Data Activity Group

Update for the CCL Work Group

Plenary Meeting

February 5-6, 2003

Data Activity Work Group Members

- Rick Becker
- Wendy Heiger-Bernays
- Jeff Griffiths
- Buck Henderson
- Nancy Kim
- Benson Kirkman
- Gary Lynch
- Ken Merry
- Graciela Ramirez-Toro
- Jamie Bartram*

* - did not participate in January

Process since December 16-17 Plenary

- Weekly 1-2 hr conf. calls (4) beginning 1/9/03
- Other conf. call participants:
  - Tom Carpenter, Karen Wirth, and selected EPA staff
  - Joanne Shatkin, Charlie Pittinger and other Cadmus staff
  - Mike Focazio, USGS
  - Steve Via, AWWA
  - Sara Litke, RESOLVE
  - Doug Owen, Malcolm Pirnie

Deliverables Scheduled for February

- Characterization of available data sources for chemicals
- Characterization of available data sources for microbes
- Draft chemical data elements desired for populating the universe
- Draft microbial data elements desired for populating the universe

Topics Discussed in January

- Defining the Universe
- Data elements for chemicals
- Data sources to build the Universe for chemicals
- A process to address emerging contaminants
- Criteria for including data in the Universe

Topics For Future Discussions

- Data elements for microbes
  - Draft data elements for pathogens developed
  - Graciela and Jeff will be providing guidance
- Data sources to build the Universe for microbes
  - “Micro commonalities: Occurrence” paper prepared
  - “Micro commonalities: Health effects” paper being prepared
- Straw criteria for including microbe data in the Universe

Materials Reviewed

- Overview of NRC Recommendations and Data Elements
- Future CCL Data Elements Evaluation Phase 1: Occurrence Elements (“occurrence commonalities”)
- Process and Criteria for Screening the "Universe" of Potential Drinking Water Contaminants to a Preliminary Contaminant Candidate List (PCCL)
- Dimensions of the Chemical Universe
- Top-Down Versus Bottom-Up Database Approaches for Defining the CCL Universe
- Data Gaps Issue Paper Outline

Materials Produced

- Draft data elements list for chemicals
- Emerging contaminants discussion proposal

Approaches for Constructing the Universe

- "Top Down" - begin with all chemicals/microbes known and/or envisioned. Reduce the list to a "manageable" Universe through some form of screening (e.g., start with the Chemical Abstract Services (CAS) database and apply filtering criteria)

- "Bottom Up" – merge or recombine discrete databases/data sources to compile a set of records with multiple criteria (e.g., merge databases for High Production Volume Chemicals with the registry of Toxic Effects of Chemical Substances)

Exhibit 1: Conceptions of the Chemical Universe

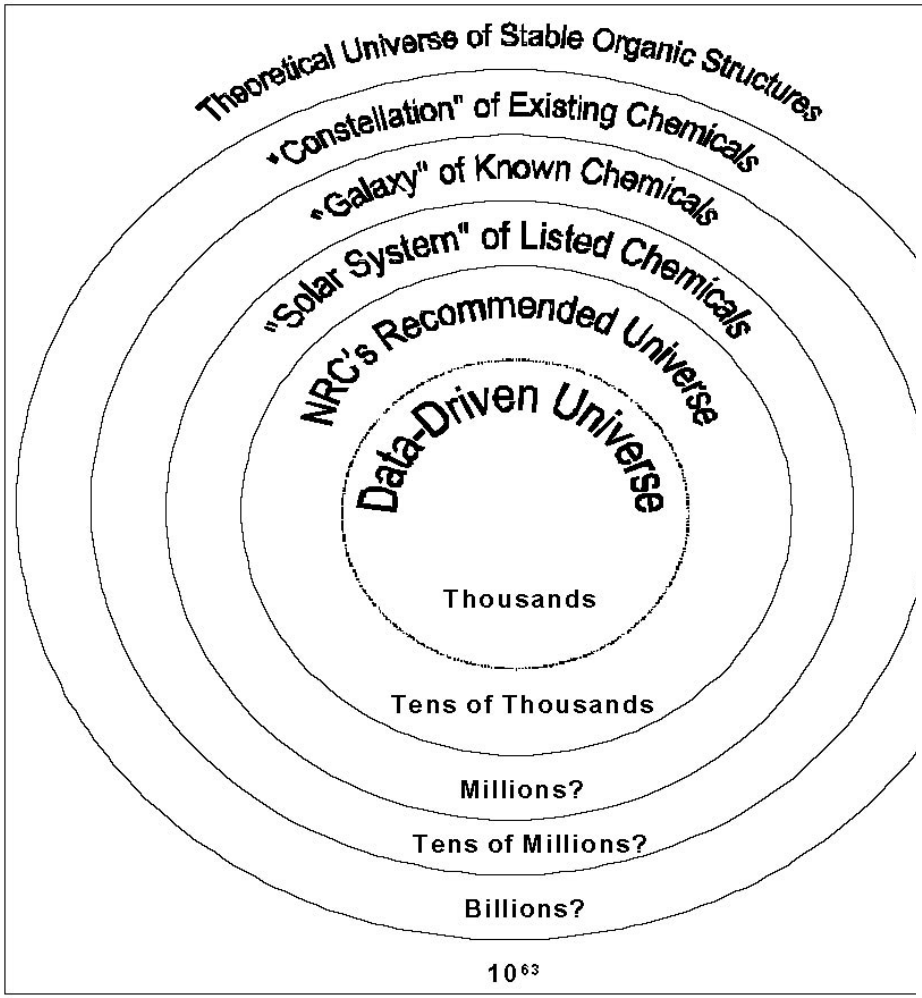

Exhibit 2: Example Databases for Dimensioning the Chemical Universe

Dimension	Approximate Size	Example Databases	Number of Chemicals Represented	Typical Information Available
Theoretical Universe	$>10^{63}$	-	-	-
Constellation of Existing Chemicals	Billions?	Unknown Chemicals - None Known - See Below	-	-
Galaxy of Known Chemicals	Tens of Millions	CAS	>44.5 Million	CASRN, structure chemical property, toxicity (more rarely)
		Beilstein	~ 8.4 Million	chemical properties, toxicity
Solar System of Known Chemicals with relevant data	Thousands to hundreds of thousands	RTECS	>150,000	toxicity
		EINECS/ELINCS	>100,00	identity, minimal property data
		EDPSD	>87,000	occurrence, toxicity
		HPV	~2,800	some chemical properties TSCA Inventory and Updates ~ 75,000 list only
		MRCK	~10,250	chemical properties, toxicity
		TSCA Inventory and Updates	~ 75,000	list only
		TSCATS	~8,000	some health data, environmental fate, exposure; HPV ~2,800 some chemical properties

Challenges with Either Approach

- Updating the Universe
  - New information is developed rapidly & reproducibility is a "moving target" – impacts transparency
- Cross-referencing
  - Unique identifiers for chemicals/microbes may not be compatible among databases
- Synonym and homolog confusion
  - It is easy to unintentionally omit data or interject redundant data because of inconsistency in identifiers for agents

Activity Group Recommendation

- Use the “bottom up” approach
  - Logistics – the sheer magnitude of the data is less and there are more data elements per record. Reduces unwarranted time and effort to search and evaluate large databases with limited and potentially irrelevant information.
  - Selectivity – records are pre-screened for inclusion in pre-existing data sources on the basis of key attributes.
  - Searchability – discrete data sources are typically designed to allow for specialized searches.
  - To date – Cadmus & WG have identified > 140 databases & data sources

Issues with "Bottom Up" Approach

- Compounds with known issues/data are more likely to be included.
- Screening criteria for including data in a specific data source may not coincide with overall goals for constructing the Universe.
- Recombined data sources are only as current and accurate as the least robust source.

Data Elements

- The group prepared a draft list of chemical data elements, including general information, health effects, and occurrence elements.

- The list can be considered a “work in progress” and will continue to be developed as data sources are reviewed and compared to evaluate commonalities.

What is an "Emerging Contaminant"?

- A potential contaminant for which:
  - Detection limits have improved so it is now being measured in quantifiable amounts
  - Information is being developed but has not been compiled into a recognized data source
- Issues:
  - How often is the Universe updated?
  - How does emerging contaminant review schedule coordinate with CCL update?

Emerging Contaminants Discussion Proposal

- The goal is to capture agents that are not sufficiently characterized at the time of the construction of the Universe
- Identification of a process that is distinct from the CCL process
- This process should probably be carried out by EPA

Next Steps

- Select chemical data sources from revised "A4: Database Review Table"
- Evaluate proposed chemical data elements to determine their availability in the data sources
- Prepare draft microbe data elements and select data sources from revised "A4: Database Review Table"
- Refine emerging contaminant definition and approach