Chemical Registries – in the Fourth Decade of Service

Before I begin, Stu, I'd like to congratulate you on being selected as the Skolnik awardee for 1999. It's very well deserved, and it's about time! Although I'm a mere rookie – a professional chemical information specialist for only 28 years -- I've been performing searches for 35 years. That makes me all the more appreciative of Stu's 40 years of toiling in the vineyards – and saving more than a few babies from the rapidly disappearing bathwater (metaphor might as well be our middle name, right Stu?).

Definition: precise definition is at the heart of every discussion or dialog, "learned" or otherwise. After all, that's the function of language – to allow interpersonal communication on an organized basis.

Fig. 1

definition 1: an act of determining

registry 3a: a place of registration
4a: an official record book

(Webster's New Collegiate Dictionary)

As I observed early on in my career, chemical information has two unique features compared to all other species of information: 1) chemical structure, and 2) chemical reactions, which include chemical structures, physicochemical entities, and vector functions.

However, chemical structure is not always well defined and even if it is, precise nomenclature is required to define chemical entities. It soon becomes obvious that neither structure nor precise chemical nomenclature is that utilitarian, especially for use by the non-technical (or non-chemical) public, but even for effective, precise communication between chemists.

Examples of nomenclature difficulties in the public sphere abound, including a very prevalent confusion of "silicon" and "silicone", a problem confounded by different spellings for either entity in some languages other than English. Word processor spell checkers don't always solve the problem of poor editing, a phenomenon on the increase. For years, I've advised educators, including high school teachers, to teach at least some chemistry from the newspapers and mass media. Hardly a day goes by without the possibility of finding some goof about science, technology, and especially chemistry. Some are real laughers, but some are stupid or even dangerous.

I recently saw a news article on problems with controlling abuse of GHB, a prominent "date rape" or "party drug". GHB, or gamma hydroxybutyrate, was banned by the FDA in 1991, but related chemical ingredients – with similar hazardous effects -- of party drugs are proliferating, including GBL – gamma butyrolactone, and BD – 1,4-butanediol (of course the product street names are considerably more exotic). Even if better nomenclature or CAS Registry Numbers were used for identification of ingredients in products, would it help? That's debatable.

For decades, chemists – always pragmatic -- have responded by developing lists of chemical compounds, with associated attributes. Such lists are chemical registries and the individual compounds can be tersely and effectively described by an ID number. Although commercial chemical catalogs still list their "registries" by nomenclature – with all of the associated problems – compounds are then ordered by the catalog number, making it a registry number.

Organizations in the business of making new chemical compounds, especially in the pharmaceutical and agricultural chemical industries, developed internal registries very early in the game. One result of these registries was that new compounds with potentially marketable activity were always referred to by the registry number – typically a two letter company code followed by 4-5 numerals -- at least until they acquired a trade name. One somewhat amusing side effect that begin appearing in the '60s was the appearance of two sets of numbers: one, a serial number, used only internally, and the second, a randomly assigned number for use on the outside. The reason: to not allow the competition to infer the level of your research activity.

Fig. 2

24,055-9 1,4-Butanediol, 99+% [110-63-4] HO(CH2)4OH
FW 90.12 mp 16 deg. … 2g … ; 100 g …

B8480-7 1,4-Butanediol, 99% [110-63-4] HO(CH2)4OH
... 1kg … ; 3kg … ; 18 kg …

Aldrich Catalog, 1998-1999

Abstracting and indexing organizations began to see the advantages of registry numbers for not only codifying information in their files, but also to allow another means of searching the files. Prominent examples are the Chemical Abstracts Service (CAS) Registry System, the Derwent List of Registry Compounds, and the Beilstein file. Every compound in the Beilstein file has a Beilstein Registry Number (BRN), as well as Beilstein System Numbers and Lawson Numbers. The BRN is strictly a serial number, but the latter two implicitly contain information on chemical composition and structure. However, Derwent Registry Numbers (DRN) are assigned only to about 2000 compounds commonly encountered in the patent literature. For the remainder of the presentation, the discussion will be centered on the CAS Registry System.

As defined by CAS (Chemical Registry Systems, P. E. Swartzentruber, Abstracts of the ACS National Meeting, CINF 38, Aug. 30, 1984; Chemical Abstracts Service Chemical Registry System: History, Scope, and Impacts, D. W. Weisgerber, JASIS, 48(4), p. 349-360, 1997), a registry system is an inventory of chemical substances with means of inputting, processing, searching, retrieving, and outputting information about the substances. Although the fundamental representation of substances is by structure, structure representation can be accomplished by: 1) nomenclature, 2) linear notation (e.g., WLN, SMILES), 3) fragmentation codes (e.g., GREMAS, Derwent), and 4) connection tables (e.g., Morgan, DARC). CAS chose the Morgan connection table method for the CAS Registry File.

The CAS Registry System was originally designed as a labor saving device in support of the indexing effort used to prepare the Chemical Abstracts database. Prior to the advent of CAS Registry in 1965, with the exception of a list of ca. 2500 common chemicals, there was no good way to determine if a compound had already appeared in the database. Each potentially new compound had to be drawn and named and the name compared to the list of index names. Since the advent of Registry II in 1968, compounds to be indexed undergo name and structure matching procedures against the Registry File. Those found are henceforth referred to by CAS Registry Number (CASRN). Those not found by either method are added to the file as new compounds and a Registry Number is generated.

The design, implementation, and performance of the CAS Registry System have been well documented over the years. However, the remainder of this presentation will focus on the use of CAS Registry Numbers for searching of a number of files and use in other inventories. At the end of the presentation, I'll make a brief digression on the topic of commercialization of academic research.

The format of CAS Registry Numbers is intentionally unique. As mentioned previously, the number is a serial number; the next available number is assigned to the next new compound to be entered into the file. (As an aside, Registry Numbers have disappeared from the file. For example, formaldehyde – CASRN 50-00-0, or compound number "5000" -- is serially the second compound in the file.) The last number is an algorithmically assigned check character. The format is from two to six numerals followed by a hyphen, followed by two numerals, followed by a hyphen, and concluding with one numeral. As indicated, 50-00-0, formaldehyde, is compound 5000, and 1746-01-6, dioxin, is compound 174,601. Compare the format to other number formats:

999-99-9999 = Social Security Number (of one of the patriarchs?)

1-612-389-8370 = telephone number (of the author, after 9/8/99)

60563-3024 = US Postal Zip +4 Code (of the author, until 9/8/99)

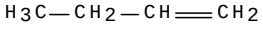
58-08-2 = CAS Registry Number (of caffeine)

Note that the format makes the numbers more readily identifiable even at a glance.

Note the example provided of a CAS Registry File record – for 1-butene:

```
L1   ANSWER 1 OF 1  REGISTRY  COPYRIGHT 1999 ACS
RN   106-98-9  REGISTRY
CN   1-Butene (8CI, 9CI)  (CA INDEX NAME)
OTHER NAMES:
CN   .alpha.-Butene
CN   .alpha.-Butylene
CN   1-Butylene
CN   Butene-1
CN   Ethylethylene
FS   3D CONCORD
DR   1735-75-7, 54366-07-3, 33004-02-3
MF   C4 H8
CI   COM
LC   STN Files:   AGRICOLA, ANABSTR, APILIT, APILIT2, APIPAT, APIPAT2,
       BEILSTEIN*, BIOBUSINESS, BIOSIS, CA, CAOLD, CAPLUS, CASREACT, CEN,
```

```
    CHEMCATS, CHEMINFORMRX, CHEMLIST, CBNB, CHEMSAFE, CIN, CSCHEM,
    CSNB, DETHERM*, DIPPR*, EMBASE, GMELIN*, HODOC*, HSDB*, IFICDB,
    IFIPAT, IFIUDB, MEDLINE, MRCK*, MSDS-OHS, NIOSHTIC, PIRA, PROMT,
    SPECINFO, TOXLINE, TOXLIT, TRCTHERMO*, TULSA, USPATFULL, VTB
      (*File contains numerically searchable property data)
Other Sources:   DSL**, EINECS**, TSCA**
    (**Enter CHEMLIST File for up-to-date regulatory information)
```



```
          7746 REFERENCES IN FILE CA (1967 TO DATE)
           114 REFERENCES TO NON-SPECIFIC DERIVATIVES IN FILE CA
          7758 REFERENCES IN FILE CAPLUS (1967 TO DATE)
            11 REFERENCES IN FILE CAOLD (PRIOR TO 1967)
```

The Registry Number is shown, followed by nomenclature – first index name, then synonyms, deleted Registry Numbers (if any), molecular formula, source of registration, files containing this Registry Number (locator field), structure, the existence of abstract references in the CAOLD file (if any), and the current number of references in the CA and CAplus files.

There are a number of reasons for deleted Registry Numbers. One is a different source of registration. Another was an early attempt to provide additional “faceted” Registry Numbers, which would link compounds similar in stereochemistry, etc. If these were encountered earlier, especially in CAS files, they are preserved in the database and a search for them will retrieve the current compound record in the Registry File.

Early in the process, a list of chemical names and Registry Numbers appeared, the TSCA Candidate List that was commonly known as the “Pre-TSCA” list. Many of the Registry Numbers therein were deleted. A successor list is the Registry Handbook—Common Names, which is published on microform. Of course, the complete list of Registry Numbers also appears in print as the Registry Handbook. Even in the current Registry System, a number of CASRN exist that have no references in bibliographic or data files. They merely register substances (usually less defined compositions or mixtures) that appear on one or more of the national regulatory lists, like the TSCA Inventory or EINECS.

Not only has the CAS Registry System become the cornerstone of the CAS indexing process, but also CAS Registry Numbers are being used for a number of other purposes. They appear in almost 60 files on the STN system, several of which are produced by CAS. A few examples are shown:

CA	Beilstein
CAplus	BIOSIS
CAOLD	HSDB
CASREACT	MEDLINE
CIN	PROMT

CASRN are used in sales (chemical catalogs, including Aldrich), transportation (including export/import), regulatory agency reporting, and disposal. In fact, they are required in many of these cases, especially the latter.

An example of use of CASRN in other files is shown for the Chemical Abstracts bibliographic files (CA files) from CAS:

```
L6   ANSWER 17 OF 191  HCA  COPYRIGHT 1999 ACS
AN   125:257263  HCA
TI   Filters for selective separation of cells or substances from blood
IN   Onodera, Hirokazu; Suemitsu, Junsuke
PA   Asahi Medical Co, Japan
SO   Jpn. Kokai Tokkyo Koho, 6 pp.
PI   JP08196627
AB   Filters for selective sepn. of cells or other substances from blood
     contain arom. ring and/or olefin chain linked to OCH2NR1COCR2(R2)(R2)
     or OCH2NR1COCH(R2)(R3) [ R1 = H or alkyl; R2 = halo; R3 = halogenated
     hydrocarbons].  As an example, polystyrene nonwoven fabrics were
     treated with a mixt. contg. N-hydroxymethyltribromoaceytamide,
     sulfolan and trifluoromethanesulfonic acid and the resultant nonwoven
     fabric was then treated with anti-human CD4 for immobilization. ...
IT   126-33-0P, Sulfolan   1493-13-6P, Trifluoromethanesulfonic acid
     17354-02-8P   91298-06-5P
     RL: NUU (Nonbiological use, unclassified); PNU (Preparation,
     unclassified); RCT (Reactant); PREP (Preparation); USES (Uses)
        (in prepn. of filters for selective sepn. of cells or other
        substances from blood)
...
```

At the dawn of the Registry Systems era, CAS Registry numbers appeared in the abstract as well as in the index phrases. However, that practice has been discontinued. Since 1987, author inspired names have been provided for Registry Numbers in CA File index term phrases.

Until recently, role qualifiers (preparation, uses, etc.) were only provided for ca. 1000 commonly cited chemical compounds and some controlled index terms (CT). The preparation role was indicated, as suffix “P”, for all Registry Numbers, where appropriate. However, beginning in 1995, an expanded list of roles is now algorithmically applied to all Registry Numbers and CTs in the CA Files. The algorithms are good, but not perfect, as indicated in this example of sulfolane. Note that sulfolane was used IN THE PREPARATION of filters – it was not prepared itself. Also note the other roles that were applied for sulfolane. The use roles are obviously accurate, but I think more information is needed to determine if sulfolane is indeed a reactant in this process rather than just a solvent.

Assignment of CASRN by the CAS Registry System is quite accurate. Less accurate assignments are usually the result of author supplied information that is insufficiently accurate. This is often the case with patents, which are often written by non-chemists. Ten years ago, I previously described the difficulties that occur with C4 (or higher) compounds. Although some misassignments were found to be made by the CA indexers, the infamous “polybutenes” reside in an information swamp (definitely not on the Information Highway). To comprehensively search these oligomers of mixed butenes, at least 40 CASRN must be used, in addition to compound names, the nomenclature for which is even less precise.

Several of the CASRN are only partially correct, and some are not accurate at all. Polybutenes are oligomers of mixed butenes, primarily of isobutylene with much smaller amounts of 1-butene and 2-butene. Those CASRN shown are in increasing order of accuracy:

9003-27-4	Isobutylene Homopolymer
9003-28-5	1-Butene Homopolymer
9003-29-6	Butene Homopolymer
26938-45-4	1-Butene/Isobutylene Copolymer
28300-07-4	1-Butene/2-Butene/Isobutylene Copolymer

The first, 1-butene homopolymer, although often used for polybutenes, is not correct and should be reserved for crystalline poly 1-butene.

The situation is even more murky for derivatives of polybutenes, including the even more legendary infamous “PIBSAs”, the majority of which don’t have CAS Registry Numbers. Unfortunately, the terminology is even more chaotic – some examples are shown:

Polybutenyl Succinate

Isobutenyl Succinic Acid

Poly(alkenyl) Succinate

Alkylated Succinic Esters

Isobutenyl Succinimide

“Maleic Acid/Anhydride, Alkylated Derivatives”

Some PIBSAs are indexed like the last example, as derivatives of maleic or succinic acids or anhydrides, both by text names or by derivative CAS Registry Numbers. A good share of the blame for this indexing nightmare can be directed to sloppy nomenclature and description by the original authors. However, the polybutenes and PIBSAs represent several cases where new, more definitive CASRN should be assigned, rather than attempting to get by with established, but inadequate CASRN.

Especially among those dealing with regulatory agency information, a number of myths about Registry Numbers and associated information have sprung up over the years.

Myth 1: CAS Registry Numbers contain information on chemical composition.

Wrong. As stated before, Registry Numbers are serial numbers – the associated chemical information is contained in the file record, including references. Some numbers in other files do contain structural information, e.g., Lawson Numbers in the Beilstein file.

Myth 2: Toxic substances have low-numbered CAS Registry Numbers.

Not necessarily. Compounds that were described often in the literature tended to acquire Registry Numbers very early in the history of the system, including many of those that are toxic. The reverse is also not true: not all "old" compounds are toxic.

Myth 3: Per its title, the TSCA Inventory only lists toxic substances.

By any realistic definition, "toxicity" both a relative concept, yet highly dependent on data (i.e., "toxic compared to what?"). Depending on the definition, all compounds are toxic, at least in some concentrations and environments. However, those compounds more commonly perceived as "toxic" are also commonly encountered in commerce and hence are on TSCA or other regulatory lists. Their hazardous potential must be established, if not already done.

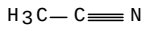
Myth 4: All CASRN in the various sources are assigned and used accurately and precisely.

Not completely true. The primary method of assigning CAS Registry Numbers with the CAS Registry System and the CA files has already been described and is accurate to a very high degree. However, other methods are used to assign Registry Numbers to compounds appearing in other files. As previously described (1994, 1995, 1996), even if CAS assigns CASRN for compounds in databases other than its own, these CASRN may have the same precision as for compound information totally within the control of CAS. Due to chemical "puns" like "ether", the precision may fall even further for CASRN assigned by algorithm. Information users should always heed caveat emptor, but should they expect a "CAS-housekeeping seal of approval". Not a bad idea.

I hope I've made it obvious that CASRN are extremely valuable in searching for chemical substances. Should CASRN be used universally for precise identification of chemical substances, including in all published articles and patents? It's tempting to say yes. Recently on the Chemical Information discussion list, someone asked about the identity of "MeCN". Several members responded with acetonitrile. One responder was reminded of an "urban myth" (probably true) that some shipper refused to accept "methyl cyanide" for shipment, but would accept acetonitrile. Would use of the CASRN help? Possibly, but if CASRN become a utility, who should be expected to administer the process?

```
L1   ANSWER 1 OF 1  REGISTRY  COPYRIGHT 1999 ACS
RN   75-05-8  REGISTRY
CN   Acetonitrile (8CI, 9CI)  (CA INDEX NAME)
OTHER NAMES:
CN   Acetonitrile cluster
CN   Cyanomethane
CN   Ethanenitrile
CN   Ethyl nitrile
CN   Methane, cyano-
CN   Methanecarbonitrile
CN   Methyl cyanide
CN   Methyl cyanide (MeCN)
FS   3D CONCORD
DR   54841-72-4
MF   C2 H3 N
CI   COM
LC   STN Files:   AGRICOLA, AIDSLINE, ANABSTR, APILIT, APILIT2, APIPAT,
       APIPAT2, BEILSTEIN*, BIOBUSINESS, BIOSIS, CA, CANCERLIT, CAOLD,
       CAPLUS, CASREACT, CEN, CHEMCATS, CHEMINFORMRX, CHEMLIST, CBNB,
```

```
      CHEMSAFE, CIN, CSCHEM, CSNB, DETHERM*, DDFU, DIPPR*, DRUGU, EMBASE,
      GMELIN*, HODOC*, HSDB*, IFICDB, IFIPAT, IFIUDB, IPA, MEDLINE,
      MRCK*, MSDS-OHS, NAPRALERT, NIOSHTIC, PDLCOM*, PIRA, PROMT, RTECS*,
      SPECINFO, TOXLINE, TOXLIT, TRCTHERMO*, TULSA, ULIDAT, USPATFULL, VTB
        (*File contains numerically searchable property data)
    Other Sources:   DSL**, EINECS**, TSCA**
        (**Enter CHEMLIST File for up-to-date regulatory information)
```



```
          21089 REFERENCES IN FILE CA (1967 TO DATE)
            289 REFERENCES TO NON-SPECIFIC DERIVATIVES IN FILE CA
          21141 REFERENCES IN FILE CAPLUS (1967 TO DATE)
             10 REFERENCES IN FILE CAOLD (PRIOR TO 1967)
```

On a final note, I'd like to make a slight digression to a more general topic. From recent articles in C&EN and the Wall Street Journal, it's become obvious that an increasing number of educational institutions have assumed control of the intellectual property aspects of academic research. In addition to the debate on various issues, including impairment of communication on research, even in-house, I'd like to present another issue for debate.

The academic community has been quite vehement on the need for discounted (or "free") information for academic purposes. Usually granted for use for teaching purposes, academic discounts should also apply to academic research, according to many faculty. At several institutions, faculty, employees, and even students are being required to sign non-disclosure agreements regarding any research performed at that institution. I submit that any information gleaned for any research so covered should be acquired at "retail", not at the steeply discounted "wholesale" rates typical of most academic discount plans. Who should pay? Adequate support of information services should be built into all funding for academic research. Any information research required for the disclosure and patenting processes required by the institutional P&L department should be funded by that department.

I realize that this is a controversial point, but I believe the approach is fair. I make this proposal as an outsider to the processes involved, and I welcome any discussion on this point.

As should be obvious from the following address, I'm in the midst of changes in my personal and professional lives. In moving to Minnesota, Gloria and I plan to more fully retire. The consultancy will remain in business, but at a reduced level. I plan to attend meetings throughout the next year, with further participation to be determined. Meanwhile, if peace and quiet appeals to you (motors are forbidden on our lake), stop in and see us if you're in the area.

Robert E. Buntrock
Buntrock Associates, Inc.
11335 300th Ave. NW
Princeton, MN 55371
612-389-8370
Fax 612-389-8371
Buntrock2@earthlink.net

Thanks again, Stu, for the invitation. Thanks also to the staff at CAS for advice and information, and to you, the audience. I'll entertain any questions.