

XML in Chemistry

D. P. Martinsen and S. R. Heller, *Organizers*

Tuesday, March 28, 2006

8:30 —53. Computational Chemistry in XML. P. Murray-Rust, H. S. Rzepa, J. A. Townsend, D. Wilson

Abstract

High-throughput computation of the structures and properties of molecules and materials is now supported by a generic infrastructure based on Chemical Markup Language (CML). By converting the input to and output from a code (such as CASTEP, GAMESS, DL-POLY, SIESTA, etc.) it is possible to chain together several operations which can process jobs automatically. This is supported by flexible dictionaries (XML) and ontologies (RDF) to represent computational processes, physical properties, strategies, parameters and algorithms. This can support coarse-grained parallelism, data mining and analysis. XMLisation is either through the additional of CML libraries to the code or transduction of legacy data (stylesheets and parsers). An important benefit is the increased detection of program errors and control of input and output quality.

9:00 —54. AnIML: A new XML-based standard format for analytical data.

M. Fiege

Abstract

Analytical instruments today are producing data in a multitude of different formats. This makes the interchange of data between systems difficult. To deal with this problem, standard formats like ANDI and JCAMP have been created in the past. Based on the experience gained with these, ASTM has started an effort to create a highly flexible yet validateable standard format based on XML that can accommodate any kind of analytical data. This presentation will give an introduction into the concepts behind AnIML, and will show how AnIML can be customized to suit special needs without breaking the standard.

9:30 —55. Chemistry publications in CML. P. T. Corbett, P. Murray-Rust, N. E. Day, J. A. Townsend, H. S. Rzepa

Abstract

Much of the semantics in a chemistry article are now supported by Chemical Markup Language (CML) describable by an XML Schema (XSD). CML can support molecules, structures, reactions and reaction schemes, spectra (including annotations) and physicochemical data. These are supported by dictionaries and lexicons (also in XML) that provide linguistic and semantic support for the markup. Manuscript components can be created either with a range of authoring tools or through linguistic processing of conventional text. The semantics in such papers can now be processed by machine leading to high-throughput information extraction. A major feature is that chemical documents will be quicker to author and have a higher quality of embedded data and structure through machine validation.

10:00 —56. Ensuring the interoperability of the Analytical Information Markup Language (AnIML). **A. Roth**, R. Jopp, P. J. Linstrom, G. W. Kramer

Abstract

AnIML (Analytical Information Markup Language) is being created by ASTM Subcommittee E13.15 to describe chromatography and spectroscopy data and metadata based on XML (eXtensible Markup Language) and its associated technologies. Once in AnIML format, analytical data can be interchanged over the web, converted to other formats, validated, or visualized in multiple formats using existing XML-based tools.

AnIML is built around a core schema that defines ways for describing almost any data. Technique Definition files are used to constrain the myriad data description mechanisms available for a given analytical technique to only those commonly accepted, to delineate the metadata items ordinarily associated with such domain data, and to permit content extension by vendors and users without changing the core schema. This presentation will describe the naming and design rules (NDRs) and other techniques being employed to ensure that AnIML is as interoperable as possible with other markup languages.

10:30 —57. Incorporating Units Markup Language (UnitsML) into AnIML (Analytical Information Markup Language). **R. Jopp**, A. Roth, P. J. Linstrom, G. W. Kramer

Abstract

Units Markup Language (UnitsML) is being developed to encode scientific units of measure using XML (eXtensible Markup Language). The development and deployment of a markup language specifically for units will allow for the unambiguous storage, exchange, and processing of numeric data, thus facilitating collaboration and the sharing of information, especially over the Internet. Incorporating UnitsML into other markup languages prevents duplication of effort and improves interoperability.

ASTM Subcommittee E13.15 is creating AnIML (Analytical Information Markup Language) to describe chromatography and spectroscopy data and metadata based on XML and its associated technologies. AnIML facilitates access to analytical data by building in descriptions of the data and metadata with delimited tags. UnitsML is being employed to handle the markup of the units information in AnIML. This presentation will describe how UnitsML is being used and how it is being incorporated into AnIML.

11:00 —58. Integration of the Chemical XML standard in Laboratory Content Management Systems. M. Burke

Abstract

Abstract text not available.

11:30 - BREAK

2:00 —64. Use of XML for analytical instrument control. A. Mutin

Abstract

There is a growing interest among analytical instrument users for multi-vendor support of their equipment in terms of instrument control, data acquisition and data processing capabilities.

Different vendors provide different software interfaces to control their instruments. Many users prefer to standardize on software to minimize validation and training costs, while keeping their hardware diverse. Because most laboratory software have limited multi-vendor support, often times when shopping for a new instrument users are burdened by a necessity to stay with one type of software.

XML-based web service embedded into an analytical instrument is a new technology that can potentially solve multi-vendor support limitations of current software. A web server equipped HPLC is directly connected to a computer network. Such system can be controlled from any PC without a need for any additional software except for a web browser such as the Internet Explorer. If laboratory software is linked with such web-service one can easily assemble systems out of multi-vendor hardware components while controlling them from the same application. In addition, the data can be interchanged between instruments, applications and databases using the Analytical Information Markup Language (AnIML) format.

2:30 —65. XML for comprehensive 2-D gas chromatography. A. Visvanathan, Q. Tao, S. E. Reichenbach, M. Li, S. Deshpande, X. Tian

Abstract

Comprehensive two-dimensional gas chromatography (GCxGC) is an emerging technology for chemical separations that provides an order-of-magnitude improvement in separation capacity, significantly greater signal-to-noise ratio, and higher-dimensional chemical ordering compared to traditional gas chromatography. Information systems are being developed to visualize, process, and analyze the complex data produced by GCxGC. The eXtensible Markup Language (XML) is powerful and flexible technology for structuring and describing data and so is especially well-suited for expressing the rich relationships that are only beginning to be discovered in GCxGC data. This paper describes the use of XML for GCxGC data, metadata, and information, including raw and processed data, peak tables, templates for chemical identification, journals and scripts with processing sequences, and formal reports. Ongoing work is evaluating XML-based technologies, such as the ANalytical Information Markup Language (AnIML), for GCxGC methods.

3:00 —66. Integrative analytics and data harmonization in TOPCOMBI. F. Gilardoni, D. Farrusseng

Abstract

Best practice data mining techniques are ineffective without high-quality data, fast and reliable access to the information and a consistent capture of data and processes. The experimental issue is addressed with an apposite methodology by the experimentalist. The second topic is more challenging because it has to cope with the disparate data structures and data exchange protocols, and usually requires a plethora of data mining and analytical tools. This heterogeneous information is overwhelming to maintain and requires tailored tools to be utilized. This drastically impacts the total cost of ownership of the Informatics infrastructure, precludes a proper dissemination of knowledge and hinders scientific breakthroughs. TOPCOMBI, a project for Nanotechnologies and Nanosciences funded by the European Commission, dedicates collegially important resources to harmonize and integrate this incongruent information issued from high-throughput platforms, instruments, and data mining. TOPCOMBI is investigating how XML schemas – existing and in development – and webservice suit the stringent requirements for data standardization, accessibility, portability and modularity with new computational techniques. Also, the consortium is exploring how the semantic and the underlying ontology defined in the XML schema can facilitate the transformation of data into tangible knowledge. We will present the work in progress and how the integrative analytics paradigm and data harmonization operate on both software and data.

3:30 —67. XML for quantum chemistry program input. G. S. Kedziora, S. R. Brozell, E. A. Stahlberg

Abstract

A new XML input format for the COLUMBUS suite of Multi-Reference Configuration Interaction (MRCI) programs will be described. This XML language, called COLUMBUS Input Meta Language (CIML), is designed to be easy for a human to prepare with a text editor as well as by the back end of a GUI. It specifies a clear and complete description of the computation that is suitable for archival. Since MRCI is generally not used as a model chemistry, CIML provides the flexibility for tailoring a MRCI calculation to a specific molecule, which often requires careful planning and exploratory runs. A corresponding program has been written that parses the CIML, produces the legacy input for the COLUMBUS programs, and provides the user with useful feedback about the calculation. The ontology of more general quantum chemistry calculations will be discussed in relation to CIML.