

Data Curation: 5% inspiration, 95% frustration (cleaning up data inconsistencies)

id

Published September 16, 2018

Citation

Willighagen, E. (2018, September 16). Data Curation: 5% inspiration, 95% frustration (cleaning up data inconsistencies). *Chem-bla-ics*. <https://doi.org/10.59350/b65kv-58g66>

Keywords

Curation, Toxicology, Nanosafety

Abstract

Slice of the spreadsheet in the supplementary info.

Copyright

Copyright © None 2018. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

chem-bla-ics

Just some bit of cleaning I scripted today for a number of toxicology end points in a database published some time ago the zero-APC Open Access (CC_BY) journal [Beilstein of Journal of Nanotechnology](#), NanoE-Tox (doi:[10.3762/bjnano.6.183](https://doi.org/10.3762/bjnano.6.183)).

The curation I am doing is to redistribute the data in the eNanoMapper database (see doi:[10.3762/bjnano.6.165](https://doi.org/10.3762/bjnano.6.165)) and thus with ontology annotation (see doi:[10.1186/s13326-015-0005-5](https://doi.org/10.1186/s13326-015-0005-5)):

```
recognizedToxicities = [
    "EC10": "http://www.bioassayontology.org/bao#BAO_0001263",
    "EC20": "http://www.bioassayontology.org/bao#BAO_0001235",
    "EC25": "http://www.bioassayontology.org/bao#BAO_0001264",
    "EC30": "http://www.bioassayontology.org/bao#BAO_0000599",
    "EC50": "http://www.bioassayontology.org/bao#BAO_0000188",
    "EC80": "http://purl.enanomapper.org/onto/ENM_0000053",
    "EC90": "http://www.bioassayontology.org/bao#BAO_0001237",
    "IC50": "http://www.bioassayontology.org/bao#BAO_0000190",
    "LC50": "http://www.bioassayontology.org/bao#BAO_0002145",
    "MIC": "http://www.bioassayontology.org/bao#BAO_0002146",
    "NOEC": "http://purl.enanomapper.org/onto/ENM_0000060",
    "NOEL": "http://purl.enanomapper.org/onto/ENM_0000056"
]
```

With 402(!) variants left. Many do not have an ontology term yet, and I [filed a feature request](#).

Units:

```
recognizedUnits = [
    "g/L": "g/L",
    "g/l": "g/l",
    "mg/L": "mg/L",
    "mg/ml": "mg/ml",
    "mg/mL": "mg/mL",
    "μg/L of food": "μg/L",
    "μg/L": "μg/L",
    "μg/mL": "μg/mL",
    "mg Ag/L": "mg/L",
    "mg Cu/L": "mg/L",
    "mg Zn/L": "mg/L",
    "μg dissolved Cu/L": "μg/L",
    "μg dissolved Zn/L": "μg/L",
    "μg Ag/L": "μg/L",
    "fmol/L": "fmol/L",
```

National Institute of Chemical Physics and Biophysics														
Characterization in the test environment														
Nanomaterial Test Summary														
Name	CAS no.	Origin (transport, synthesis method)	Shape	Size (size distribution)	Diameter, nm	Length, nm	Possible aggregation	Surface area, m ² /g	Other characteristics	Media	Method (method of characterization)			
SiO2@SiO2	555-68-0	Shenzhen Shengtian Port Co., Ltd.	rods	n/a	10	300	30-0.251	0.0-0.712	MWA	Algal media OECD 201	DLS	919	MWA	-1.25
						70								
						1700								
							30-0.451	0.0-0.701	MWA					
									MWA					
										General media ISO/IEC 111	TGA	74.9	MWA	-1.10

Slice of the spreadsheet in the supplementary info.

chem-bla-ics

```
"mmol/g": "mmol/g",
"nmol/g fresh weight": "nmol/g",
"µg Cu/g": "µg/g",
"mg Ag/kg": "mg/kg",
"mg Zn/kg": "mg/kg",
"mg Zn/kg d.w.": "mg/kg",
"mg/kg of dry feed": "mg/kg",
"mg/kg": "mg/kg",
"g/kg": "g/kg",
"µg/g dry weight sediment": "µg/g",
"µg/g": "µg/g"
]
```

Oh, and don't get me started on actual values, with endpoint values, as ranges, errors, etc. That variety is not the problem, but the lack of FAIR-ness makes the whole really hard to process. I now have something like:

```
prop = prop.replace(", ", ".")
if (prop.substring(1).contains("-")) {
    rdf.addTypedDataProperty(
        store, endpointIRI, "${oboNS}STATO_0000035",
        prop, "${xsdNS}string"
    )
    rdf.addDataProperty(
        store, endpointIRI, "${ssoNS}has-unit", units
    )
} else if (prop.contains("±")) {
    rdf.addTypedDataProperty(
        store, endpointIRI, "${oboNS}STATO_0000035",
        prop, "${xsdNS}string"
    )
    rdf.addDataProperty(
        store, endpointIRI, "${ssoNS}has-unit", units
    )
} else if (prop.contains("<")) {
} else {
    rdf.addTypedDataProperty(
        store, endpointIRI, "${ssoNS}has-value", prop,
        "${xsdNS}double"
    )
    rdf.addDataProperty(
        store, endpointIRI, "${ssoNS}has-unit", units
    )
}
```

chem-bla-ics

```
)  
}
```

But let me make clear: I can actually do this, add more data to the eNanoMapper database (with [Nina](#)), only because the developers of this database made their data available under an Open license (CC-BY, to be precise), allowing me to reuse, modify (change format), and redistribute it. Thanks to the authors. Data curation is expensive, whether I do it, or if the authors of the database did. They already did a lot of data curation. But only because of Open licenses, **we only have to do this once**.