Expansion of DSSTox: Leveraging public data to create a semantic cheminformatics resource with quality annotations for support of U.S. EPA applications

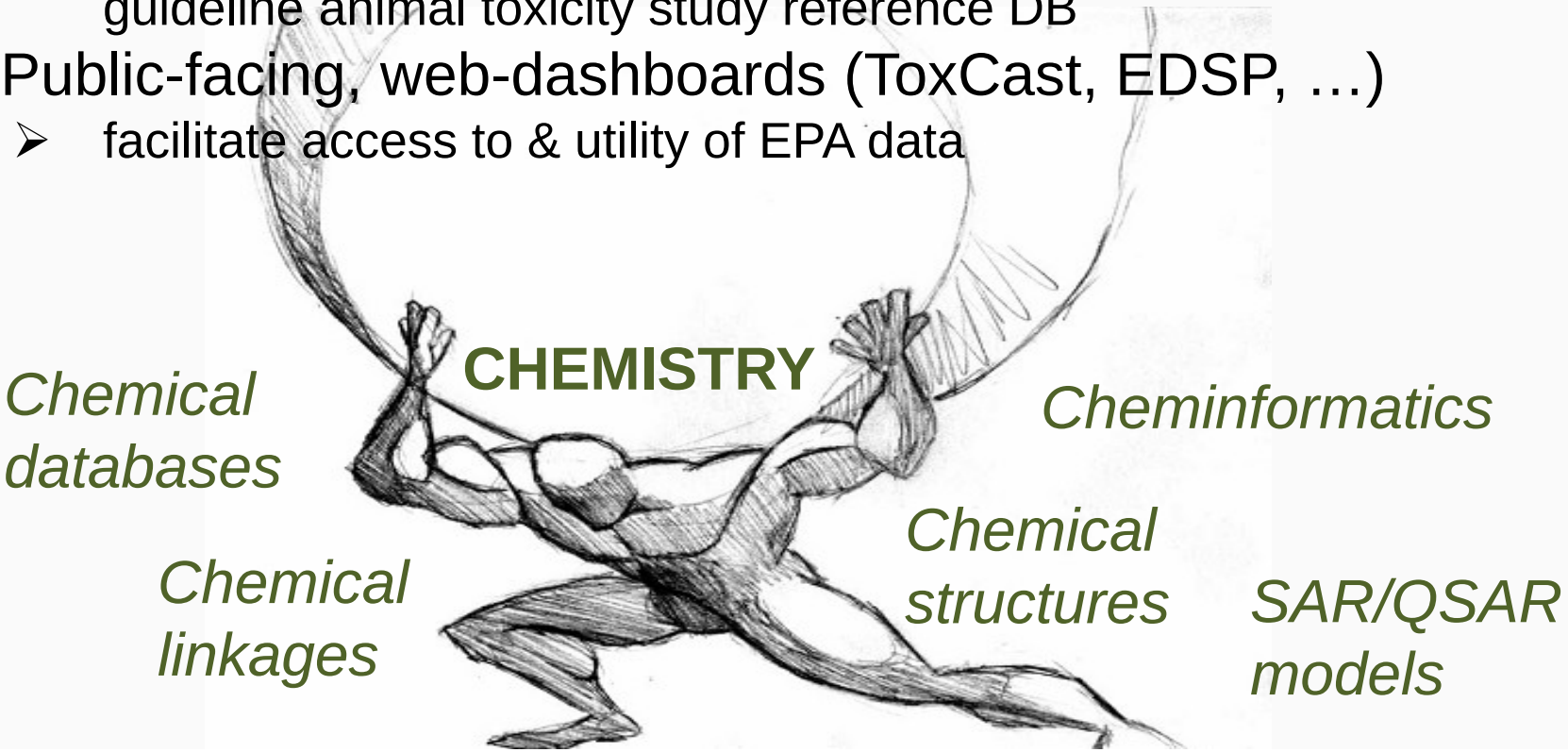
Ann Richard, Chris Grulke, Inthirany Thillainadarajah, Antony Williams, David Lyons, Jeffrey Edwards*

National Center for Computational Toxicology
U.S. Environmental Protection Agency, RTP, NC

March 13-17, 2016
ACS Spring Meeting, San Diego, CA

EPA's National Center for Computational Toxicology (NCCT)

- ToxCast (EPA) & Tox21 (Multi-Agency)
  - screening >3800 (ToxCast) to >10K (Tox21) environmentally relevant chemicals across 10's to 100's of HTS assays
- ACToR, ExpoCast, CPCat, ToxRef DB
  - meshed CAS lists, product-use database, exposure models, guideline animal toxicity study reference DB
- Public-facing, web-dashboards (ToxCast, EDSP, …)
  - facilitate access to & utility of EPA data



Current NCCT activities

- Building a new cheminformatics architecture and public-facing chemical dashboard
- Aiming to be the host for curated chemistry across EPA
- Integrating NCCT and agency resources –both internal *and* external
- Designing tools and “services” to underpin other EPA research efforts
  - *Predictive modeling*
  - *Read-across*
  - *Non-targeted screening (linking to analytical data)*

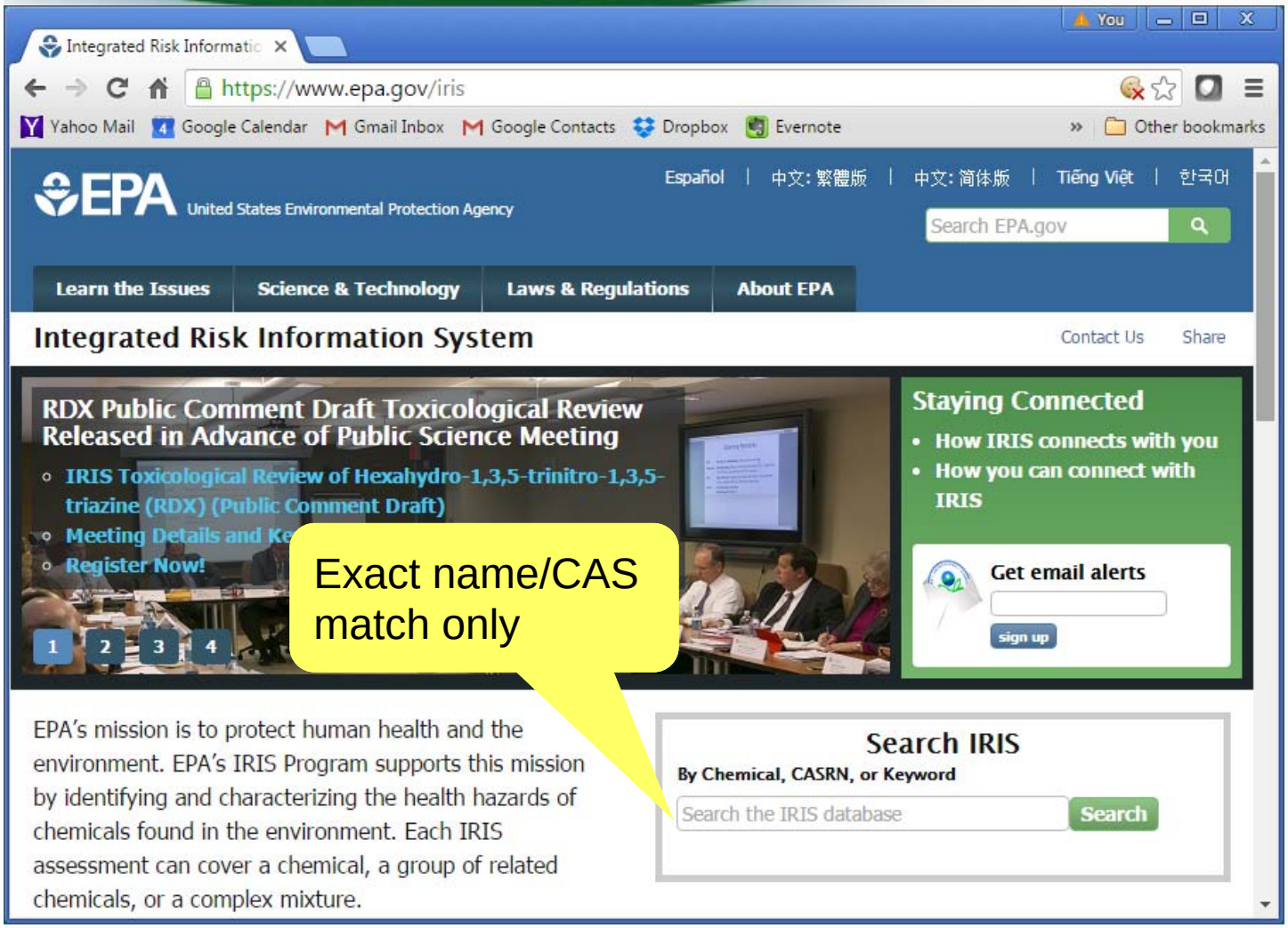
Why build yet another public chemistry database and web-interface?

Improve chemistry linkages across EPA Resources

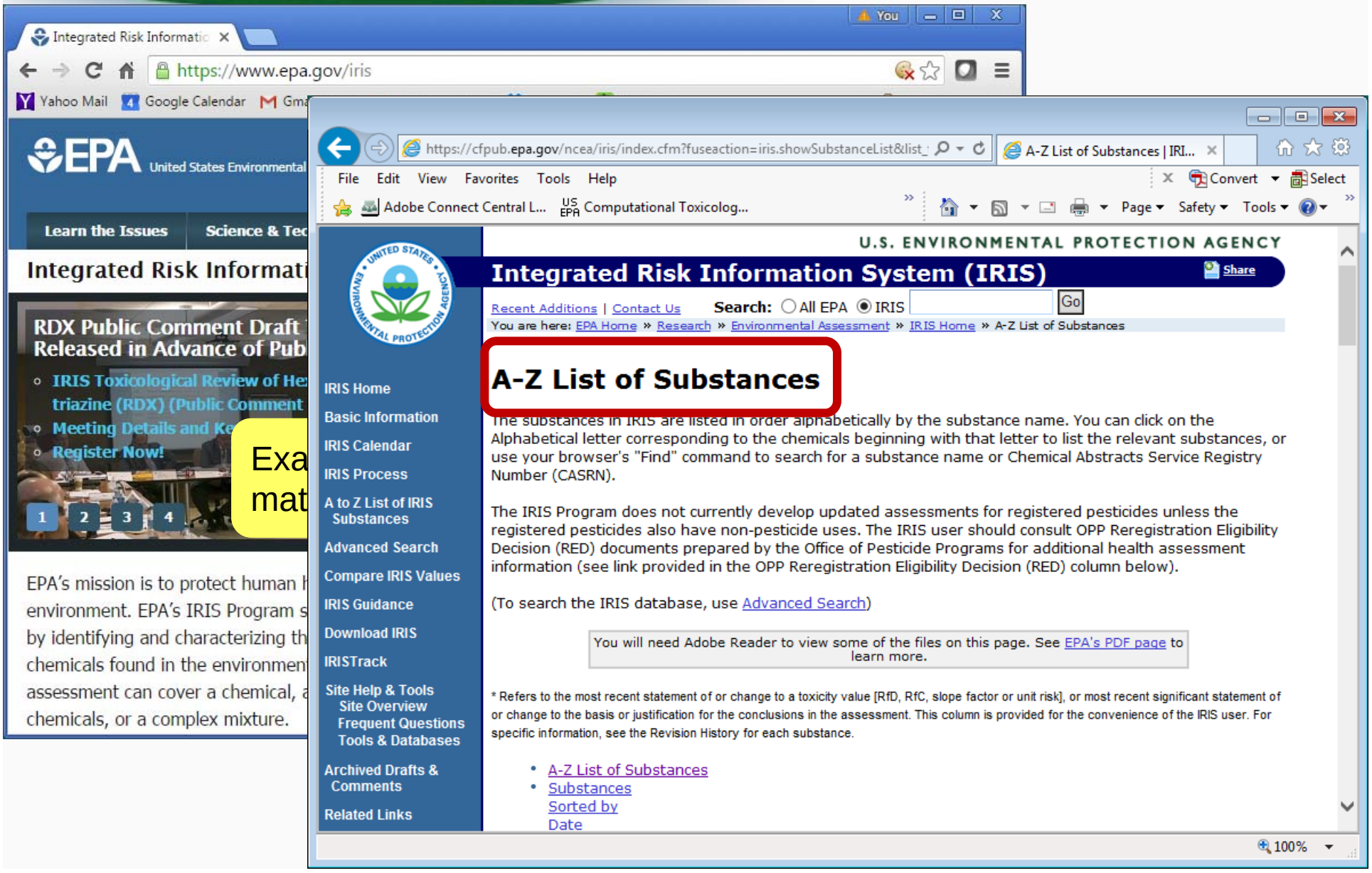
- NCCT ToxCast / EDSP Dashboards
- NCCT ACToR, ExpoCast, CPCat
- EPA Chemical Safety & Sustainability (CSS) research
- EPA Substance Registry System (SRS)
- EPA Integrated Risk Information System (IRIS)
- EPA Office of Pesticides (Inerts, Active Ingredients)
- EPA Endocrine Disruption Screening Program (EDSP)
- EPA Office of Water's Chemical Contaminants List
- EPA Office of Pollution Prevention & Toxics ChemView
- EPA EcoTox, WebICE, etc.

Lots of CAS-name lists !!

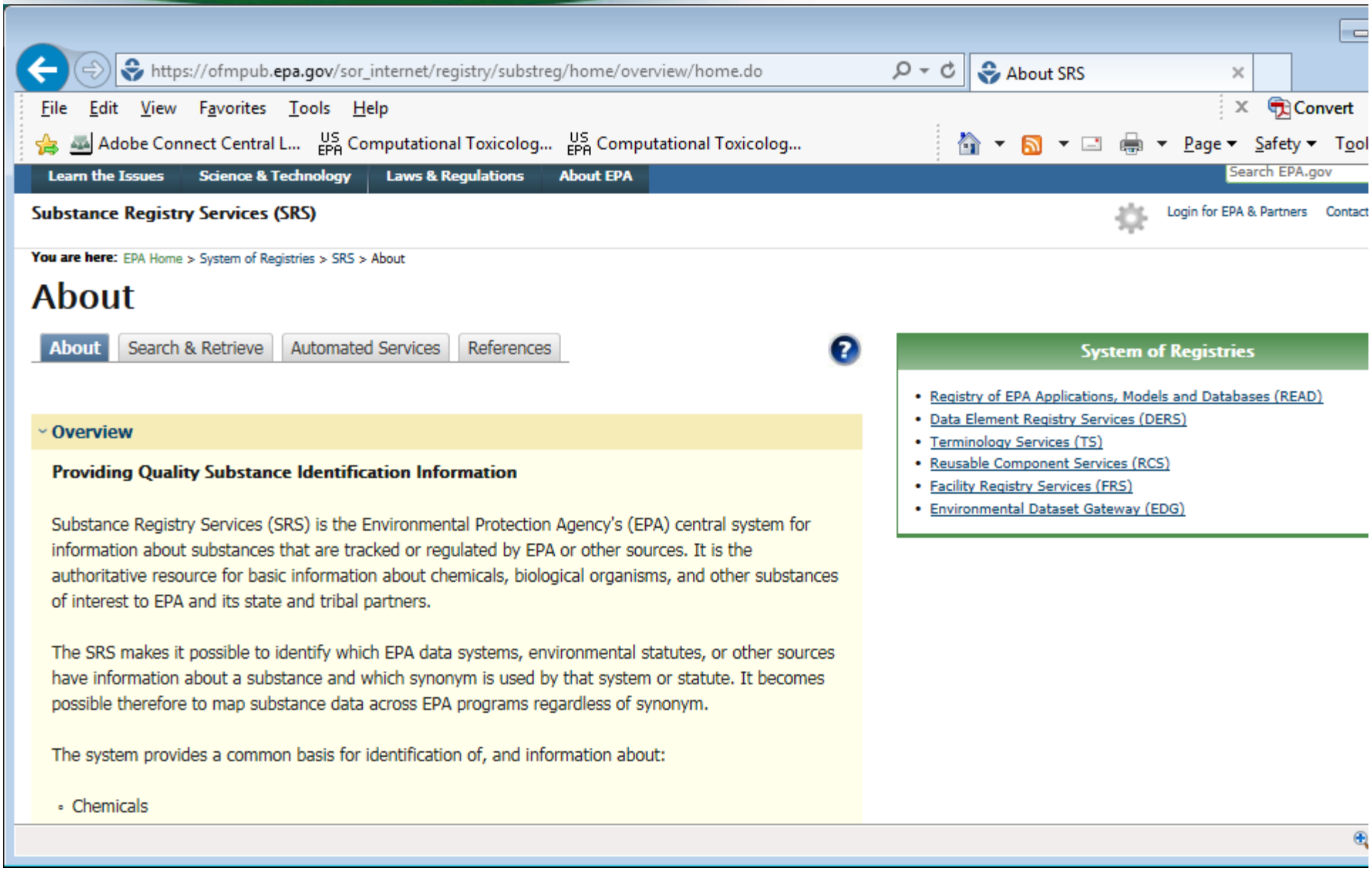
Using Chemistry to improve linkages across EPA Resources



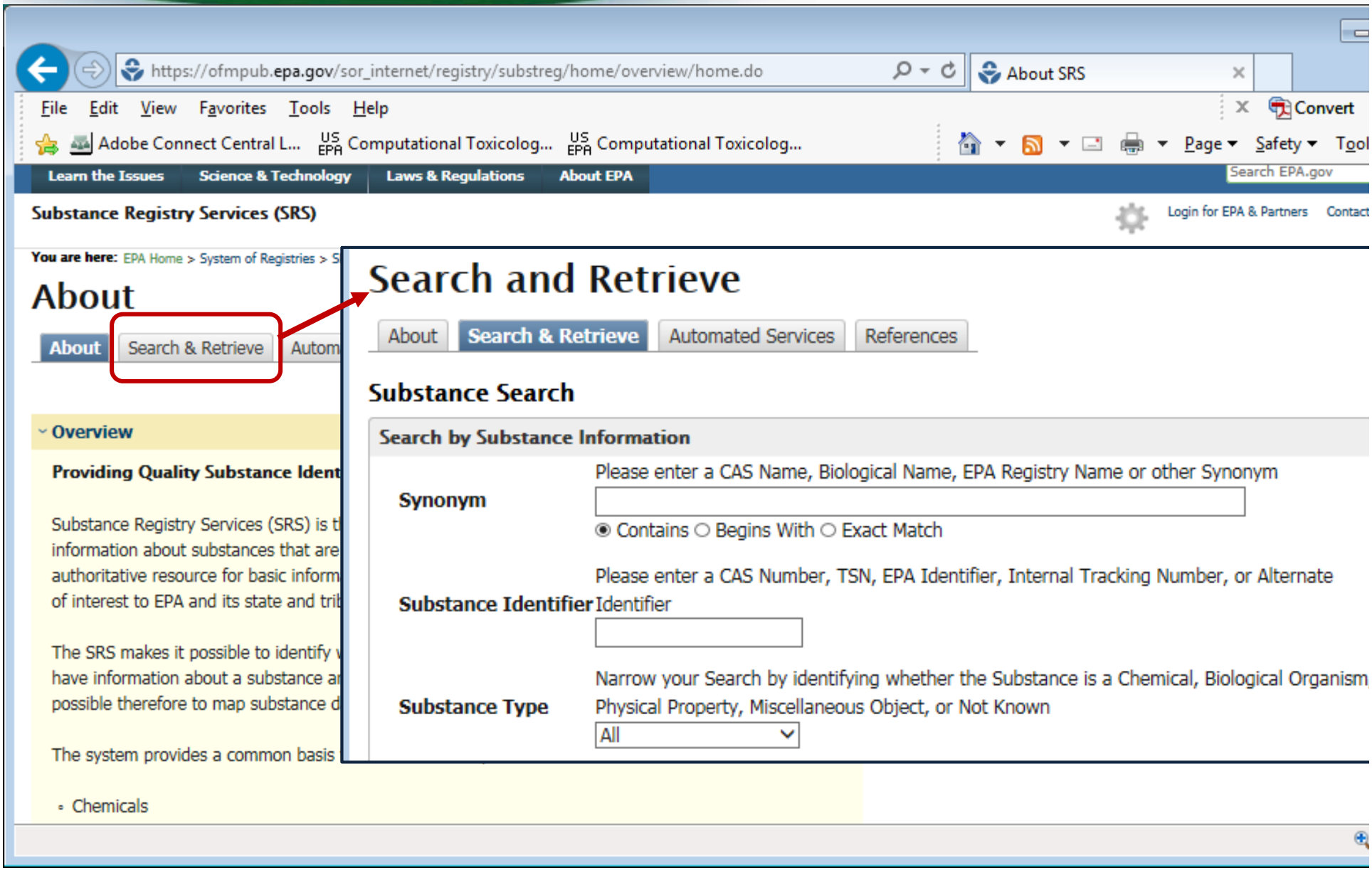
Using Chemistry to improve linkages across EPA Resources



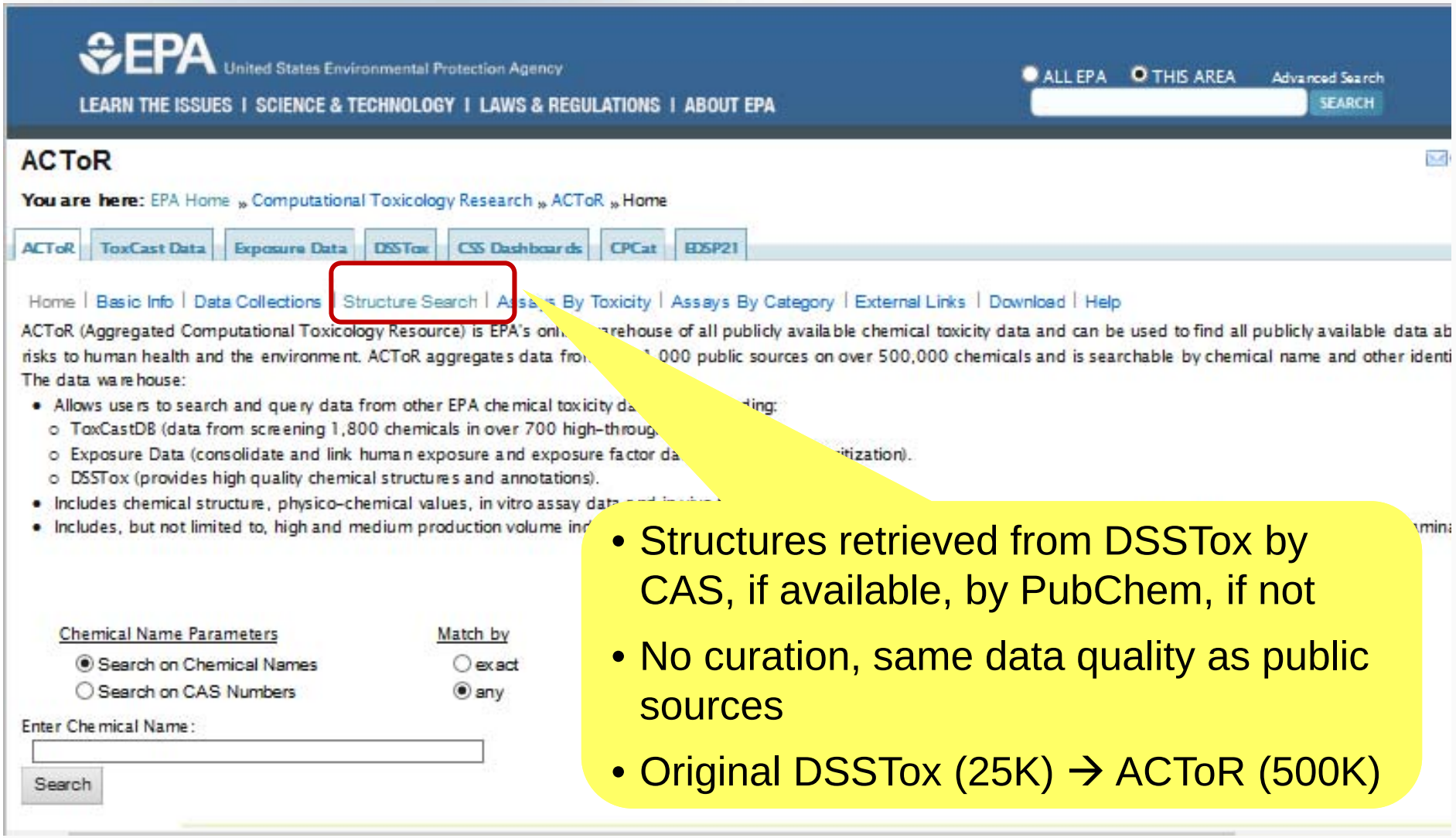
Using Chemistry to improve linkages across EPA Resources



Using Chemistry to improve linkages across EPA Resources



EPA ACToR – CAS List Meshing



- Structures retrieved from DSSTox by CAS, if available, by PubChem, if not
- No curation, same data quality as public sources
- Original DSSTox (25K) → ACToR (500K)

Why build yet another public chemistry database and web-interface?

- Focus on environmental chemicals
- Focus on data quality
- Better support EPA research & chemical regulatory programs

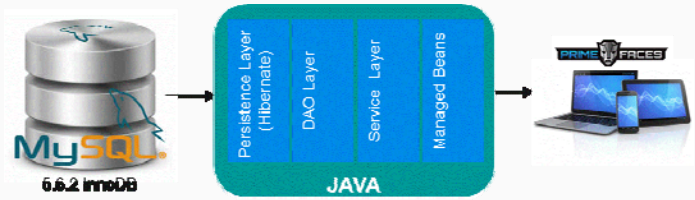
DSSTox Update

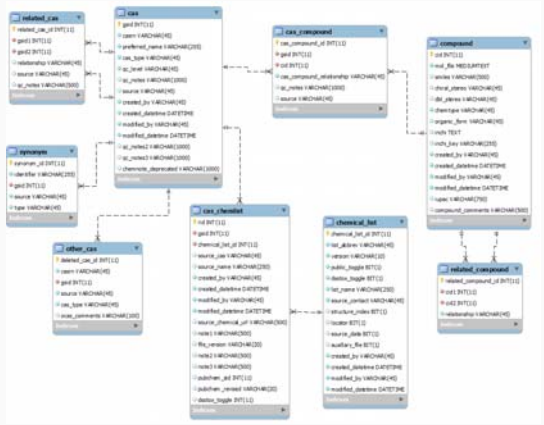
DSSTox_v1

- Manually curated 25K substance records
- EPA-focus, environmental tox datasets
- Emphasis on accurate CAS-name-structure annotations
- Public resource for high-quality structure-data files (SDF)

DSSTox_v2

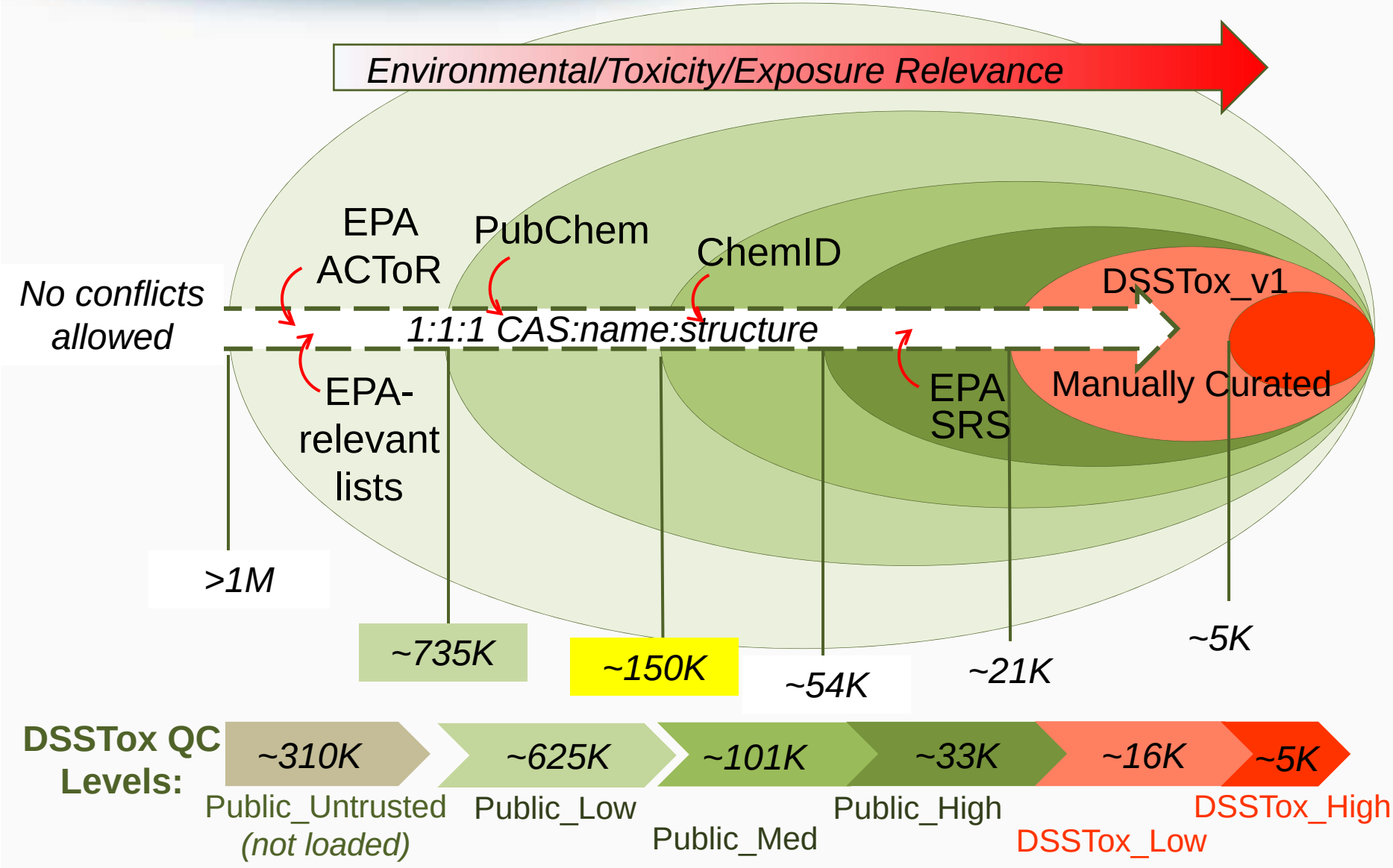
- Convert DSSTox tables to MySQL
- Develop curation interface & cheminformatics workflow
- Expand chemical content
- Web-services & Dashboard access





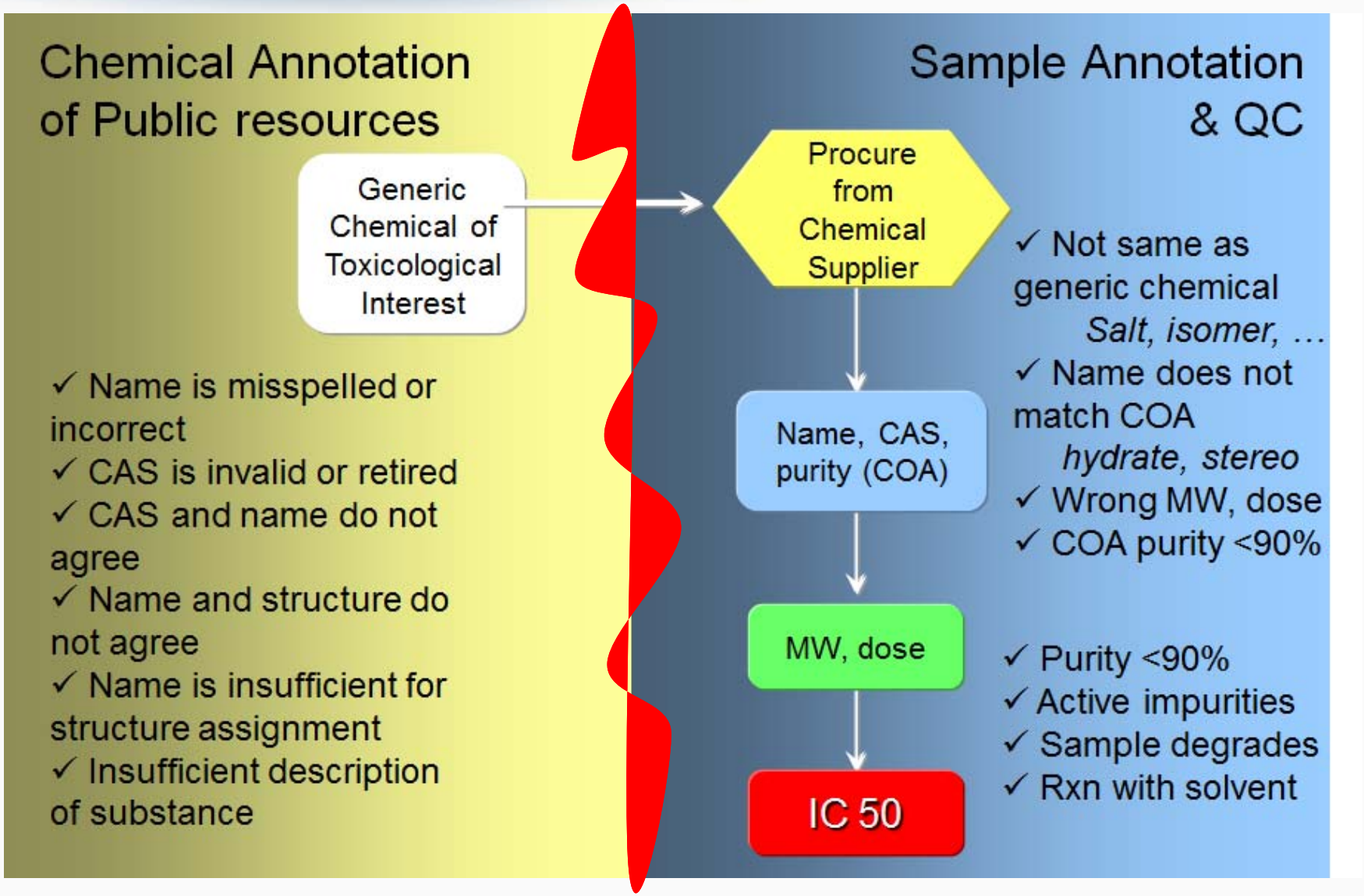
How did we go from 25K to 735K chemicals while addressing EPA's needs & quality concerns?

Building DSSTox_v2

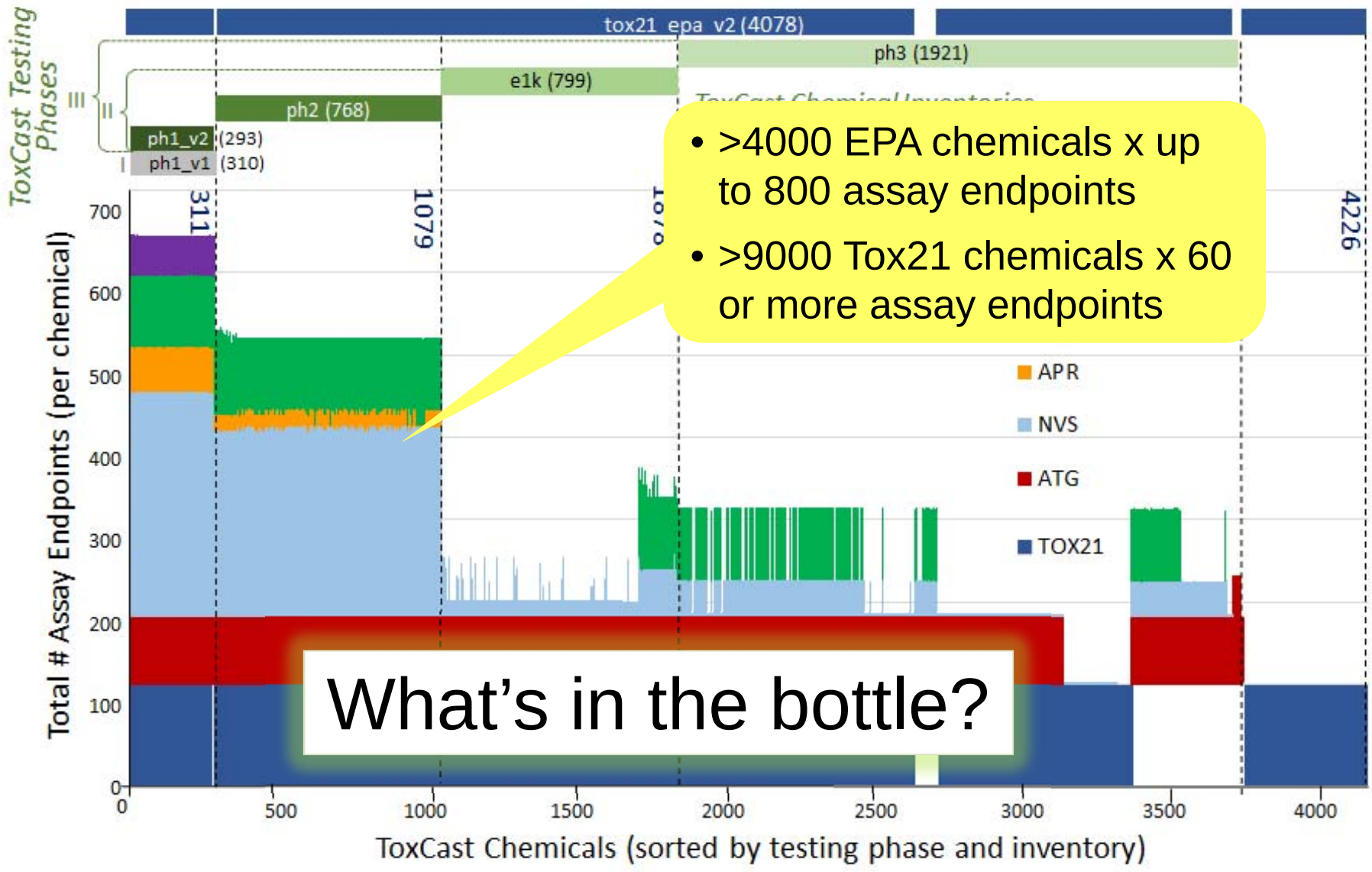


Why the concerns for data quality?

Chemical errors associated with toxicity information *(ACS, 2009)*



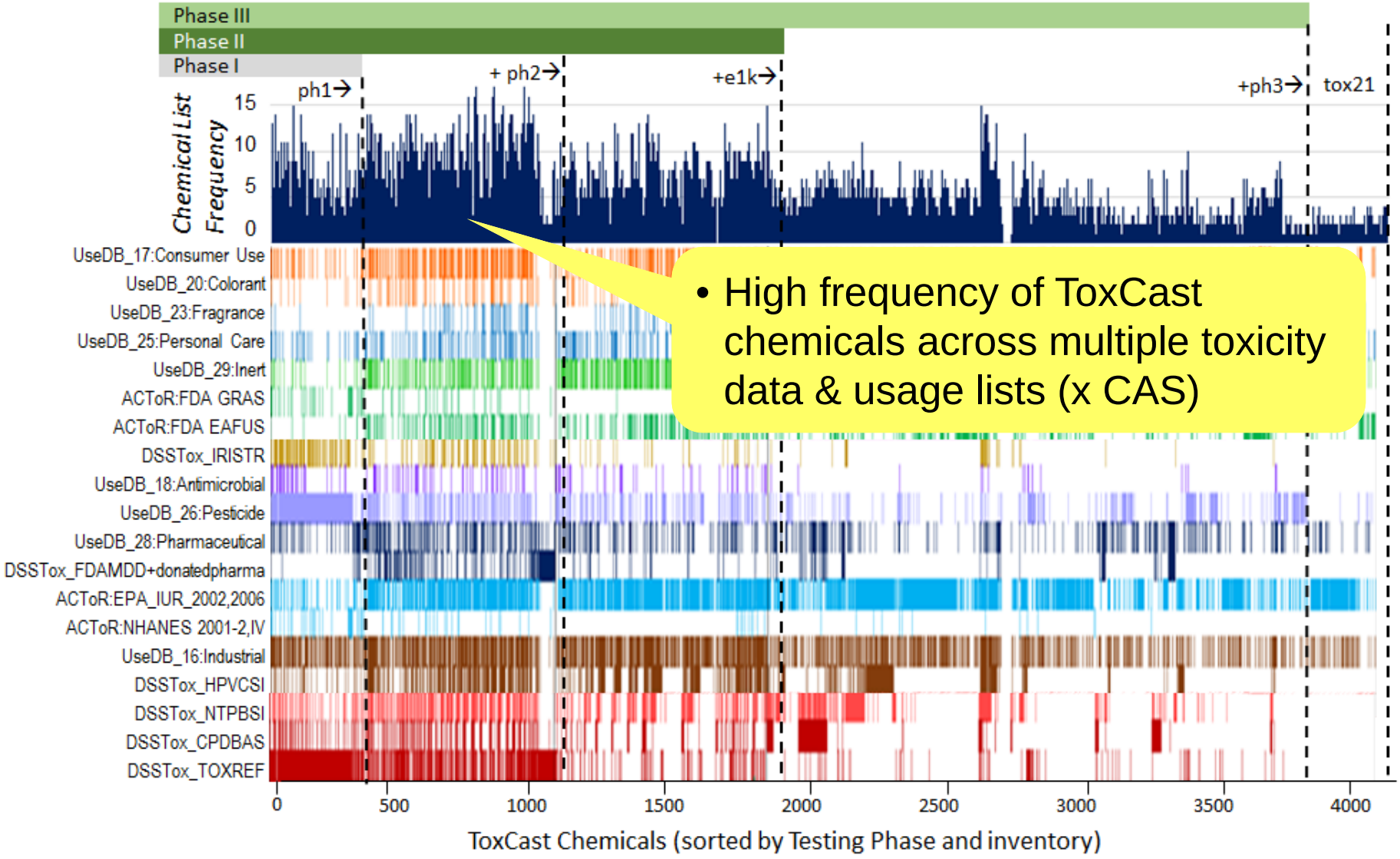
ToxCast chemical x assay counts summary (Jan 2016)



- >4000 EPA chemicals x up to 800 assay endpoints
- >9000 Tox21 chemicals x 60 or more assay endpoints

What's in the bottle?

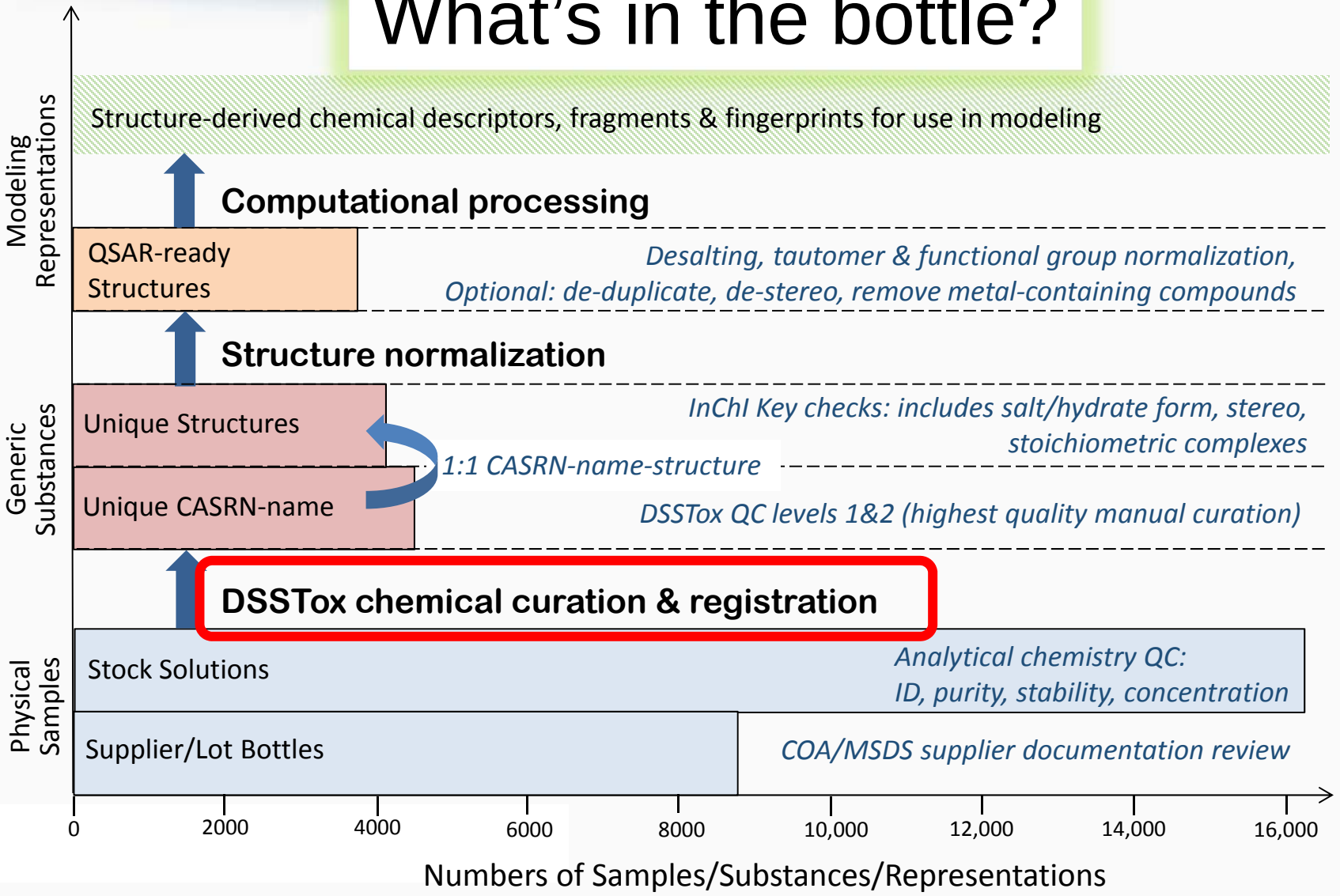
Tox Data & Use List Coverage of ToxCast Chemicals



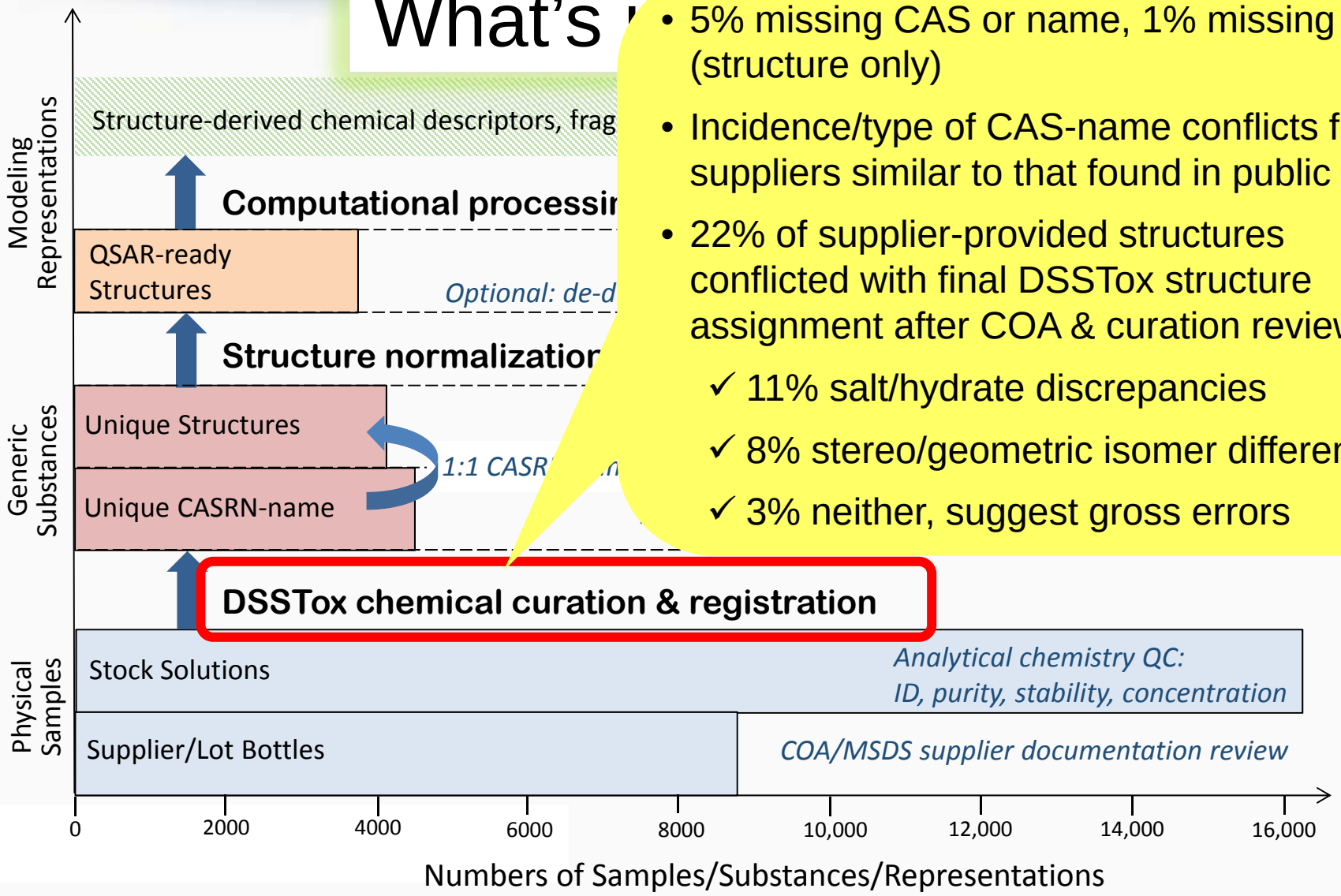
- High frequency of ToxCast chemicals across multiple toxicity data & usage lists (x CAS)

ToxCast Chemical Library: Quality Control Steps

What's in the bottle?

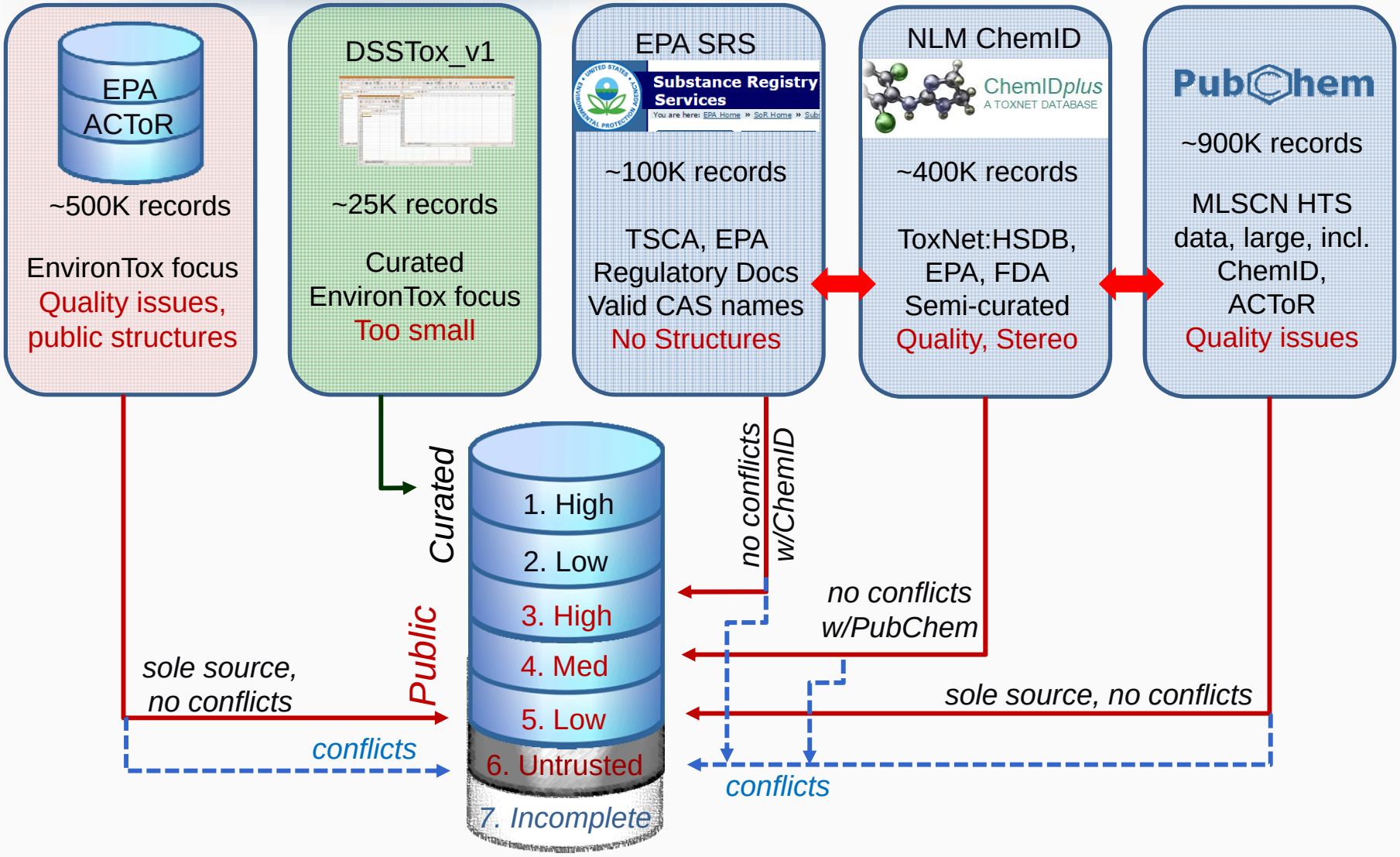


ToxCast Chemical Library: Quality Control Steps

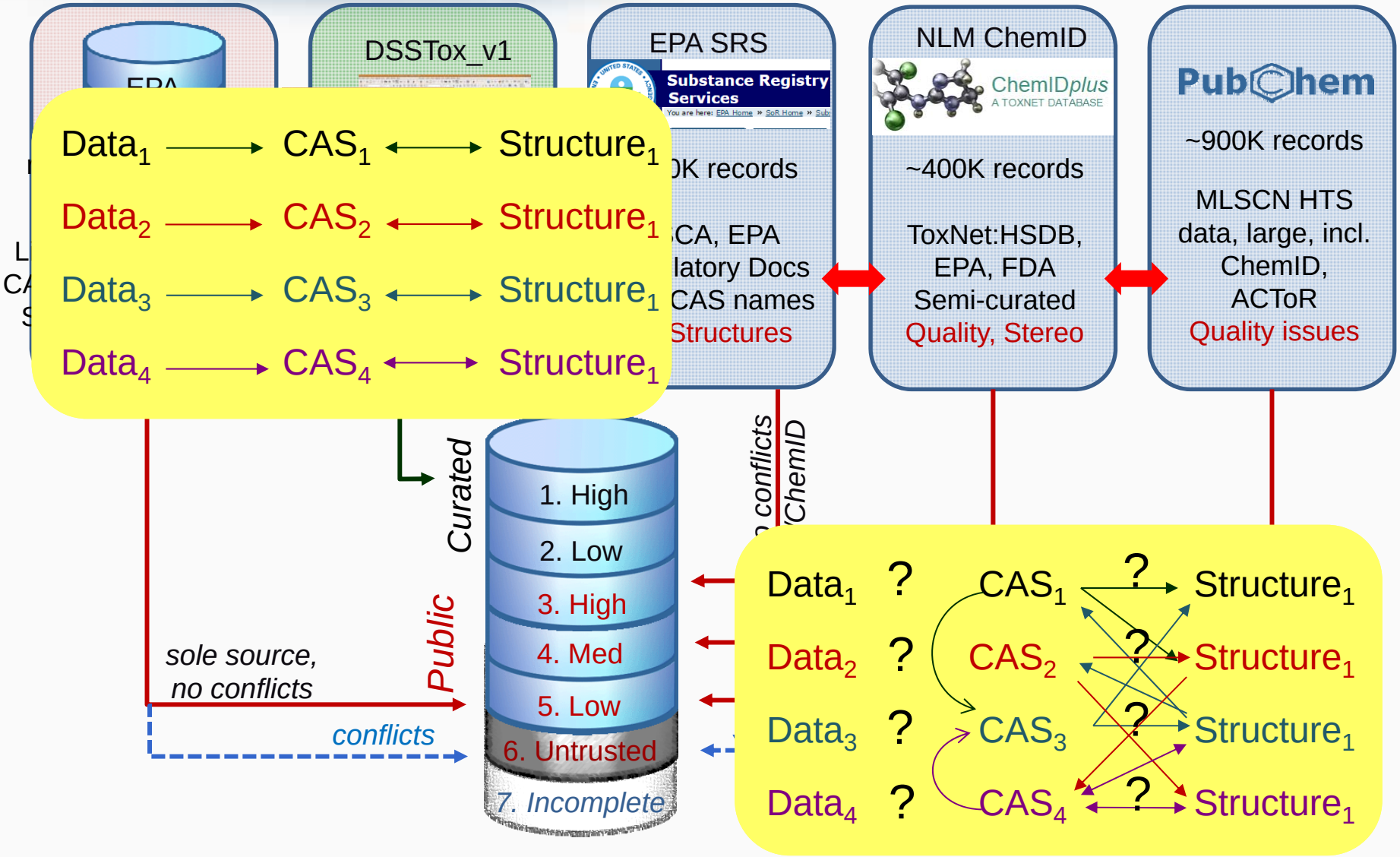


- 5% missing CAS or name, 1% missing both (structure only)
- Incidence/type of CAS-name conflicts from suppliers similar to that found in public lists
- 22% of supplier-provided structures conflicted with final DSSTox structure assignment after COA & curation review
  - ✓ 11% salt/hydrate discrepancies
  - ✓ 8% stereo/geometric isomer differences
  - ✓ 3% neither, suggest gross errors

DSSTox_v2 CAS-Structure Sources: QC levels



DSSTox_v2 CAS-Structure Sources: QC levels



DSSTox_v2 Construction

Data source load order:

1) DSSTox_v1 (~22K)
   - ✓ 1:1 CAS-structure mappings
   - ✓ Assign NOCAS_GSID
   - ✓ Related CAS & structure mappings (e.g., NOCAS, mixtures)

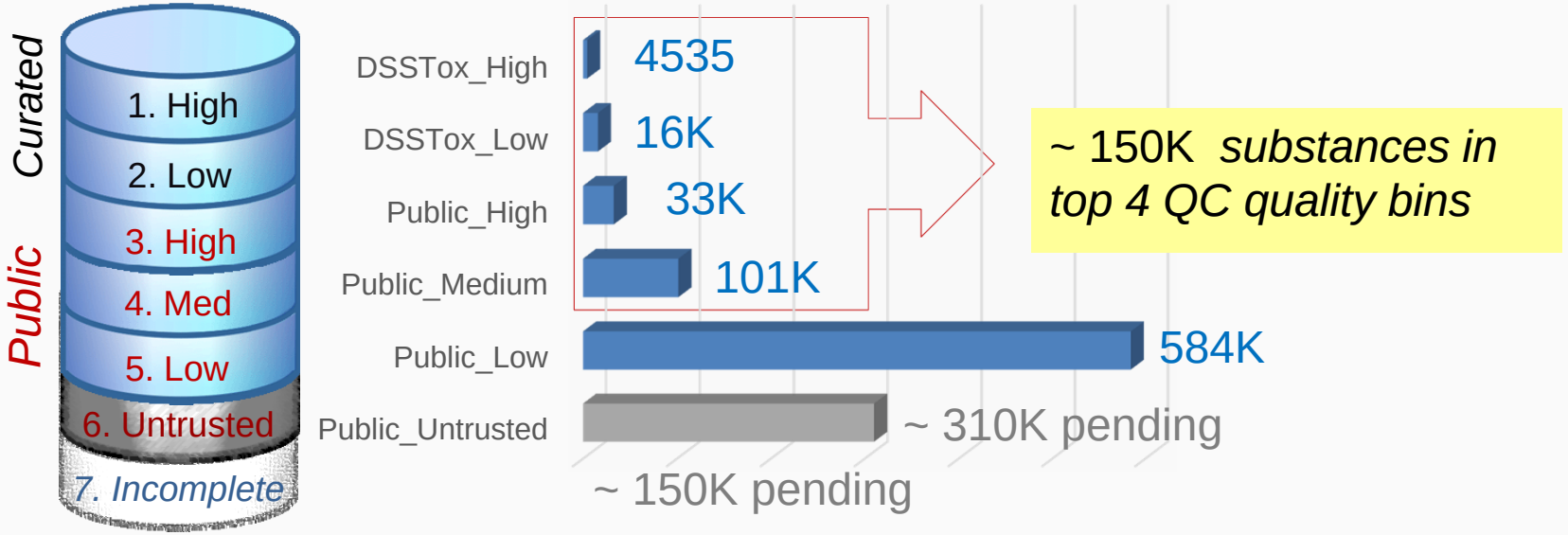
2) EPA SRS (~77K)
   - ✓ *systematic name → structure conversion*
   - ✓ *internal CAS-structure conflicts (12.5%)*
   - ✓ *ChemID conflicts (24% of 30K overlaps)*
   - ✓ *DSSTox conflicts (8% of 6200 overlaps) → queue for curation*

3) ChemID (~400K)
   - ✓ *internal CAS-structure conflicts (4.5%)*
   - ✓ *PubChem conflicts (45% of 225K overlaps) … OUCH!!*
   - ✓ *DSSTox conflicts (11% of 2300 overlaps) → queue for curation*

4) And so on …

DSSTox_v2 Totals

QC Level Totals (12Jun2015)



QC Levels

DSSTox_High:	Hand curated - highest confidence
DSSTox_Low:	Hand curated and confirmed using multiple public sources
Public_High:	Extracted from EPA SRS and confirmed to have no conflicts in ChemID and PubChem
Public_Medium:	Extracted from ChemID and confirmed to have no conflicts in PubChem
Public_Low:	Extracted from ACToR or PubChem (single source)
Public_Untrusted:	Postulated, but found to have conflicts in public sources

e.g., structure mapping collision

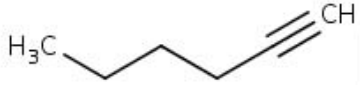
ChemIDplus
A TOXNET DATABASE

Start New Query | Modify Query | Search History | Show Query
Switch to Summary View

Substance Name: Hexyne
RN: 26856-30-4
InChIKey: CGHIBGNXEGJPQZ-UHFFFAOYSA-N

Molecular Formula
C6-H10

Molecular Weight
82.145



Na +

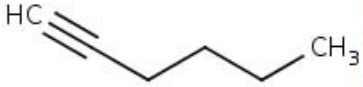
ChemIDplus
A TOXNET DATABASE

Start New Query | Modify Query | Search History | Show Query
Switch to Summary View
Go to summary view

Substance Name: 1-Hexyne
RN: 693-02-7
InChIKey: CGHIBGNXEGJPQZ-UHFFFAOYSA-N

Molecular Formula
C6-H10

Molecular Weight
82.145



Na +

3D

e.g., list curation problems

Raw file → *Initial manual review, clean-up* → *store source CAS-names*
curator checks mapping
curator registers 32

Functional Use DB:

List of 176 CAS chemicals indicated for use as "Flame Retardants"

17 CAS, no name

9 invalid CAS (manual fix)

10 duplicate names, different CAS

23 duplicate CAS, different names

152 cleaned list, no duplicates

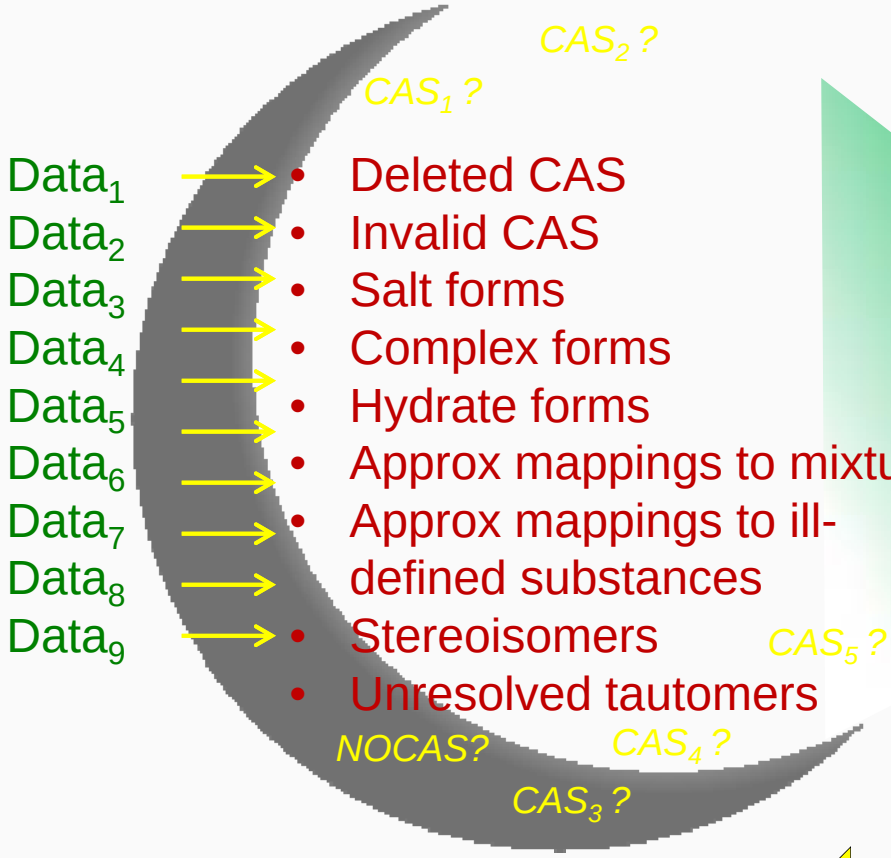
121 CAS map to DSSTox_v2

32 CAS not found

15% CAS error or no name, 19% conflicts,18% CAS not found,
80% of cleaned list map to existing DSSTox_v2 content

e.g., non-targeted MS screening of environmental samples

CAS-Name-Structure "Sphere of Confusion"



many:1

Monoisotopic Mass

Formula

Parent structure
(no stereo, desalted)

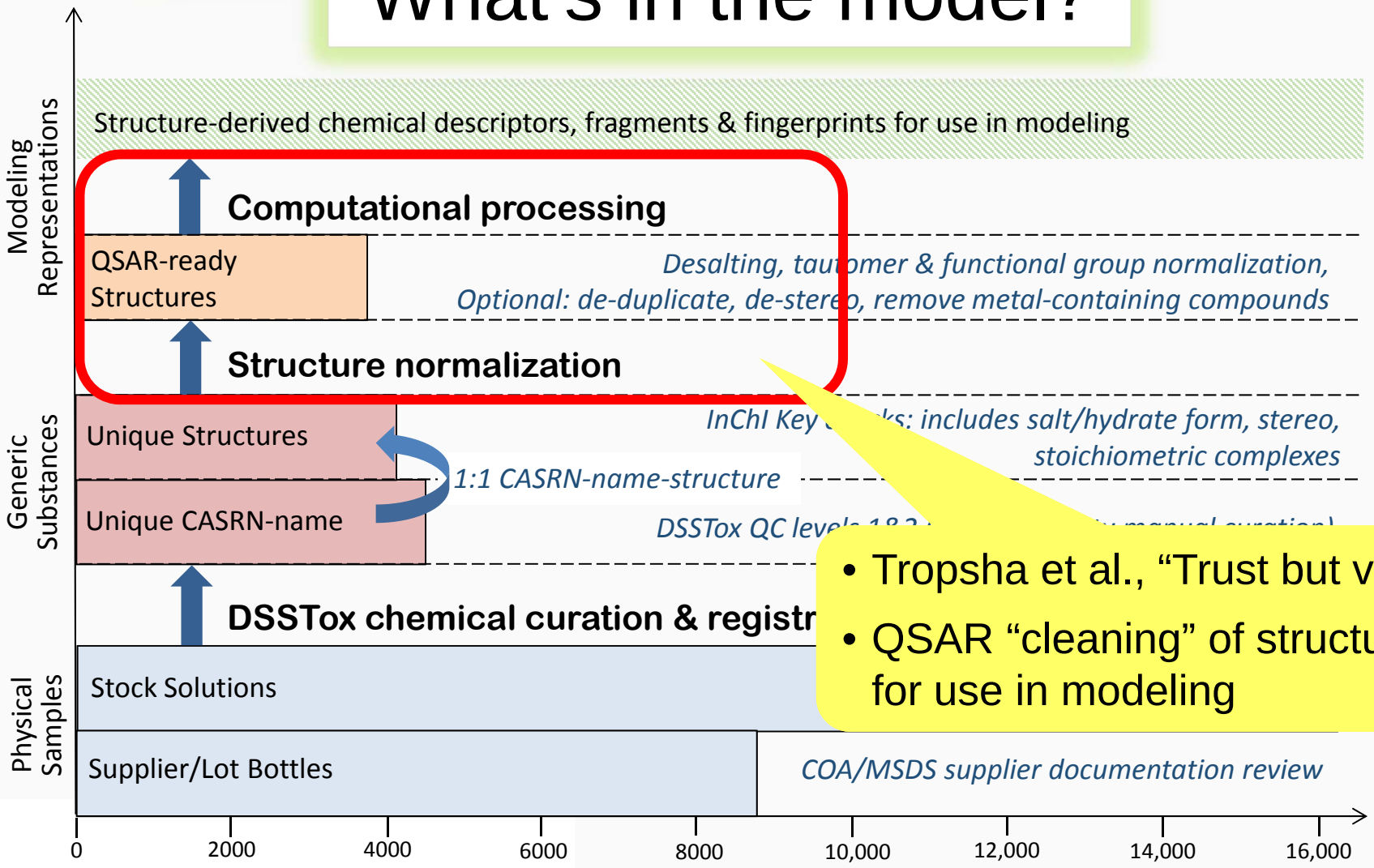
Valid CAS-substance?

DSSTox_v2 Database & Cheminformatics Layer

- Resolve CAS-structure mappings for accurate data mapping
- Collapse sphere to collect all data at parent structure-formula level

ToxCast Chemical Library: Quality Control Steps

What's in the model?

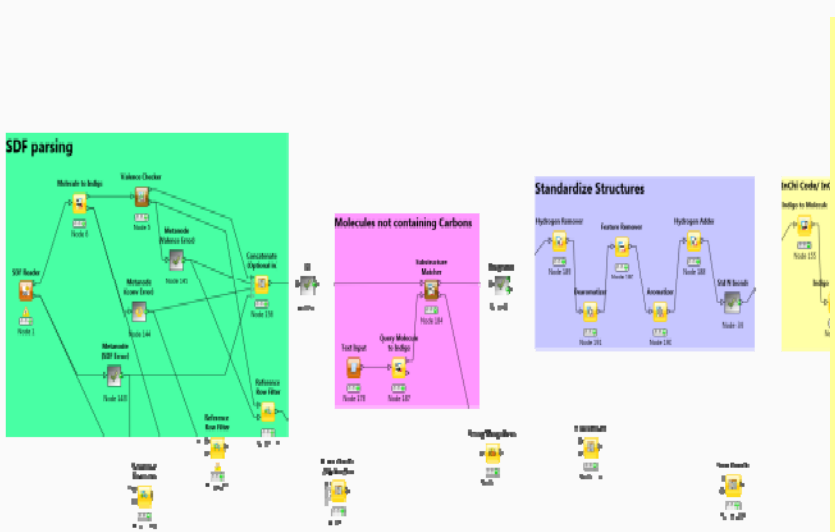


- Tropsha et al., "Trust but verify"
- QSAR "cleaning" of structures for use in modeling

KNIME structure-“cleaning” workflow

https://www.knime.org/knime

- Combine community approaches to structure processing
- Develop a flexible workflow to be used by EPA and shared publicly
- Process DSSTox files to create “QSAR-ready” structures



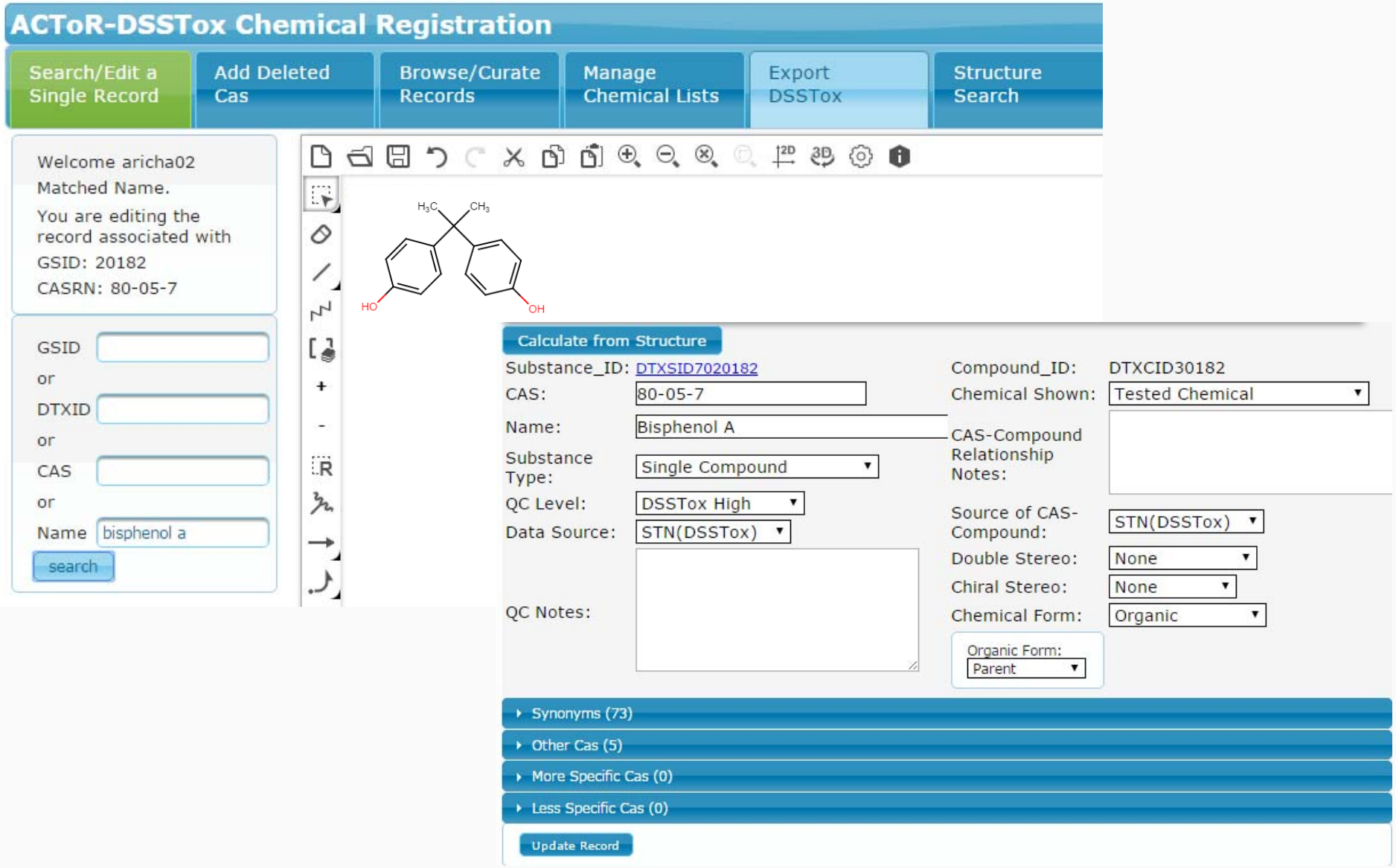
- ✓ Parse SDF, remove fragments
- ✓ Explicit hydrogen removed
- ✓ Dearomatization
- ✓ Removal of chirality info, isotopes and pseudo-atoms
- ✓ Aromatization + add explicit hydrogen atoms
- ✓ Standardize Nitro groups
- ✓ Other tautomerize/mesomerization
- ✓ Neutralize (when possible)

Publicly available cheminformatics toolkits in KNIME:

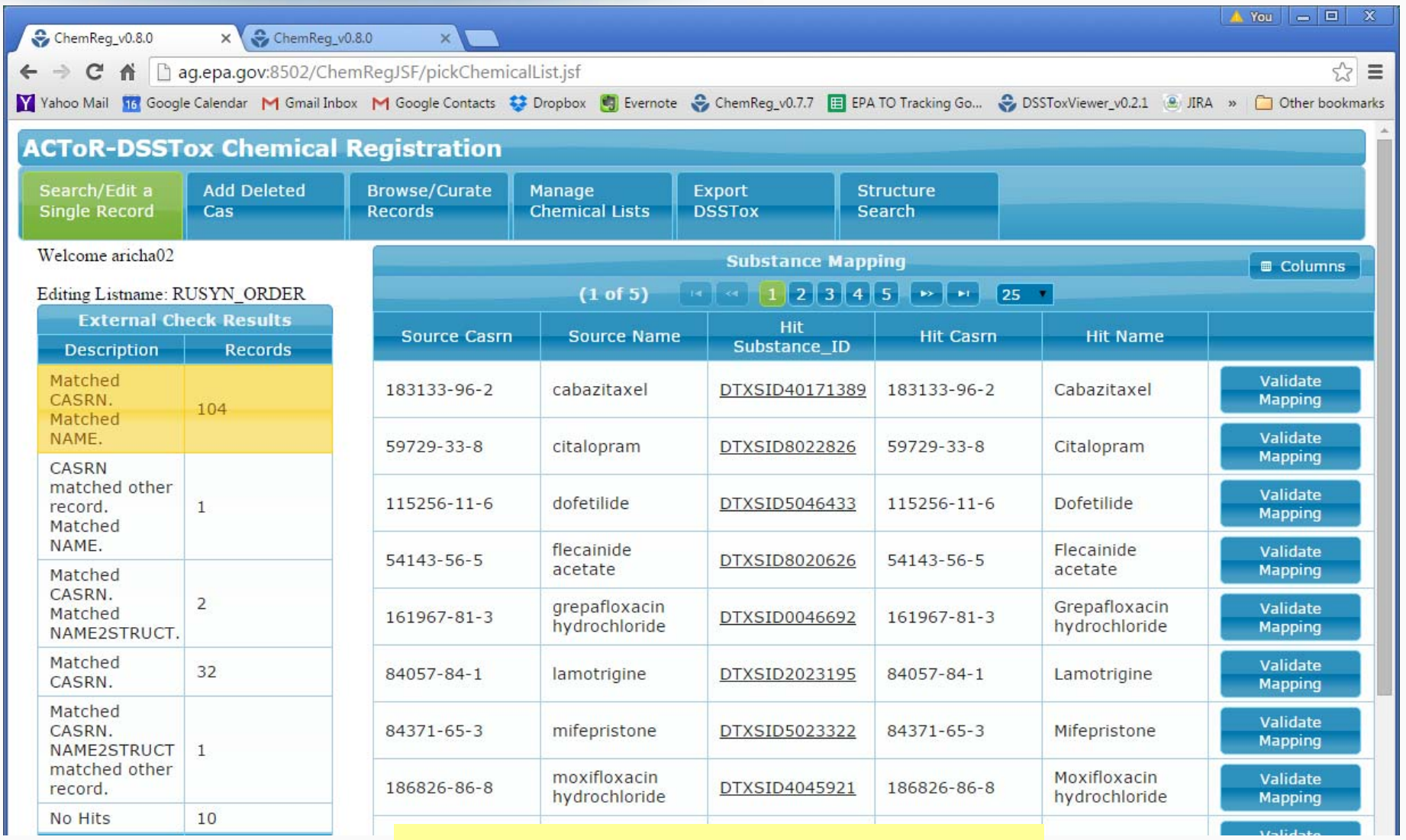
Indigo RDKit CDK Pharmaceutical Data Exploration Laboratory

Slide courtesy of K. Mansouri

DSSTox Chemical Registration Tool

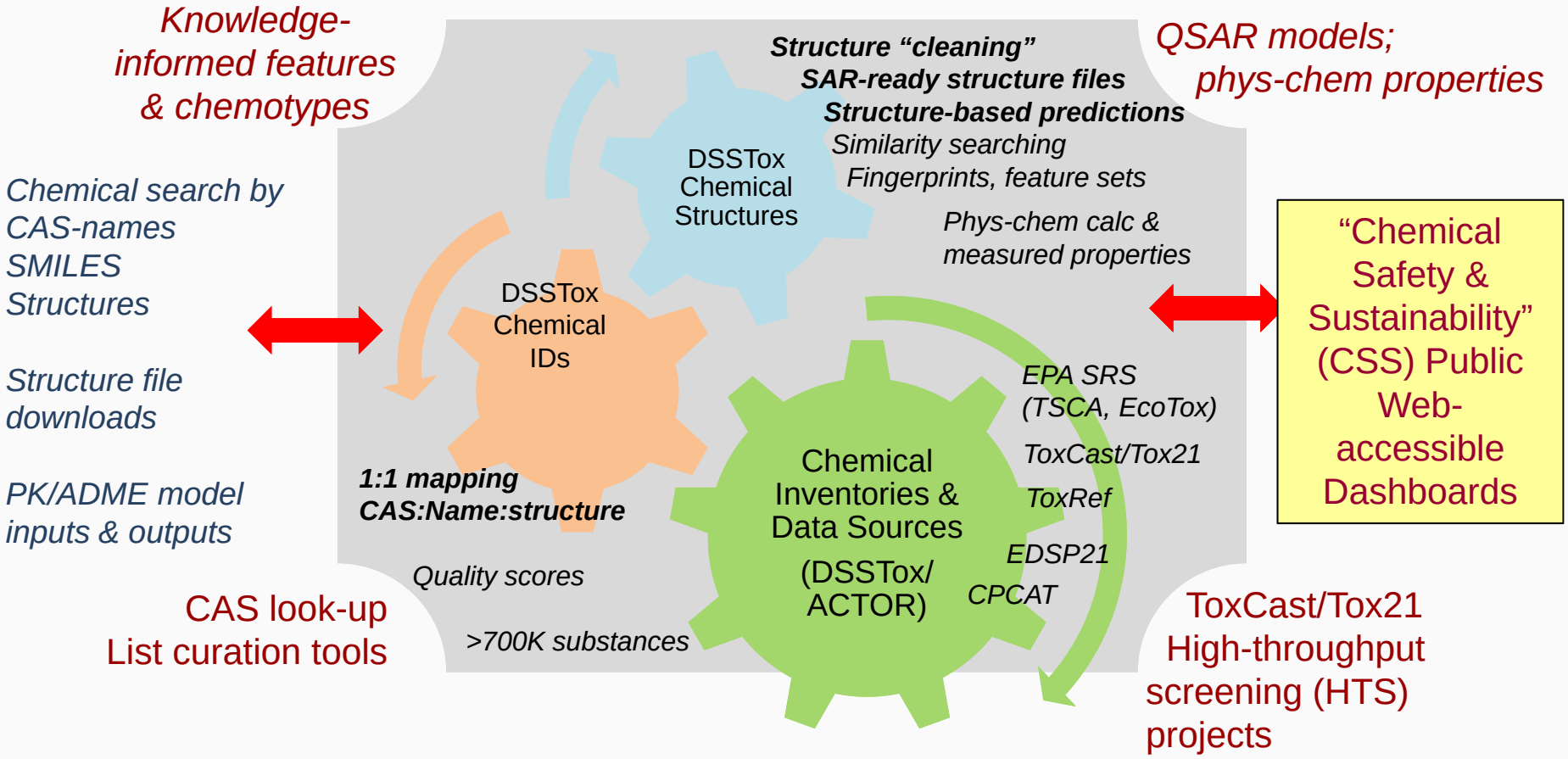


DSSTox List Curation Tool

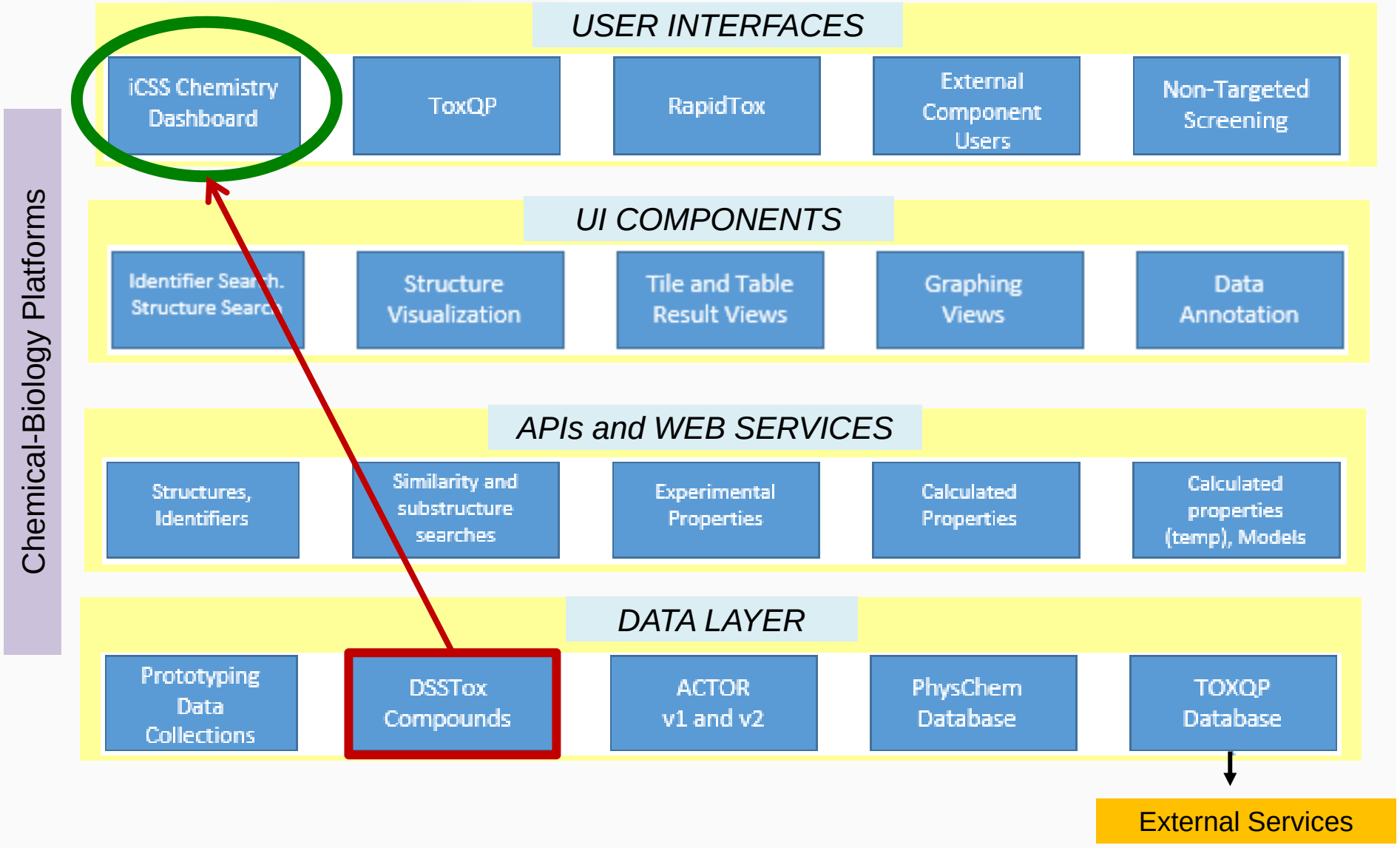


Conflicts binned to facilitate curation

Chemistry Foundation to Support Multiple NCCT & EPA Projects

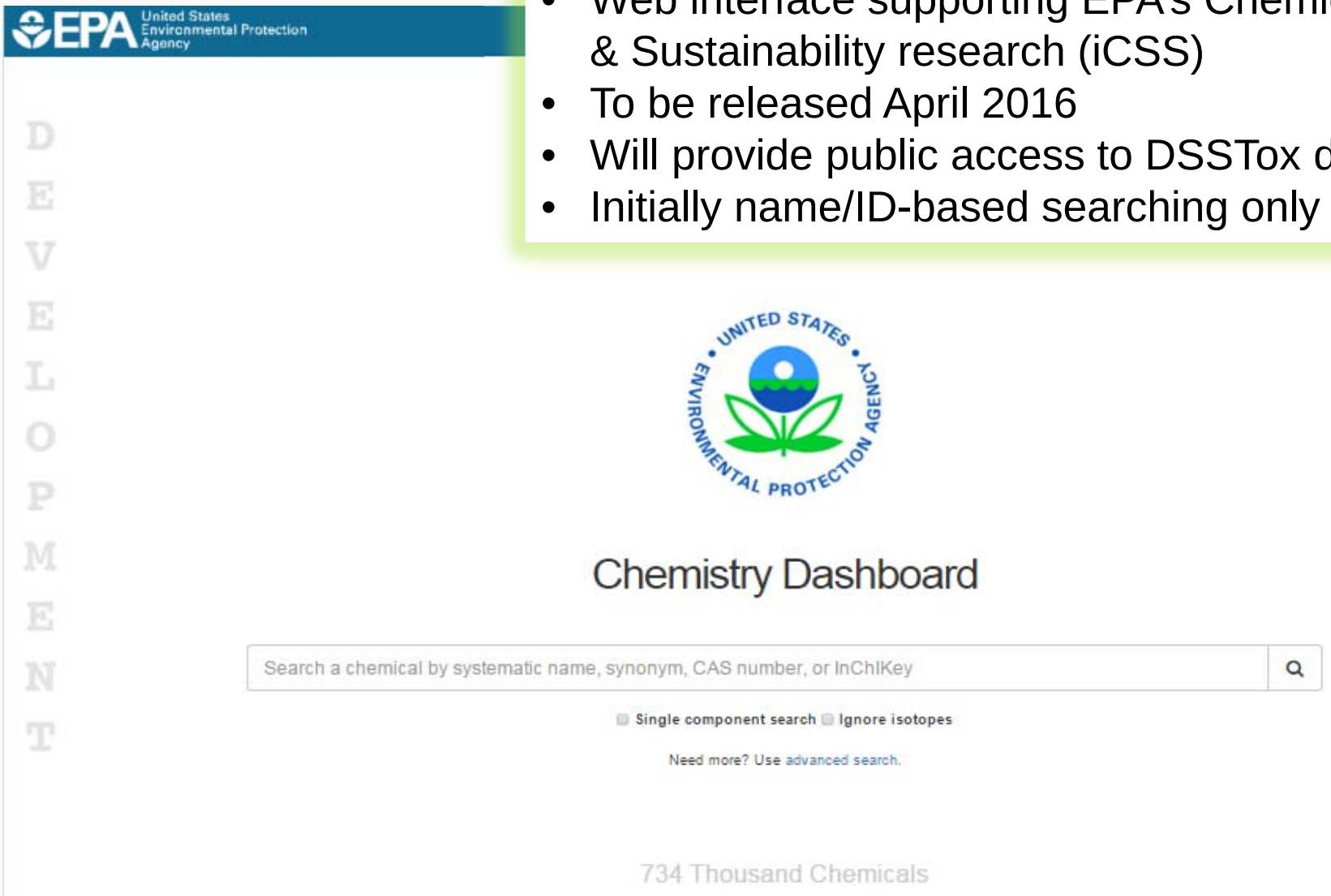


Building the Chemistry Software Architecture

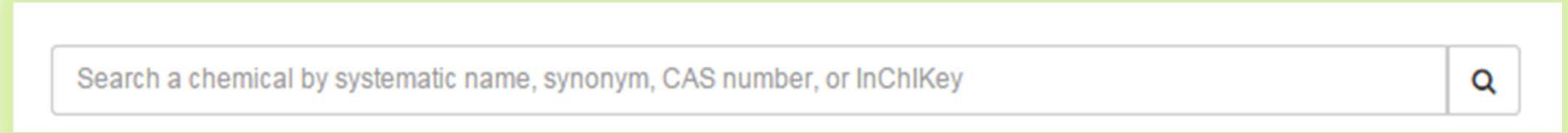


iCSS Chemistry Dashboard

- Web interface supporting EPA's Chemical Safety & Sustainability research (iCSS)
- To be released April 2016
- Will provide public access to DSSTox db
- Initially name/ID-based searching only …but



Componentizing API for Resolving Chemical Names/IDs



- Look familiar?
  - *ChemSpider, PubChem, ChEMBL, NCI Resolver …*
- ONE reusable EPA component API that searches for Names and CAS Registry Numbers…
- … and SMILES, and InChIKeys, and recognizes invalid CAS Numbers, and converts SMILES when they aren't in DB, etc. etc.

iCSS Chemistry Dashboard Releasing in April 2016

EPA United States Environmental Protection Agency

EPA United States Environmental Protection Agency

Search Chemical

DEVELOPMENT

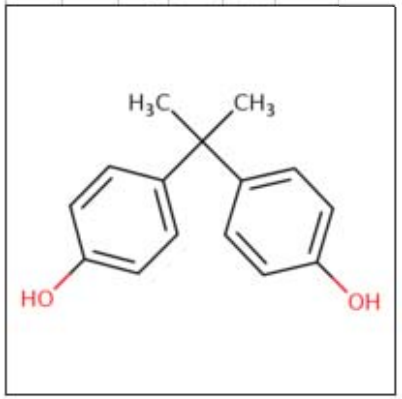
Bisphenol A

80-05-7 DTXSID7020182

Formerly DSSTox GSID, new unique public substance ID, essential for RDF/semantic web applications

Searched by Synonym: Found 1 result for 'bisphenol A'.

2D 3D



Intrinsic Properties

Molecular Formula: C15H16O2 Search in DSSTox

Average Mass: 228.291 g/mol

Monoisotopic Mass: 228.11503 g/mol

Structural Identifiers

Citation

Chemical Properties | External Links | Synonyms | PubChem Biological Activities | PubChem Articles | PubChem Patents | Comments

CSV Excel

Property	Average (Exp.)	Range (Exp.)	Average (Pred.)	Range (Pred.)
Solubility	0.001 (1)	0.0005257 to 0.0005257	0.38 (2)	0.003675 to 0.7565
Melting Point	154.929 (7)	153.0 to 158.0	144.033 (3)	131.8 to 158.0
Boiling Point	200.0 (1)	200.0 to 200.0	348.95 (2)	334.4 to 363.5
LogP	3.357 (3)	3.32 to 3.431	3.524 (3)	3.205 to 3.727
Atmospheric Hydroxylation Rate	N/A	N/A	0.0 (1)	4.237e-11 to 4.237e-11

About Contact ACToR DSSTox Privacy Accessibility Help

Build once, use many times

Chemical Input: CCNC1=NC(NC(C)C)=NC(Cl)=N1

(Retrieved using standard InChIKey)

Chemical Input: 0000001912-24-9

(Retrieved using numeric identifier)

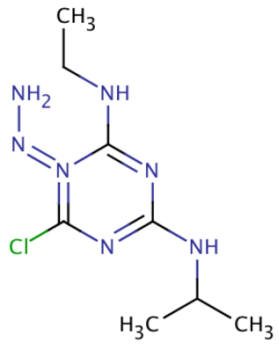
Warnings: Leading zeros stripped

Chemical Input: 1912-24-7

Warnings: Identifier not found in DSSTox, Unable to resolve identifier

Chemical Input: CCNC1=NC(NC(C)C)=NC(Cl)=N1=NN

Warnings: Identifier not found in DSSTox



Chemical Input: MXWJVTOOROXGIU-UHFFFAOYSA-N

(Retrieved using standard InChIKey)

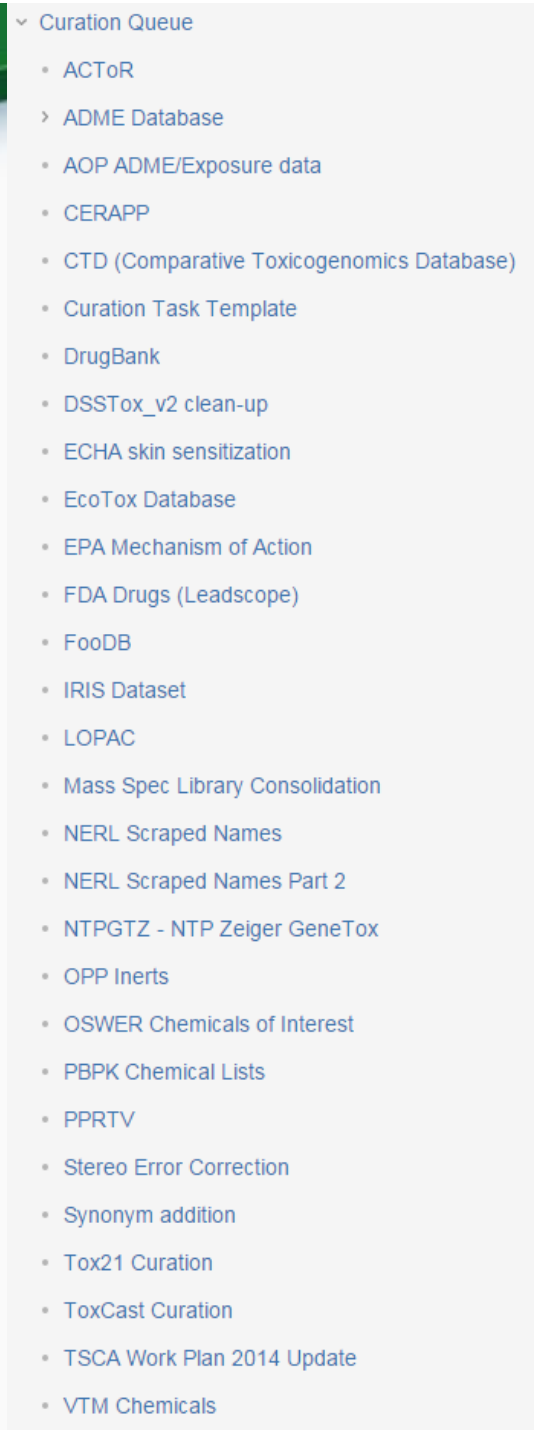
- Componentizing functionality for reuse
- Adopting a common development framework

Synonym Sourcing & Chemical Name Searching

- Entering high-low quality chemical name synonyms for DSSTox (NO COLLISIONS ALLOWED!)
- 500,000 “synonyms” (names + IDs) loaded and growing
  - DSSTox historical resources (Done)
  - ChemID-Plus (Done)
  - CAS Common Chemistry (Done)
  - *ChEBI (to come)*
  - *ChemSpider Validated Synonyms (to come)*
- “Caffeine Fix” (used by ChemSpider & others) for partial & fuzzy name searching

Where are we now?

- Public iCSS Chemistry Dashboard to be released April 2016
- DSSTox External IDs created for RDF linkages wherever DSSTox content is used publicly
  - e.g. Atrazine:
  - DTXSID9020112 (DTXSID.cksum.0.GSID)
- Curation continues…
  - *Prioritized EPA and ACToR lists*
- Coming over next year:
  - *Structure/substructure/similarity searching*
  - *List searching, structure downloads*
  - *Expanded webservices, open data & APIs*

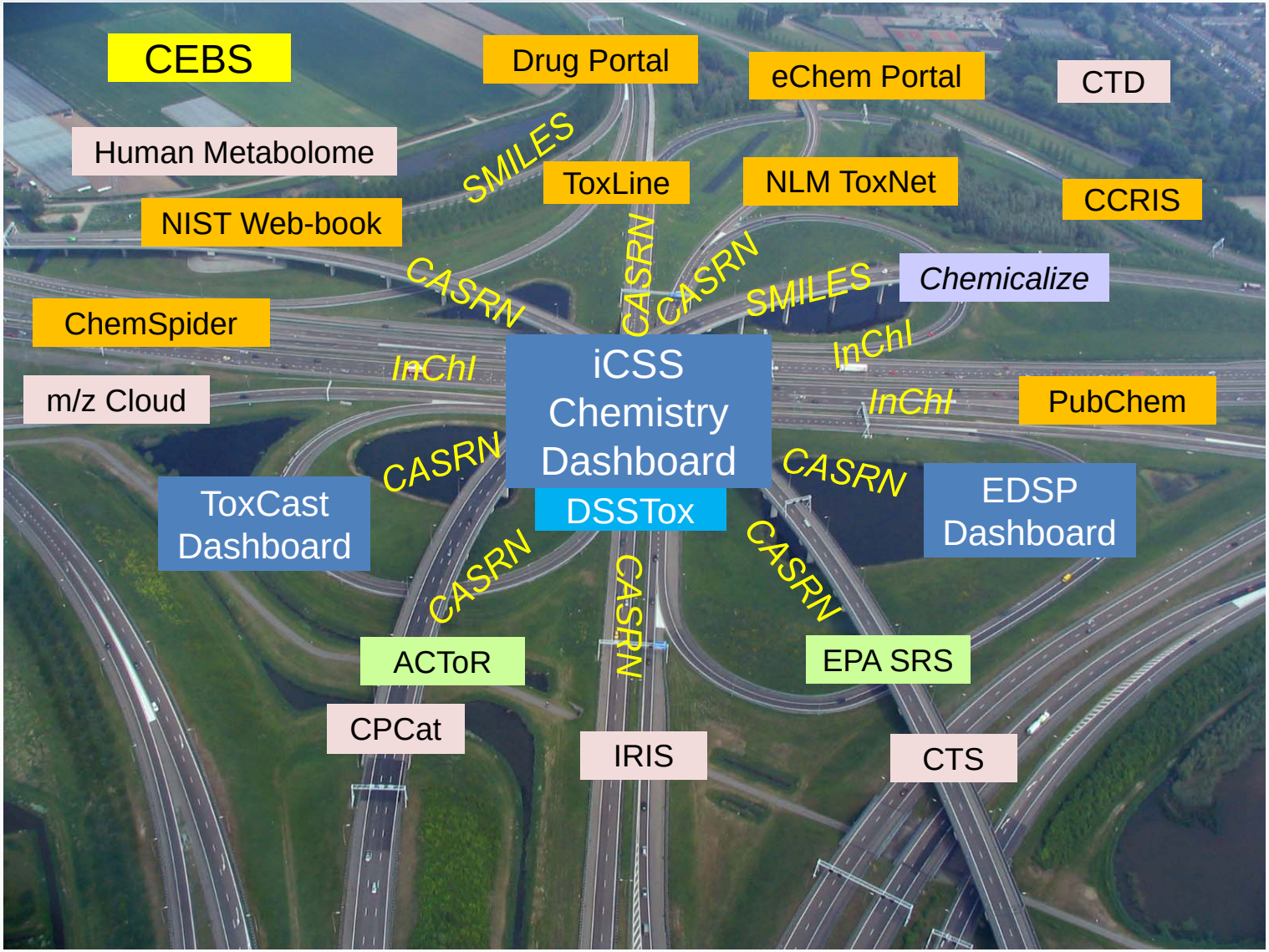


How can others contribute & benefit?

- Chemically indexed content is crucial!
- Encourage use of DSSTox content, chemical list registration tool, and use of quality metrics for chemical annotations of data
- Greater linkages, use of web-services, standardized chemistry resources, cheminformatics
- But chemical indexing, linkages, and annotations are NOT enough

→ Quality of chemical-data annotations needs to be improved!!!

A possible future for EPA …



Acknowlegements

NCCT Contributors to DSSTox & Chemistry Dashboard

- Chris Grulke
- Indira Thillainadarajah
- Antony Williams
- Richard Judson
- Kamel Mansouri
- Jayaram Kancherla
- Dave Lyons
- Marty Wolf