



Species Dictionary

A Proposal

K. W. Smith

Proposal for Dictionary of Species

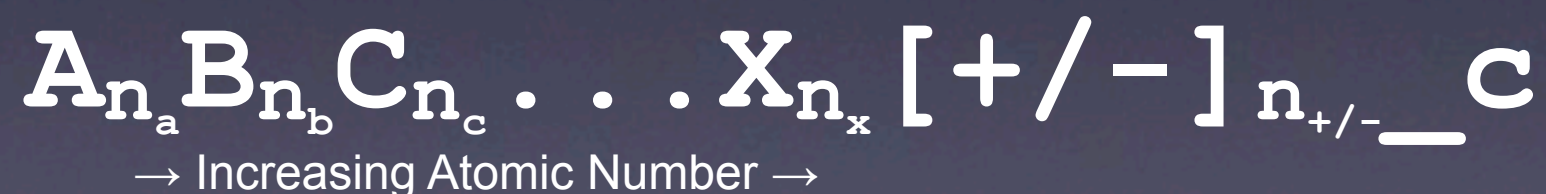
VAMDC must be able to merge queries from multiple databases

Therefore must have a common language for species

Proposal to create master list of molecules, identified by a unique Molecule ID

Create a Molecule ID by:

- Choosing a human readable stoichiometric formula (arranged in ascending Atomic Number order) with charge
- Counter (_c) appended to identify specific variant (e.g. isomers, isotopologues, conformers)



- n value is omitted if the number of atoms (or charges) = 1
- The counter suffix currently has no particular meaning, thus represents no particular property of the Atom or molecule

Molecule ID + Stoichiometric Formula



- Could also store another stoichiometric formula (also ordered by ascending atomic number) that explicitly names the isotopes of the particular constituent atoms
- Isotopes (if specified) are preceded by the atomic mass in brackets, including (2)H for deuterium and (3)H for tritium

E.g.:

Water (H₂O):

Molecule ID = **H2O_1**, Stoichiometric Formula = **H2O**

Deuterated Water (HDO):

Molecule ID = **H2O_2**, Stoichiometric Formula = **(2)HHO**

Aliases

- Related to each Molecule ID is one or more Aliases
- Aliases might include Structural Formulae, Other Names, etc
- Can be used to help identify the Molecule ID

E.g.:

Molecule ID = **H4C2O2+_1**

Stoichiometric Formula = **H4C2O2+**

Aliases:

COOCH4+ (structural formula)

Acetic acid ion (other name)

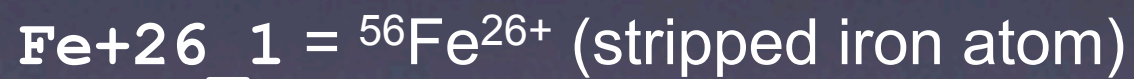
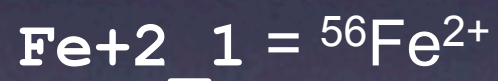
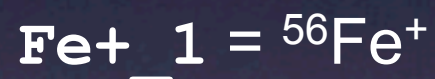
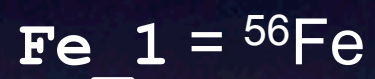
Methyl formate ion (other name)

C2H4O2+ (other name)

Elements

- All the elements can be represented using this scheme, though note that each ion would be represented as a separate entity

E.g.:



etc...

UMIST Examples

- All 739 identifiable species converted into this format (using a Python script)

Example Queries (using MySQL)

molecule_id	stoichiometric_formula	data_origin	data_origin_id
HCN_1	HCN	UMIST	58
HCN_2	HCN	UMIST	61
HCN_3	(2)HCN	UMIST	73
HCN_4	H(13)CN	UMIST	458
HCN_5	H(13)CN	UMIST	459
HCN_6	HC(15)N	UMIST	468
HCN_7	(2)HCN	UMIST	489

Note: The first & fifth molecules in this list are actually HNC, not HCN

UMIST Examples

Query the Aliases list (for Hydrogen isocyanide, HNC)

molecule_id	stoichiometric_formula	name	alias_type
HCN_1	HCN	HNC	structural
HCN_1	HCN	Hydrogen isocyanide	other
HCN_5	H(13)CN	HN(13)C	structural
HCN_5	H(13)CN	Hydrogen isocyanide	other

Issues

- Representation of Stable Isotopes in Stoichiometric formulae: Do we need to always specify the isotope even for the most common? - e.g. ^{79}Br to ^{81}Br (both stable isotopes) is roughly 50:50.
- The number after the underscore '_' has no physical or chemical meaning.
- The '+' sign cannot be used in an XSAMS URI.
- Completely different molecules are listed under the same parent stoichiometric formula. E.g. HCN and HNC.
- Some databases contain completely stripped nuclei - e.g. Fe^{26+} . Will the number of "Molecule" IDs become unmanageable?
- Some databases contain 'unidentified' species. E.g. Charged and neutral "grains" in UMIST. How do we represent these?

Outlook

- Extend conversion scripts to add species from other databases - CDMS, BASECOL, HITRAN, etc
- Deliver Species Tables to other database owners for VAMDC prototypes
- Investigate the possible alternative use of InChI numbers