Enhancing the Interoperability of the Ecotoxicology (ECOTOX) Knowledgebase via Mapping to Existing Controlled Vocabularies and Ontologies

Jennifer H. Olker
U.S. Environmental Protection Agency
ORD/CCTE/GLTED
Duluth, MN

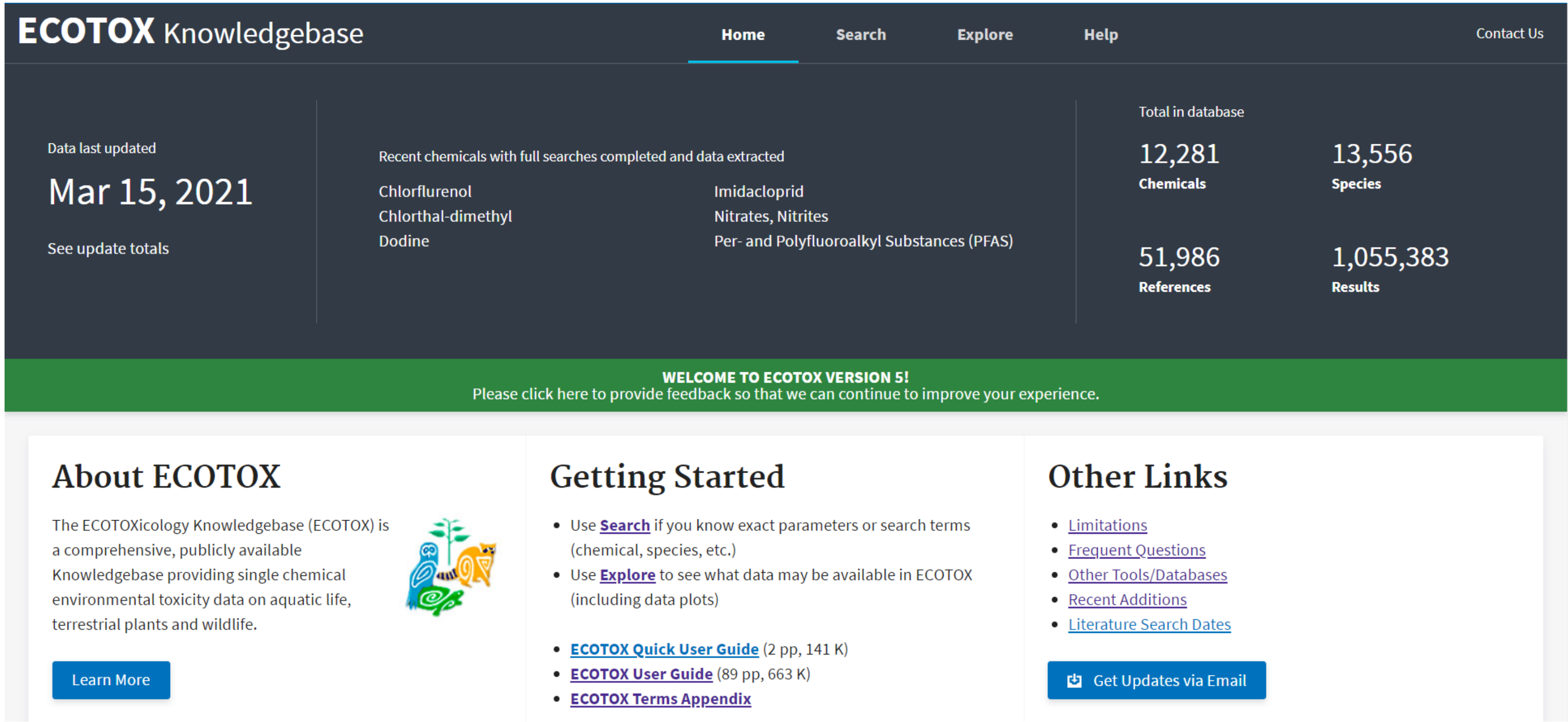
OpenTox 2021 Virtual Conference
20-24 September 2021

Office of Research and Development
Center for Computational Toxicology and Exposure

The views expressed in this abstract are solely those of the authors and do not represent the policies of EPA. Mention of trade names or commercial products should not be interpreted as an endorsement by EPA.

What is the ECOTOXicology Knowledgebase?

- 30+ years of reported single chemical toxicity effects data on aquatic and terrestrial organisms
- >1 million test results from >51,000 references
- Controlled vocabulary based on ecotoxicological literature



ECOTOX Knowledgebase

Data last updated

Mar 15, 2021

See update totals

Recent chemicals with full searches completed and data extracted

Chlorflurenol	Imidacloprid
Chlorthal-dimethyl	Nitrates, Nitrites
Dodine	Per- and Polyfluoroalkyl Substances (PFAS)

Total in database

12,281 Chemicals	13,556 Species
51,986 References	1,055,383 Results

WELCOME TO ECOTOX VERSION 5!
Please click here to provide feedback so that we can continue to improve your experience.

About ECOTOX

The ECOTOXicology Knowledgebase (ECOTOX) is a comprehensive, publicly available Knowledgebase providing single chemical environmental toxicity data on aquatic life, terrestrial plants and wildlife.

Getting Started

- Use **Search** if you know exact parameters or search terms (chemical, species, etc.)
- Use **Explore** to see what data may be available in ECOTOX (including data plots)

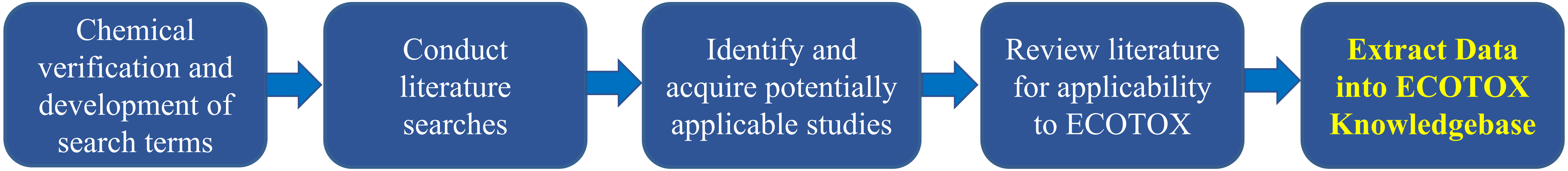
- **ECOTOX Quick User Guide** (2 pp, 141 K)
- **ECOTOX User Guide** (89 pp, 663 K)
- **ECOTOX Terms Appendix**

Other Links

- Limitations
- Frequent Questions
- Other Tools/Databases
- Recent Additions
- Literature Search Dates

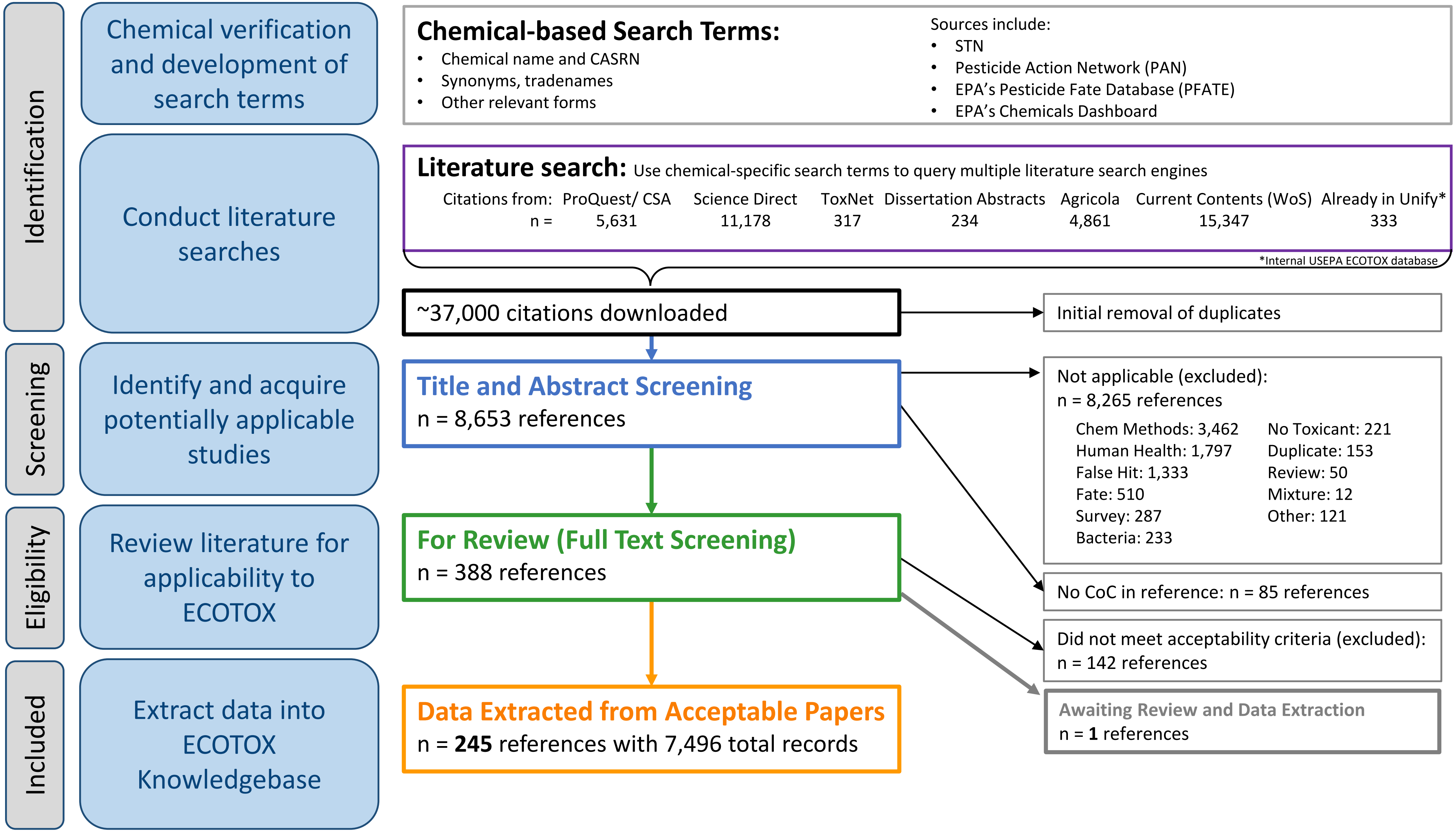
www.epa.gov/ecotox

ECOTOX Data Curation Pipeline



Chemical verification and development of search terms
Conduct literature searches
Identify and acquire potentially applicable studies
Review literature for applicability to ECOTOX
Extract Data into ECOTOX Knowledgebase

ECOTOX Pipeline: Systematic Process/Data Curation



Chemical verification and development of search terms

Chemical-based Search Terms:

- Chemical name and CASRN
- Synonyms, tradenames
- Other relevant forms

Sources include:

- STN
- Pesticide Action Network (PAN)
- EPA's Pesticide Fate Database (PFATE)
- EPA's Chemicals Dashboard

Conduct literature searches

Literature search: Use chemical-specific search terms to query multiple literature search engines

Citations from:	ProQuest/ CSA	Science Direct	ToxNet	Dissertation Abstracts	Agricola	Current Contents (WoS)	Already in Unify*
n =	5,631	11,178	317	234	4,861	15,347	333

*Internal USEPA ECOTOX database

~37,000 citations downloaded

Initial removal of duplicates

Identify and acquire potentially applicable studies

Title and Abstract Screening
n = 8,653 references

Not applicable (excluded):
n = 8,265 references

Chem Methods: 3,462	No Toxicant: 221
Human Health: 1,797	Duplicate: 153
False Hit: 1,333	Review: 50
Fate: 510	Mixture: 12
Survey: 287	Other: 121
Bacteria: 233	

No CoC in reference: n = 85 references

Review literature for applicability to ECOTOX

For Review (Full Text Screening)
n = 388 references

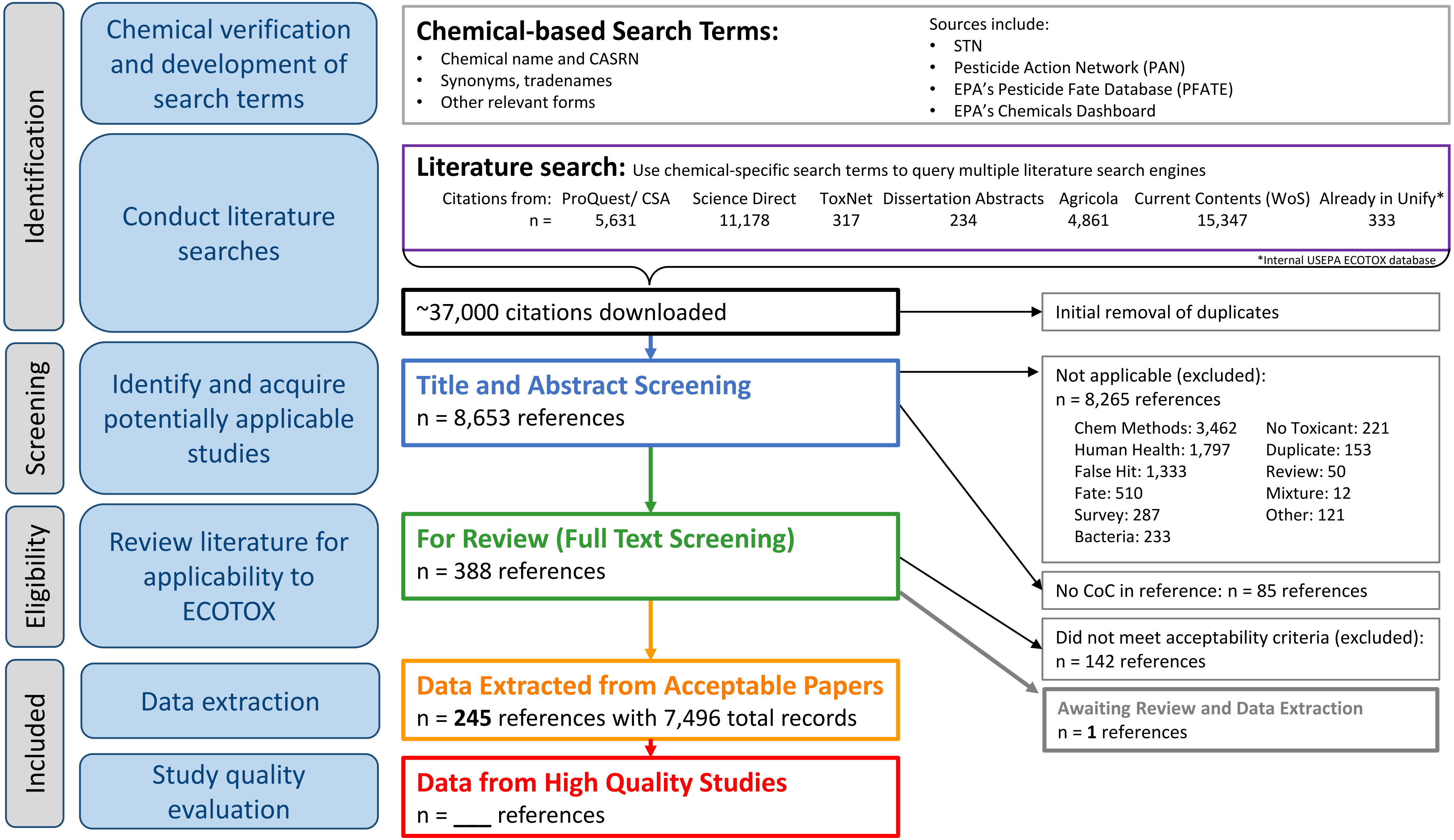
Did not meet acceptability criteria (excluded):
n = 142 references

Awaiting Review and Data Extraction
n = **1** references

Extract data into ECOTOX Knowledgebase

Data Extracted from Acceptable Papers
n = **245** references with 7,496 total records

ECOTOX Pipeline: Systematic Process/Data Curation



Identification

Chemical verification and development of search terms

Chemical-based Search Terms:

- Chemical name and CASRN
- Synonyms, tradenames
- Other relevant forms

Sources include:

- STN
- Pesticide Action Network (PAN)
- EPA's Pesticide Fate Database (PFATE)
- EPA's Chemicals Dashboard

Conduct literature searches

Literature search: Use chemical-specific search terms to query multiple literature search engines

Citations from:	ProQuest/ CSA	Science Direct	ToxNet	Dissertation Abstracts	Agricola	Current Contents (WoS)	Already in Unify*
n =	5,631	11,178	317	234	4,861	15,347	333

*Internal USEPA ECOTOX database

~37,000 citations downloaded

Initial removal of duplicates

Screening

Identify and acquire potentially applicable studies

Title and Abstract Screening
n = 8,653 references

Not applicable (excluded):
n = 8,265 references

Chem Methods: 3,462	No Toxicant: 221
Human Health: 1,797	Duplicate: 153
False Hit: 1,333	Review: 50
Fate: 510	Mixture: 12
Survey: 287	Other: 121
Bacteria: 233	

Eligibility

Review literature for applicability to ECOTOX

For Review (Full Text Screening)
n = 388 references

No CoC in reference: n = 85 references

Did not meet acceptability criteria (excluded):
n = 142 references

Awaiting Review and Data Extraction
n = **1** references

Included

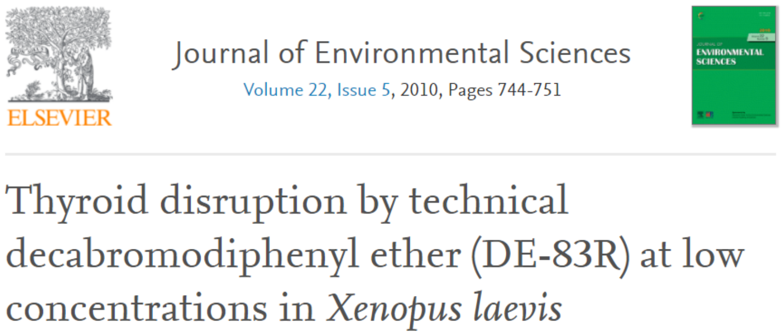
Data extraction

Data Extracted from Acceptable Papers
n = **245** references with 7,496 total records

Study quality evaluation

Data from High Quality Studies
n = ___ references

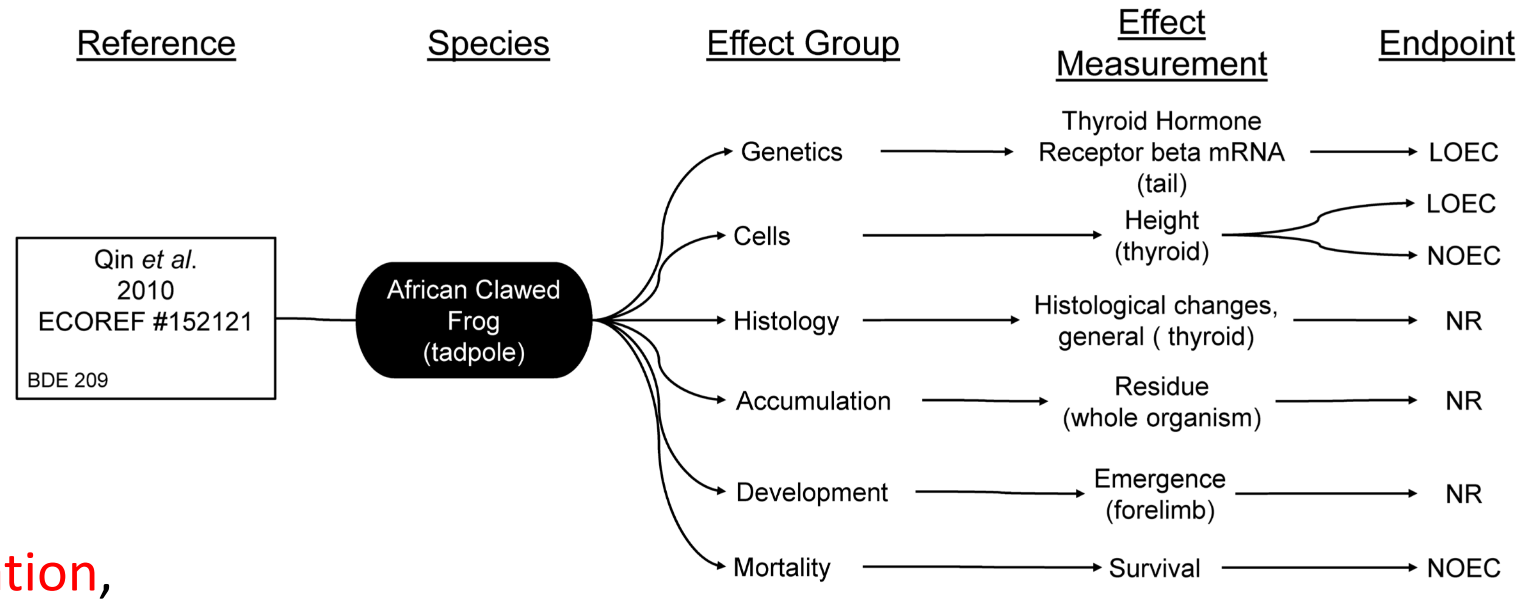
Example



ELSEVIER

Journal of Environmental Sciences
Volume 22, Issue 5, 2010, Pages 744-751

Thyroid disruption by technical decabromodiphenyl ether (DE-83R) at low concentrations in *Xenopus laevis*



Reference	Species	Effect Group	Effect Measurement	Endpoint
Qin *et al.* 2010 ECOREF #152121 BDE 209	African Clawed Frog (tadpole)	Genetics	Thyroid Hormone Receptor beta mRNA (tail)	LOEC
		Cells	Height (thyroid)	LOEC NOEC
		Histology	Histological changes, general (thyroid)	NR
		Accumulation	Residue (whole organism)	NR
		Development	Emergence (forelimb)	NR
		Mortality	Survival	NOEC

NOEC = No Observed Effect Level LOEC = Lowest Observed Effect Level NR = Not Reported

Additional Study and Result Fields

- ***Chemical***: Grade, Purity, Formulation, Radiolabel, Analysis, Carrier
- ***Test Organism:*** Life stage, Age, Source, Sex (Gender), Initial and Final Weight
- ***Study Design and Test Conditions:*** Location, Habitat, Media Type, Exposure Type, Test Method, Test Type, # of Doses, Doses, Experimental Design, Control, Initial sample size, Duration (Study, Exposure, Observed), Application Frequency, Water Quality or Soil Parameters

- ***Test Results***: Concentration, Concentration Type, Effect, Effect Measurement, Endpoint, Response Site, Effect Percent, Statistical Significance, Trend, Other Effects.

Red fields follow ECOTOX Vocabularies: https://cfpub.epa.gov/ecotox/help.cfm?sub=term-appendix

The Challenge: Data Accessibility and Interoperability

- Regulatory mandates require safety assessments for more chemicals, faster
- Demand for easy, efficient access to essential toxicology literature so that end users can rapidly identify critical data

Decision-makers and Researchers

From This:

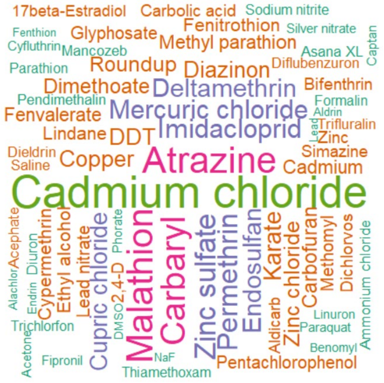
To This:

- Requires interoperability across tools and databases
  - Provide end users easy access to wealth of information
  - Increases efficiency
- New applications for data
  - Development of Adverse Outcome Pathways

The Challenge:
Data Accessibility and Interoperability

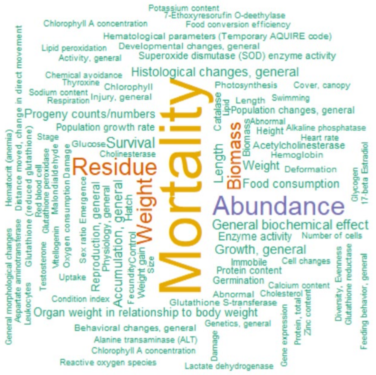
- ECOTOX contains an immense amount of single chemical toxicity data
- Extensive ECOTOX-specific vocabulary

12,382 Chemicals



13,598 Biological Species

6,209 Biological Effects



1,000s of Study and Result Terms

- **Chemical details** (e.g., Grade, Formulation, Verified Conc.)
- **Study design** (e.g., Duration, Test Method, Media Type Exposure Type)
- **Species characteristics** (e.g. Life stage, Source, Gender)
- **Result parameters** (e.g. Response Site, Endpoint)

The Solution:

- Harmonizing identifiers for species and chemicals
- Mapping ECOTOX terms to ontology classes

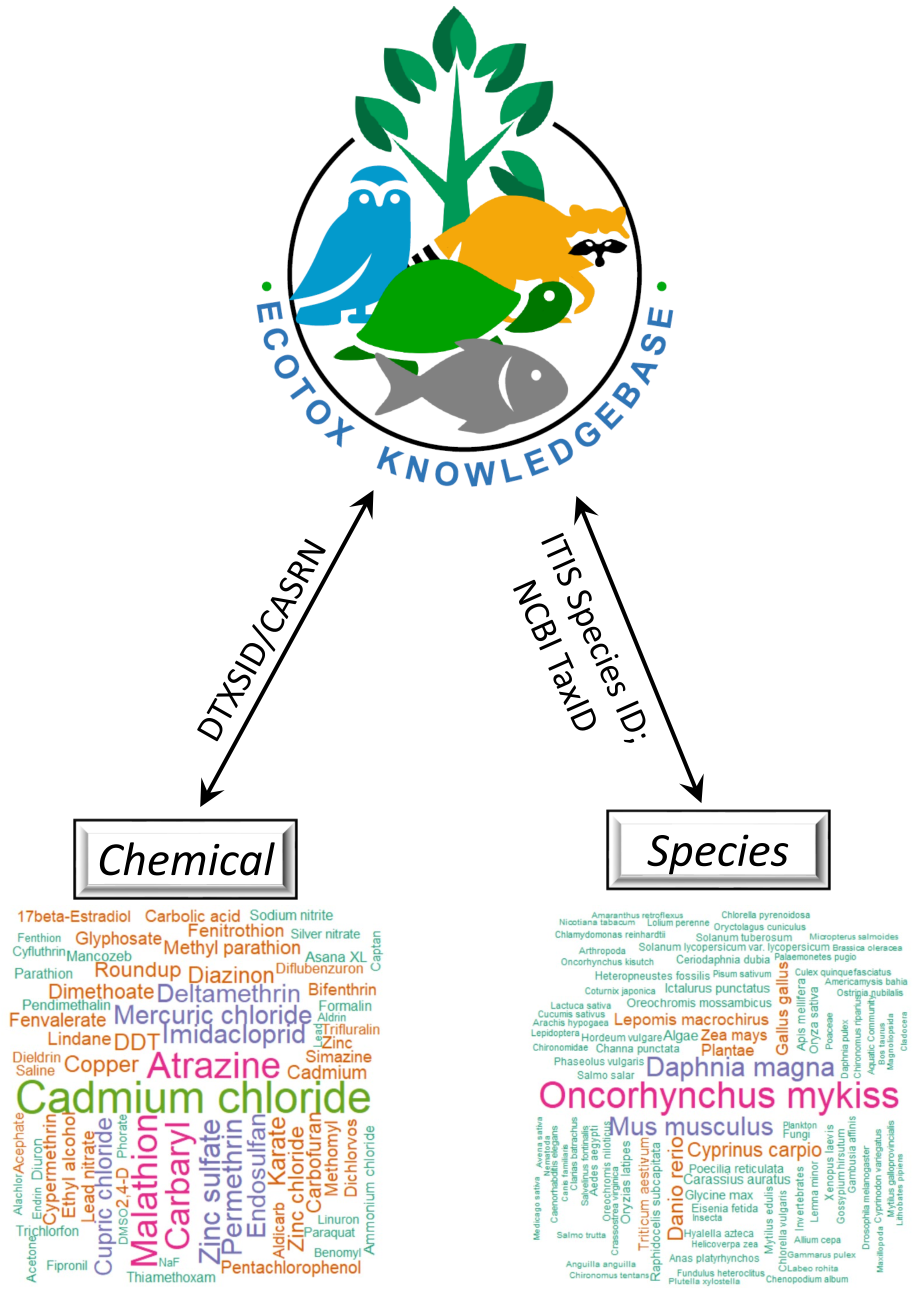
Harmonized Identifiers

Chemicals:

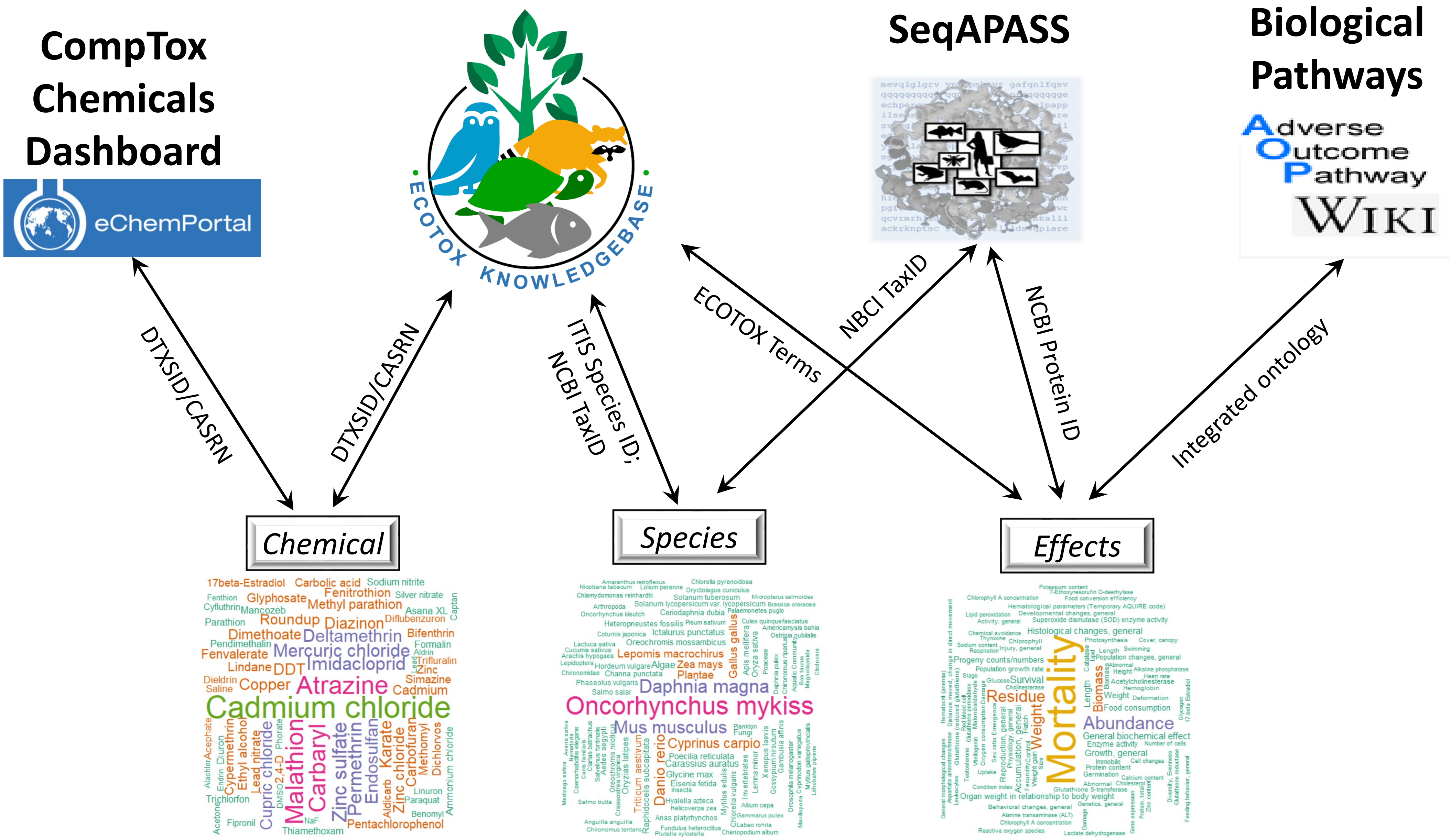
- CAS Registry Numbers (CASRNs)
- DSSTox Substance IDs (DTXSIDs)
  - Mapping current list
  - Registration of future additions

Species:

- USGS taxonomic serial numbers from the Integrated Taxonomic Information System (ITIS)
- NCBI taxonomic identifiers (taxids)
- Adding Taxonomic Hierarchy



Connections Across Databases and Tools



CompTox Chemicals Dashboard

SeqAPASS

Biological Pathways

DTXSID/CASRN

DTXSID/CASRN

ITIS Species ID; NCBI TaxID

ECOTOX Terms

NBCI TaxID

NCBI Protein ID

Integrated ontology

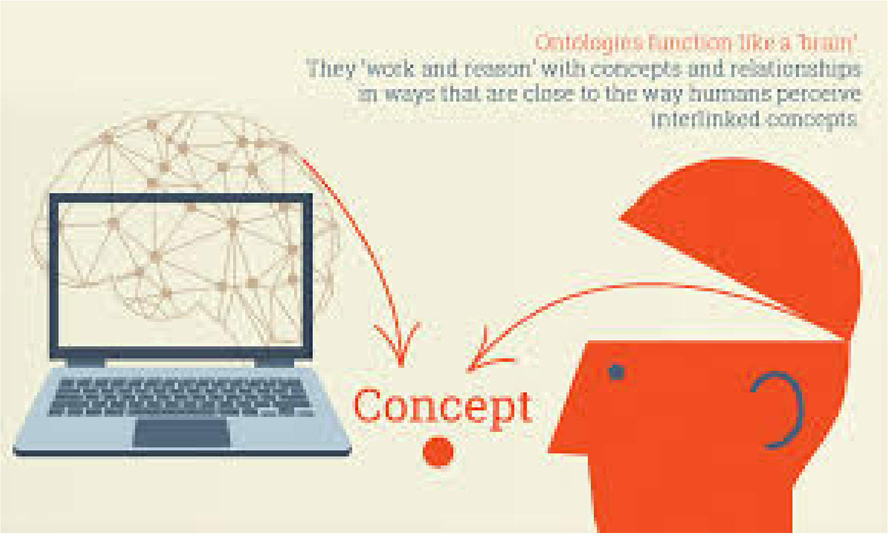
Chemical

Species

Effects

9

Mapping to Ontology Classes


Ontologies function like a 'brain'
They 'work and reason' with concepts and relationships
in ways that are close to the way humans perceive
interlinked concepts
Concept

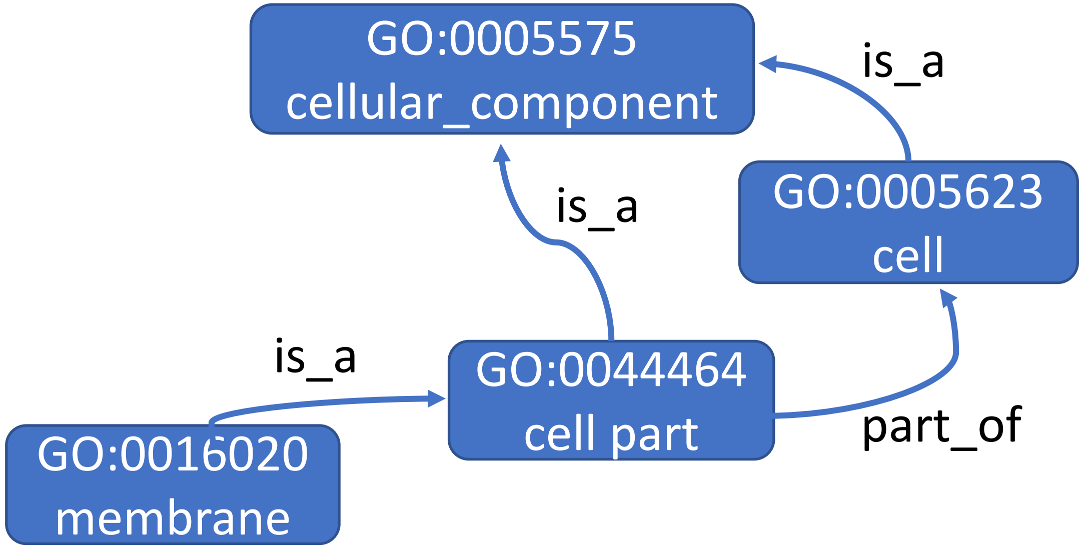
<u>An ontology is</u>:
Formal naming and definitions of a set of concepts within a domain and the relationships among them.

Both machine-readable and human interpretable.

Represent knowledge in sharable and reusable format, and can also add <u>new</u> knowledge about the domain.

Depicted as a graphical relationship

<u>Gene Ontology (GO) example</u>



GO:0005575
cellular_component

is_a

GO:0005623
cell

is_a

is_a

GO:0044464
cell part

part_of

GO:0016020
membrane

https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

Mapping to Ontology Classes

Development of BioPortal Lookup Tool - (led by Kellie Fay):

- BioPortal Ontology Browser (https://bioportal.bioontology.org)
- Java-based tool that uses REST API Services
- Makes use of BioPortal's Annotator and Recommender features
- Output from individual ontologies, including preferred class name, synonyms, definition, parent class, etc

- **Manual review for quality of mapping**
  - Appropriate context evaluated with textual definition, synonyms, parent class, etc.
  - Applicability scores developed and applied

Initial BioPortal Mapping Results

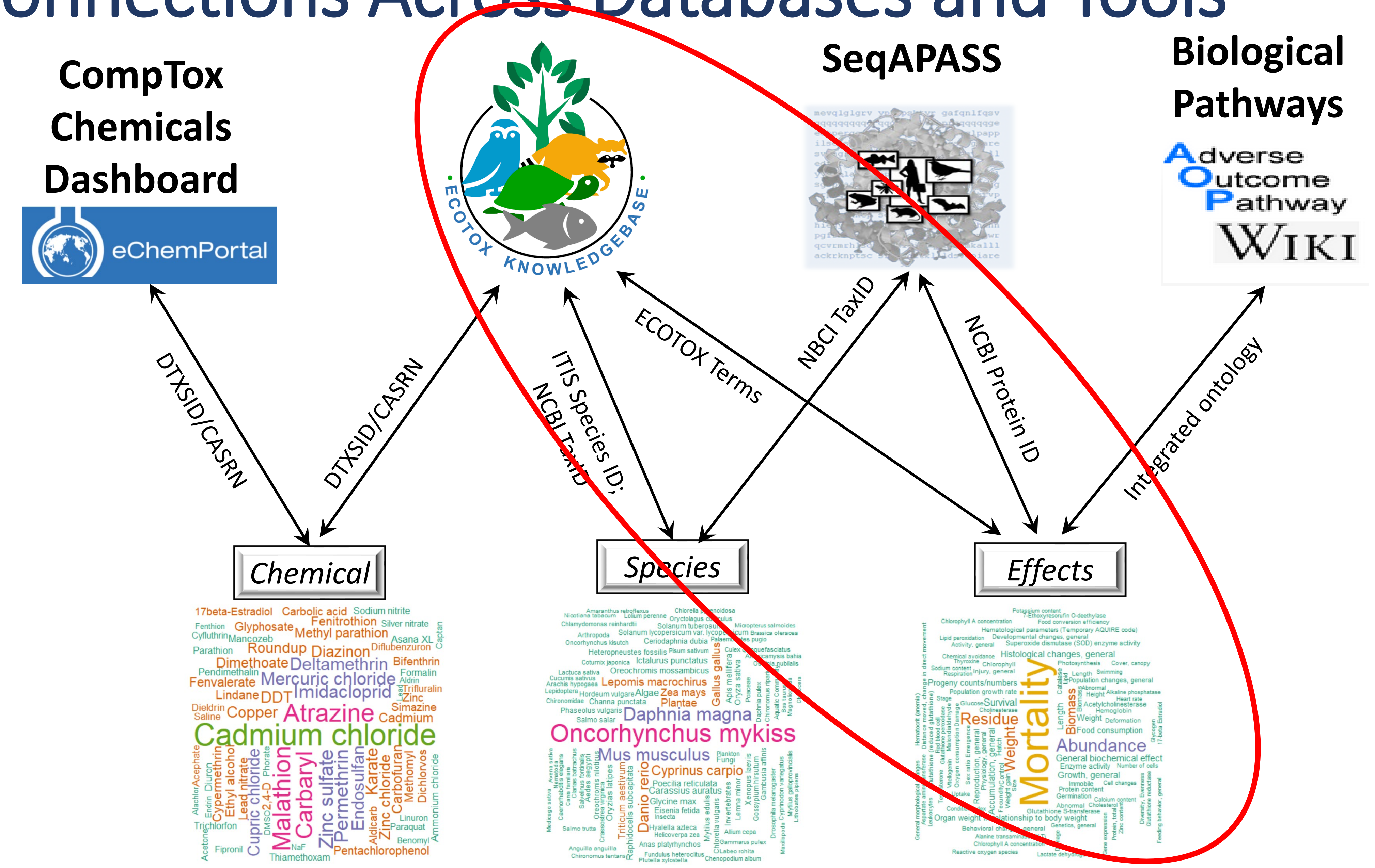
Successes:

- 3526/8626 = 41% were "successfully" mapped to at least one ontology class
- Some categories were 100% successful (Trend, Weight, Substrate, Seasons)
- 1991/4090 = 49% of Effect Measurement codes successfully mapped

Challenges:

- Several major categories of unmapped terms
  - Complex terms or representation (e.g., 'Concurrent control', 'Adult(s)', 'Mitotic abnormalities, micronuclei')
  - Ecotoxicology/Ecology terms which do not appear to be present in existing ontologies (e.g., 'LC50', 'froglet', 'instar')
  - Messenger RNA effect measurements (n = 1383 at time of mapping)

Connections Across Databases and Tools



Gene and Protein Mapping

Goal: Incorporate gene and protein IDs for relevant ECOTOX effect measurements

- Species-specific IDs (e.g., UniProt ID, NCBI accession numbers)

Approach: Use maintain general ECOTOX effect measurement terms and definitions to map to available resources from NCBI, UniGene, and UniProt

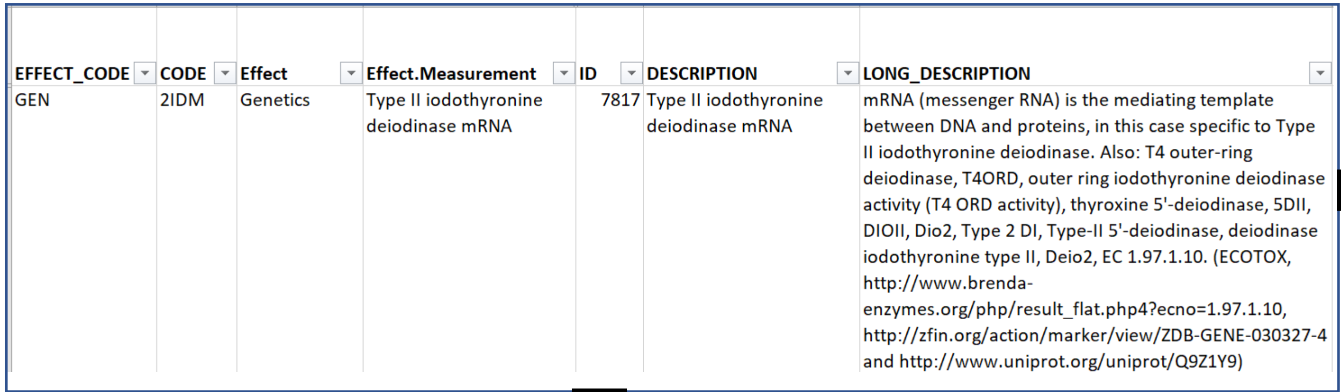
- Gene symbols included in definitions used to query NCBI Homologenes
- Map to species-species NCBI Gene IDs

Status: *In progress*

- Pilot with subset of zebrafish toxicity results
- Initial mapping successful for >50% of genes with detailed descriptions
- Additional term details and accession numbers for mapping review

Gene and Protein Mapping

ECOTOX – Effect Measurement Terms

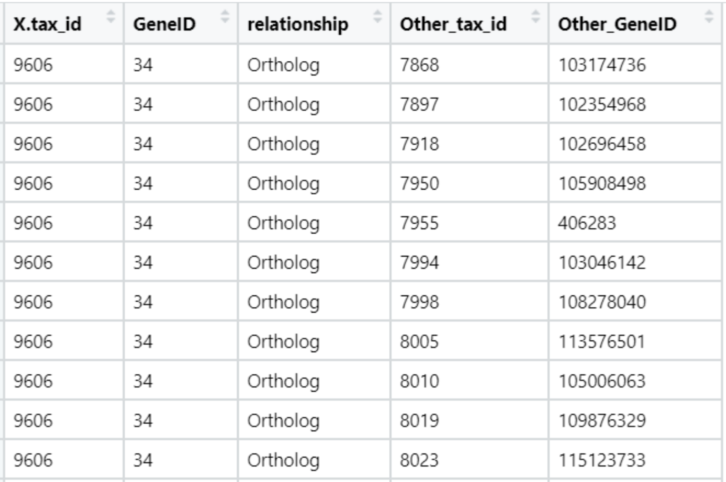


EFFECT_CODE	CODE	Effect	Effect.Measurement	ID	DESCRIPTION	LONG_DESCRIPTION
GEN	2IDM	Genetics	Type II iodothyronine deiodinase mRNA	7817	Type II iodothyronine deiodinase mRNA	mRNA (messenger RNA) is the mediating template between DNA and proteins, in this case specific to Type II iodothyronine deiodinase. Also: T4 outer-ring deiodinase, T4ORD, outer ring iodothyronine deiodinase activity (T4 ORD activity), thyroxine 5'-deiodinase, 5DII, DIOII, Dio2, Type 2 DI, Type-II 5'-deiodinase, deiodinase iodothyronine type II, Deio2, EC 1.97.1.10. (ECOTOX, http://www.brenda-enzymes.org/php/result_flat.php4?ecno=1.97.1.10, http://zfin.org/action/marker/view/ZDB-GENE-030327-4 and http://www.uniprot.org/uniprot/Q9Z1Y9)

'Reference' Species

NCBI gene_orthologs table

- 9606 (human)
- 7955 (zebrafish)
- 9031 (chicken)
- 9615 (dog)
- 9685 (cat)
- 9823 (pig)
- 9913 (cow)
- 10090 (mouse)

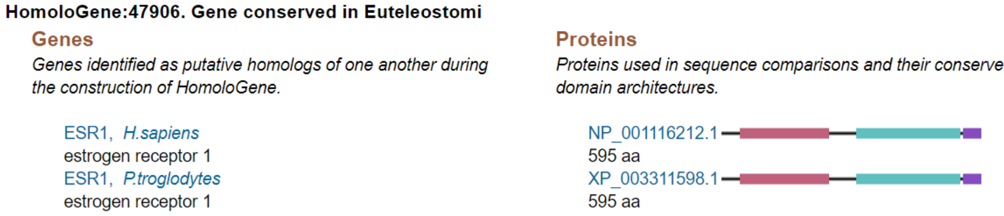


X.tax_id	GeneID	relationship	Other_tax_id	Other_GeneID
9606	34	Ortholog	7868	103174736
9606	34	Ortholog	7897	102354968
9606	34	Ortholog	7918	102696458
9606	34	Ortholog	7950	105908498
9606	34	Ortholog	7955	406283
9606	34	Ortholog	7994	103046142
9606	34	Ortholog	7998	108278040
9606	34	Ortholog	8005	113576501
9606	34	Ortholog	8010	105006063
9606	34	Ortholog	8019	109876329
9606	34	Ortholog	8023	115123733

Other_tax_id includes ~300 species

Gene/Protein Clusters

HomoloGene



HomoloGene:47906. Gene conserved in Euteleostomi

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

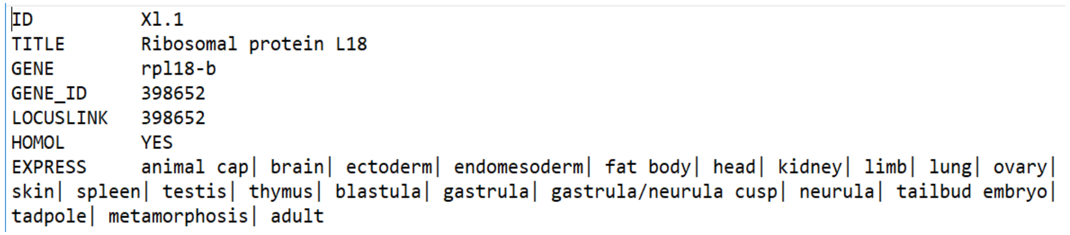
ESR1, *H.sapiens*
estrogen receptor 1
ESR1, *P.troglodytes*
estrogen receptor 1

Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

NP_001116212.1
595 aa
XP_003311598.1
595 aa

UniGene



```
ID          X1.1
TITLE       Ribosomal protein L18
GENE        rpl18-b
GENE_ID     398652
LOCUSLINK   398652
HOMOL       YES
EXPRESS     animal cap| brain| ectoderm| endomesoderm| fat body| head| kidney| limb| lung| ovary|
skin| spleen| testis| thymus| blastula| gastrula| gastrula/neurula cusp| neurula| tailbud embryo|
tadpole| metamorphosis| adult
```

Entrez GeneIDs for each Species

Effect Measurement	GeneID	Symbol	UniProt Entry
Thyroid Hormone Receptor beta mRNA	30607	thrb	Q9PVE4
Transcription factor 3a mRNA	30310	tcf3a	Q90491
Transthyretin (prealbumin, amyloidosis type I) mRNA	449556	ttr	B8JLL8
Type II iodothyronine deiodinase mRNA	352937	dio2	Q6PBR8

Pathways

reactome

Diseases

OMIM

Continuing Work…

Harmonized Identifiers:

- Routine updates for Species and Chemical IDs
- Addition of gene and protein IDs

ECOTOX Controlled Vocabulary:

- Updating and simplifying terms
  - Address redundance, non-standard, overly-specific terms
- Establishing hierarchical relationships

Mapping to Ontology Classes:

- Refresh the mapping of updated ECOTOX terms
- Focus on OBO (Open Biological and Biomedical Ontology) domain ontologies
- Further development and maintenance of BioPortal Lookup Tool?

Anticipated Applications

- Enhanced ECOTOX querying

- Text mining to inform systematic review
- Text mining to support semi-automated data extraction

- Computational predictions of chemical effects
  - Adverse Outcome Pathway development
  - Phenotypic profiles of chemical toxicity (e.g., Wang et al. 2018, https://doi.org/10.1016/j.tox.2018.11.005)

- Development ECOTOX application ontology

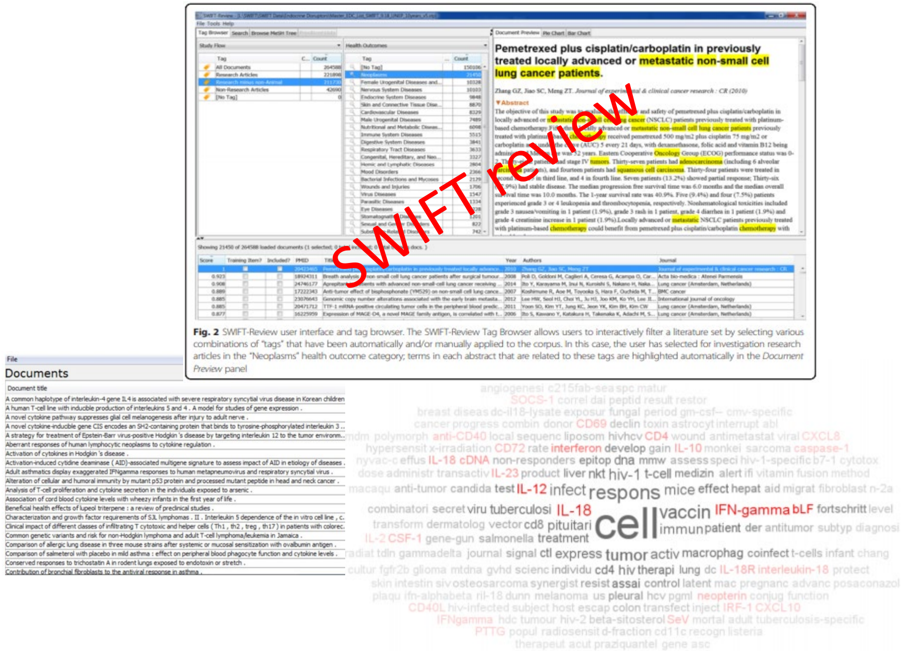
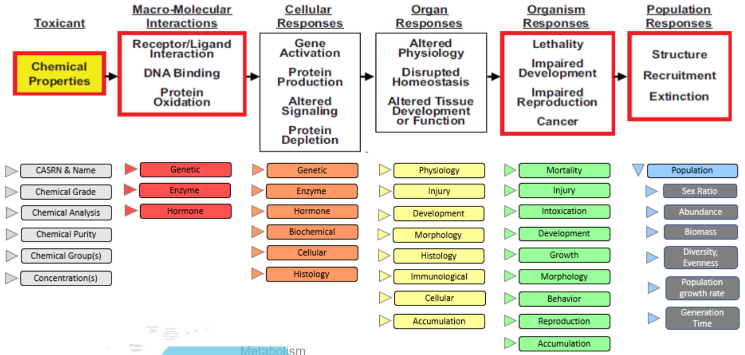
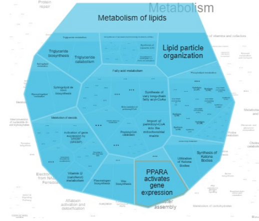

SWIFT review
Pemetrexed plus cisplatin/carboplatin in previously treated locally advanced or metastatic non-small cell lung cancer patients.
Documents
cell

Fig. 2 SWIFT-Review user interface and tag browser. The SWIFT-Review Tag Browser allows users to interactively filter a literature set by selecting various combinations of "tags" that have been automatically and/or manually applied to the corpus. In this case, the user has selected for investigation research articles in the "Neoplasms" health outcome category; terms in each abstract that are related to these tags are highlighted automatically in the *Document Preview* panel




Metabolism
Metabolism of lipids
Lipid particle organization
Triglyceride biosynthesis
Triglyceride catabolism
PPARA activates gene expression

Acknowledgements

U.S. EPA ORD, CCTE

Carlie LaLone*

Rong-Lin Wang*

Colleen Elonen*

Dale Hoff*

U.S. EPA OCSPP, OPPT

Kellie Fay*

General Dynamics Information Technology

Michael Skopinski*

Travis Karschnik*

Anne Pilli*

Brian Kinziger

Thank you!

olker.jennifer@epa.gov

*Coauthors