Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language

Nico Adams,*,† Jerry Winter,‡ Peter Murray-Rust,† and Henry S. Rzepa§

Unilever Centre for Molecular Science Informatics, University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom, Unilever HPC R&D Port Sunlight, Physical and Chemical Insights Group, Quarry Road East, Bebington, Wirral, CH63 3 JW, United Kingdom, Department of Chemistry, Imperial College London, SW7 2AZ, United Kingdom

Received June 25, 2008

Polymers are among the most important classes of materials but are only inadequately supported by modern informatics. The paper discusses the reasons why polymer informatics is considerably more challenging than small molecule informatics and develops a vision for the computer-aided design of polymers, based on modern semantic web technologies. The paper then discusses the development of Polymer Markup Language (PML). PML is an extensible language, designed to support the (structural) representation of polymers and polymer-related information. PML closely interoperates with Chemical Markup Language (CML) and overcomes a number of the previously identified challenges.

INTRODUCTION

Polymers are ubiquitous in the modern world and are of growing importance in materials science. The world production of polymers is consistently increasing and novel application areas are constantly being developed. Furthermore, the advent of high-throughput experimentation and parallelization,[1−4] as well as novel synthesis[5,6] and processing[7] approaches, has triggered a profound change in the way polymer science is being carried out. The increasing speed with which new polymers can be synthesized and screened almost automatically generates the need for sophisticated informatics tools to manage the vast amounts of data that can now be produced and to develop "design-rules" (structure−property relationships) for polymers. It is surprising therefore, that, unlike small molecule informatics, the field of polymer informatics is woefully underdeveloped and almost nonexistent. One explanation for the scarcity of informatics solutions that are suitable for the particular problems posed by polymers is inherent in the nature of polymers themselves.

In conventional chemical informatics, the use of three metaphors, namely, the "molecular formula", the "structure diagram" and the "connection table" are tremendously powerful tools not only for the description of molecular structure, but also for attempting to relate that structure to the macroscopic properties of a pure bulk sample of the corresponding molecule. Traditionally, chemical informatics has embraced the central paradigm of (synthetic) chemistry: the preparation of pure and therefore well-defined compounds and the investigation of their behavior. For synthetic polymers, however, these metaphors loose a significant part of their meaning, as polymers are mixtures or ensembles of macromolecules, all of which have slightly different architectures and therefore (assuming that the central dogma of chemistry holds) also slightly different properties. For synthetic polymers, molecular weight distributions are almost unavoidable and even extremely well-behaved polymerizations, carried out under carefully controlled conditions, lead to polymers with polydispersity indices (PDIs) greater than 1 (very controlled living polymerizations achieve PDIs of around 1.03, see, for example, ref 8). An intrinsic structural variability therefore exists in synthetic polymers and what is observed as a physical property of a macroscopic polymer sample, in reality represents the average over the properties associated with each of the macromolecules in a given ensemble. This introduces considerable fuzzyness into the description of a polymer and leads to the breakdown of both the "pure molecule" and the connection table paradigmata. Apart from differences in length, side reactions in polymerizations often lead to branching, cross-linking, the introduction of unsaturations at the chain end, variations in tacticity, etc. It is usually possible to detect these phenomena using spectroscopic or other characterization techniques, but we are generally unable to pinpoint the particular macromolecular members of the ensemble in which these occurred, which increases the fuzzyness of the "polymer" concept even further.

While for some polymers (e.g., polystyrene) we have a fairly good understanding of how the atomistic structure of the constituent macromolecules can be represented, we are unable to draw a detailed atomistic picture of the structure of other polymers (e.g., Bakelite, phenol/formaldehyde resin). In these cases, we may know how the polymer was made, but we only have a partial understanding (or no understanding at all) of the final product. This tremendous complexity led Wilks to comment in the late 1990s, that "a polymer should have a unique, unambiguous structural representation, even if it fails to depict fully the complete structure of the polymer."[9] Effectively, this comment advocates a reductionist/abstracted approach to the representation of polymer structure. Unfortunately, this is precisely what has happened in the past and even now represents the state of the art.

* To whom correspondence should be addressed. E-mail: na303@cam.ac.uk.
† University of Cambridge.
‡ Unilever HPC R&D Port Sunlight.
§ Imperial College London.

10.1021/ci8002123 CCC: $40.75
Published on Web 11/08/2008

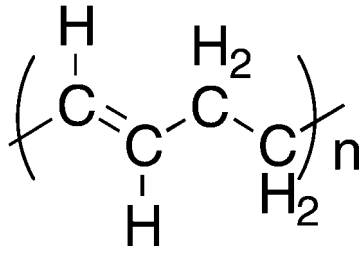


Figure 1. Constitutional repeating unit structure of polybutadiene.

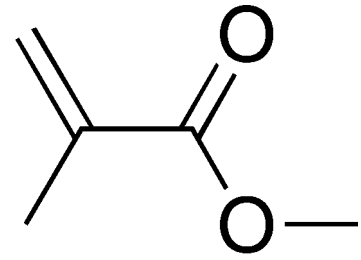


Figure 2. Structure of methacrylic acid, methyl ester.

Current State of the Art in Polymer Representation. In information systems in common use today, polymers are either registered using a string (name), which only has limited chemical information, or using an idealized and abstracted graphical representation (i.e., a constitutional repeating unit (CRU)). Either representation has specific associated problems, which we will briefly discuss.

Name-Based Representations. There are generally three different approaches to naming polymeric materials: (a) the invention of trivial names/identifiers, (b) the construction of the name based on the component monomers of the polymer (source-based representation, e.g., polystyrene) or (c) the construction of a name based on the polymer constitutional repeating unit (structure-based representation, e.g., poly(1-phenylethylene)).

While some trivial names for polymers are derived from monomers or CRUs which do hold a limited amount of chemical information, others often are tradenames or names derived from the inventor's name, which have entered common usage (e.g., Bakelite) and it is not always possible to draw inferences as to the structure of the polymer on the basis of a trivial name. As such, we will not discuss trivial names further.

Which of the other possible name-based representations of a polymer is chosen in an information system, depends on the different nomenclature philosophies used in the chemical domain, and there is no general agreement as to which representation is preferable. As an example, the Chemical Abstract Service (CAS) will register the polymer whose CRU is depicted in Figure 1 as "1,3-butadiene, homopolymer",[10] whereas the International Union of Pure and Applied Chemistry (IUPAC) will allow registration under a number of source-based and structure-based names such as "polybutadiene" (source-based), "poly(but-1-ene-1,4-diyl)" (structure-based), "1,4-polybutadiene" (source-based), or "poly(buta-1,3-diene)" (source-based).[9] 1,4-Polybutadiene is in essence a source-based name (polybutadiene), but it provides additional information concerning the structure of the polymer: 1,3-butadiene can polymerize by both 1,2- and 1,4-addition.

While most name-based representations have a certain amount of chemical information contained in them, the information is either ambiguous (e.g., polybutadiene; no indication concerning the location of the double bond in the diene monomer and hence no indication of the structure of the resulting polymer) or, as in the case of source-based representations, make it potentially difficult for a machine to infer the structure of the macromolecule associated with a particular monomer. The profusion of possible names for a chemical entity also has the potential to create information silos across multiple information systems: if multiple systems use different names for the same polymer, mapping the information contained in these systems to the same concept is almost impossible in the absence of a dictionary.

Finally, the construction of any chemical name (ignoring trivial names) is accomplished through the application of nomenclature rules. However rules-based systems change over time and are therefore subject to historical (dis-)continuities: the compound depicted in Figure 2 is registered as "methacrylic acid, methyl ester" in the eighth Collective Index (CI) of Chemical Abstracts and as "2-propenoic acid, 2-methyl, methyl ester" in the ninth CI.[9] Again, historical discontinuities in rule-based indexing systems have the potential to lead to "walled-off" information.

Graphical/Structure-Based Representations. The alternative to using a name-based representation for polymers has thus far involved the use of a structural sketch. In this context, "structural representation" refers to the use of chemical structure diagrams as a graphical metaphor for a connection table and should not be confused with the "structure-based names" representation discussed above. Indexing systems such as CAS register polymeric compounds either based on the structures of the component monomers or using the structure of the constitutional repeating unit as an abstracted representation. The component monomer approach clearly holds the least information about the corresponding polymeric product: while such an approach states which monomers the polymer was made from (and sometimes this is all that can be said about a polymer), it does not provide any further information such as the degree of polymerization, endgroups, etc. Although the resulting polymer structure can often be deduced by a human, it causes significant problems when the structure is to be inferred by a machine, particularly in cases, where there is a large structural change from monomer to resulting polymer (in ring-opening polymerizations, for example, large structural shifts are observed in going from a monomer to the corresponding macromolecule). In other words, polymer knowledge is, at best, present implicitly.

The constitutional repeating unit is an abstracted and reduced representation of a macromolecule. As such, it holds more information about the corresponding polymeric product than the component monomer representation (e.g., degrees of polymerization are usually reported, as are endgroups). However, only the idealized skeletal architecture of the macromolecular constituents of the polymer is provided and information about side reactions leading to, for example, branching or chain-end unsaturations are not usually accounted for. Beyond these more general challenges, another problem arises. When considering the expanded CRU of polybutadiene (Figure 3), for example, it becomes obvious that several valid repeating units can be constructed. To decide upon a preferred repeating unit, a set of rules needs to be developed and applied. As discussed above, rules are

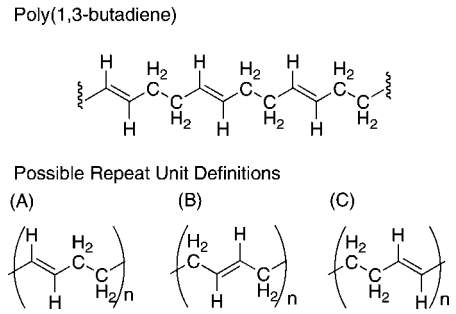


Figure 3. Possible repeating unit definitions of polybutadiene.

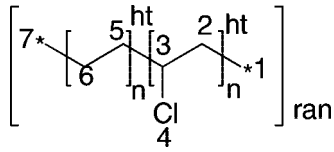


Figure 4. Sgroup representation of a random copolymer.

subject to change over time and could lead to information compartmentalization.

One notable exception to these challenges, is a proprietary technology by Symyx (previously MDL Information Systems Inc.), which does not require construction rules for repeating units. MDL has developed the concept of "Sgroups", collections of molecular fragments for the representation of polymers, which, in combination with the "flexmatch" search technology, avoid the problem of multiple repeating units.[9]

A typical Sgroup representation of a random copolymer is given in Figure 4. Repeating units in Sgroups are enclosed by square brackets, with the subscript "n" being placed to the right of the closing bracket. A superscript, also to the right of the closing bracket indicates the orientation of the repeating units (hh = head-to-head, ht = head-to-tail, eu = either unknown, left justified). Bonds crossing the brackets enclosing the repeating unit indicate known connectivity. Furthermore, the crossing bonds in a "ht" arrangement in the above example imply that the repeating units can either connect to themselves or to another repeating unit because it is a random polymer. This type of labeling also prevents mismatching of bonds, with mismatched bond types only allowed to be hh. When searching for a particular polymer by repeating unit, MDL's "flexmatch" search cyclizes all possible repeating unit representations, creating molecules which are "phase-shifted" with respect to each other, meaning that all structures look identical. Sgroups also form the basis of the MACCS-II Substance module from MDL, which allows not only for the granular representation of polymers, but also for the representations of formulations, impurities and nonstoichiometric mixtures.[11]

The discussion presented so far has only scratched the surface of polymer nomenclature and registration. However, the fundamental problem with almost all types of representations is, that they attempt to describe an inherently fuzzy concept in terms of an inappropriate metaphore, namely a (reduced) connection table. Apart from these more fundamental concerns, any representation of polymers should also take into account other factors, such as the presence of impurities or additives or postprocessing steps. All of these can have a tremendous impact on the properties of a polymeric sample and are, with the exception of Sgroup technology, not usually accounted for. Moreover, commercial manufacturers will often only provide information about the method of manufacture of a polymer and some of its properties, without providing detailed information concerning the chemical structure and composition of the material (usually for commercial reasons). This information, too, should be captured in a machine readable form.

There are examples of the successful application of the connection table approach for property prediction. Van Krevelen[12] has developed a substantial collection of group contributions for the prediction of polymer properties, whereas Bicerano's system makes use of simple 2D descriptors to calculate a set of properties, which are, to a large extent, identical to those that can be calculated from group contributions.[13] While the connection table metaphore has therefore had some successes, it is clearly limited and does not deal with many of the phenomena outlined above. However, more often than not, this is precisely what would be needed to establish meaningful correlations between the structure and composition of a (bulk-)polymer and its physicochemical properties.

Chemistry and the Web. Current developments on the world wide web are starting to have a marked impact on how chemistry and chemical data is structured, stored, distributed, and presented.[14] While the version of the web, which most users experience today is mainly one of documents, which are intended for consumption by a human reader and are interconnected by hyperlinks, the web is currently evolving toward a semantic web of data,[15] which will allow machines to discover data, understand its meaning and to autonomously act on it. In a typical scenario, a polymer chemist wishing to design a new polymeric entity against a given requirements profile could be envisaged to deploy an agent (a piece of software, which acts on behalf of a user) to collect information and data about polymers which either completely match or approximately match the requirements profile or to infer required information from "the cloud" (e.g., finding, retrieving and combining information on the web, in-house resources, and open and proprietary databases) and to return a list of suggested polymers or polymer architectures to the researcher. In practice, this means that polymer information not only needs to be discoverable, but also needs to be endowed with well-defined meaning. The semantic web is therefore a vision of machine-readable data, which can be used for automation, integration, and reuse across different applications, as well as a vision of intelligent agents, which can retrieve and manipulate relevant information. The technical foundations necessary to accomplish this vision are, among others, Extensible Markup Language (XML),[16] the Resource Description Framework (RDF),[17] and Web Ontology Language (OWL).[18] The power of the semantic web is beginning to be leveraged across the sciences and also in chemistry.[19−23] A particular example, in this context, is the development of Chemical Markup Language (CML), which was designed to provide functionality for atomic, molecular, crystallographic, spectroscopic, and reaction information.[14,24−28] The language is extensible, thus allowing other areas of chemistry to be covered as well. CML has now been widely adopted by the chemical informatics community and support for the language has also been implemented

in commercial products such as ChemDraw, as well as open source software, such as X-DrawChem.

However, like most other formal languages used to describe chemical structure, Chemical Markup Language makes heavy use of the connection table metaphore to codify molecular information. The above discussion has already made it clear that in the area of polymer science, the connection table is an inadequate structural representation. For this reason, we have developed Polymer Markup Language (PML) as an extension to CML, which specifically supports the peculiarities of polymeric materials and is completely interoperable with CML.

Relationship of PML to CML. We have published definitive papers in CML approximately once a year over the last 5 years. In each case there has been a specific domain need for a further specification, but each event has also given the opportunity to incorporate general developments in XML and related technologies. Each publication has introduced one or more new general features into the language, which are applicable beyond the bounds of the specific domain. CMLSpect,[26] for example, showed how the `convention` attribute could be used to define and distinguish different approaches to the practice and formalism of recording and exchanging spectral data.

In this paper, we introduce a few new terms into the markup vocabulary but tackle several emerging problems, which have only become tractable as markup technology has developed. These are all required for polymers but are also applicable beyond this particular domain. They are discussed in detail below:

• CML as computable language. Hitherto CML has taken a simple declarative approach where the document represents the structure of the information. For polymers we require a language which includes computational instructions, which might either produce new (CML) documents or modify the existing one. In PML `<join>` is an example of a computable element, and `countExpression` is an example of a computable attribute.

• Free content model. In the early versions of CML, the DTD (and later the Schema) were regarded as an authoritative statement as to whether the document was valid or invalid. In particular, the *content model* specified what children an element could or must contain, often dictating the precise order. It has become clear that chemistry is too rich to be bounded by this, and so, CML is developing as a model where only the element vocabulary, but not the content model, is specified.

• Variability. Variability acknowledges that a concept is known to vary (for example the degree of polymerization). In principle it is possible to give quantitative expressions for distribution functions of quantities. This may be applied not only to numeric quantities but also for chemical ones such as the presence of functional groups.

• Uncertainty. Here we acknowledge that the quantity or concept is wholly or partially unknown. For example, when considering the cross-linking of polymers, it may be known that a branch from one chain links to a different chain, but precisely which one is unknown.

• References. The `ref` attribute represents a pointer to an object, which can either be called by value or called by reference. There are frequent requirements to use a standard set of building blocks for polymers and call-by-value is the most common construct.

• Parameterization. Because PML can be either implicitly or explicitly computed it is possible to create documents with *free variables* (i.e., symbols representing numeric or chemical quantities). One approach we have used is to embed the computable language XSLT[29] into specifications of PML and to evaluate expressions in a lazy manner. For example we have used `<molecule> <xsl:value-of select= "$tor123"/> </molecule>` to inject a user-specified value of the torsion into the description of a polymer using the `<xsl:param>` mechanism. The concept of a computable declarative chemical language is sufficiently important that we shall formalize it elsewhere. The free content model likewise represents the first step in an important development in CML to create a grammar from which complex objects can be constructed but where the semantics are defined by the creator. It is likely that the convention attribute will play a central role in this definition. The parametrization also requires a loosening of the CML specification requiring that simple content can be replaced by an evaluatable expression. It is likely that we shall look to newer XML languages (e.g., RELAX-NG, Schematron and RDFS/OWL) to express some of these constraints and it would be premature to list specific syntax here.

We note that the use of computable functions and free variables can be used to define two separate concepts. With unbound variables it represents a precise statement of the grammar of a system: $R_1-C(=O)OH$ represents "any carboxylic acid". It can be deduced that the molecule contains a carbonyl and a hydroxy group. It could be deduced that the molecule could form an ester with an alcohol. With bound variables it represents one or more bound systems: $R_1-C(=O)OH$; (R_1 = Me, Et) represents acetic acid, propanoic acid, or both. The unbound system still carries important and precise information, even though not substituted by values. For example, the equations $f(x) = x*x$ and $g(y) = \sin(y)$ lead to $g(f(z))$ because $\sin(x*x)$, even though no values of x and y have been proposed. In a similar manner, we expect that it will be possible to reason over free variables in CML constructs.

An example of a computable process is the serial elongation of an AB copolymer chain. If a growing polymer has probabilities pA, and pB of adding A or B and pT or terminating, then this can be expressed in PML. However if we apply these probabilities to a given chain, we create a new chain and can repeat this process until termination. This is an example of a Markov chain (Figure 8) and shows how the description of a polymer (free variables) is also a recipe for computation (bound variables). We can at every step represent the growing polymer chain by a CML object and the transition probabilities by the fragment. This represents an algorithm for generating polymers in a stochastic manner which can be elaborated to allow for depletion of reactants or addition of new materials. Similarly the computation of torsion angles in a growing polymer can be adapted to respond to the size and conformation of the current chain. In this way, the successive PML snapshots represent complete knowledge of the state of the system and are therefore independent of the program used to compute them.

Table 1. PML Element Subset Definitions

CML element	definition
<fragment>	The element <fragment> is a container for a <molecule>, potentially to be joined to other fragments. In addition there may be fragmentLists, which represent branches from the molecule. There may also be a <join> child, which is normally only found if there is a @countExpression.
<fragmentList>	A <fragmentList> is a container for one or more <fragment>elements and <join > elements. The normal content model for this element is <fragment>, <join>, <fragment>.
<join>	The <join> element will normally use atomRefs2 to identify two r atoms (i.e., elementType="r" that should be joined. The atoms to which the r atoms are attached are then joined by a new bond, and the r groups are then deleted. It is currently an error if these atoms already have a connecting bond.

Design Criteria for Polymer Markup Language. Because PML is an XML dialect, many of the goals formulated for XML also apply to Polymer Markup Language. These are

1.PML shall support a wide variety of applications.
2.It shall be easy to write programs, which process PML documents.
3.PML documents should be human legible and reasonably clear.
4.4.The PML design should be prepared quickly.
5.The design of PML should be formal and concise, with conciseness tempered by scientific need.
6.PML documents shall be easy to create.
7.Terseness in PML documents is of minimal importance because explicit semantics are preferable to abbreviated approaches. Apart from these general goals, PML has the following specific goals:
1.PML shall provide a granular and normalized view onto a polymer and polymer information.
2.PML should be based on CML and reuse CML components where appropriate.
3.PML should interoperate with other mature STM markup languages.
4.PML should be namespace aware and tools processing PML should support this.
5.PML should support current polymer informatics wherever possible.
6.If more than one convention is in use, both should be allowable using the PML convention attribute.
7.PML shall have explicit support for computational applications.

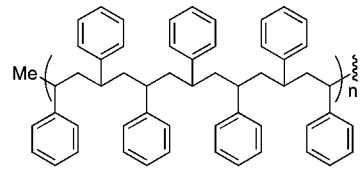
PML Shall Provide a Granular and Normalized View onto a Polymer and Polymer Information. We have already discussed that polymers are inherently fuzzy because they constitute ensembles of macromolecules of differing connection tables. Furthermore, while it is possible to draw an abstracted connection table (e.g., representing polystyrene using an atomistic representation of the constitutional repeating unit of one of the macromolecules in the ensemble) for some polymers, for others this is impossible, and we can only talk about the monomers the polymer was prepared from. Yet we might wish to compare a polymer for which an atomistic picture can be drawn to one where this is not the case. Polymer Markup Language must enable this comparison to be made and must therefore provide a level of normalization that is more coarse-grained than the atomistic picture, yet allow the mapping of the coarse-grained representation onto a fully atomistic description in cases where this is appropriate. PML allows data to be associated with the polymer representation at the atom, fragment, molecule, and molecular ensemble level, and as such, PML represents a "normal form" of the polymer description. We are not claiming that PML is a canonical representation.

PML Should Be Based on CML and Reuse CML Components Where Appropriate. The initial development of PML is mainly concerned with supporting polymer chemistry and physics and ignores the area of polymer processing altogether, although it may well transpire in the future that aspects of PML are useful for the processing community too. Because CML already covers significant parts of chemistry, its components should be reused wherever appropriate and will interoperate with the new vocabulary of PML.

PML Should Interoperate with Other Mature STM Markup Languages. Increasingly, the most valuable scientific work occurs at the interfaces of individual disciplines, with polymers, for example, finding increasing applications in the nanomedicine and pharmaceuticals arena.[30−35] For this reason, a researcher should be able to readily combine polymer (related) information with information and data from other fields. Interoperability with other STM markup languages is therefore mandatory.

Polymer Markup Language Element Set. PML makes heavy use of the following elements <fragment>, <fragmentList>, and <join>, which have since been incorporated into CML. The formal definitions of the elements are given in Table 1. The <fragment> element is central to PML and is a container for a <molecule>, potentially to be joined to other fragments. The CML element <molecule>, in turn, is a container for atoms, bonds and submolecules, along with properties, such as crystal and nonbuilt in properties. <molecule> need not represent a chemically meaningful molecule. It can contain atoms with bonds (as in the solid state), and it could simply carry a name (e.g., "taxol") without formal representation of the structure. It can contain "submolecules", which are often discrete subcomponents. Furthermore, <molecule> can contain an element to contain data related to the molecule. Within this, in turn, can be string/float/integer and other nested lists. Normally molecule will not contain <fragment> or <fragmentList>.

<fragment> and <join> PML has required the development of a language component to represent molecular




```
<?xml version="1.0" encoding="UTF-8"?>
<fragment convention='cml:PML-basic'
      xmlns='http://www.xml-cml.org/schema'
      xmlns:g='http://www.xml-cml.org/mols/geom1'>
      <fragment id="f0">
            <fragment>
                  <molecule ref='g:me' />
            </fragment>
            <join moleculeRefs2="PREVIOUS NEXT" atomRefs2="r1 r1">
                  <torsion>90</torsion>
            </join>
            <fragment countExpression="*(6)">
                  <join moleculeRefs2="PREVIOUS NEXT" atomRefs2="r1 r1">
                        <torsion>90</torsion>
                  </join>
                  <fragment>
                        <molecule ref="g:ch">
                              <join moleculeRefs2="PARENT CHILD"
                                    atomRefs2="r3 r1">
                                    <torsion>90</torsion>
                                    <molecule ref="g:benzene" />
                              </join>
                        </molecule>
                        <join moleculeRefs2="PREVIOUS NEXT" atomRefs2="r2 r2">
                              <torsion>90</torsion>
                        </join>
                        <molecule ref="g:ch2" />
                  </fragment>
            </fragment>
      </fragment>
</fragment>
```

Figure 5. Simple PML document describing a styrene heptamer.

fragments which are covalent parts of larger systems. (Noncovalent association is already covered by nested <molecule> elements). The fragment elements have wider use beyond polymers and can be used for

- Generic molecules (with substitutable free variables).
- Markush systems where the free variables can be algorithmically substituted, either exhaustively or stochastically.
- Libraries of components for assembly into larger systems (because fragments can be nested or evaluated recursively).

Further discussion is omitted here although examples are available in the JUMBO distribution. A fragment-based approach to molecular construction was independently reported by Sankar and Aghila.[36]

Construction of a Simple Homopolymer. Figure 5 shows a simple PML document describing a styrene heptamer. The molecule is constructed starting from a root <molecule ref="g:me"/>, in this case a methyl group. PML is, of course, fully namespace aware, and hence, we can use a shorthand and look up the full CML connection table for the methyl group or any of the other fragments in other documents (xmlns:g="http://www.xml-cml.org/mols/geom1"). The root is then joined to the contents in the next <fragment> container, namely, a −CH− fragment (<molecule ref ="g:ch"/>), which, in turn is connected to a $-CH_2-$ (<molecule ref="g:ch2"/>) and a phenyl fragment (<molecule ref="g:benzene"/>). The fragment thus defines one complete repeating unit, although this is coincidental and not required by the language. In principle the user is free to choose any type of chemical fragment in the polymer construction process, such as multiples of repeat units, structural units that do not correspond to the traditional understanding of repeating units, or even whole macromolecules themselves (e.g., in their role as macroinitiator). In the present example, the <fragment> element carries a countExpression attribute, which indicates to PML processing software to iterate over the <fragment> another six times. The countExpression attribute is therefore computable, in that it provides instructions on how to produce new CML/PML documents or alter existing ones. Furthermore, the countExpression as a generating function can evaluate either deterministically or stochastically, thus opening the door to modeling ensembles of macromolecules (e.g., molecular weight distributions etc.). Computability represents a departure from the hitherto declarative nature of XML/CML.

Instructions concerning how the fragments are to be joined are contained in a computable element, namely the <join> element. <join> uses the atomRefs2 attribute to identify dummy atoms of type "r_n" in those fragments, which are to be joined together. Once identified, the atoms, to which the r dummies are attached, are joined by a new bond, and the dummies are deleted. In the present example, the r_1-group of the methyl fragment is joined to the r_1-group of the methylene fragment (atomRefs2="r1 r1") as illustrated in Figure 6. Because the methyl group is the first fragment, it is identified as the parent fragment in the moleculeRefs2 attribute (moleculeRefs2="PREVIOUS NEXT"), and the methylene fragment is identified as the next fragment because it follows the methyl group. The moleculeRefs2 attribute makes reference to two distinct molecules and is, in principle, available for any reference to molecules but normally will be the normal reference attribute on the join element. The <torsion> element describes the torsional angle between four distinct atoms. It is important to note, that in CML usage, these atoms do not need to be formally bonded. In our particular example, the phenyl group

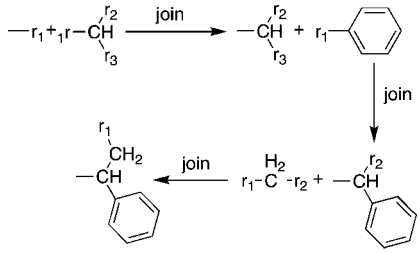


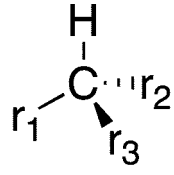
Figure 6. Pictorial representation of the joining instructions contained in the PML description of a polystyrene heptamer (Figure 5).

is modeled as a child of the methylene group, which is indicated by the moleculeRefs2="PARENT CHILD" statement.

In regular macromolecules, monomers are normally connected in a head-to-tail configuration. However, particularly for vinyl polymers, head-to-head or tail-to-tail arrangements also need to be considered. It should be apparent by now that these situations can be dealt with easily, by simply defining appropriate <join> operations in the PML document. Should the user wish to make head- and tail-atoms or groups explicit, this can be achieved via the use of the CML label element or attribute.

Tacticity. When considering polymers such as polystyrene, tacticity is an important consideration because it has a significant influence on polymer properties. Polymer Markup Language deals with tacticity by using appropriately constructed fragments during the polymer building process. To construct (one enantiomer of) isotactic polystyrene, for example, only one styryl-enantiomer (relatively all R or relatively all S) is used (with the chirality being defined for the −CH− fragment), whereas syndiotactic polymers make use of alternating R and S enantiomers and for atactic polymers, the R or S enantiomers are selected at random during the build process. The chirality in fragments, in turn, is defined by using the <atomParity> element of CML. This element defines the stereochemistry around an atom center and essentially follows the Molecular Information File (MIF) format,[37] by using four distinct atoms to define chirality. These can be any atoms, though in principle they are bonded. The value of <atomParity> is a signed number (either 1 or −1 for either enantiomer and 0 if two or more atoms are coincident or the configuration is planar). An example of the use of <atomParity> is given in Figure 7.

Construction of a Statistical Copolymer. The construction of a statistical copolymer is significantly more complex than that of the simple homopolymer example shown above, but is nevertheless handled extremely well by PML. The PML document in Figure 8 exemplifies how a hypothetical statistical copolymer can be described. The polymer is constructed using a number of major fragments, in the present document Cl, acetyl, and EE and BB. The fragments EE and BB, in turn, are composed of smaller fragments (eo, eB, eE, po) in terms of a <fragmentList>. These, in turn, carry an attribute role= "MarkushMixture", indicating that the <fragmentList> describes a Markush structure. Furthermore, the fragments in the MarkushMixture <fragmentList> are annotated with probabilities (<scalar dictRef="cml:ratio" dataType="xsd:double">foo</scalar>), which represent the probabilities of the particular fragment following the preceding one in the chain. As discussed above, the construction of the chain is described by a Markov model. In the example in Figure 8, we start the construction of a hypothetical macromolecule (<fragment id="0">), by placing a Cl headgroup, which is subsequently connected to a member of the MarkushMixture EE. EE consists of the fragments eo, eE, and eB, and the probabilities of the headgroup being connected to any one of these fragments is given by the probabilities associated with that group. In the least likely case, the headgroup will be connected to an eo fragment, followed by the terminating acetyl group, which ends the polymerization and leads to the formation of a small molecule. In the most likely case, the headgroup will be connected to an eE fragment, which itself consists of an eo fragment connected to another EE fragment, that is, an ensemble of fragments with different probabilities of forming




```
<?xml version="1.0"?>
<molecule xmlns="http://www.xml-cml.org/schema" id="ch">
 <atomArray>
  <atom id="r1" elementType="R"/>
  <atom id="a2" elementType="C"/>
   <atomParity atomRefs="r1 r2 r3 a7">1</atomParity>
  <atom id="r2" elementType="R"/>
  <atom id="r3" elementType="R"/>
  <atom id="a7" elementType="H"/>
 </atomArray>
 <bondArray>
  <bond atomRefs2="r1 a2" order="1"/>
  <bond atomRefs2="a2 r2" order="1"/>
  <bond atomRefs2="a2 r3" order="1"/>
  <bond atomRefs2="a2 a7" order="1"/>
 </bondArray>
</molecule>
```

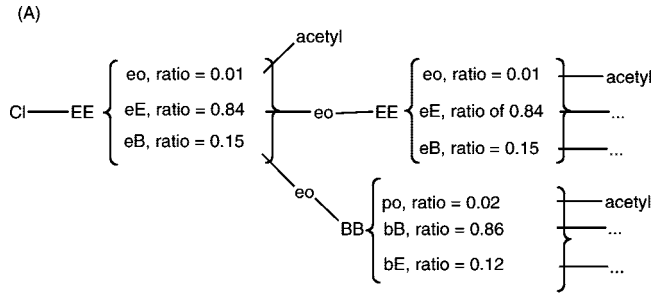
Figure 7. Using <atomParity> to define the chirality of a −CH− fragment.

```
<fragment  xmlns='http://www.xml-cml.org/schema'
  xmlns:g='http://www.xml-cml.org/mols/geom1'>
  <fragmentList>
    <fragment id='acetyl'>
      <molecule ref='g:acetyl'/>
    </fragment>
    <fragment id='cl'>
      <molecule ref='g:cl'/>
    </fragment>
    <fragment id='eo'>
      <molecule ref='g:eo'/>
    </fragment>
    <fragment id='po'>
      <molecule ref='g:po'/>
    </fragment>
    <fragment id='eE'>
      <fragment ref='eo'/>
      <join atomRefs2='r2 r1' moleculeRefs2='PREVIOUS NEXT'/>
      <fragment ref='EE'/>
    </fragment>
    <fragment id='eB'>
      <fragment ref='eo'/>
      <join atomRefs2='r2 r1' moleculeRefs2='PREVIOUS NEXT'/>
      <fragment ref='BB'/>
    </fragment>
    <fragment id='EE'>
      <fragmentList role='markushMixture'>
      <fragment ref='eo'>
        <scalar dictRef='cml:ratio'
dataType='xsd:double'>0.01</scalar>
      </fragment>
      <fragment ref='eE'>
        <scalar dictRef='cml:ratio' dataType='xsd:double'>0.84</scalar>
      </fragment>
      <fragment ref='eB'>
        <scalar dictRef='cml:ratio' dataType='xsd:double'>0.15</scalar>
      </fragment>
      </fragmentList>
    </fragment>
    <fragment id='bE'>
      <fragment ref='po'/>
      <join atomRefs2='r4 r1' moleculeRefs2='PREVIOUS NEXT'/>
      <fragment ref='EE'/>
    </fragment>
    <fragment id='bB'>
      <fragment ref='po'/>
      <join atomRefs2='r4 r1' moleculeRefs2='PREVIOUS NEXT'/>
      <fragment ref='BB'/>
    </fragment>
    <fragment id='BB'>
      <fragmentList role='markushMixture'>
      <fragment ref='po'>
        <scalar dictRef='cml:ratio' dataType='xsd:double'>0.02</scalar>
      </fragment>
      <fragment ref='bB'>
        <scalar dictRef='cml:ratio' dataType='xsd:double'>0.86</scalar>
      </fragment>
      <fragment ref='bE'>
        <scalar dictRef='cml:ratio' dataType='xsd:double'>0.12</scalar>
      </fragment>
      </fragmentList>
    </fragment>
  </fragmentList>

  <fragment id='f0'>
    <fragment ref='cl'/>
    <join atomRefs2='r1 r1' moleculeRefs2='PREVIOUS NEXT'/>
    <fragment ref='EE'/>
    <join atomRefs2='r2 r1' moleculeRefs2='PREVIOUS NEXT'/>
    <fragment ref='acetyl'/>
  </fragment>
</fragment>
```

Figure 8. PML document describing the construction of a random copolymer.

a bond to the preceding eo fragment. In this scenario, the polymerization continues. A final possibility is to connect the headgroup to a po, bB, bE fragment, which together form the `BB fragmentList`. Connection to a po fragment will



(B)

	EE	BB	termination
EE	0.84	0.15	0.01
BB	0.12	0.86	0.02
termination	0	0	0

Figure 9. (A) Pictorial representation of the joining procedures contained in the PML description of a hypothetical random copolymer (Figure 8). (B) The corresponding Markov model.

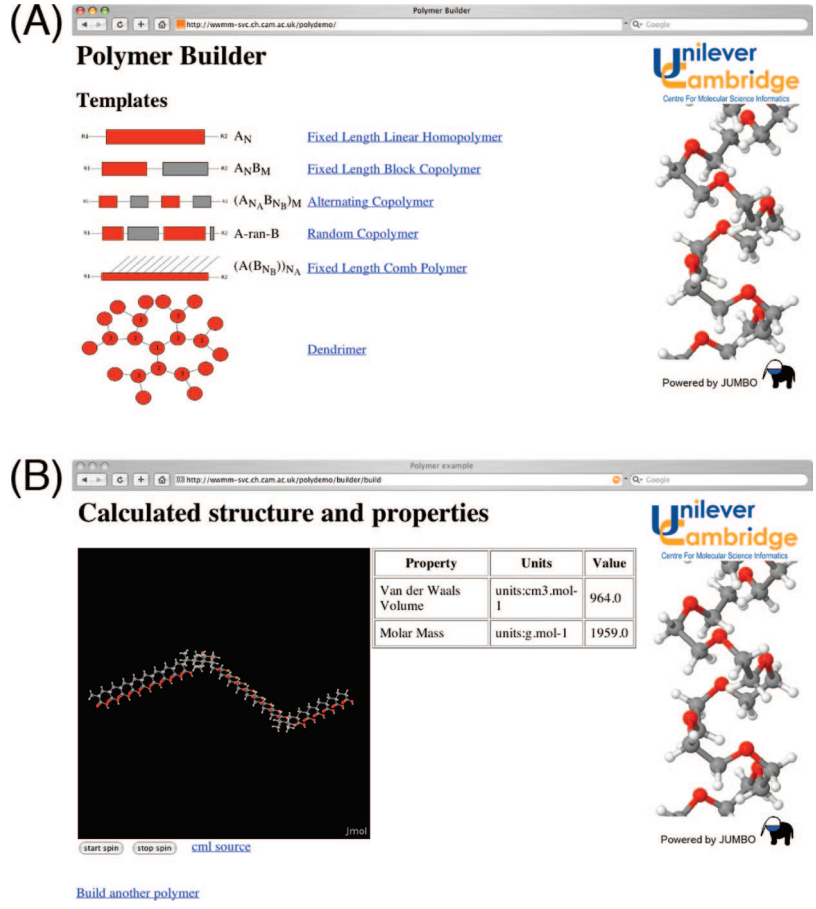


Figure 10. The Cambridge Polymer Builder. Copyright University of Cambridge, 2008.

effectively lead to the termination of the polymerization, whereas connection to bB or bE continues the polymerization, that is, the situation here is analogous to the one encountered for the EE fragment. A visualization of the building process is shown in Figure 9a. Figure 9b gives the corresponding Markov model. A combination of the Markov model described above and the `countExpression` attribute can also be used to model the law of mass action.

In general, PML is capable of describing all major polymer structural motives such as homopolymers, block, alternating, and random copolymers, combs, and other hyperbranched polymers, as well as dendrimers. Furthermore, cyclic and cross-linked polymers can also be described, albeit only in 2D at present. Unfortunately, modeling polymer networks represents a significant challenge at the moment. Polymer networks are composed of "highly ramified macromolecules, in which essentially each constitutional unit is connected to each other constitutional unit and to the macroscopic phase boundary by many permanent paths through the macromolecule."[38] While it is perfectly possible to describe and

represent polymer networks at the level of PML, the elaboration of this representation into a fully atomistic document will most likely lead to the construction of an "infinite molecule". PML documents describing network polymers ultimately describe topologically multidimensional extension rules and thus behave not unlike crystallographic unit cells. An attempt to map the PML document to an atomistic representation is therefore most likely to result in a nonterminating "*in silico* polymerization" which will only stop when the computer's memory is full: at any one time during the building of the atomistic document, more chains will be extended than terminated.

DISCUSSION

Polymer Markup Language is innovative in a number of ways and represents a new approach to the representation of polymers. The language is semantically completely explicit and allows polymers to be represented at various levels of certainty in a fully consistent manner. As an example, it is possible to represent an ill-defined system such as a phenol/formaldehyde resin in exactly the same way in which a well-defined polymer such as polystyrene could be represented. This is achieved through coarse-graining the description of the polymer, while preserving the possibility for mapping the coarse representation onto an atomistic description. At the level of PML, however, the descriptions are consistent, which, in turn allows for the comparison of polymers at different levels of certainty. Furthermore, because of the extensibility of the language, the fragments used in the construction process can carry a wide range of (computable) annotations such as values for group contributions[12] or measures of reactivity and probability, which can, for example, be used to model competing reactive centers. The language also allows the law of mass action to be taken into account. Finally, we have introduced the concept of computable elements and attributes into PML, which is a novel concept for normally purely declarative markup languages.

To support PML programmatically, we have added a module to JUMBO[39] (an XML infrastructure toolkit), which is capable of reading PML documents, expanding them to the greatest level of certainty and creating connection tables where possible and exemplified in the Cambridge Polymer Builder (Figure 10). The polymer builder supports both deterministic and stochastic models and can vary chain lengths, branching and chemical functionality according to the description in the PML template. Furthermore, it can also use fragments containing three-dimensional coordinates to build exemplars of macromolecules. We have not currently addressed the building of condensed phases. In summary, Polymer Markup Language represents a significant advance in comparison to other known polymer representation systems.

SUMMARY AND CONCLUSIONS

The advent of high-thoughput and combinatorial paradigms as routine tools in materials and polymer science, coupled with shortening innovation cycles in both industry and academia and the increasingly interdisciplinary nature of scientific research result in increasingly data driven science, which needs powerful informatics support. Furthermore, the Internet is currently impacting significantly on how we store, view, manipulate, and distribute scientific data.

From an informatician's point of view, polymers are fuzzy and ill-defined materials, which lead to the failure of "traditional" data models developed for small molecule informatics when applied to polymers. To address this situation, we have developed Polymer Markup Language, which implements a coarse-grained and "normalized" representation of polymers, while allowing the definition of a fully atomistic representation, where appropriate. This particular representation overcomes many of the problems associated with traditional polymer representation systems and advances our progress toward the web of data.

ACKNOWLEDGMENT

The authors acknowledge Dr Ian Stott (Unilever Research Port Sunlight), as well as David Jessop and Nicholas England (Cambridge), for helpful discussions during the preparation of this manuscript.

REFERENCES AND NOTES

(1) Meier, M. A. R.; Schubert, U. S. Selected successful approaches in combinatorial materials research. *Soft Matter* **2006**, *2*, 371–376.
(2) Guerrero-Sanchez, C.; Paulus, R. M.; Fijten, M. W. M.; de la Mar, M. J.; Hoogenboom, R.; Schubert, U. S. High-throughput experimentation in synthetic polymer chemistry: From RAFT and anionic polymerizations to process development. *Appl. Surf. Sci.* **2006**, *252*, 2555–2561.
(3) Zhang, H.; Hoogenboom, R.; Meier, M. A. R.; Schubert, U. S. High-throughput experimentation in polymer chemistry. *Trans. Mater. Res. Soc. Jpn.* **2004**, *29*, 319–324.
(4) Zhang, H.; Hoogenboom, R.; Meier, M. A. R.; Schubert, U. S. Combinatorial and high-throughput approaches in polymer science. *Meas. Sci. Technol.* **2005**, *16*, 203–211.
(5) Wiesbrock, F.; Hoogenboom, R.; Leenen, M.; Van Nispen, S. F. G. M.; Van der Loop, M.; Abeln, C. H.; Van den Berg, A. M. J.; Schubert, U. S. Microwave-assisted synthesis of a 42-membered library of diblock copoly(2-oxazoline)s and chain-extended homo poly(2-oxazoline)s and their thermal characterization. *Macromolecules* **2005**, *38*, 7957–7966.
(6) Guerrero-Sanchez, C.; Hoogenboom, R.; Schubert, U. S. Fast and "green" living cationic ring opening polymerization of 2-ethyl-2-oxazoline in ionic liquids under microwave irradiation. *Chem. Commun.* **2006**, 3797–3799.
(7) Adams, N.; Moneke, M.; Gulmus, S. A.; Chenouf, D.; Rehahn, M.; Schubert, U. S. Combinatorial compounding. *Mater. Res. Soc. Symp. Proc.* **2006**, *894*, 171–179.
(8) Ma, H.; Melillo, G.; Oliva, L.; Spaniol, T. P.; Englert, U.; Okuda, J. Aluminum alkyl complexes supported by [OSSO] type bisphenolato ligands: synthesis, characterization and living polymerization of rac-lactide. *Dalton Trans.* **2005**, 721–727.
(9) Wilks, E. S. Polymer nomenclature and structure: a comparison of systems used by CAS, IUPAC, MDL and DuPont. 1. Regular single-strand organic polymers. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 171–192.
(10) *Chemical Abstracts Index Guide 1997*; Chemical Abstracts Service: Columbus, 1997.
(11) Gushurst, A. J.; Nourse, J. G.; Hounshell, W. D.; Leland, B. A.; Raich, D. G. The substance module: the representation, storage and searching of complex structures. *J. Chem. Inf. Comp. Sci.* **1991**, *31*, 447–454.
(12) van Krevelen, D. W. *Properties of Polymers*; 3rd ed.; Elsevier: Amsterdam, 2003.
(13) Bicerano, J. *Prediction of Polymer Properties*; 3rd revised ed.; Marcel Dekker Ltd: New York, 2002.
(14) Murray-Rust, P.; Rzepa, H. Chemical markup, XML, and the world-wide web. 1. Basic principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928–942.
(15) Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–44.
(16) Bray, T.; Paoli, J.; Sperberg-McQueen, C. M.; Maler, E.; Yergeau, F.; Cowan, J. Extensible Markup Language (XML) 1.1 (2nd ed.). http://www.w3.org/TR/REC-xml/ (accessed July 10, 2007).
(17) Manola, F.; Miller, E., RDF Primer. http://www.w3.org/TR/rdf-primer/ (accessed July 10, 2007).

(18) McGuiness, D.; van Harmelen, F. OWL web ontology language overview. http://www.w3.org/TR/owl-features/ (accessed May 12, 2007).
(19) Casher, O.; Rzepa, H. S. SemanticEye: A semantic web application to rationalize and enhance chemical electronic publishing. *J. Chem. Inf. Model.* **2006**, *46*, 2396–2411.
(20) Taylor, K. R.; Gledhill, R. J.; Essex, J. W.; Frey, J. G.; Harris, S. W.; De Roure, D. C. Bringing chemical data onto the semantic web. *J. Chem. Inf. Model.* **2006**, *46*, 939–952.
(21) Taylor, K. R.; Essex, J. W.; Frey, J. G.; Mills, H. R.; Hughes, G.; Zaluska, E. J. The semantic grid and chemistry: Experience with Comb*e*Chem. *J. Web Semant.* **2006**, *4*, 84–101.
(22) Frey, J. G.; Hughes, G. V.; Mills, H. R.; Schraefel, M. C.; Smith, G. M.; de Roure, D. Less is more: lightweight ontologies and user interfaces for smart labs. In *Proceedings of the UK e- Science All Hands Meeting*; Nottingham, U.K., 2004; pp 500−507.
(23) Frey, J. G.; de Roure, D.; Schraefel, M. C.; Mills, H. R.; Fu, H.; Peppe, S.; Hughes, G.; Smith, G.; Payne, T. R. Context slicing the chemical aether. In *Proceedings of the First International Workshop on Hypermedia and the Semantic Web*; Nottingham, U.K., 2003; http://eprints.ecs.soton.ac.uk/8790/ (accessed Sept. 15, 2008).
(24) Holliday, G. L.; Murray-Rust, P.; Rzepa, H. S. Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. *J. Chem. Inf. Model.* **2006**, *46*, 145–157.
(25) Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L. Chemical markup, XML, and the world wide web. 5. Applications of chemical metadata in RSS aggregators. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 462–469.
(26) Murray-Rust, P.; Rzepa, H. S. Chemical markup, XML, and the world wide web. 4. CML schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757–772.
(27) Gkoutos, G. V.; Murray-Rust, P.; Rzepa, H. S.; Wright, M. Chemical markup, XML and the world-wide web. 3. Toward a signed semantic chemical web of trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124–1130.
(28) Murray-Rust, P.; Rzepa, H. S. Chemical markup, XML and the world-wide web. 2. Information objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1113–1123.
(29) Clark, J. XSL Transformations (XSLT). http://www.w3.org/TR/xslt (accessed Aug. 04, 2008).
(30) Cuchelkar, V.; Kopecek, J. Polymer-drug conjugates. *Polym. Drug Delivery* **2006**, 155–182.
(31) Kataoka, K.; Kwon, G. S.; Yokoyama, M.; Okano, T.; Sakurai, Y. Block copolymer micelles as vehicles for drug delivery. *J. Controlled Release* **1993**, *24*, 119–132.
(32) Ranquin, A.; Versees, W.; Meier, W.; Steyaert, J.; van Gelder, P. Therapeutic nanoreactors: Combining chemistry and biology in a novel triblock copolymer drug delivery system. *Nano Letters* **2005**, *5*, 2220–2224.
(33) Malmsten, M. Soft drug delivery systems. *Soft Matter* **2006**, *2*, 760–769.
(34) Qiu, L. Y.; Bae, Y. H. Polymer architecture and drug delivery. *Pharm. Res.* **2006**, *23*, 1–30.
(35) Schmaljohann, D. Thermo- and pH-responsive polymers in drug delivery. *Adv. Drug Delivery Rev.* **2006**, *58*, 1655–1670.
(36) Sankar, P.; Aghila, G. Ontology aided modeling of organic reaction mechanisms with flexible and fragment based XML markup procedures. *J. Chem. Inf. Model.* **2007**, *47*, 1747–1762.
(37) Allen, F. H.; Barnard, J. M.; Cook, A. P. F.; Hall, S. R. The molecular information file (mif):core specifications of a new standard fomat for chemical data. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 412–427.
(38) Gold, V.; Loening, K. L.; McNaught, A. D.; Shemi, P. *Compendium of Chemical Terminology: IUPAC Recommendations*; 2nd ed.; Blackwell Scientific: London, 1997.
(39) Zhang, Y.; Murray-Rust, P.; Dove, M. T.; Glen, R. C.; Rzepa, H. S.; Townsend, J. A.; Tyrell, S.; Wakelin, J.; Willighagen, E. L. JUMBO−An XML infrastructure for eScience. In *Proceedings of the UK e-Science All Hands Meeting*; Nottingham, U.K., 2004, 930−933.

CI8002123