ORIGINAL: www.iupac.org/symposia/conferences/CIandXML_jul02/index.html

CAS/IUPAC Conference on Chemical Identifiers and XML for Chemistry

The Pfahl Executive Education and Conference Center and The Blackwell at The Ohio State University, Columbus, Ohio, USA
July 1, 2002

> Meeting Program and Abstracts

This conference will bring together experts to survey current activities in the research and development of chemical substance representations and identifiers, including both nomenclature and computer-based structural descriptions, and of chemical markup language.

The conference is designed for researchers and developers working in the areas of chemical identifiers and chemical markup language and chemical information specialists, database producers, and others who have an interest in or utilize chemical substance information.

Speakers for the conference include:

- Jonathan Brecher (CambridgeSoft Corporation): From chemical name to structure: finding a noodle in the haystack
- Alexander Lawson (MDL Information Systems GmbH): Nomenclature practice and post-Postman factors
- Peter Murray-Rust (Department of Pharmaceutical Sciences, University of Nottingham): The chemical semantic web: a common infrastructure for chemistry

- Henry S. Rzepa (Department of Chemistry, Imperial College of Science): The vision of a chemical semantic web
- Stephen E. Stein (Physical and Chemical Properties Division, NIST): The IUPAC Chemical Identifier
- Matthew J. Toussant (CAS): CAS chemical identifier systems
- Antony J. Williams (Advanced Chemistry Development): Unifying chemical nomenclature standards - The roundabout of names and structures
- Janusz L. Wisniewski (MDL Information Systems GmbH): Computer-based naming service for very large chemical databases: from AutoNom in the Beilstein File to AutoNom in the ISIS system
- Stephen E. Stein (Physical and Chemical Properties Division, NIST): An XML namespace for IUPAC

The conference is being organized by David W. Weisgerber, CAS (retired), and questions about the conference can be directed to him.

The conference will begin with a welcoming reception on the Sunday evening preceding the meeting. The one-day conference will be held in the Pfahl Executive Conference Center on The Ohio State University campus and conclude with a banquet on Monday evening. An optional visit to Chemical Abstracts Service will be offered to the attendees on Tuesday morning, July 2, 2002.

Hotel accommodations will be provided by The Blackwell, a new upscale hotel located adjacent to the Pfahl Executive Conference Center.

Registration
There is **no fee** associated with attending this conference but you must be registered in advance.

ORIGINAL www.iupac.org/symposia/conferences/CIandXML_jul02/program.html

Meeting Program and Abstracts

9:00 - Introduction

9:10 - **Matthew J. Toussant** (CAS):

CAS chemical identifier systems

Abstract: Chemical compounds, their syntheses, their properties, and their applications, are the core of chemistry. Recording, storing, and retrieving information on chemical substances have been paramount to the progress of chemistry. The challenge to CAS has been to provide its users and itself with efficient and effective means of identifying substances reported in the world's chemical literature. This presentation will describe the components of the CAS chemical identification systems with a primary focus on the CAS Registry System. The foundation of the CAS Registry is the computer-based connection table with its three-dimensional structure representation. Complementing the structure-based representations are chemical substance names, including names systematically assigned by CAS according to a set of rigorously based rules, plus other systematic and semi-systematic names and trade names compiled from the chemical literature. Linking each set of structure and nomenclature information for a particular substance is the CAS Chemical Registry Number, a concise and unique identifier that has become widely used as a standard for chemical substance identification. A recently introduced thesaurus capability by CAS provides users with additional links to specific substance information from general subject and class terms. The CAS MARPAT service provides a means of identifying generic substances reported in the patent literature, thus complementing and extending the range of substance identification systems offered by CAS.

9:50 - **Stephen E. Stein***, Dmitrii Tchekhovskoi, Steve Heller (Physical and Chemical Properties Division, NIST):

The IUPAC Chemical Identifier

Abstract: IUPAC has long been a recognized source of rules for naming chemical substances. However, names generated by these rules are designed primarily for human communication and are not optimal digital representations of chemical identity. In view of the ever-increasing volume of digital communication in chemistry, IUPAC has undertaken a program to establish a digital signature for a compound derived algorithmically from its digital structure representation (connection table). It is hoped that this IUPAC Chemical Identifier (IChI) will one day become an accepted standard representation of chemical substances.

> link to corresponding IUPAC project > www.iupac.org/project/2000-025-1-800

Following discussions at IUPAC meetings, a test version of the Identifier was developed at NIST and distributed in March, 2002. The principal objective of this version is to begin a community-wide discussion of the form of the Identifier.

The first implementation of IChI was designed for covalently bonded structures only. It processes an input connection table in three steps:

1. Normalization - all structural information unnecessary for identification is ignored. This, for example, eliminates ambiguities arising from different representations of pi-electrons, such as occur in the depiction of aromatic and zwitterionic structures.
2. Canonicalization - each unique atom is given a unique label. This is a mathematical procedure applied individually to distinct "layers" that describe connectivity, tautomerism, isotopes, stereochemistry (presently includes sp^3 and Z/E) and charge.
3. Serialization - a string of characters derived from labels produced by canonicalization. This generates the observable output form of the IChI, which may be viewed as a series of ordered connection tables, one for each "layer".

In addition to current features of the IChI, this discussion will examine still unresolved structure representation issues as well as ideas for extension to other classes of chemical compounds.

> download pdf file of this presentation (pdf file - 255KB) or
view slides at <http://www.hellers.com/steve/pub-talks/columbus-702/frame.htm>

10:50 - **Alexander J. Lawson** (MDL Information Systems GmbH):

Nomenclature practice and post-Postman factors

Abstract: The history of the communication of concepts (as influenced by the technical development of the available medium) was dramatically summarized by Postman in the 1980's. The phases of relative importance of the graphic, spoken and written traditions have been a constant companion to the development of civilisation in general, always involving deep consequences for the societies involved. The current general trend in the technically developed world involves an accelerated transition to the graphic representation at the expense of the spoken word in particular.

This general phenomenon can be argued to apply also to the learned sciences, none more so than mainstream organic chemistry.

Some possible consequences and opportunities for the specialist field of chemical nomenclature will be explored, with particular emphasis on organic chemistry.

11:30 - **Jonathan Brecher** (CambridgeSoft Corporation):

From chemical name to structure: finding a noodle in the haystack

Abstract: Of all ways to identify a chemical, the one with the longest history and widest use is the simple chemical name. On the one hand, the broad acceptance of chemical names brings several advantages, including that they are convenient and easy to use in many environments. On the other hand, those strengths bring with them several serious drawbacks when chemical names are used as chemical identifiers -- what is the molecular structure of "glucose", let alone "sugar"? This presentation will highlight the state of the art in interpreting textual chemical names to produce chemical structure diagrams. Practical uses of such automated conversions will be demonstrated, with special emphasis on the strengths and weaknesses of using chemical names as chemical identifiers.

> download pdf file of this presentation (pdf file - 1.02MB)

1:45 - **Antony J. Williams** (Advanced Chemistry Development):

Unifying chemical nomenclature standards - the roundabout of names and structures

Abstract: Systematic Nomenclature is predisposed to software generation since rules-based systems are ideal tasks for computers to handle. In an ideal world there would be a single static, non-language specific systematic nomenclature accepted by chemists and in general usage. With general acceptance, rigorous application of systematic rules would produce fully reversible chemical names from which chemical structures could be generated. Of course there are multiple systematic nomenclature systems and chemical names found in the

literature often are only close approximations to the correct names. The ability to generate systematic names from structures would lead us to conceive that systematic names can be reversed to structures using software. Of course this is possible with trivial names, synonyms, IUPAC names and CAS Index names being able to be reversed to the structure based world. Systematic naming and its reversal are by their very nature demanding of quality and the adherence to naming standards is a true test during software development. This issue will be reviewed during this examination of systematic nomenclature software development.

2:25 - **Janusz L. Wisniewski** (MDL Information Systems GmbH):

Computer-based naming service for very large chemical databases: from AutoNom in the Beilstein File to AutoNom in the ISIS system

Abstract: Design and practical implementation of algorithms and routines in the worldwide first computer-based system for generation of the systematic IUPAC-sanctioned nomenclature directly from connection tables of organic compounds is discussed. Detailed overview of the performance, accuracy, and reliability of the system is presented. Practical issues and obstacles encountered and solved during inclusion of the program package into the established traditional production of large chemical databases such as Beilstein Handbook at the beginning, the Beilstein Database later and finally of the current MDL Crossfire system are described. The seamless integration of the nomenclature software into a company compound database registration and production using the ISIS platform is discussed. Advantages of the AutoNom TT package as DLL for DBMS independent general Naming Services are illustrated and analyzed.

> Link to this presentation

3:25 - **Henry S. Rzepa** (Department of Chemistry, Imperial College of Science):

The vision of a chemical semantic web

Abstract: The increasing trend enabled by the Web is of fusion between the sources of primary data (Instruments, modelling and simulation, databases) and

the repositories of terms, dictionaries and peer-reviewed publications in a multi-disciplinary environment. Much of this fusion currently has to be achieved with a significant injection of "human perception", both on the part of the creators and authors of the information and knowledge, and of second and tertiary publishing resources. Part of this process involves establishing "trust" and common semantics within a domain such as chemistry. XML is essentially a remarkably powerful infra structure built up over the last six years which provides a set of guidelines for introducing appropriate elements of machine processing capability to the overall process, and where the previously expensive need to create software and tools to achieve this is ameliorated by the ability to re-use a vast communal toolkit. These themes will be illustrated in the context of a chemical vision of Berners-Lee's Semantic web, in which the "datument" (data+document) and the information objects it contains plays a central role. In such datuments, the chemistry can be identified via specific namespaced components such as CML (Chemical markup language) seamlessly integrated with other components such as MathML, STMML, SVG. Demonstrations of these concepts will be included in the presentation.

> Link to this presentation
<http://rzepa.ch.ic.ac.uk/talks/chemidxml/>

4:05 - **Peter Murray-Rust** (Unilever Centre for Molecular Informatics, University of Cambridge):

The chemical semantic web: a common infrastructure for chemistry

Abstract: An XML infrastructure for a domain must be built from carefully designed components, which have an Open architecture and which can interoperate. The components describe agreed subdomains of the discipline, such as molecules, reactions, spectra, computations, analytical and theoretical. They are defined by XML Schemas, which take advantage of re-use of existing designs, e.g. reactions can be based in part on molecular components. Schemas allow very precise validation of chemical information objects - it is possible to prevent invalid input to systems - and this will lead to a degree of quality missing in current practice but essential for the Semantic Web. A key resource is metadata which must be systematised for Chemistry and must use a universal architecture. Metadata is required for discovery of resources, validation, and for descriptions and annotation (e.g. the history of a piece of chemical information

as it passes through the community). Allied to this is an XML-based query system designed for chemical applications.

We shall describe an Open system, including a toolset, on which groups can layer their applications.

4:45 - **Stephen E. Stein** (Physical and Chemical Properties Division, NIST):

An XML namespace for IUPAC

Abstract: IUPAC has long served as a source of standard terminology in Chemistry. However, these "standards" have generally been expressed in conventional publications for use in conventional publications. A project intended to transform a portion of these standard definitions into an IUPAC "XML namespace" to aid the digital transmission of chemical information is under consideration within IUPAC. The scope, goals, methods and challenges of this potential project will be discussed.

> download pdf file of this presentation (pdf file - 553KB)

> For more about XML in Chemistry, see *Chem. Int.* July '02, p. 3