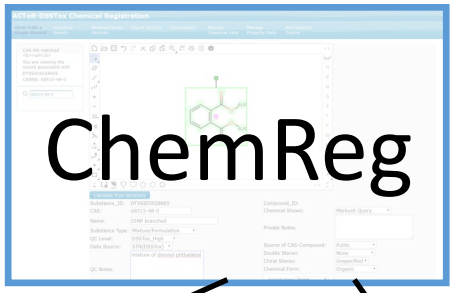


I need ToxCast Results…



EPA United States Environmental Protection Agency





Modeling Scripts

DSSTox

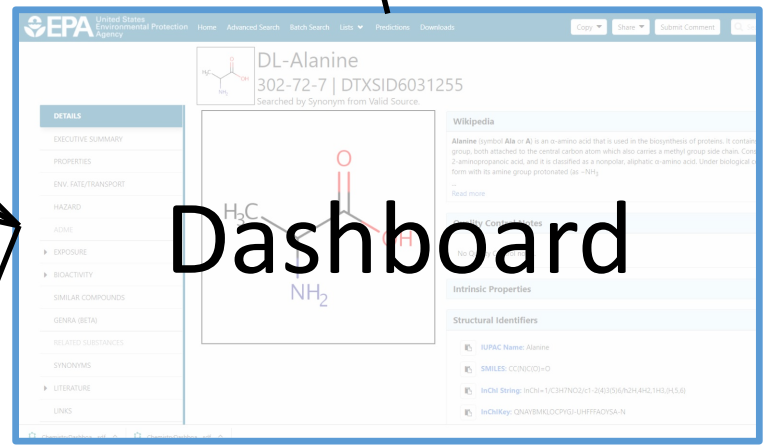
ChemTrack

ChemProp

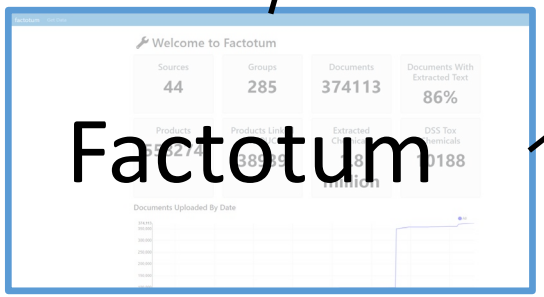
ToxVal Load Scripts

ToxVal

InVitroDB

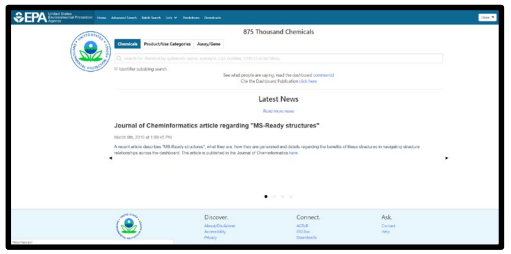


TCPL



CPDat

Proposed EPA-ORD Data Structure



Production Application Layer

Virtual or Physical Application Databases – Data Structured to efficiently support our external production applications

Presentation Layer
(Fit-for-Purpose)

ORD Data Lake

Data Analytics/ Learning

Generalized data model providing integration across the domains for scientific exploration

Integration Layer

ChemTrack App
ChemTrack

ChemReg Application
DSSTox
ChemProp

TCPL R Package
InVitroDB

Toxicity QC App
ToxRef DB

Facto
CPDat

Model + Pred App
Models

Transactional Layer

Raw data processing and data curation, quality tools

EPA United States Environmental Protection Agency

Conclusions

- The Comptox Chemicals Dashboard is great
- The desire to make our data available resulted in a web of tight coupling
- This tight coupling inhibits improvement and flexibility
- Separation of concerns in different data layers adds complexity and overhead but removes restrictions
- APIs provide greater flexibility for maintaining backward compatibility than direct database querying

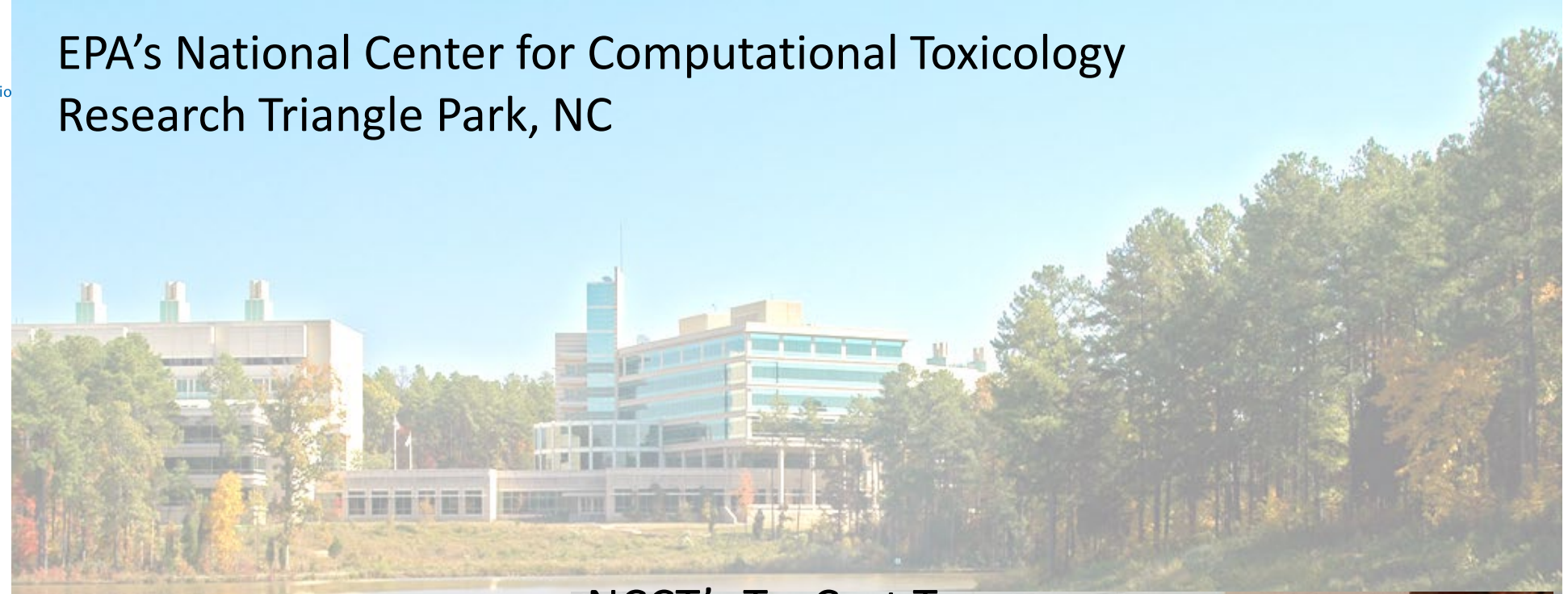
Acknowledgements

Software Development
Jeff Edwards
Jeremy Dunne

Modeling Software
Kamel Mansouri
Todd Martin

DSSTox
Inthirany Thillainadarajah
Sakuntala Sivasupramaniam
Ann Richard

EPA's National Center for Computational Toxicology
Research Triangle Park, NC



NCCT's ToxCast Team

EPA United States Environmental Protection Agency

Questions?