2014-06-10

Resources for Chemical Substance Data: "Small" data and Big Data

Li, Ye; Lalwani, Leena

https://hdl.handle.net/2027.42/107434

http://creativecommons.org/licenses/by-nc-sa/3.0/us/

Resources for Chemical Substance Data : "Small" Data and Big Data

Ye Li, Shapiro Science Library, University of Michigan, liye@umich.edu

Leena Lalwani, Art, Architecture, and Engineering Library, University of Michigan, llalwani@umich.edu

Introduction

Chemical substance data, such as physical/chemical/biological properties and spectra, are essential for researches in many fields including chemistry, physics, material science, chemical engineering, biochemistry, pharmaceutical and medical science etc. In this study, we

- surveyed 40 major resources providing access to chemical substance data
- evaluated their scope and usability
- identified their potential applications in "Big Data" research.

Scope

Resources

Representative primary, secondary, and tertiary resources as well as emerging data repositories and databases with durated data sets

Chemical substances

Small organic and inorganic molecules, metal, alloy, and polymers. (Macro biopolymers such as proteins and DNA are not our focus.)

Strengths of Resources for Chemical Substance Data

- Long history of indexing and curating with many reference works and databases covering data published from 18th or early 19th century
- Critically evaluated data available via many traditional reference works and databases
- Emerging efforts in converting or mining existing data sources into reusable data for both human and computer
- Natural organization of data around one entity - chemical substance

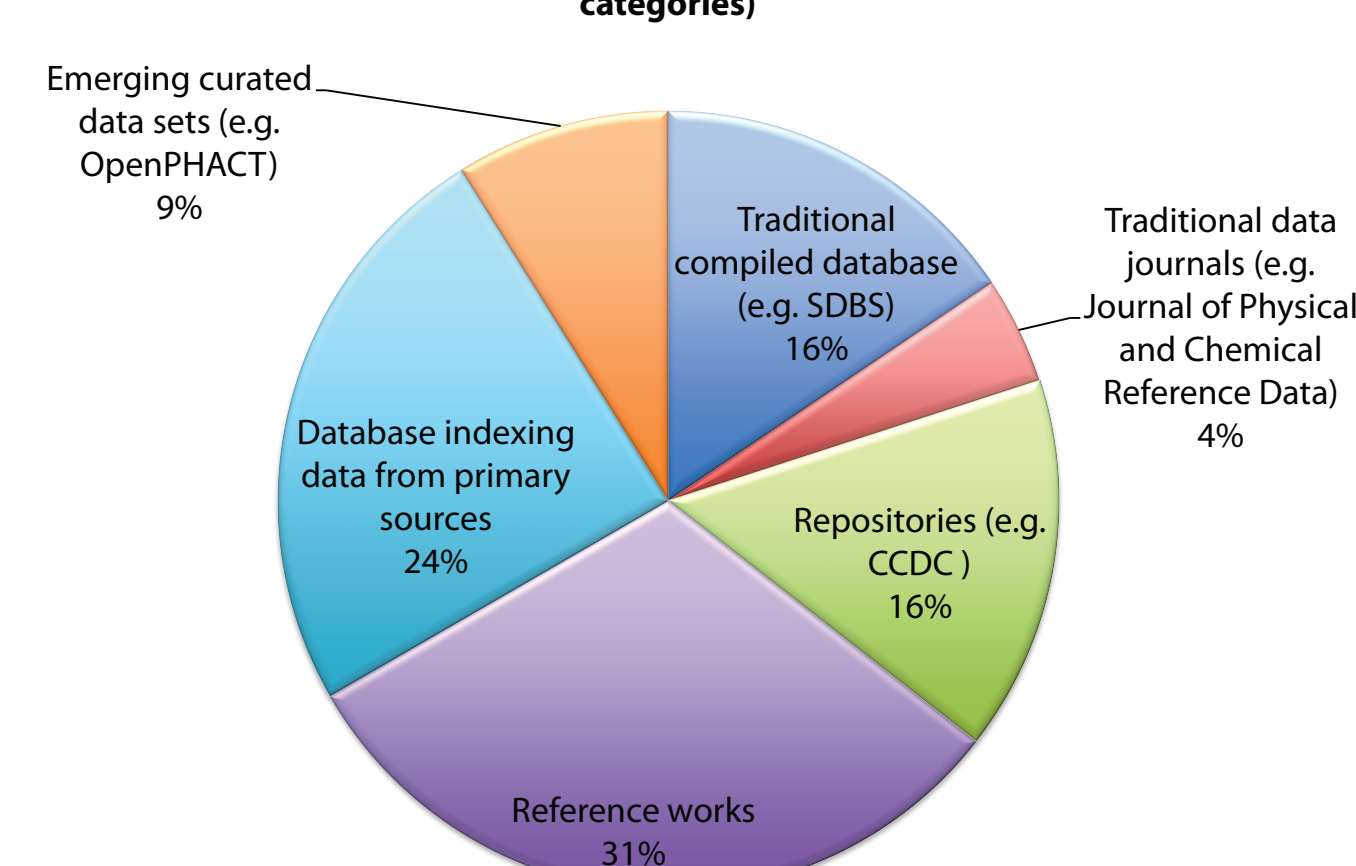## Challenges in Finding and Reusing Chemical Substance Data

- Minimal data sharing tradition
  - Long-tail and individual / small lab centered science
  - Dominating proprietary resources due to strong tie to industry
- Traditional databases / resources based on scattered data in primary literature
  - Mostly subscription-based and limited for mining
  - No sufficient metadata and provenance for reuse

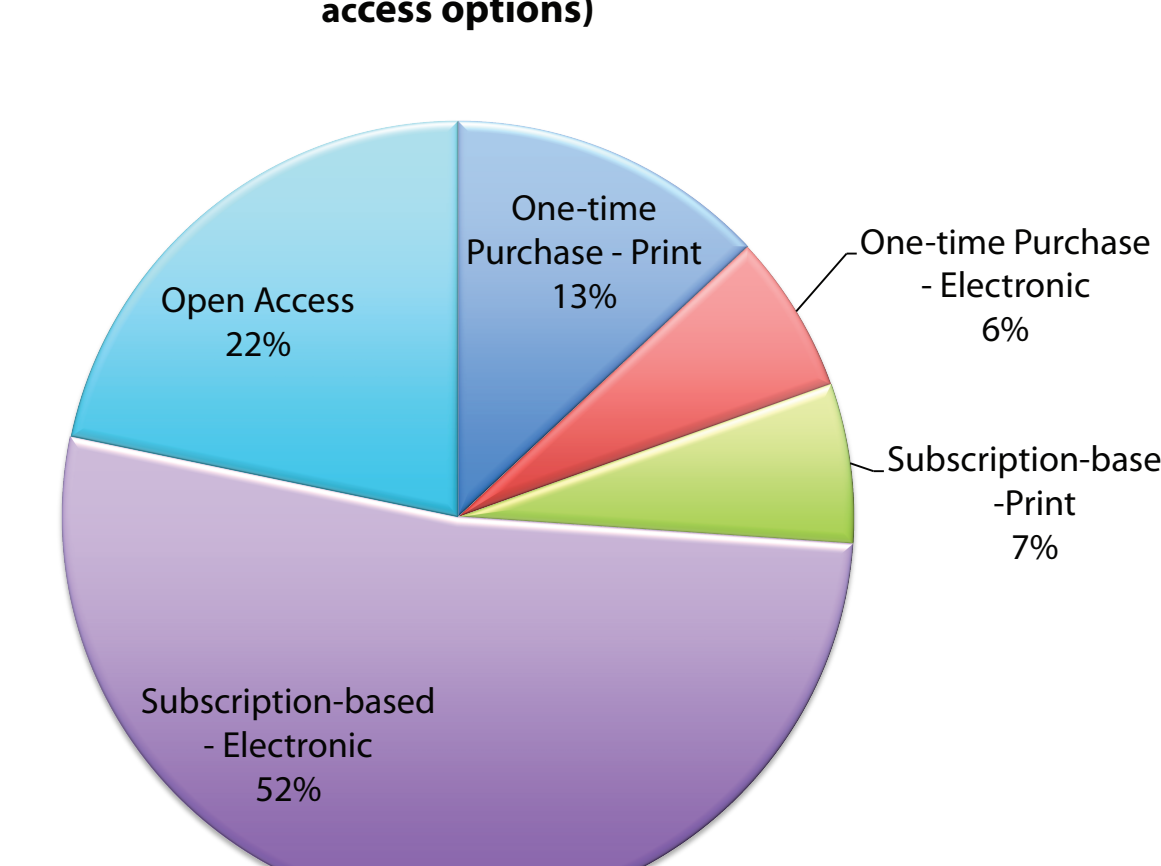## Basic Statistics of Resources for Chemical Substance Data Surveyed

- Range of time established: earliest 1771 and latest 2013
- Time span of covered materials: mid- to late- 18th century to present
- Most resources contain experimental data; and 15 out of 40 also contain calculated or predicted data
- 14 out of 40 resources enables chemical structure search while only 2 of them enables Markush structure search
- 7 out of 40 resources provide sufficient metadata for reuse while 22 of them provide limited metadata
- 4 out of 40 resources provide API services while 8 of them provide knowledge discovery tools on top of the databases

Data Source	Identification	Physical property	Crystallographic data	Spectra / Spectral data	Thermodynamic data	Bioactivity	Ecological data	Economic data	Use (e.g. safety)	Application	Quantum chemical data	Reaction data
(Subtotal)	23	29	12	11	15	8	6	6	11	11	10	9
ASM Alloy Center	X	X										
ASM Failure Analysis Center												
ASM Handbook	X	X										
ASM Micrograph Center	X											
ASM Phase diagrams	X	X										
Bilbao Incommensurate Structures Database			X									
BioRad Spectral Literature Libary				X								
Cambridge Structural Database (CSD)			X									
Chemsafe												
ChemSpider	X	X	X	X	X	X	X	X	X	X	X	X
CRC Handbook of Chemistry & Physics Online		X					X		X			
Crystallography Open Database (COD)			X									
DETHERM		X	X		X							
DIPPR 801					X							
e-EROS (Encyclopedia of reagents for organic synthesis)	X	X							X			X
Handbook of Inorganic Substances 2012		X	X								X	
International critical tables of numerical data, physics, chemistry, and technology	X	X	X		X							
Inorganic Crystal Structure Database (ICSD)			X									
Journal of Chemical and Engineering Data		X			X						X	
Journal of Physical and Chemical Reference Data		X		X	X						X	
Kirk-Othmer encyclopedia of chemical technology	X	X			X		X	X	X	X	X	
Lange's Handbook of Chemistry	X	X		X	X					X	X	
Merck Index Online	X	X				X			X	X		X
NIST Chemistry Book	X	X		X	X							
NIST Standard Reference Data	X	X	X	X	X	X	X	X	X	X	X	X
OpenPHACTS	X	X				X						
Pearson's Crystal Data (PCD)		X	X									
Perry's Chemical Engineers' Handbook		X			X					X		
PubChem	X	X				X						
Reaxys	X	X		X	X	X	X	X	X	X	X	X
SciFinder	X	X		X	X	X		X	X	X		X
SciGlass database	X	X										
Spectral Database for Organic Compounds	X			X								
Springer Materials	X	X		X	X				X		X	
SureChEMBL	X	X				X			X	X		
The International Centre for Diffraction Data (ICDD)		X	X									
The Metals Database (CRYSTMET®)			X									
The Wood database	X	X								X		
Ullmann's encyclopedia of industrial chemistry	X	X			X		X	X	X	X	X	
Wiley Chemical Database and Spectral Libraries	X	X		X								X

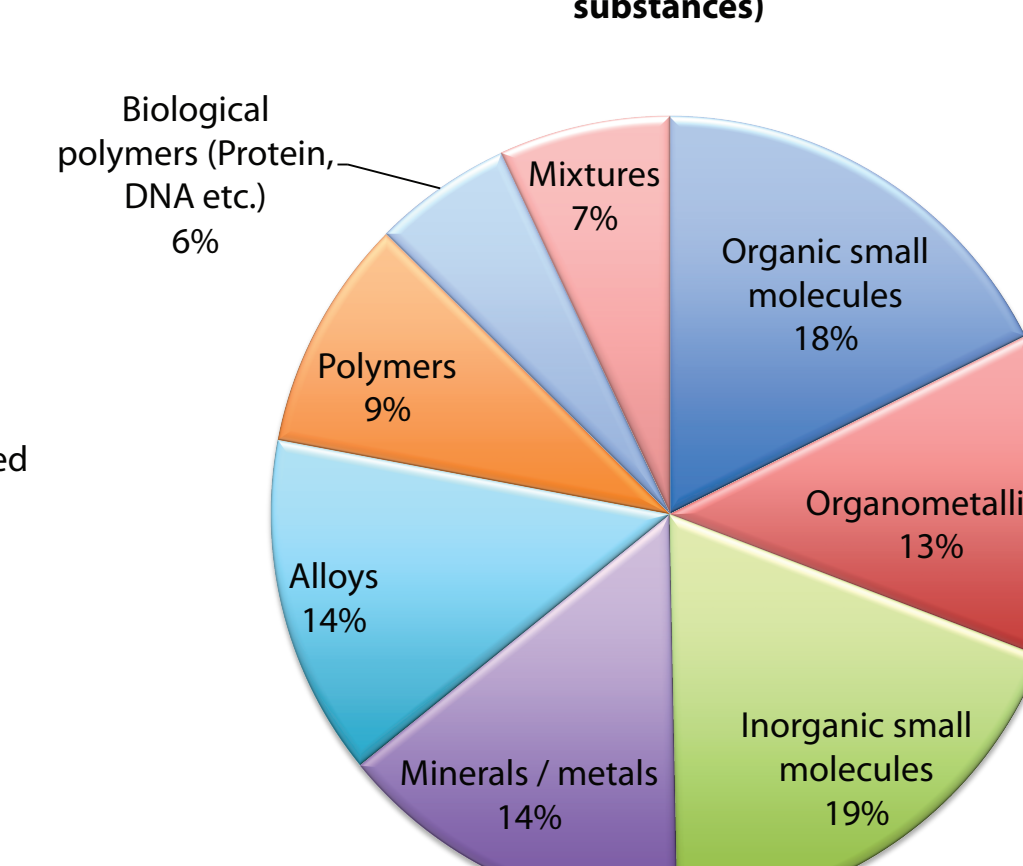

Types of Surveyed Resources for Chemical Substance Data
(Total number of resources: 40, some resources may belong to two categories)

Emerging curated data sets (e.g. OpenPHACT) 9%; Traditional compiled database (e.g. SDBS) 16%; Traditional data journals (e.g. Journal of Physical and Chemical Reference Data) 4%; Repositories (e.g. CCDC) 16%; Reference works 31%; Database indexing data from primary sources 24%

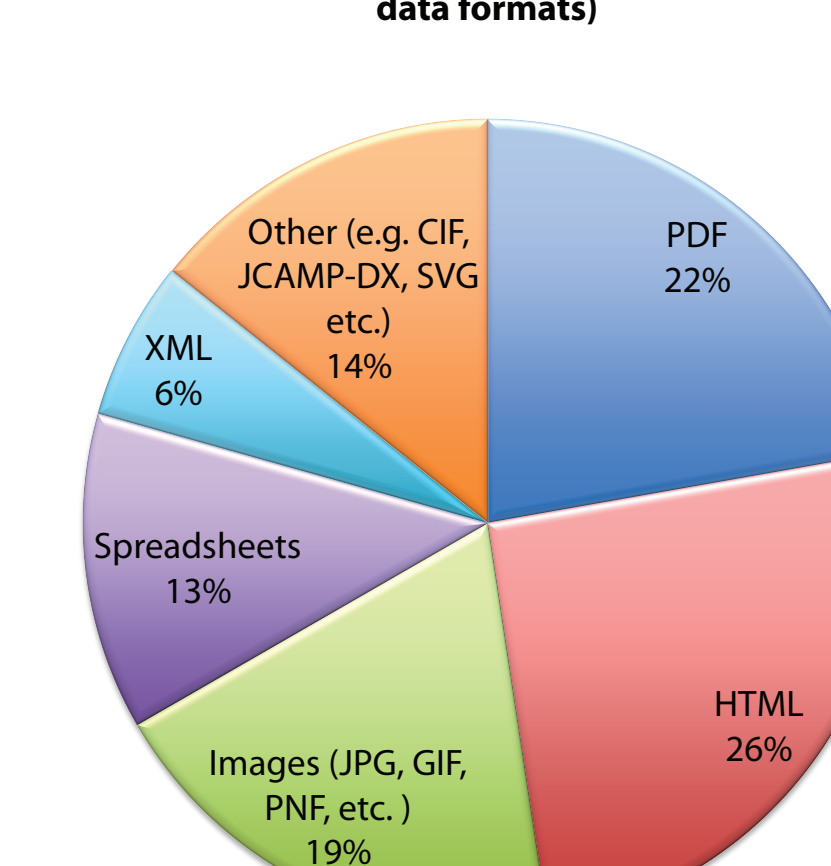Available Access Options for Resources Surveyed
(Total number of resources: 40, some resources may have more than one access options)

One-time Purchase - Print 13%; One-time Purchase - Electronic 6%; Subscription-based - Print 7%; Subscription-based - Electronic 52%; Open Access 22%

Types of Substances Covered in Resources Surveyed
(Total Number of Resources: 40, each resource covers one or more types of substances)

Biological polymers (Protein, DNA etc.) 6%; Mixtures 7%; Organic small molecules 18%; Organometallics 13%; Polymers 9%; Inorganic small molecules 19%; Alloys 14%; Minerals / metals 14%

Data Formats Available from Resources Surveyed
(Total number of resources: 40; One resource may provide multiple types of data formats)

Other (e.g. CIF, JCAMP-DX, SVG etc.); PDF 22%; XML 6%; HTML 26%; Images (JPG, GIF, PNF, etc.) 19%; Spreadsheets 13%; Images 14%

Discussion

- Many resources, especially reference works and some traditional data journals, contain critically evaluated data and associated methodology, which can be used to develop best practices for recording metadata and provenance of specific types of data
- Resources provide data for mining and reuse were mostly focusing on crystallographic data but are now expanding to spectral data and other types of data
- Resources are organized around substance, specific property, or application purpose. Many resources allow customized output organized in different ways
- Data types summarized in the table above can be coverd at different levels in different resources. E.g. Identification data could be very specific with structure, all kinds of registry numbers etc. like in SciFinder or ChemSpider but could also focuses on the micro-visualization like in ASM Micrograph Center
- Chemists, cheminformaticians, chemistry librarians, and computer scientists can work together to improve chemical susbtance data sharing through building new infrastructure and repository as well as converting existing data sources into reusable format