Machine-processable representation and applications of the Globally Harmonized System

247th ACS National Meeting
Dallas, Texas, USA

Dr Mark Borkum (m.i.borkum@soton.ac.uk)
16 March 2014
Introduction

• Original developer of myExperiment VRE.
• PhD projects (supervised by Frey and Coles) included:
  – oreChem (also known as “BorChem”);
    • Now extended to “Planning & Enactment” of any “formal process”, e.g., recipes, arts, etc.
  – RDF representation of IUPAC Colour Books;
  – RDF representation of RSC ChemSpider database;
    • $2 \times 10^7$ records (over $1.25 \times 10^9$ RDF triples)
  – ...and many more!
Overview

• Problem Statement
• Methodology
• Results
• Discussion
• Conclusion and Future Work
Risk Assessment

• Risk assessment is the determination of the risk associated with a given hazard.

• In the context of chemistry research, risk assessment determines the risk associated with chemical hazards, e.g.,
  – Explosive,
  – Flammable,
  – Carcinogenic,
  – etc.
COSHH

- The Control of Substances Hazardous to Health (COSHH) Regulations 2002 are a UK statutory instrument.

- COSHH requires employers to protect their employees from the effects of exposure to hazardous chemical substances.
  
  – (The main mechanism for this is risk assessment.)
<table>
<thead>
<tr>
<th>SUBSTANCE NAME</th>
<th>PHYSICAL FORM</th>
<th>QUANTITY</th>
<th>NATURE OF HAZARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>liquid</td>
<td>1000ml</td>
<td>None</td>
</tr>
<tr>
<td>Dextrose</td>
<td>solid</td>
<td>&lt;20 g</td>
<td>Possible irritation to eyes and skin</td>
</tr>
<tr>
<td>Caffeine</td>
<td>solid (tea)</td>
<td>&lt;1 g</td>
<td>Harmful &amp; swelling, induce vomiting</td>
</tr>
<tr>
<td>Milk</td>
<td>liquid</td>
<td>&lt;1000ml</td>
<td>No particular hazards</td>
</tr>
</tbody>
</table>

**NATURE OF PROCESS**

Liquid extraction of caffeine, followed by combination with dextrose to produce a sweet drink.

Is there a less hazardous substance? **No**

If so, why not use it?

**CONTROL MEASURES REQUIRED**

*(Local exhaust ventilation, personal protection, etc.)*

**No specific measure required**
COSHH Risk Assessment Methodology

0. Plan scientific experiment.

1. Enumerate collection of chemical substances.

2. Discover relevant health and safety information.

3. Interpolate assessment form template.
GHS

- The Globally Harmonized System for Classification and Labelling of Chemicals (GHS) is a UN system.
  - Designed to supersede the various C&L systems that are currently in use around the world.
CLP Regulation

• In the EU, the GHS is implemented as the CLP Regulation (see: Regulation (EC) No 1272/2008).
  – Came into force in January 2009.
  – Mandatory from June 2015.
Problem Statement

• For humans, performing a risk assessment has two main drawbacks:
  – Less time spent performing experiments; and,
  – Potential for human error, e.g.,
    • Incomplete enumeration of chemical substances;
    • Use of incorrect information;
    • Transcription errors; and,
    • Tacit assumption of “common” elements.
Let’s Automate!
Automation Strategy

• Given a description of a scientific experiment:
  1. Discover and resolve all chemical identifiers;
  2. Aggregate relevant health and safety information; and,
  3. Feed the results through a template.
Methodology

1. Formalise content of CLP Regulation document:
   - Annex I ⇒ Definitions
   - Annexes III—V ⇒ Instances (C&L elements)
   - Annex VI ⇒ Instances (classifications for approx. 3000 chemical substances and mixtures)

2. Implement Web application.

3. Profit!
Resource Description Framework (RDF)

http://www.w3.org/TR/rdf11-primer/example-graph.jpg
“ghs” Ontology
“Flammable solid. Category 1”
“EC (No) 007-001-01-2”

<table>
<thead>
<tr>
<th>Index No</th>
<th>International Chemical Identification</th>
<th>EC No</th>
<th>CAS No</th>
<th>Classification</th>
<th>Labelling</th>
<th>Specific Conc. Limits, Misation</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>007-001-01-2</td>
<td>ammonia ...%</td>
<td>215-847-8</td>
<td>1336-21-8</td>
<td>Skin Corr. 1B</td>
<td>GHS05</td>
<td>Aquatic Acute 1</td>
<td></td>
</tr>
</tbody>
</table>

- **Hazard Class and Category Code(s):**
  - Skin Corr. 1B
  - Aquatic Acute 1

- **GHS05**
  - STOT SE 3
  - H314

- **Aquatic Acute 1**
  - H400

- **Danger**
  - C ≥ 5%

- **H335**
  - hasHazardCategory

- **Skin Corr. 1B**
  - hasSubstance

- **ammonia ...%**
  - hasHazardClass

- **Warning**
  - hasSignalWord

- **Aquatic**
  - hasHazardClass

- **GHS09**
  - hasHazardPictogram
“(EC) No 001-001-00-9”
Results

• Dataset is available at:

• Provides descriptions of approx. 3000 chemical substances and mixtures.
  – However, on its own, the dataset is not very practical (for use in production environments).
  – The critical observation is that, for many researchers, chemical substances are procured.

• In Autumn 2013, we approached a major supplier...
Benzene
analytical standard

CAS Number 71-43-2  |  Empirical Formula (Hill Notation) C₆H₆  |  Molecular Weight 78.11
Bellatine Registry Number 960212  |  EC Number 200-753-7  |  MDL number MFCD00003000
PubChem Substance ID 24847564

Safety Information

<table>
<thead>
<tr>
<th>Symbol</th>
<th>GHS08</th>
<th>GHS02, GHS07, GHS08</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal word</td>
<td>Danger</td>
<td></td>
</tr>
<tr>
<td>Precautionary statements</td>
<td>P201-P210-P301 + P310-P305 + P351 + P338-P306 + P313-P331</td>
<td></td>
</tr>
<tr>
<td>Personal Protective Equipment</td>
<td>Eyes,呼吸道,Face,全面口罩,呼吸面罩,防毒口罩,多用途组合呼吸器面罩(US), 型号 ABEK (EN14387) 呼吸面罩滤器</td>
<td></td>
</tr>
<tr>
<td>Hazard Codes (Europe)</td>
<td>F,T</td>
<td></td>
</tr>
<tr>
<td>Risk Statements (Europe)</td>
<td>45-46-11-36/38-48/23/24/25-65</td>
<td></td>
</tr>
</tbody>
</table>

Documents

Bulk Quote-Order Product
MSDS
Certificate of Analysis
Enter Lot No.

FT-IR Raman
FT-NMR
Pressure-Temperature Nomograph
Similar Products
Structure Search
“12540 FLUKA”

@base <http://www.sigmaaldrich.com/catalog/product/> .
@prefix chemaxiomprop: <http://www.polymerinformatics.com/ChemAxiom/ChemAxiomProp.owl#> .
@prefix chemdomain: <http://www.polymerinformatics.com/ChemAxiom/ChemDomain.owl#> .
@prefix ghs: <http://xmlns.com/ghs/0.1/> .
@prefix sial: <http://www.sigmaaldrich.com/ns#> .

<fluka/12540> a chemdomain:NamedChemicalSpecies;
  chemdomain:hasIdentifier [ a sial:ChemicalNumber;
    chemdomain:hasValue "042802"],
    [ a chemdomain:Name;
    chemdomain:hasValue "Benzene"],
    [ a chemdomain:CASNumber;
    chemdomain:hasValue "71-43-2"],
    [ a chemdomain:SMILES;
    chemdomain:hasValue "c1ccccc1"],
    [ a chemdomain:MolecularFormula;
    chemdomain:hasValue "C6H6"];
  ghs:hasHazardCategory <http://id.unece.org/ghs/hazard_categories/Flam_Liq_2>,
    <http://id.unece.org/ghs/hazard_categories/SPEC_RE_1>,
    <http://id.unece.org/ghs/hazard_categories/Skin_Irrit_2>,
    <http://id.unece.org/ghs/hazard_categories/Carc_1A>,
    <http://id.unece.org/ghs/hazard_categories/Asp_Tox_1>,
    <http://id.unece.org/ghs/hazard_categories/Eye_Irrit_2>,
    <http://id.unece.org/ghs/hazard_categories/Muta_1B> .
Nitric acid
(ACS reagent, 70%)

Red square = Chemical substance
Red circle = Mixture part
Orange circle = Chemical identifier
Blue square = GHS hazard category
Demo
List of Technologies

- Web technologies:
  - JavaScript Object Notation (JSON)
  - Representational State Transfer (REST)

- Semantic Web technologies:
  - Resource Description Framework (RDF)
  - SPARQL Query Language for RDF (SPARQL)
  - Web Ontology Language (OWL)

- Programming languages:
  - JavaScript
  - Ruby

- Libraries:
  - handlebars.js
  - jQuery / jQuery UI
  - RDF.rb
  - Sinatra
Software Architecture
1) Load Web application

“Generate” button is disabled
2) Search for products by name

Dynamic search with auto-completion
2) Search for products by name

Selected product is added to list

Link to Sigma-Aldrich product page
3) Accept terms and conditions

“Generate” button is enabled
The University requires that assessment forms are printed and signed.
Toggle visibility of classification and labelling elements

Integration with purchasing platform
Discussion
Stakeholders
The Wrong Question

- Who is to blame when the end-user obtains and relies upon the “correct” answer to the “incorrect” question?

  - User  Can I drink this poison?
  - System  Yes
  - User  Is it safe to drink this poison?
  - System  No
Conclusion
Conclusion

- Data modelling is trivial (really!)

- The hard problems are:
  - Correctness and consistency;
  - Interoperability;
  - Provisioning for transparency/accountability; and,
  - Politics.

- These problems cannot be solved by “yet another API”.
  - Instead, provide accessible data (with provenance) in a format that is amenable to discovery and reuse.
Future Work

• Approach UN/EC to host dataset.

• Enhanced input for service, either:
  – List of “chemical identifier—phase—quantity” triples; or,
  – Whole “plan” for scientific experiment.

• Full-blown Web application with template designer.

• RDF formalisation of Sigma-Aldrich product catalogue.
Acknowledgements

• University of Southampton: Doug Akins, Jeremy Frey
• Sigma-Aldrich: Jane Murray, Judith Pruss