Chemical Hazard Data Commons Working Paper[1]

Speeding & Improving Assessment through Data Sharing

by Tom Lent

Originally published February 6, 2014
Revised March 1, 2014

Undertaking the hazard assessments and alternatives assessments that underlie the summaries described in Section 2 can be a challenging project. A large number of different pieces of information are required, including toxicological studies, physicochemical properties, health endpoint classifications, chemical synthesis pathways, residuals, and modeling studies. These may be accessible on the web or in paper sources but are scattered in many different places in diverse and incompatible formats, sometimes behind expensive paywalls. Appendix 1 provides a list of the data sources that we are surveying for this project. Due to the particular characteristics and focus of each data source, no single source supports searching for all the necessary data elements, for all types of materials to be assessed in one place. Furthermore, data gaps abound. One database may have very comprehensive information on target chemicals with specific determinate molecular structure, but may not address mixtures and polymers. Another may have good coverage of polymers but contains only a small portion of the information needed and may be poorly structured.

The result of this information is that the process of gathering the data needed for a chemical hazard assessment is difficult and resource-intensive. That makes chemical hazard assessment expensive and makes it challenging to ensure that assessments are complete, consistent and up to date .

The Data Commons project aims to confront this challenge by identifying the diversity of informational and organizational resources on the web that constitute different pieces of this critical knowledge base, integrating access through a common Portal, and using the power of open collective crowd sourcing activities to bridge data gaps, improve linkages and facilitate peer review.

Substance Identification

Different systems store chemical data are under a wide variety of names, synonyms and registration numbers. Information also may be stored by reference to a chemical group or near related substances. Connecting these disparate elements before searching for hazard data ensures a comprehensive search.

- **Synonyms** – A single chemical (such as formaldehyde) may be identified by many different names: systematic naming conventions (methanal), formula (CH2O, HCHO), numeric identifiers (50-00-0, 200-001-8, LP8925000) and commercial brand names (Formadon, Formalaz, Forma-Ray).
- **Related substances** – Occasionally a large number of different CAS numbers or other names or identifiers enumerate very closely related substances that share the same hazard characteristics. They might differ by their origin, or specifics of their synthesis, or subtle form differences that don’t necessarily affect hazard (e.g., the rutile, anatase and brookite crystalline forms of titanium dioxide).
- **Forms -** Conversely, some physical differences between otherwise molecularly identical substances can be very significant from a hazard perspective. Nanoparticulate forms, and

[1] See Lent, Tom, et al, *Toward Safer Products: Accelerating Change with a Chemical Hazard Data Commons* for an overview of the Chemical Hazard Data Commons project at https://commons.healthymaterials.net

crystalline substances that become airborne and inhalable at certain size dimensions, are examples of physical forms with hazards different from other forms of their corresponding molecular substance. Current categorization schemes rarely differentiate among these forms.

- **Groups** – Hazards for a chemical may be identified through association with a group of compounds with a common element (e.g., all organic arsenic-based compounds) instead of by identification of each chemical on its own.

In the Data Commons we anticipate providing a search procedure that the user can initiate with a numeric identifier, systematic name or another synonym. Search results will be provided for the requested substance and its related group. Results will also indicate if there are specific forms of a substance that may have further hazard implications. Section 4 addresses in more detail how the Data Commons will help aggregate access to the numeric identifiers of different systems, facilitate the grouping of similar compounds, and differentiate forms for hazard identification by creation of a hazard-based identification system.

Data access, integration and collection

Once the relevant names and identifiers are established, there are several approaches that the Data Commons can take to retrieve and aggregate chemical-specific information from different information sources, each with different timelines:

- **URI-based click-through access –** The Data Commons could provide click-through access to specific chemical pages on select partner sites that have URI-based navigation.[2] This approach will generally provide the user navigation from a Data Commons portal to the partner site web page in a separate tab or window. Some of the sites that offer this approach include USEPA's ACToR, NLM's ChemIDPlus, OECD's eChemPortal (which provides direct links to results in 28 major databases around the world including the massive ECHAECHA REACH database), HBN's Pharos and ATSDR's Toxic Substance Portal. Appendix 1 (Chemical Hazard Information Nodes) indicates which information nodes we have researched to date which include URI-based searches.
- **API based integration**[3] – Sites that offer API access to structured data could be integrated into the Data Commons portal display and searches. Some of the sites that offer this include PubChem and ChemSpider. US EPA has indicated an interest in developing this functionality for ChemView. Appendix 1 (Chemical Hazard Information Nodes) includes an indication of which information nodes which we have researched include API services.
- **Create a crowdsourced repository –** For information from sources that do not offer any direct machine access to pages or data, the Data Commons may host its own repository in the Platform for the information and accept and curate submissions. This is of course a longer-term strategy requiring the building of more functions and establishing the crowdsource social network to populate it. This functionality is powerful, not restricted to simply pulling from pre-identified sources. The Data Commons Platform can be a repository for references to important information culled from any source. When a member of the Data Commons Community encounters an important new study or other relevant news, they will be able to easily link a reference and its abstract and their own interpretive commentary to that chemical or chemical group in the Commons with keyword indexing to increase retrieval options for future reference. The user can also link this entry to discussions about that chemical or group in the Summary areas for the benefit of the larger Data Commons Community.

[2] Referring to where the web address is based upon a universal indicator such as the CASRN of the substance

[3] API – Application Programming Interface – allows interaction of multiple computer systems. In this case it allows another computer application to make use of some of the data in the Pharos Chemical and Material Library

** *Data Commons Recommendation:* Develop a tool to search for a chemical or group across multiple databases. Phase 1 – provide click-through access to pages of systems with URI-based searches. Phase 2 – provide integrated aggregation display of data from systems with APIs. Phase 3 – develop scraping and crowd-sourced processes for hosting data that is not accessible by API. Develop smart search functions to suggest related chemicals, forms and groups to search as well.

Data types

Key information that may be needed for a Chemical Hazard Assessment includes the following information on each substance assessed::

- **Assessments & Listings**
  - **Full chemical hazard assessments** – An assessor can save considerable time and resources by building on previous efforts to assess a chemical, updating outdated reviews and extending to fill data gaps rather than starting from scratch. Assessments of similarly structured chemicals can also help fill data gaps for the chemical under study. IC2's Chemical Hazard Assessment Database currently hosts links to public GreenScreen® assessments and QCATs[4] and HBN will be adding thse assessments to Pharos shortly.
  - **Authoritative hazard listings** – Listings of scientifically assessed associations between chemicals and health endpoints from authoritative governmental, professional or non-governmental organizations can obviate the need for the assessor to research the scientific literature for a chemical to determine the hazard level for an endpoint. Pharos currently provides all listings required for the GreenScreen® List Translator, and more. US NLM's ChemIDplus and the IC2 State Priority Chemicals Resource Source List provide listing compilation as well.
  - **Threshold listings** - Listings of reporting, warning and prohibition levels[5] for chemicals can provide guidance to the assessment process for which chemicals to pursue. The automotive industry's GADSL[6] is one such listing and the NIH's Hazardous Substance Data Bank (HSDB) summarizes US government-regulated thresholds for consumer products (CPSC), worker exposure (OSHA), and many others.
  - **Alternatives** assessments of a chemical in comparison against other functional alternatives will also provide useful hazard information, as well as guidance on alternatives. The EPA's Design for Environment lists alternatives assessments performed under this program and EU's Subsport[7] portal has collected alternatives assessments and case studies.
- **Endpoint**:
  - **Published studies:** The core of a full hazard assessment involves reviewing the outcomes of peer reviewed published scientific studies that evaluate the effect of a chemical on different human and environmental health endpoints. These may be clinical, environmental or epidemiological studies. The Data Commons can facilitate consolidated

[4] Quick Chemical Assessment Tool (QCAT) is a streamlined version of the GreenScreen assessment methodology.

[5] Reporting thresholds are those concentrations above which the ingredient must be disclosed. Warning thresholds are those above which a health warning must accompany the product. Prohibition thresholds are maximum concentration levels allowed in a certain type of product

[6] Global Automotive Declarable Substance List (GADSL)

[7] Substitution Support Portal (SUSBSPORT) www.subsport.euwww.subsport.eu

access to a variety of governmental databases that list relevant studies, such as EPA's ChemView, NLM's HSDB and PubChem. Structured templates can be very useful for making scientific report results accessible in a consistent, retrievable fashion. OECD has developed templates as guides for structuring data for REACH reporting of results of a test on a chemical to determine its properties or effects on human health and the environment (e.g., hydrolysis, skin irritation, repeat dose toxicity, etc.)[8] US EPA is working similarly to structure the data that is aggregated in ChemView. The Data Commons will work to leverage these efforts to structure data for interoperability throughout the Data Commons. The Data Commons also anticipates developing relationships with open publishing sources such as PLOS to develop structured approaches to cataloguing published studies relevant to hazard assessment to facilitate searching. Data Commons access to scientific studies could begin as click-through access to traditional repositories of abstracts and grow over time to include structured searches of open access articles.

  - **Study assessments:** In order to fully support the assessment community and make assessments more understandable and easily renewed over time, the Data Commons can support a more structured approach to the organizing the underlying science and its interpretation. This would involve capturing any values that are extracted from a study of one or more particular endpoints (e.g., an LD50 for aquatic toxicity), a hazard level interpretation (High/Medium/Low) for the specific endpoint, a measure of confidence in the results, and which protocol (e.g., GreenScreen, DfE, etc.)was used to assess and set the hazard level.
  - **Estimation:** When scientific studies are lacking for an endpoint, some protocols such as the GreenScreen® have procedures for using estimations to fill the resulting data gaps.
    - **Modeling** involves using quantitative structure activity relationship (QSAR) methods to apply statistical tools to correlate descriptors of the molecular structure and other properties of chemicals with biological activity. There are several tools available including the US EPA's EPISuite and Oncologic. The Data Commons could support documentation of modeling methods for use in assessments.
    - **Analogs** involve finding very similar chemicals for which there have been relevant studies to fill data gaps. These may be chemically similar based on chemical structure or biologically similar base on metabolic breakdown. A number of databases such as ChemSpider, PubChem, and ChemIDplus can provide search results ranked by structural similarity. The Data Commons related chemical ID work could facilitate compiling results from parallel searches and cross-referencing them with hazard-focused databases. This could help identify and document the analogs for which relevant toxicological data is available.
- **Life cycle**
  - **Feedstocks and process chemistry** – Understanding potential hazards from the life cycle requires characterizing manufacturing methods and different chemical synthesis pathways to identify the source materials from which a material is made, the chemicals that are used, their role in the process (e.g., monomer, catalyst, etc. and some characteristics of the process (such as whether it is open or closed). This analysis can help identify hazards for workers in the manufacturing facilities as well as communities in

[8] OECD REACH templates www.oecd.org/ehs/templates/oecdharmonisedtemplates.htm OECD REACH templates www.oecd.org/ehs/templates/oecdharmonisedtemplates.htm

the vicinity of manufacturing and processing facilities and along transportation corridors. The Pharos Chemical and Material Library has undertaken substantial work to aggregate significant elements of process chemistry for chemicals used in the Building Product Library and to characterize their role. The Data Commons can consolidate access to a variety of sources of information on manufacturing methods, including the NLM HSDB, ATSDR Toxic Substance Portal, and IUCLID Data Sheets. It can also provide a repository to collect crowd-sourced manufacturing information.

- **Functional use -** The role of the chemical, and context in which it is used, are valuable pieces of data for identifying exposures and finding alternatives. Section 5 addresses how the Data Commons project proposes to handle classification of functional use. This can be seeded by conversion of existing functional use schemas but ultimately will require significant amounts of data collection and curation to capture permutations for each industry and application. This will require harvesting functional uses indicators from assessments and process chemistry analysis within the Data Commons and also from collaboration with product disclosure efforts outside of the Commons, such as the Health Product Declaration. Compilations of basic functional use information include ACToR, HSDB, PubChem, ChemView, and the ECHA REACH registered substances database. Converting all of this into the multi-attribute system anticipated in Section 5 will require a serious amount of work, which is probably only practical to do in a crowdsourcing environment.
- **Physical properties** – Physical properties (molecular weight, boiling point, vapor pressure) may be an important part of a hazard assessment to help inform potential pathways of exposure, such as inhalation of vapors. The Data Commons can draw from excellent compilations of physical properties, such as PubChem and ChemSpider, to consolidate listings of physical properties. For those properties missing from the main databases, searches of government assessments and industry data such as Product Safety Assessments from manufacturers[9] may be required. This would be another good project to crowdsource.
- **Residuals -** The feedstock and process chemistry characterization is also key to predicting residuals that may end up in final products. In Pharos, HBN identifies significant process chemistry for chemicals used in products listed in its Building Product Library and currently assumes that monomers, catalysts, contaminants and non-reactive additives in the manufacture are potential residuals, meaning they may unintentionally show up in the final product, while feedstocks and intermediates and reactive additives are not. Additionally, the HSDB provides a list of impurities for some chemicals. The Data Commons can use crowdsourcing techniques to help more precisely identify the roles a substance plays in a process. It can serve as a reference repository for scientific studies that can help identify process chemistry which remains as a residual in products, at what levels, and when those levels exceed established levels of concern.
- **Transformation products -** Determining what substances may be formed from the degradation or other changes of the substance over time, and with exposure to the environment, is important for understanding the full exposure potential. Different forms of transformation considered by the GreenScreen® assessment protocol include biodegradation, oxidation, hydrolysis and photolysis. The GreenScreen® provides a wealth of resources for identifying potential transformation products.[10] The Data

[9] Such as Dow Product Safety Assessments

[10] GreenScreen Chemical Hazard Assessment Procedure V1.2 , Annex VI - Sources for Identifying Feasible and Relevant Transformation Products http://www.greenscreenchemicals.org

Commons could investigate providing tools to automate searching for the identified substance and its related substances in these sources to the extent that they provide URI navigation or API services. The Data Commons can provide a registry of potential degradation products found, the conditions under which they are formed and references to scientific studies that document this.

Biomonitoring – Results from studies of chemicals in blood and tissue samples, while not directly related to the hazard assessment, can give indications of the prevalence of a chemical in the population and hence the scope of exposure. The Data Commons could integrate information from the CDC'sCDC's National Biomonitoring Program information and the State of California's Biomonitoring Program and provide a reference repository to aggregate abstracts and references to other biomonitoring studies of.

** *Data Commons Recommendation*: Target the following information sets: Full chemical hazard assessments, authoritative hazard listings, threshold listings, alternatives assessments, published endpoint studies, study assessments, modeling and analog studies, feedstocks and process chemistry, functional use categorization, physical properties, residuals, transformation products and biomonitoring studies.

** *Data Commons Recommendation*: Support efforts to develop standardized templates and XML dictionaries for communicating structured information between computer systems