Data Extraction: 40 Chemicals from 18 Sources

Report for the NDWAC CCL Work Group

Plenary Meeting

September 17-18, 2003

Attribute Scoring Data Set

- Purpose: extract and organize data for testing proposed attribute scoring approaches for PCCL to CCL classification
- 40 chemicals with a range of data availability
- 18 data sources with a range of data types and formats
- Identify data extraction issues

Attribute Scoring Data Set - 10 chemicals randomly selected from:

- CCL Universe Example Data Set
  - Gate 1
  - Gate 4
  - Chemicals With No Health Effects or Occurrence Data/Info
- National Reconnaissance of Emerging Contaminants (NREC)
  - Now in CCL Universe Example Data Set

Attribute Scoring Exercise Chemicals

Gate 1	Gate 4	Outside Gates	NREC
1,3-Dichlorobenzene	Carbonyl sulfide	(E)-2-Hexenyl butyrate	17a-Estradiol
Boron	1,2-Dibromo-2,4-dicyanobutane	2-Propanol, 1-(tert-dodecylthio)-	5-Methyl-1H-benzotriazole
Chloroethane	2,3,7,8-Tetrachlorodibenzofuran	alpha-Damascone	bis-Phenol A
Dicamba	Aluminum oxide	C.I. Pigment yellow 119	Cimetidine
n-Butylbenzene	Ethylene	Dimethyl trisulfide	Diethylphthalate
Methane, dibromo-	Ethane, 1-chloro-1,1-difluoro-	Flamprop	Equilin
Hexachlorobutadiene	Heptachlorodibenzo-p-dioxin	Isobutyric acid	Lincomycin
Zinc	Isocyanic acid, methyl ester	Naphthalene, 1,2,3,4-tetrahydro-	Phenanthrene
Metolachlor	Phosgene	Phthalide, 6-(dimethylamino)-3,3-bis[p-(dimethylamino)phenyl]-	Tetracycline
Vanadium	Diazomethane	Sodium acid pyrophosphate	Warfarin

Occurrence Sources

- Agency for Toxic Substances and Disease Registry (ATSDR) - Internet HazDat
- United States Geological Survey (USGS) - National Water Quality Assessment (NAWQA) Program
- USGS - National Reconnaissance of Emerging Contaminants (NREC)
- NLM - Hazardous Substance Databank (HSDB)

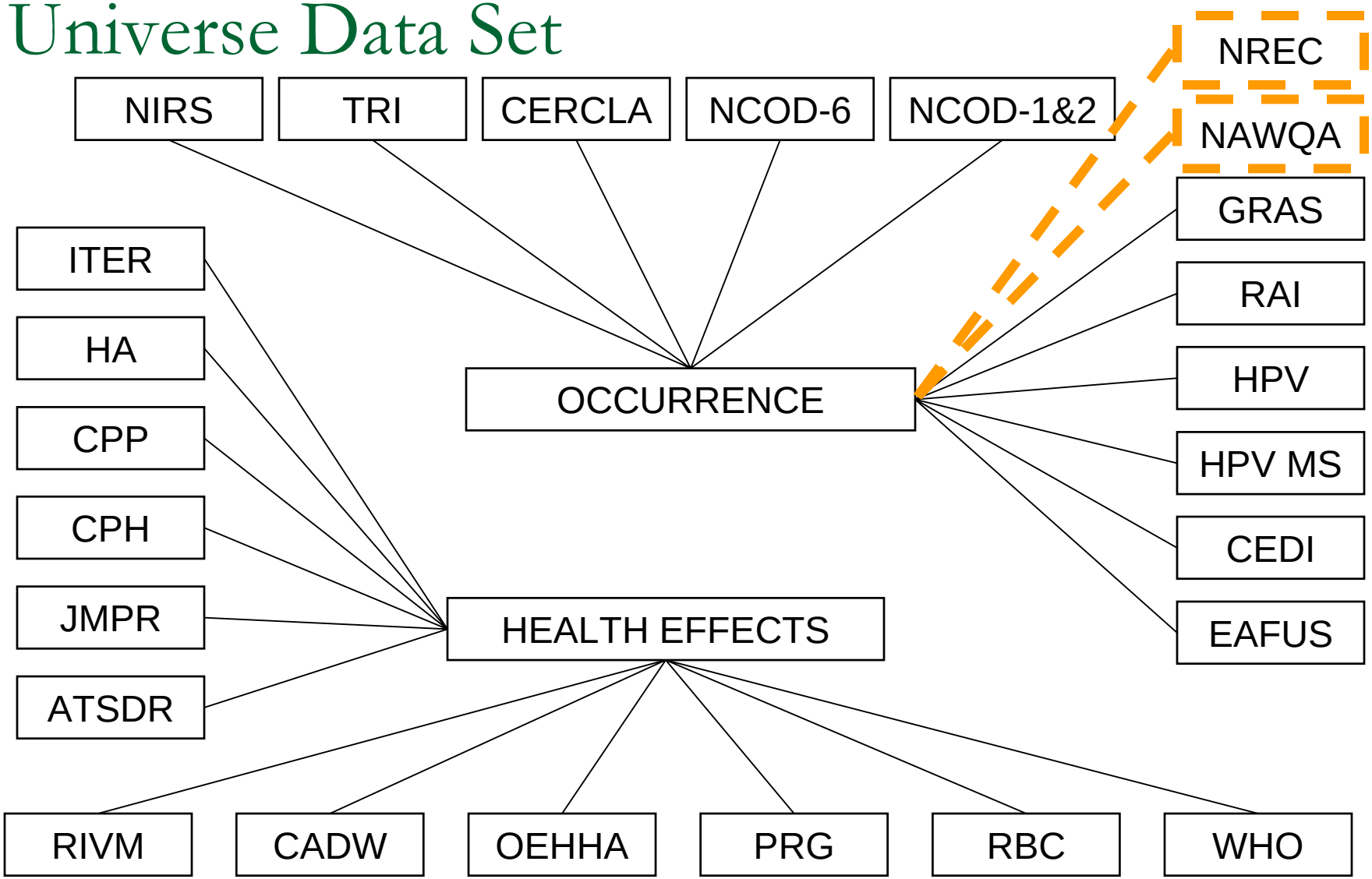
Chemical Property Sources

- National Library of Medicine (NLM) - ChemIDplus
- National Center for Manufacturing Sciences - Solv DB
- Syracuse Research Corporation (SRC) - PHYSPROP database
- NLM - Hazardous Substance Databank (HSDB)

Health Effects Sources

- ATSDR - Toxicological Profiles
- NLM - Chemical Carcinogenesis Research Information System (CCRIS)
- International Programme on Chemical Safety (IPCS) - Concise International Chemical Assessment Documents (CICADs)
- NLM - Developmental and Reproductive Toxicity (DART)
- NLM - Genetic Toxicity (GENETOX)
- NLM - Hazardous Substance Databank (HSDB)
- USEPA - Integrated Risk Assessment System (IRIS)
- IPCS - Joint Expert Committee on Food Additives (JECFA)
- IPCS - Joint Meeting on Pesticide Residues (JMPR) - Monographs
- National Toxicology Program (NTP) - Toxicity & Health/Safety Reports
- Risk Assessment Information System (RAIS) - Toxicity Factors (and supporting data)
- Registry of Toxic Effects of Chemical Substances (RTECS)
- NLM - TOXLINE

25 Data Sources from Example CCL Universe Data Set



Data/Information Downloading

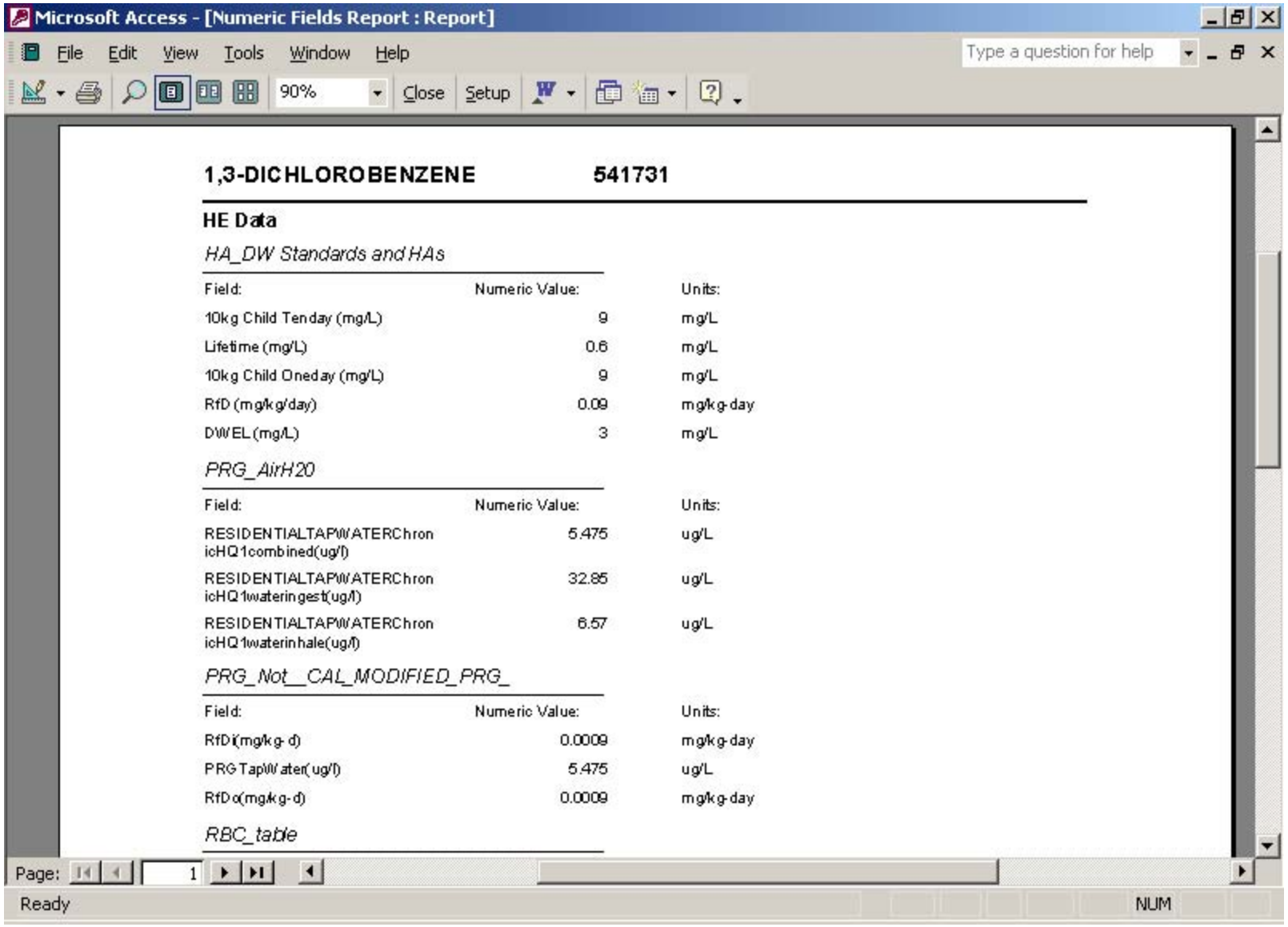
- Tabular Sources (e.g., NAWQA)
  - Downloaded to MS Excel
  - Developed summary statistics
  - Imported to MS Access
- Monographic Sources (e.g., JECFA)
  - Text and data cut and pasted into Access as textual memo fields
- Bibliographic Sources (e.g., TOXLINE, DART)
  - Downloaded to Endnote

Types of Data Elements Obtained

- Health Effects
  - RfD, SF, UR, LO(A)EL, NO(A)EL, LD_{50}
  - Supporting study data (e.g., dose, duration)
  - Absorption, excretion, metabolic data (not obtained)
  - Inhalation-based data (not obtained)
- Occurrence
  - Obtained/developed summary statistics
    - Mean, maximum, ranges, frequency of detection
  - Obtained production/use information
- Chemical property data
  - e.g., solubility, Henry's Law
  - Half-lives (not found)
  - Production (not found - except HPV list)

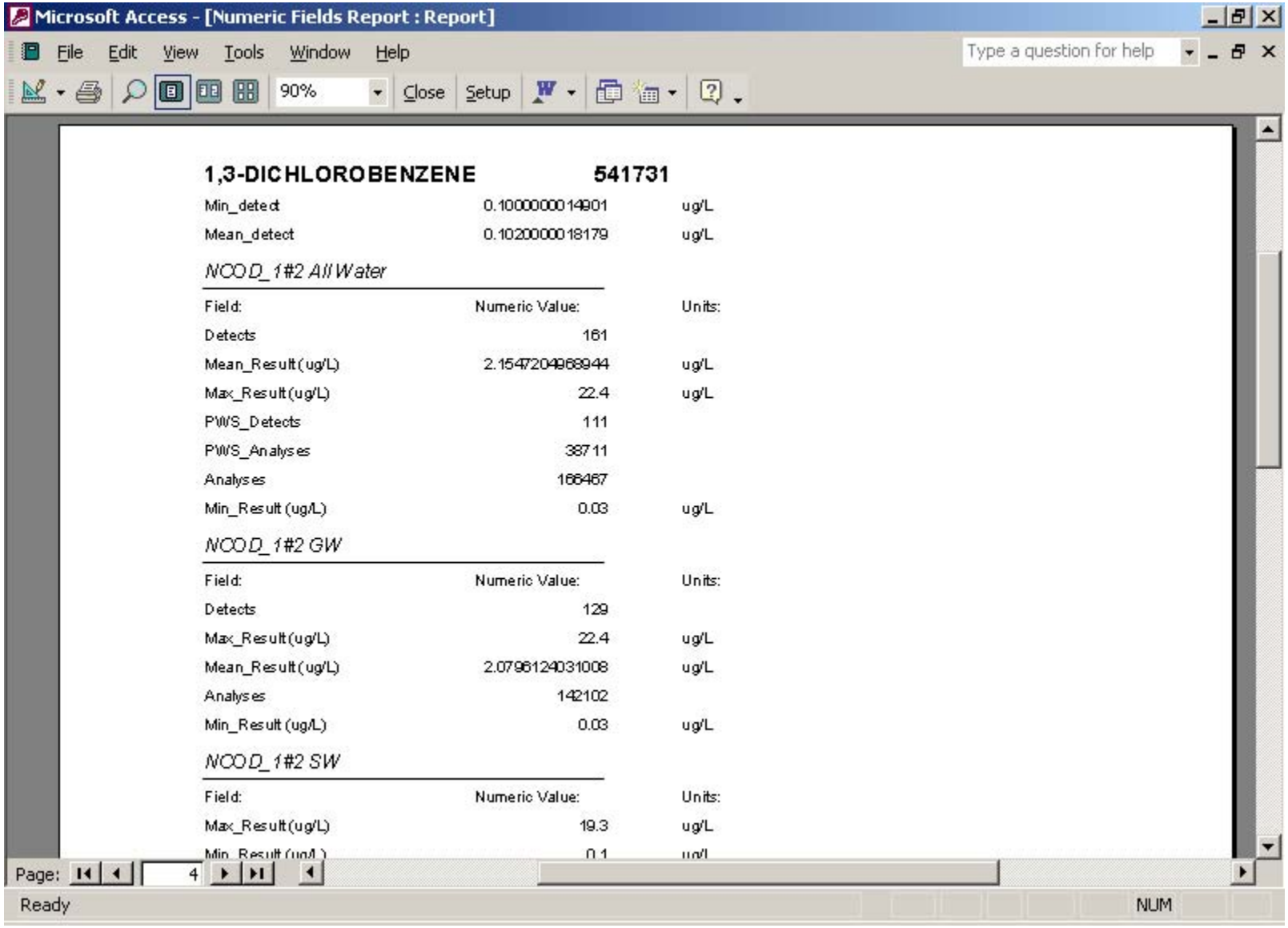
Example of Data -1

1,3-Dichlorobenzene - Health Effects Data



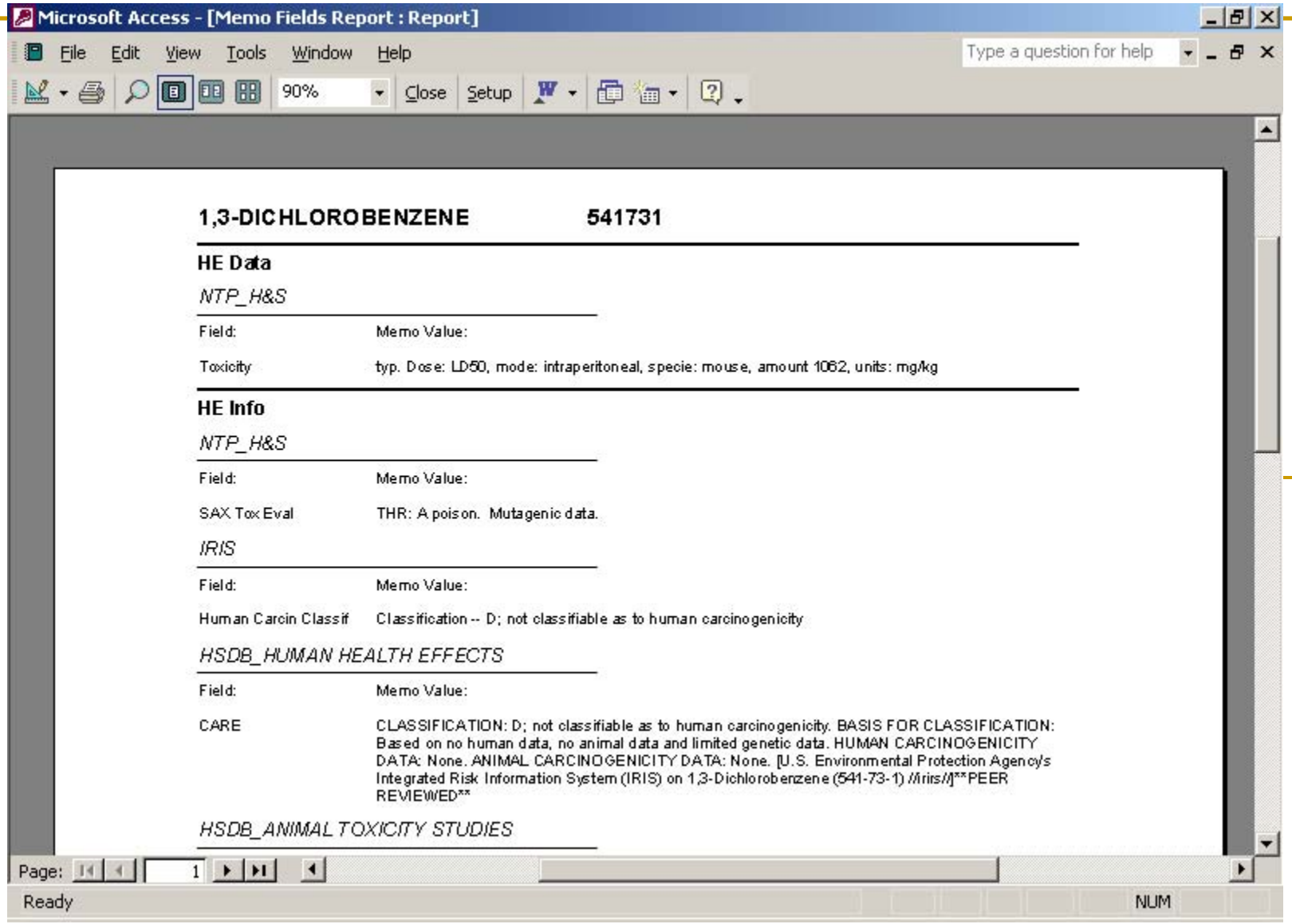
Example of Data -2

1,3-Dichlorobenzene - Occurrence Data

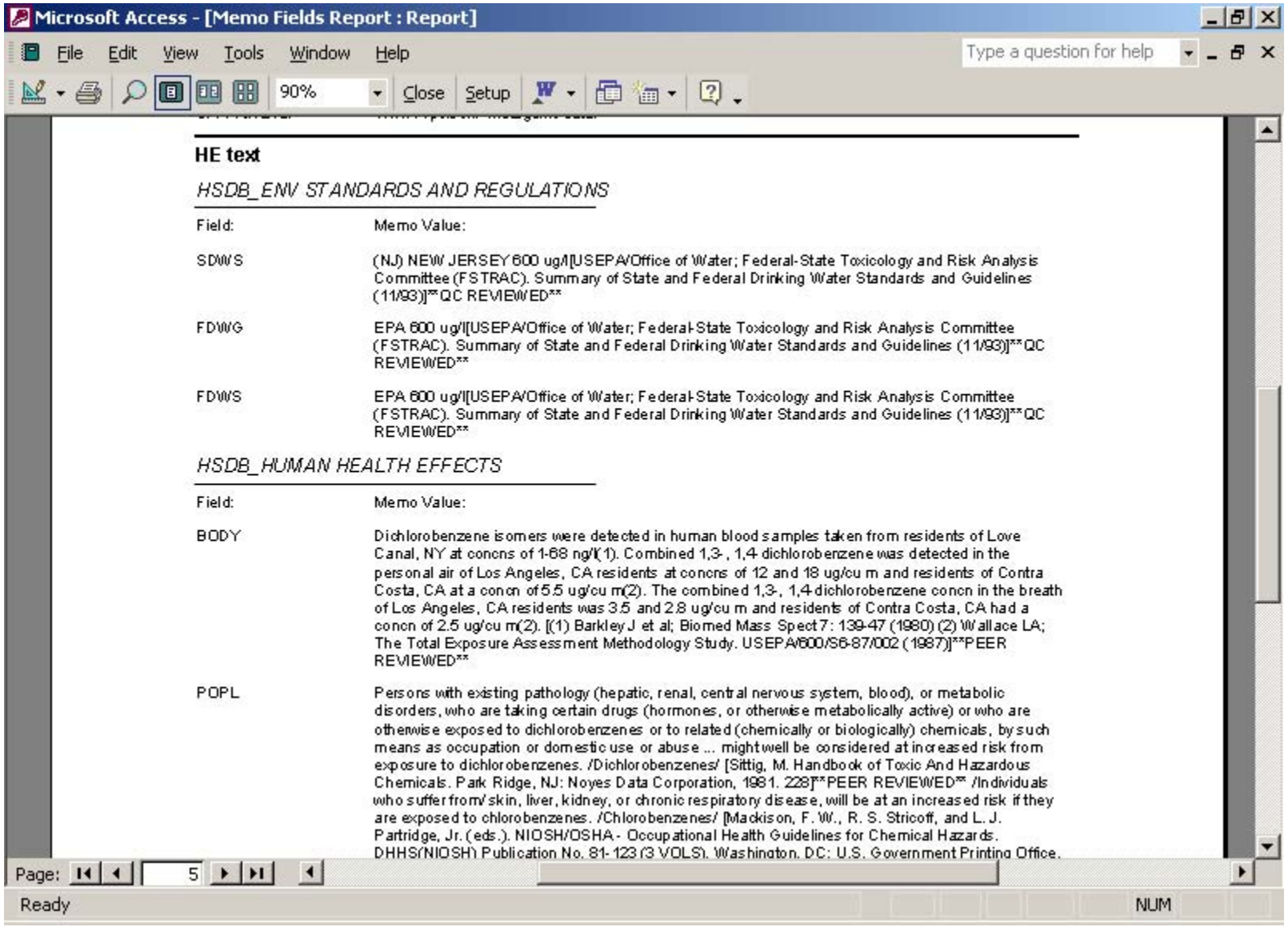


Example of Data -3

1,3-Dichlorobenzene – Data from text sources



Example of Data -4
1,3-Dichlorobenzene – Data from text sources (continued)



Data Extraction Level-of-Effort

Varies greatly

- Extraction method
- Tabular, monographic, bibliographic
- Complexity of format
- Number of chemicals in source data/information
- Number of data elements in source

TYPE	Data Extraction/ formatting time	Example
Tabular	0.25 to 5 days	SRC Physprop
Mono-graphic	1 to 5 days	HSDB
Biblio-graphic	1 to 10 days	DART

Example - Registry of Toxic Effects of Chemical Substances - RTECS

- RTECS is a bibliographic source with a twist: reports data values from various studies
- Each RTECS field is designated with a tag at the start of the line, and a tag for the start of different sections
- Wrote a program to gather specific data
- Data of interest [e.g. Lowest Observable Adverse Effect Level (LOAEL) from multiple dose studies] was automatically entered into tables

Issues with RTECS Importing

- Not all rows began with a tag (e.g. study title)
- Some sections contained extraneous data
- Data was within text, though in a standardized format
- Not every file had the same group headers or the same fields for each entry – some missing
- Units are as cumulative dose, rather than customary daily dose

Solutions to the RTECS Importation

- A parsing program was customized to:
  - Recognize that non-tagged rows went with the previous field
  - Recognize when fields or groups were not provided
- Text fields with data were parsed using string functions
  - Followed a particular pattern 99% of time
  - Remaining values were extracted manually

Data Extraction Issues: Compiling Monographic Data and Information

- Textually-formatted data/information not well suited for entry in a tabular format
  - Data contained in textual passages
  - Requires individual data source programs
  - We're learning – programmers have had successes segregating values from text
  - Continue to evaluate and consider alternatives
  - Requires careful review post processing

Data Extraction Issues: Elemental and Inorganic Contaminants

- Metals
  - Data/information for elemental and inorganic forms is voluminous
  - Various oxidation states
    - Differences in toxicity (e.g., Cr^{+3}, Cr^{+6})
    - Differences in chemical properties (e.g., Na^{0} vs. Na^{+})
- Analytical Methods
  - Speciate by oxidation state
  - Do not report individual inorganic compounds
    - May not match CAS # for compounds

Lessons Learned on Data Extraction

- Demonstrated it is feasible to extract and develop data
- Level of Effort to obtain data ranges from hours to days
  - An option for the more difficult text sources is a placeholder table with candidate identifiers, letting the user know information on candidate exists in the source
- Developing programs to obtain data from text sources
  - Requires flexibility in programs to account for exceptions to patterns
  - May still require some manual entry, but will be limited to a set of sentences rather than entire file

Recommendations/Next Steps

- Develop a hierarchical approach for CCL data gathering
  - Avoid gathering duplicate/extraneous data
  - Update knowledge on elements for each contaminant before going to additional sources
  - Identify desired elements in each source in advance
- Continue to develop and test parsing programs to obtain the data needed
- Develop approach for inorganics
- Create and track a consistent approach