Strategy for Developing Literature Search Strings to Identify Publications Containing Environmental Fate & Transport and Physical-Chemical Property Data

Amina Wilkins*, Danielle Moore*, Marcy Card†, Nerija Orentas†, Hillary Hollinger*, Ryan Jones*, Kristina Thayer*, Elaine Cohen Hubal*, Peter Egeghy^, Cory Strope†, Kristan Markey#, John Wambaugh^

*U.S. EPA, ORD, Center for Public Health & Environmental Assessment †U.S. EPA, OCSPP, Office of Pollution Prevention & Toxics #U.S. EPA OCSPP, Office of Science Coordination & Policy ^U.S. EPA, ORD Center for Computational Toxicology & Exposure

www.epa.gov

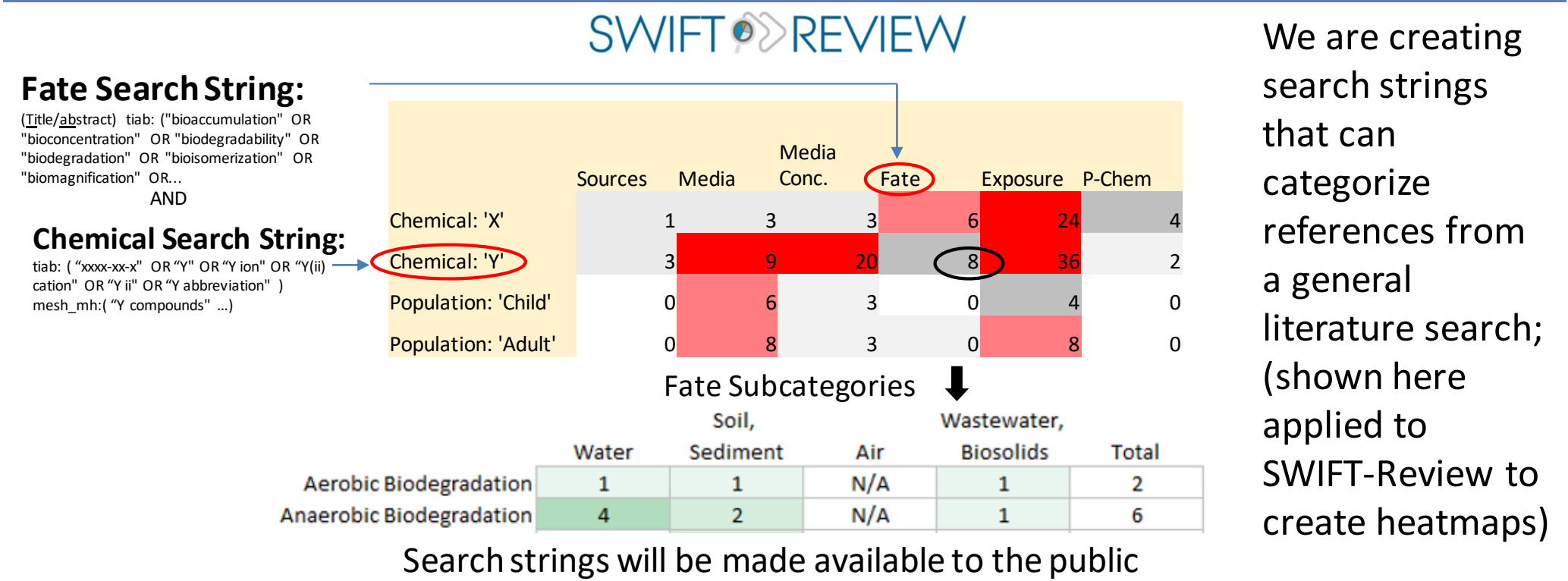Amina Wilkins I Wilkins.Amina@epa.gov I 202-564-1224

1. Background

- Literature search strings are being created to facilitate development of TSCA Risk Evaluations.
- For the first 10 risk evaluations, discipline-specific literature searches did not identify all relevant literature. A second round of searches was performed to improve this data gathering step with targeted literature searches related to the chemicals regardless of discipline.
- For the next 20 chemicals, a broad search of peer-reviewed literature was completed for each chemical. New p-chem and modified fate search strings were needed to filter those results.

2. Purpose

- Provide chemical agnostic search strings that identify references containing fate and p-chem information (that can be used in PubMed, Web of Science (WoS), SWIFT-Review (see Section 5), etc.) to retrieve the greatest number of on-topic references while decreasing the number of off-topic references.
- Facilitate the systematic reviews (SR) of scientific evidence needed for transparent chemical assessments.
- Utilize skilled librarians to develop and test search strings; to help standardize approach for systematic review.

3. End Product

Fate Search String:
(Title/abstract) tiab: ("bioaccumulation" OR "bioconcentration" OR "biodegradability" OR "biodegradation" OR "bioisomerization" OR "biomagnification" OR... AND

Chemical Search String:
tiab: ("xxxx-xx-x" OR "Y" OR "Y ion" OR "Y(ii) cation" OR "Y ii" OR "Y abbreviation") mesh_mh: ("Y compounds" ...)

We are creating search strings that can categorize references from a general literature search; (shown here applied to SWIFT-Review to create heatmaps)

Search strings will be made available to the public

4. Approach

Assembled a team of agency exposure and fate experts and librarians who provided a prioritized list of relevant keywords and identified 'on-topic' references to create training and test sets.

Methods for crafting search strings included both a traditional approach (iterative between the librarian and subject experts), and an approach utilizing keyword analysis tools (KAT & SWIFT-Review's fingerprinting features) and on-topic reference lists.

5. Tools

SWIFT REVIEW

A free software program that allows reviewers to collaboratively screen, tag and transparently document results.

- Fingerprinting feature returns the root words and frequency of occurrence among a selected set of references.

ICF Keyword Analysis Tool

KAT is a software tool that searches a bibliographic file in RIS format for the occurrence of a specified keyword(s) within a reference's title or abstract.

- Match and no-match EndNote files: Provides the number of occurrences of each keyword within each individual reference
- Provides statistics on the number and percentage of documents across entire reference set containing each keyword
- Provides the number of documents in which the keyword uniquely occurs

6. Strategy

Physical-Chemistry String

1. Keyword selection: Librarian started with a list of keywords provided by fate experts, and discussed variations of terms to account for differences (e.g., synonyms, etc).

2. Database testing: Crafted various search strings using 'physical form' keywords and applied them to WoS and PubMed, then included additional terms to target 'physical state/properties' and 'physical chemistry'. These keyword additions decreased the number of references resulting from the 'physical form' keyword search. Fate scientists tested the draft string, requested additional terms, and then re-tested.

3. SWIFT Testing: Once the search strings were determined, they were translated into SWIFT-Review syntax and tested against several projects in SWIFT to ensure efficiency.
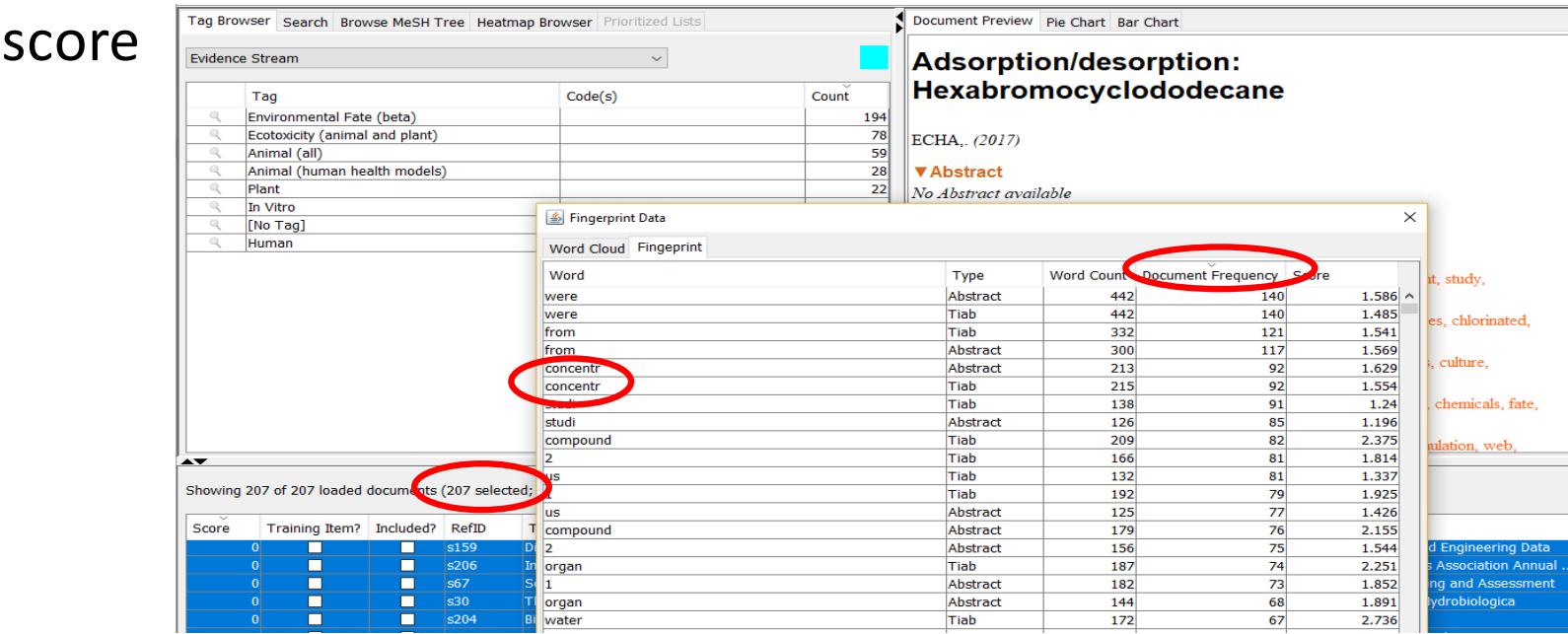
Fate Search String

1. Identified list of on-topic fate references and keywords
   Fate references: 207 high-confidence, on-topic
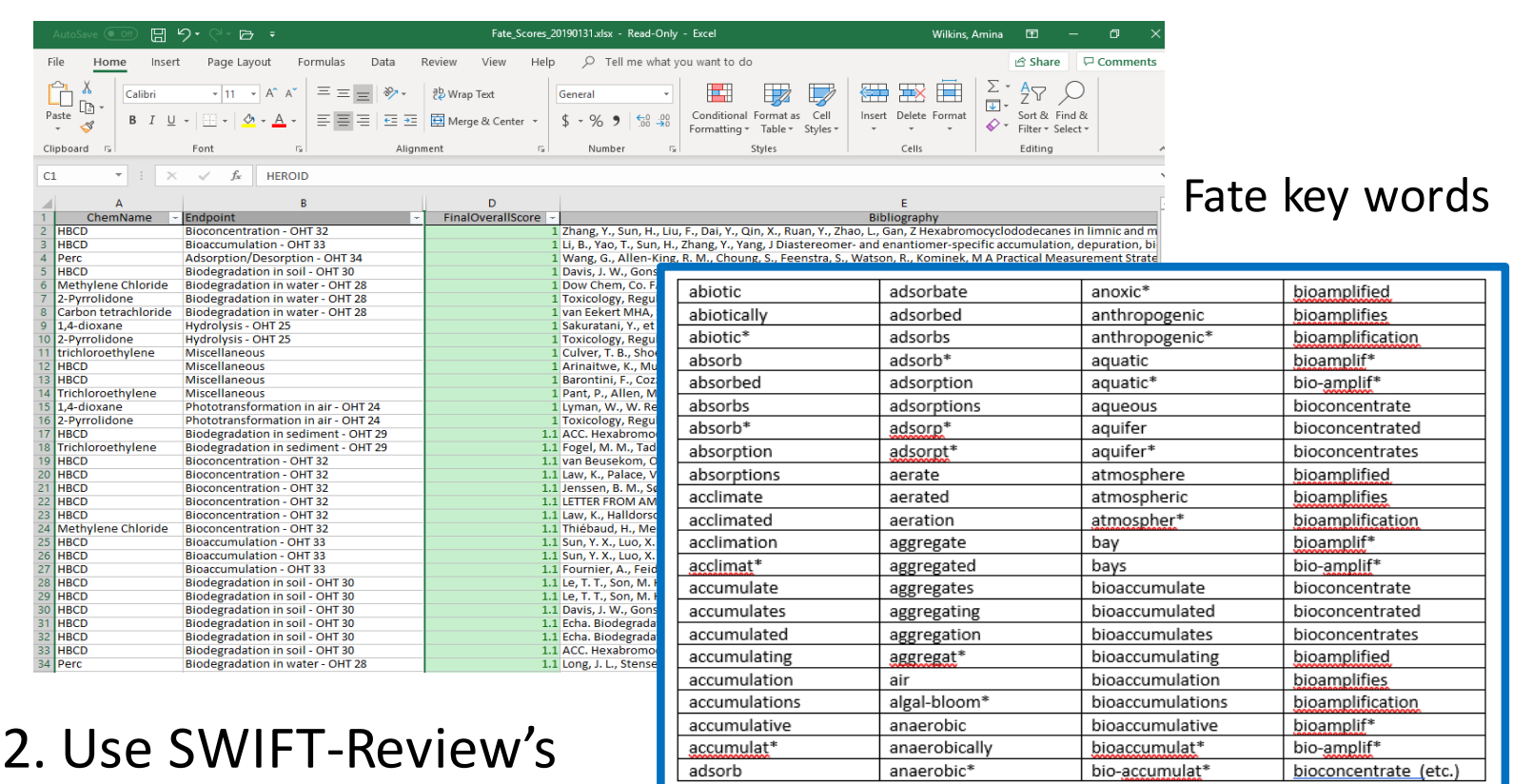
Fate key words

2. Use SWIFT-Review's *Fingerprint* to get a list of all words/roots occuring in titles and abstracts along with word counts, document frequency, and score

3. KAT is used to analyze the distribution of keywords in the combined title and abstract of literature input in .RIS format

Use an Excel file with macros to run KAT

Inputs specified in Value column:
- RIS file name
- List of keywords separated by semicolons (431 terms)

Click Run

4. KAT Output: EndNote Match File (shows keyword occurrence in each reference) .CSV File (Helps identify most useful and necessary terms)

Realization: Use KAT to create a "minimal" search string: Identify terms that are *necessary* or *unambiguous to* attempt to retrieve > # on-topic references and < # off-topic references

Application of Fate Search Strings: The 10,800 o- and p-dichlorobenzene references found by searching the chemical name were reduced to 2,600 by applying the fate strings. SWIFT Active Screener machine learning feature identified that ~12% of the 2,600 references (approximately 350 were on-topic references. Fate expert confirmed satisfactory results; 'extraneous' refs were fate-related just not the very specific type desired (biodegradation and/or groundwater mobility studies). Months later, terms to identify destruction of a chemical via incineration were added.

7. Results & Evaluation

	Broad Search String (152 terms)	The "Minimal" Search String (35 terms)
Fate — PubMed.gov	Finds 98 out of 98 on-topic references. Retrieves nearly 12.8M references	Finds 97 out of 98 on-topic references (PM missing 1 abstract). Retrieves 1.18M references
WEB OF SCIENCE	Finds 158 out of 159 on-topic references. Retrieves nearly 20.2M references	Finds 152 out of 159 on-topic references (WoS missing 7 abstracts). Retrieves 2.9M references

P-Chem Strings

tiab: (("physical form" OR "physical state" OR "physical chemistry" OR "physical properties") AND ("solid" OR "liquid" OR "gaseous state" OR "color" OR "scent" OR "odor" OR "odour" OR "smell")) OR "melting point" OR "boiling point" OR "density" OR "vapor pressure" OR "vapor density" OR "water...

Fate Strings

tiab:("bioaccumulation" OR"bioconcentration" OR "biodegradability" OR "biodegradation" OR "bioisomerization" OR "biomagnification" OR"biotransformation" OR "dechlorination" OR "degradation" OR "dehalogenation" OR "fate" OR "food web" OR "groundwater" OR "hydrolysis" OR...

Conclusion: The minimal search string found most key references while retrieving significantly fewer off-topic references compared to the broad search term.

National Academy of Sciences Meeting, Toxic Substances Control Act (TSCA) Systematic Review, June, 2020, Washington, D.C.