Engineering Polymer Informatics: Towards The Computer-Aided Design of Polymers

Nico Adams, Peter Murray-Rust*

Unilever Centre for Molecular Science Informatics, University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW (United Kingdom) (Facsimile: +44 (0)1223 763076, Email na303@cam.ac.uk)

Keywords

chemical markup language, polymer informatics, polymer markup language, polymer ontology, structure-property relation, structure

Abstract

The computer-aided design of polymers is one of the holy grails of modern chemical informatics and of significant interest for a number of communities in polymer science. The paper outlines a vision for the *in silico* design of polymers and presents an information model for polymers based on modern semantic web technologies, thus laying the foundations for achieving the vision.

1. Introduction

Polymers are ubiquitous materials in our modern world and have found use in diverse application areas such as packaging, the delivery of drugs[1-4] and genes[5-7] and as ingredients in many formulations such as inkjet inks, personal care products (e.g. shampoos, hairsprays) and others. One of the interesting features of polymers is that they are less heavily regulated than small molecules[8] and can sometimes even be used as functional substitutes.

For these reasons, the rapid discovery, development and optimisation of (novel) polymeric entities is of high importance as has been evidenced by the development of high-throughput and combinatorial methods both in polymer synthesis and screening[9-12] and processing.[13] One component which has been notably absent from the high-throughput vision so far, is the use of informatics tools for the computer-aided design of polymers and polymeric systems, although a number of attempts have been reported in the past.[14-16] Given the relative importance of this class of materials, though, as well as the increasingly data-driven nature of polymer research, both the "rational" design of polymers and the development of a sophisticated polymer informatics should be high up on the agenda of polymer scientists. The use of informatics is also mandated by the often complex nature of the problem: when attempting to develop polymer pharmaceuticals, for example, not only does "the polymer chemistry need to be right", (*i.e.* in the case of a conjugate, a polymer can be connected to the active ingredient, has a given phase behaviour, responds to external stimuli such as pH, heat etc.) but it also needs to have the "right" absorption, distribution, metabolism, excretion and toxicology (ADMET) profile. The polymer scientist, therefore, is confronted with a highly complex, non-linear and multivariate

problem, which requires the confluence of data and knowledge from diverse and variable sources.

Modern (small molecule) design has recognized this fact and medicinal chemistry, for example, routinely combines bioinformatics (which, in turn, brings together data and knowledge from genomics, proteomics, structural biology etc.), chemoinformatics (quantitative structure-activity relationships (QSAR), molecular modelling, data mining) and data from combinatorial and high-throughput experimentation in the design process. The task, in every case, is often similar: bioinformatics aims to establish a correlation between sequence, structure and function, whereas chemoinformatics aims to develop the correlation between chemical composition, structure and (bulk) property. Polymer informatics therefore should enable the polymer scientist to do the same thing: it should allow the polymer scientist to either correlate the composition and structure of a polymer with its physicochemical and other properties, or help to develop a hypothesis as to which chemical features a polymer must contain in order to achieve a certain physical behaviour. Again, this is a complex and multivariate problem, for which sophisticated informatics is absolutely necessary.

Apart from the increasing importance of informatics for polymer science, the internet in general is currently radically changing how we structure, handle, present and exchange information. Whereas the current body of the world-wide-web is mainly a web of documents interconnected by hyperlinks and primarily used by humans discovering information in those documents, the web is currently evolving to a semantic web[17] of data, in which machines not only are able to discover information and the meaning of information, but also to act on it. In a typical scenario a polymer scientist whishing to design or discover a polymeric entity against a certain

requirements specification would deploy a software agent (a piece of software which acts on behalf of a user) to collect information and data concerning a certain polymer or polymers from the web, in-house resources, proprietary and open databases *etc.*. Once collected, the agent would reconcile the data against the requirements specification and use existing quantitative-structure property models or other rules to infer properties not directly discovered on the web. In a final step then, the agent would present the user with a list of polymers, which potentially fulfil all or most of the user specified requirements. In practice, this means that polymer information needs to be discoverable as well as structured and endowed with well-defined meaning, which allows software agents to carry out well-defined tasks. The semantic web is therefore a vision of machine-readable data, which can be used for automation, integration and re-use across different applications, as well as a vision of intelligent agents, which can retrieve and manipulate relevant information.

The technological foundations necessary for realising this vision are currently being developed and depend on a number of specifications such as eXtensible Markup Language (XML),[18] XML Schema,[19] the Resource Description Framework (RDF),[20] RDF Schema,[21] Web Ontology Language (OWL)[22] as well as logic, proof and trust. These technologies are interdependent on each other and can be arranged in layers (Figure 1), with each layer being progressively more specialized and complex. In developing polymer informatics, we make use of most of these specifications.

2. Polymer Informatics

The central dogma of chemoinformatics is that the structure of a molecule determines its properties and that, given a structure it is, in principle, possible to predict the

resulting properties of a molecule. In some cases this can be done using calculations based on the physics and chemistry of the system. In others, one has to rely on patterns deduced from existing knowledge and to try and implement these in heuristic and statistical approaches. In the latter case, particularly, it is important to have as much high quality data as possible and to have a clear informatics formulation of the structures and the properties, frequently described as metadata and/or ontologies.

In principle it is also possible to predict the properties of a polymer, if all the structures of its component macromolecules were available. In practice, however, this is considerably harder than for "small-molecules" because:

- The nature of a given polymer is often not fully understood. We may know how it was made, but not necessarily everything about the final product. Alternatively we may have physical and chemical data on the product, but not know in detail how it was made.
- Even given full knowledge of the polymer, there is intrinsic variability in the structure.
- Because of the variability and uncertainty in polymers, the traditional methods used to describe small molecules do not extend easily to polymers.
- Although a considerable amount of data on polymers is published, it is often widely scatted and heterogeneous and there is very little systematization of metadata and ontologies. Properties are often constrained by other quantities, which are sometimes assumed as defaults rather than being explicit.

In developing formal representations and tools for polymer informatics we therefore have to address the problems of uncertainty, variability, and imprecision. We need new informatics methods based on ontologies and markup languages, and software that is capable of using these. We need greater access to communal data and metadata

so these and other methods can be rigorously tested, and we now explore these issues in detail.

2.1 The Challenging Nature of Polymer Information

Small molecule informatics is in essence a solved problem. A number of methods and technologies exist to represent molecules to a machine in multiple dimensions (0 - 3D), ranging from trivial and systematic names and brutto formulae to line notations such as the "simplified molecular input line entry specification"[23] (SMILES) and the International Union of Pure and Applied Chemistry's (IUPAC) International Chemical Identifier[24] (InChI) and to full connection tables in a plethora of formats, such as mol, pdb or Chemical Markup Language [25-28] (CML). These representations are normally constructed on the basis of results derived from modern analytical chemistry, which can be successfully used to elucidate the structure and therefore the "connection table" of small molecules.

While chemists are accustomed to think of both small molecules and polymers as "substances", *i.e.* a particular kind of matter with uniform properties, there is a profound difference between the two, which causes confusion and difficulties for the chemical information scientist. Unlike substances composed of well-defined small molecules of usually identical structure, polymers consist of ensembles of macromolecules, all of which have slightly different architectures (in the simplest case only differing by length, in more complicated cases showing extensive branching or cross-linking) and therefore slightly different properties.[29] Physical quantities commonly referred to as "polymer properties" do not relate to a pure substance with a unique connection table, but are averages over structurally diverse ensembles of macromolecules. Molecular weight distributions in classically prepared synthetic

polymers are unavoidable – even the most controlled polymerisations lead to polydispersity indices (PDIs) larger than 1 (very controlled living polymerisations achieve PDIs of around 1.03 (see, for example, reference [30]). Furthermore, even modern analytical tools do not allow for the "connection table" of all of the constituent macromolecules in an ensemble to be determined, which makes the accurate description of a polymer in terms of the structures of its constituent macromolecules impossible and introduces a significant fuzziness of concept. The latter, in turn, breaks the transition from structure to property, which traditional chemical informatics is trying to make.

2.1.1 Representation of Polymers

The fuzziness of concept discussed above can be found right across polymer science and probably nowhere more so than in the representation of polymers to machines (*e.g.* in databases etc.). Typically, polymers are represented in information systems using either a name (a text string) or an idealised/abstracted or reduced structural description (an idealised connection table, a graphical representation) or a combination of both. Both types of representations have their particular problems.

2.1.1.1 Name-based representations.

Name-based representations are normally constructed either from the component monomers of a polymer (source-based representations) or from the repeating unit (structure-based representation) and frequently trivial names are still in use. Each of these representations has merits and disadvantages and there is no general agreement in the polymer science community, as to which representation is preferable. Furthermore, the form which the name based representation will take, depends on the

different nomenclature philosophies used across chemistry. As an example, consider the representation of the polymer with the repeat unit structure depicted in Figure 2. The Chemical Abstracts Service (CAS) will register the polymer as "1,3-butadiene, homopolymer"[31] whereas IUPAC allows the use of "polybutadiene" (IUPAC source based), "poly(but-1-ene-1,4-diyl)" (IUPAC structure based), "1,4-polybutadiene" (IUPAC semisystematic name) or "poly(buta-1,3-diene)" (IUPAC source based).[32] In addition to the different representation conventions (source-based/structure-based), these examples also illustrate the inversion of names for registration purposes (CAS), as well as the inconsistent use of brackets. Furthermore, each nomenclature and registration system has its own historical continuity - as the system evolves, naming conventions and therefore registrations change. The CAS 8th collective index (CI) name for poly(ethylene terephthalate) (Figure 3), for example, is poly(oxyethyleneoxyterephthaloyl), whereas the 9th CI name is poly(oxy-1,2-ethanediyloxycarbonyl-1,4-phenylenecarbonyl) (at the time of writing, Chemical Abstracts is in the 15th CI period). However, many chemists continue to use old nomenclature or even trivial names in their daily work: "methyl methacrylate" is still the preferred representation for a particular monomer molecule, rather than "methacrylic acid, methyl ester" (8th CI) or even "2-propenoic acid, 2-methyl-, methyl ester" (9th CI). It is not merely enough for rules and conventions to exist and to be implemented in a closed system such as the Chemical Abstracts: they also need to be adopted by a significant number of practicing chemists to be useful.

While the plethora and complexity of possible name-based representations may, at worst, be confusing to the human chemist, it causes significant problems for the information scientist and the computer. Firstly, it may lead to multiple registrations of the same compound in a database, which, in turn, often results in only partial retrieval

of information associated with the same concept: unless one remembers to search for polybutadiene as well as all other possible representations of the same substance (taking into account both synonyms and historical continuity), one may not all the desired information. Even more gravely, the scenario outlined above requires a software agent to retrieve information about a polymer from different sources (e.g. physico-chemical properties database, toxicology database) and to subsequently unify the information. The unification process is essentially a mapping procedure, which requires software to recognize concepts as equivalent: while a chemist may be able to recognize, that the labels “poly(but-1-ene-1,4-diyl)” and “poly(buta-1,3-diene)” refer to equivalent concepts, this would be impossible for a machine if it had to exclusively rely on name based representations alone.

2.1.1.2 Graphical representations.

An idealized or abstracted structural sketch can also be used to represent polymers. “Structural” in this context 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 representations discussed above. When examining the polymer shown in Figure 4, it becomes evident that several valid repeat unit structures can be drawn (the possible repeat units A, B and C are “phase-shifted” with respect to each other) and therefore no unambiguous definition of a representation is possible in the absence of further specifying guidelines. In order to determine the preferred representation, a set of rules has to be developed and adopted by the chemical community. IUPAC defines an elaborate set of rules based on seniority of subunits, the “direction of citation” etc..[33] In this context, it is important to remember, that although we are discussing the choice of the preferred repeat unit in terms of a graphical

representation, these rules also influence the construction of polymer names, where the name is structure-based. Further rules are used to refine these constructs.

From the point of view of an information scientist, this raises problems similar to the ones discussed for name-based representations: the rules governing a rule-based system must be accepted and followed if a consistent and unambiguous representation of polymers is to be achieved. Each of these systems, however, also exists in time and is therefore subject to change, which introduces added layers of complexity. The complexity is further increased, when several competing nomenclature systems are available, which essentially multiply the problems discussed so far.

The discussion presented here has only focussed on simple linear polymers and even for those it has barely scratched the surface. Nomenclature and registration systems for polymers have been extensively reviewed by Wilks and others and the reader is referred to the literature for further information.[32,34-38]

A paper, published in the early 1990s commented that "*Just the mention of the word "polymer" has been known to strike fear into the hearts of mere mortals and certainly, at the least, a sense of apprehension, if not foreboding to an information researcher.*"[37] Sadly, the situation has not changed significantly over the last decade.

2.1.2 Sources of Polymer Information

In a set of introductory remarks at an ACS symposium on the retrieval of polymer information, Metanomski remarked in the late 1970ies, that it "*is extremely important to have an easy and reliable access to the numerical data (preferably evaluated and verified) as well as to a variety of properties […].*"[39] The two main concepts in this remark, namely "access" and "evaluated/verified data" remain as pressing and unfortunately unaddressed as they were almost two decades ago.

2.1.2.1 Access

We have already discussed the fact, that polymer science is becoming increasingly data-centric, with high-throughput and combinatorial approaches being adopted as main-stream tools in the laboratory. However, the way in which science has chosen to report and archive its results generally leads to fragmentation, inaccessibility and the development of knowledge silos.

The majority of polymer (-related) data originates from a small number of sources, namely scientific publications, theses and data compilations. In order to be able to extract data and mine these sources, they first need to be accessed by a machine. There are a number of obstacles to access, such as the physical availability of data (is it available electronically or as a paper copy on a library shelf, non-destructive document formats and copyright considerations. The requirement for the electronic availability of data and documents is obvious, if a software agent is to discover information. Although more and more institutions now require theses and dissertations to be reposited as a condition of granting a degree, this is still far from universal and a significant number are archived on a paper-only basis by libraries.

However, even if available electronically, the format, in which the document is available, is critical. Most science papers and theses are either authored in LaTeX[40] or other text processing systems such as Microsoft Word[41] or Open Office and are subsequently – more often than not - converted to portable document format (pdf) for printing, distribution and repositing. The conversion to pdf, however, often destroys vital scientific information: the process converts text to a set of graphical objects without semantics, *i.e.* without well-defined relationships between them. For example, the information concerning superscripts and subscripts (which could identify chemical

formulae) is lost. Furthermore, the resulting graphical objects, cannot be processed further by computers in a data extraction/mining exercise and have to be converted back to text. As, at this stage, a significant amount of important information has been destroyed during the initial conversion process, the back-conversion yields unsatisfactory results such as jumbled data tables and formulae, which are difficult to interpret for both human and machine (Figure 5). In the context of our vision for polymer informatics, in which a software programme automatically detects and gathers data and information, this clearly presents a major obstacle. The most machine-friendly ways of transmitting and storing information is plain text, which is augmented with a form of text-based markup (such as LaTeX, HTML and XML documents), as information transmission here is usually lossless. Furthermore, closed proprietary formats also present problems for long-term storage and archival, particularly if the software required to access them, no longer exists.[42]

Beyond these more technical considerations, the structure of a document also needs to be taken into account when considering access to data. The main form of communication in the chemical sciences is the scientific paper (and to a lesser extent the thesis), which typically intersperses (polymer) data with free text, thus effectively forming a "datument," (*dat*a + doc*ument*) albeit an unstructured one.[43] It is difficult for a machine to automatically discover chemical information in collections of unstructured documents, as these are inevitably semantically poor. A typical example of a sentence that could be found in an unstructured datument could be: "poly(styrene) has a glass transition temperature of 99 °C". Without the availability of structuring metadata or a significant amount of "information archaeology", a machine has little chance to discover that the concept "poly(styrene)" refers to a polymer and "glass transition temperature" to a polymer property which, in turn, usually has an

associated value and a unit. If, however, concepts, values and units could be marked up as such in a machine discoverable way, this information could be extracted and made available for further processing. Markup of this type as part of the text would convert the unstructured datument to a fully structured one.

2.1.2.2 Copyright Considerations

Beyond the more technical barriers to data access, copyright considerations complicate data availability even further. Copyright law was originally conceived to protect property rights of an author and to regulate the use of an expression of an idea or of information. The 1988 UK Copyright, Designs and Patents Act states that "copyright is a property right which subsists in […] (a) original literary, dramatic, musical or artistic works, (b) sound recordings, films or broadcasts, and (c) the typographical arrangement of published editions."[44] This formulation makes a clear distinction between expression of an idea or information and the idea itself. When publishing a polymer science paper containing data about one or more polymers in a commercial journal, therefore, what the publisher owns is not the data as such or any new facts, which have been discovered, but rather the particular expression of these results in the paper.

However, publishers currently appear to attempt to copyright scientific data by attaching copyright statements to both papers and corresponding supplementary data. In the best possible case, this only gives the impression that the data is copyrighted, in the worst possible case, it is an attempt by the publisher to appropriate data, which is then "re-sold" to the scientific community *via* journal subscription fees. Appending copyright statements to supporting information, *i.e.* information, which is almost

entirely data (fact), certainly obfuscates the situation and potentially deters from use of the data for scientific purposes.

The part of the scientific community, whose work is mainly data driven, has long since recognised this as a significant obstacle to further progress, resulting in an increasingly vocal open data/open access movement. One manifestation of this is the Budapest Open Access Initiative Declaration, which defines open access to literature as meaning "*its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself*."[45] In the context of the polymer informatics vision outlined above, the phrases "crawl them for indexing" and "pass them as data to software" are of particular importance. If an author wishes to confer such usage rights to the public, it is imperative to make data and access available together by issuing an appropriate licence. The Creative Commons (CC) Foundation aims to enable copyright holders, to transfer some or all of their copyright to the public, by providing a number of different licences, which cover a broad range of usage scenarios.[46] While CC licences were mainly conceived and intended for the artistic domain, a significant number of scientists, some "hybrid open access" publishers as well as full open access publishers (*e.g.* Public Library of Science) make use of various forms of creative commons licences. As some of the provisions in CC licences are not entirely appropriate for scientific endeavour, the Science Commons project came into existence in 2005 in order to provide licences and policy tailored to scientific work.[47] Another notable effort to provide open data and information specifically in the area of chemistry, is the PubChem database,[48] which contains over

10 million compound structures and thousands of datasets, including some polymer data. Table 1 provides an overview over sources of polymer data and information together with notes on accessibility.

2.1.2.3 Data Curation

Data curation is an important and often neglected aspect when developing collections of or information systems for polymer data. A significant number of polymer properties are often dependent on factors, which are independent of the precise chemical nature of the constituent macromolecules, but very dependent on factors such as measurement methods and conditions, pressure *etc*.. The glass transition temperature (T_g), for example, is formally dependent on quantities such as pressure, molecular mass, tacticity and cross-linking, etc..[49] For low molecular weight polymers, T_g increases with increasing polymer molecular weight until it reaches an upper limit and becomes essentially invariant to further increases in molecular weight. Furthermore, the glass transition temperature is usually determined by observing a thermodynamic quantity associated with temperature. Popular measurement methods to determine the quantity are Differential Scanning Calorimetry (DSC) or Thermomechanical Analysis (TMA). In the case of DSC, a change in heat capacity as a function of temperature is measured, whereas TMA determines dimensional changes (length and thickness) of a sample (dynamic mechanical thermal analysis evaluates changes in modules), which is quite different from observing changes in heat capacity. Consequently, the experimental values determined by these two techniques usually differ by several Kelvins. Furthermore, the presence of additives can also change the T_g of a sample. For these reasons, simply reporting the glass transition temperature of a polymer without the necessary "metadata" (*e.g.*

measurement method, heating rate etc.) is of only limited value. Unfortunately, this is usually the case in data compilations such as "Polymers: A property database".[50] The PolyInfo database[51] as well as the Polymer Handbook[52] attempt to supply this data, although a significant amount of "digging" is usually required.

Another aspect of curation concerns error checking: not only are measurement errors unavoidable in (experimental) science, but often - and certainly in the case of polymer science - data compilations are developed by manual abstraction from the primary literature. This, in turn, means that typographical errors invariably occur. Taking the above example of the glass transition temperature and plotting the values for T_g for all polymers found in the PoLyInfo against their corresponding melting points (T_m), it becomes evident, that for some polymers T_g is higher than T_m. This is, of course, nonsensical and suggests that either of the two values could be erroneous. It is, in principle, relatively easy for a machine to perform this kind of error checking, provided the data is accessible and machine comprehensible.

2.2 Engineering Polymer Informatics

The discussion so far makes it clear, that before we can even begin to approach "computer-aided polymer design" in any meaningful way, the appropriate data structures need to be put into place. For polymers, this means developing a combination of access to structured and meaningful data and sets of rules, which allow a computer to reason over these rules (Figure 6).

We have already alluded to the fact that structured documents can be prepared by utilizing a suitable markup language. Markup languages combine the text of a document and further information about the text (usually referred to as metadata). Text and metadata are normally intermingled and often the metadata is hidden from

human view, but accessible to machines and available for processing. Markup languages have a long tradition in informatics and fall into three main classes: presentational, procedural and descriptive markup languages. The most commonly encountered descriptive markup language is HTML[53] (HyperText Markup Language), followed by XML[18] (eXtensible Markup Language). XML allows for arbitrary structure to be added to documents through the use of tags. Tags can be user-defined and are employed to annotate text and other sources of information. Furthermore, they can be processed by machines.

Unfortunately, markup alone is not sufficient to enable a machine to autodiscover information: the arbitrary *(i.e.* user-defined) nature of markup provides structure, but does not define the "meaning" of the structure to the machine. The latter is achieved by using the Resource Description Framework[20,21] (RDF) and the Web Ontology Language (OWL),[22] both of which are layered on top of XML.

RDF makes statements about resources in the form of "triples." These are almost human language *subject – predicate – object* statements. A resource, in internet terminology, is an entity that can be named or addressed or handled (Figure 7). The simple example in Figure 7 shows two resources, namely "poly(styrene)" and "polyolefin" connected *via* the predicate "isA". This is the simplest RDF graph possible. All the components of a triple are uniquely identified by a universal resource identifier (URI), which means that anyone can define new concepts and relationships. While RDF allows simple assertions of the type we have just described above, OWL extends RDF's expressivity by adding first order description logic, thus allowing relationships between classes (disjointedness), cardinality, equality and symmetry of properties to be described. OWL was designed with computational reasoning and inferencing in mind. Both RDF and OWL are used to develop ontologies, *i.e*. "formal

explicit specifications of a shared conceptualisation",[54] which define both concepts and the relationships between concepts.

2.2.1 Chemical Markup Language

XML is the technology of choice for preparing structured documents. Because XML is, as the name suggests, extensible, a number of dialects have been created, which are useful for marking up chemical information. The most relevant of these is Chemical Markup Language[25-28,55,56] (CML). Other markup languages of importance for chemistry and polymer science include Analytical Markup Language[57] (AnIML) and ThermoML,[58] a markup language for thermochemical and thermophysical property data.

CML was designed to manage all kinds of molecular information, such as structures, spectra and general analytical data, but also crystallographic and computational data.

As an example, let us consider the molecular structure of the styrene monomer. Its connection table (information about the arrangement and connectivity of atoms) can be expressed in CML as shown in Figure 8. The document contains a set of tags ('elements') such as `<molecule>`, `<atomArray>`, [59], [60] and `<bondArray>`. Each of these acts as a data container in that they enclose data and/or other elements. Some of the elements (`<atom>` and `<bond>`) in the document have further attributes such as "`elementType`", "`id`", "`atomRefs`" and "`order`". Coordinates can, of course, also be included, although they have been omitted from the example in Figure 8 in the interests of readability. The attributes provide further information about the element: `<atom elementType ="C"/>`, for example, specifies that an XML element describing an atom is referring to a carbon atom. A connection table expressed in this way is semantically completely explicit and specifies the structure

and meaning of all of the data occurring in the document. This is in sharp contrast to other ways in which this type of information is traditionally encoded, such as the mol file format (Figure 1). Both the CML and the mol document hold identical information. In contrast to the CML file, however, the mol format contains implicit semantics. CML do only handle complete molecules, but can also be used to describe molecular fragments. In principle, this opens the door to building up molecules from a library of smaller fragments, by "concatenating" CML documents. Another approach that allows the development of molecules from molecular fragments was recently presented by Sankar *et al.*.[61]

Scientific information in free text such as papers or theses can be marked up in a similar way. Table 2 shows the first sentence of the abstract of ref. [62] in plain text and marked up in an inline notation, which is a mixture of SciXML and a technology developed by our group in Cambridge. In the present example, chemical entities such as "oleic acid" and "magnetite" are marked up as chemical entities (`type="CM"`) with further attributes specifying the relevant SMILES and InChI string. The important point here, is that because of the markup, a machine now "understands" that oleic acid is a chemical entity. Furthermore, because of the presence of a SMILES string or an InChI, a meaningful chemical structure is associated with a chemical name. The structure can be retrieved by a machine and processed, or further information can be associated with it. The markup also contains an attribute `cmlRef="cml1"`, which refers to a full CML connection table at the end of the marked up document, which has been truncated in the example presented in Table 2.

2.2.2 Polymer Markup Language

Polymers are substances, which are fundamentally different from well-defined molecular entities and any markup language attempting to describe polymers and polymer structures must take into account the associated peculiarities. Furthermore, the language must also adhere to the formal requirements of the XML specification.

To this end, we have developed Polymer Markup Language (PML) as an extension of CML. The language addresses the following polymer-relevant considerations: (a) the composition of a given polymer, (b) the structure of the polymer, (c) the record of a computational process, (d) the physical properties of a substance or material, (e) metadata associated with experiments and arising from annotation and (f) reactions and other chemical processes. We have explicitly excluded polymer processing (*e.g.* compounding etc.) from the language, although it may well be found later on, that aspects of PML are useful from a processing point of view. While the full specification of PML will be published elsewhere, the requirements for the language can be summarized as follows: (1) PML should be based on CML and (re-)use CML components where possible, (2) PML should interoperate with other mature scientific, technical and medical markup languages, (3) PML should be fully namespace aware, (4) implicit semantics in PML should be avoided wherever possible, (5) PML shall be able to address the ensemble nature of polymers (especially distributions), (6) PML shall address structural phenomena often encountered in polymers, such as ambiguous repeat units, tacticity, double bond isomerism, macromonomers, (7) PML shall be able to describe all commonly encountered polymer structural motives such as homopolymers, copolymers (statistical, alternating, block), post-treated polymers, branched polymers (combs, hyperbranched systems) and cross-linked polymers.

We have chosen to construct polymers from small molecular fragments expressed in CML. Polymer Markup Language then, holds the instructions concerning how to assemble the fragments into macromolecules and macromolecules into ensembles and thus polymers.

Figure 9 shows a simple PML document describing a poly(styrene) oligomer molecule (heptamer). The molecule is assembled by specifying a root element, in this particular example, the dummy atom R. In the PML document, this atom is specified by <molecule ref =″g:dummy″/>. The "g:" is a shorthand (a namespace prefix defined at the start of the document by the line xmlns:g="http://www.xml-cml.org/mols/geom1") and makes reference to another document, containing the definition of a dummy atom in a full CML connection table analogous to the document described in Figure 8. The root is then joined to the contents of the fragmentList container, namely a -CH- fragment (<molecule ref =″g:ch″/>), which, is in turn joined to a -CH_2- (<molecule ref =″g:ch2″/>) and a C_6H_5- (<molecule ref =″g:benzene″/>) fragment. The contents of the fragmentList, which, in this case, is coincidental with the repeat unit, are subsequently added another 6 times (countExpression="*(7)") to the RCHPhCH_2- fragment we have just constructed, to make up the heptamer. In this context it is worth noting, that the attribute countExpression represents a generating function for integers, which can simple, deterministic or stochastic. Instructions on how the fragments are to be joined, are contained in the <join> element, which specifies the bond order of the newly created bond, together with a torsional angle (<torsion> element) and information about which fragments are being linked to. To carry out the joining operation, the <join> element makes use of the atomRefs2 attribute, to identify

dummy atoms of type r_x, which are to be joined together. Once identified, the atoms to which the r_x dummies are joined, are connected by a new bond and the dummy atoms are deleted. In the present example, the r_1 group of the R dummy fragment is joined to the r_1-group of the methylene fragment (`atomRefs2=" r1 r1"`) with a bond order of 1 (`order="1"`). As the dummy R is the first fragment in the molecule, it is identified as the "parent" fragment in the `moleculeRefs2` attribute and the methylene fragment as the "next" fragment, as it follows R. The general semantics of `atomRefs2` is, that it makes reference to two different atoms.

Polymer Markup Language represents a completely new approach to the representation of polymers. Firstly, it is semantically completely explicit and allows polymers to be represented at various levels of certainty in a completely 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 poly(styrene) could be represented. In the latter case, we may be able to expand the representation into a connection table, whereas this may not be possible for the phenol/formaldehyde system. 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, components of polymers can carry a wide range of annotations such as group contribution values for polymer properties [49,63] or measures of reactivity, which can be used to model competing reactive centres. Moreover, it also allows phenomena such as the law of mass action to be taken into account when constructing a polymer. All of this represents a significant advance in comparison with other known polymer representation systems. We have added a module to JUMBO [64] (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 (exemplified in the Cambridge Polymer Builder, Figure 10). It supports deterministic and stochastic models and can vary chain lengths, branching and chemical functionality as described in the PML template. It can also use fragments with 3D coordinates to build exemplars of polymer chains. We have not currently addressed the building of condensed phases.

2.2.3 Polymer Ontology

The discussion so far has already established, that markup alone is not sufficient to generate structured and meaningful documents and that "meaning" is provided by ontologies, which we have previously defined as "formal explicit specifications of a shared conceptualisation." In other words, an ontology attempts to model concepts contained in a knowledge domain together with the relationships between these concepts. So far, only very few attempts have been made to construct formal ontologies for chemistry. An example of an early attempt is the work by Gordon, who, in a set of papers, considered the syntax, semantics and history of structural formulae as well as the semantic and formal attributes (such as transformations, tautomerism etc.) encountered in chemistry.[65-67] These efforts led to a formalized language for relational chemistry. Slightly later, van der Vet described logical construction rules for the concepts "pure substance", "phase" and "heterogeneous system" as the basic framework required for the construction of further chemically relevant concepts.[68]

The most widely used and prominent chemical ontology is the European Bioinformatics Institute's "Chemical Entities of Biological Interest" (ChEBI) ontology.[69,70] The ontology combines information from three different sources, namely COMPOUND,[71] the Chemical Ontology (CO) and IntEnz.[72] ChEBI has

been prepared in the OBO ontology language (but can be translated into OWL) and contains ontological associations, which specify chemical relationships ("chloroform isA chloroalkane"), biological roles and applications of the molecule. Other EBI ontologies currently in the development phase, are REX[73] and FIX,[74] which model physicochemical processes (REX) and methods (FIX). Further ontologies modeling chemical structure,[75] laboratory processes,[76-78] and chemical reactions have also been reported.[79]

For the purposes of polymer informatics, ontologies have several uses. First and foremost, an ontology serves to share a common understanding of the information structure of a domain between people and software agents. In the initial scenario discussed in this paper, a software agent was despatched to collect data about a polymer from various sources. This can only be done successfully, if all of the sources visited by the agent share and use the same ontology. This will guarantee that a computer is able to recognize that the concept "poly(styrene)" found in source A is equivalent to the concept "poly(vinyl benzene)" found in source B. Apart from knowledge sharing, ontologies also enable knowledge re-use. Similar to the example of explicit (CML) *versus* implicit (mol) semantics in describing molecular structure, ontologies make domain knowledge explicit. One weakness of relational databases, which are often used to build polymer information systems, for example, is the fact that domain assumptions are often hard-coded into the database. This usually makes alterations or extensions difficult and should a major revision be necessary, the system often has to be re-coded. Explicit domain assumptions are easier to revise and do not usually require a complete system re-build. Finally, ontologies allow the separation of declarative from procedural knowledge. An ontology can make statements about the nature or properties of a polymer, but cannot usually express a

process that specifies how a polymer is transformed or altered. Such procedural knowledge is most often encoded in algorithmic form as part of a computer programme, which, in turn, utilizes the assertions contained in an ontology. If the algorithm is sufficiently generic, re-use over different ontologies will be possible. Furthermore, ontologies, once constructed to a given standard, can be re-used by other researchers in their particular knowledge domain, integrated with other ontologies or otherwise extended.

We have prepared a general domain ontology for polymers, which is mainly based on existent IUPAC terminology. The ontology covers the most commonly used polymer concepts and the relationships between them and will be supplemented by further, more specialised, ontologies in the future. Figure 11 shows a graphical representation of top-level concepts and selected subsumption relationships with lower-level concepts. Specifically, the arrows denote "`isA`" relationships, *i.e.* a regular macromolecule `isA` macromolecule, which, in turn, `isA` molecule, which `isA` thing. Many other types of relationships exist (even between top-level concepts), but are not shown for reasons of clarity.

The top-level classes of the ontology are "`StructuralElement`", "`Molecule`", "`ReactionElement`", "`Substance`", "`Transformation`" and "`ValuePartition`", the latter being a modelling artefact. The classes `Molecule`, `Substance` and `Transformation` represent the particular paradigm and domain of chemistry, which also applies to polymer chemistry. `StructuralElement` and `ReactionElement` contain concepts, which are necessary for the description of aspects of molecular structure, such as "`Endgroup`", "`Branchpoint`", "`StereoBlock`", "`ChainTransfer`" *etc*.. The class molecule contains subclasses such as "`MacroMolecule`", "`OligomerMolecule`", "`FreeRadical`" and

"MonomerMolecule", which themselves can be subdivided further (*e.g.* subclasses of MacroMolecule are: "RegularMacromolecule", "BlockMacromolecule", "Macroinitiator" *etc.*. Although not depicted here, top level concepts are connected through a number of properties such as "isComposedOf", *e.g.* the class Substance isComposedOf some members of the class Molecule.

According to IUPAC, a macromolecule is a "molecule of high relative molecular mass, the structure of which essentially comprises the multiple repetition of repeat units derived, actually or conceptually, from molecules of low relative molecular mass."[80] One property of the class "MacroMolecule", which arises from the IUPAC definition is the "hasMolecularMass" property, which, in turn, carries a value (restriction) "high". Another, more complex property is, that the polymer has a repeat unit, which, in turn, has a certain multiplicity and is composed of a monomer molecule (e.g. a molecule of low molecular mass). The latter property is slightly more difficult to model (N-ary relationship) and for the purposes of our definition will be simplified to state that a polymer "hasStructuralElement" with a value of "Chain". Furthermore, although this is not contained in the formal IUPAC definiton, a domain expert might wish to assert, that the polymer has another structural element "Endgroup". The classes Chain and Endgroup, in turn, are subclasses of StructuralElement and defined appropriately. The ontological description of the concept MacroMolecule in OWL code is given in Figure 12.

The concept Polymer can then simply be defined in terms of its constituent MacroMolecules: a property "isComposedOf" with the restriction ∃ (some value from) "MacroMolecule" is asserted for the class "Polymer".[80] In this

way, knowledge can be codified quickly and complex knowledge systems can be developed.

2.2.4 Natural Language Processing and Text Mining

Having established, that markup in the form of Chemical/Polymer Markup Language and polymer ontologies expressed in RDF/OWL are indispensable for the generation of structured and meaningful polymer documents whose contents can be accessed and used by software agents, the question remains how the markup can be incorporated into those documents in an efficient manner. Incorporation can only happen during the time of writing or, alternatively, *a posteriori*.

The generation of valid markup is a non-trivial process, when the task has to be carried out by a human. Ideally, its generation should only involve a minimal learning curve, which, in turn, means that existing and familiar authoring paradigms should be used and that tool support is required. In practice, this could mean that when a structure is drawn by a chemist using a standard drawing tool and embedded in a document, the corresponding CML is autogenerated and also embedded (invisibly). Similarly, software could parse documents at the time of writing, identify chemical entities and ontology terms, generate the relevant markup and incorporate it into the document. In unclear situations, the user is prompted for further information/clarification. Tools such as the ones envisioned here do not currently exist, although their creation will be very much part of our future research endeavour.

This leaves the incorporation of markup into a corpus of scientific literature *a posteriori*. Given the sheer volume of already available literature and the ever-increasing number of papers contributed every year, the only feasible way of semantically enriching the scientific literature is to use natural language processing

(NLP). NLP is related to both linguistics and artificial intelligence research and is concerned with the machine understanding of (human) natural language. One of its goals is the extraction of structured, well-categorized information and data from essentially unstructured sources. While the use of natural language processing in the biological sciences is relatively advanced and a significant number of both commercial and open-source tools are available,[81-85] the same is not currently true for chemistry, although several efforts have been reported in the past.[86-92] To address this situation, Corbett and Murray-Rust reported the development of the OSCAR 3[93] as part of the SciBorg system[94] for the deep parsing and analysis of scientific texts. Oscar 3 accepts plain text or HTML as input, which is then passed to a recognizer module, which, in turn, identifies chemical names (trivial, semi-systematic and systematic), acronyms, ontology terms and other abbreviations. The system subsequently attempts to assign a structure to a recognised chemical name and produces a marked-up document in enhanced SciXML, which incorporates all annotations while preserving all other markup data that may have been present in the source text. The marked-up abstract shown in Table 2 was generated automatically by OSCAR 3.

OSCAR currently recognizes most polymer names as chemical entities and some polymer-related concepts (Figure 13), but is, as yet, unable to assign a meaningful structure to a recognized polymeric entity (if the name is source-based, the structure of the corresponding monomer is usually recognized). We are currently working on expanding OSCAR's functionality, to cope with the peculiarities of polymers, such as different possible representations (source-based vs. structure based), the recognition and representation of the structure of block- and random copolymers etc.. Once in place, this should facilitate the extraction of polymer structural information and

polymer data from unstructured data sources, and thus move us closer to the vision for polymer informatics discussed in the introduction of this paper.

3. Summary and Conclusions

The advent of increasingly "high-speed" and "high data" experimental paradigms in polymer science, coupled with ever shortening innovation cycles in both industry and academia as well as the increasing interdisciplinarity of research, result in increasingly data-driven science, which, in turn needs sophisticated informatics support.

However, access to polymer data is currently impeded by fuzzy concepts, fuzzy nomenclature and either fuzzy access rights or by enclosing data in walled gardens. Furthermore, all data models which have so far been used to deal with polymer information, have essentially been informed by small molecule informatics, which is not appropriate for the particular requirements of polymers.

To address this situation, we have developed a polymer information model consisting of the components CML Fragments, Polymer Markup Language (PML) and polymer ontologies. All of these components are built using light-weight semantic web technologies and allow extreme flexibility in terms of how polymer information is handled, stored, searched and retrieved. Our information model makes even relatively fuzzy information machine-discoverable and comprehensible, thus bringing science closer to realizing the vision of computer-aided polymer design. Moreover, the technology outlined in this paper will contribute to the development of the chemically intelligent semantic web and thus assist in breaking down the artificial barriers that currently surround scientific information and data.

References

[1] V. Cuchelkar, J. Kopecek *Poly,. Drug Deliv.* **2006**, 155.

[2] K.Kataoka, G.S.Kwon, M.Yokoyama, T.Okano, Y.Sakurai *J. Contr. Rel.* **1993**, *24*, 119.

[3] M. Malmsten *Soft Matter* **2006**, *2*, 760.

[4] D. Schmaljohann *Adv. Drug Deliv. Rev.* **2006**, *58*, 1655.

[5] H. C. Kang, M. Lee, Y. H. Bae In *Nanotechnology in Therapeutics*;page 131, 2007.

[6] D. Putnam *Nature Materials* **2006**, *5*, 439.

[7] E. Wagner, J. Kloeckner *Adv. Polym. Sci.* **2006**, *192*, 135.

[8] R. L. Keener, L. A. Jourdan, E. D. Weiler *Regulatory Toxicol. Pharmacol.* **1999**, *29*, 319.

[9] F. Wiesbrock, R. Hoogenboom, M. A. M. Leenen, M. A. R. Meier, U. S. Schubert *Macromolecules* **2005**, *38*, 5025.

[10] R. Hoogenboom, M. W. M. Fijten, U. S. Schubert *J. Polym. Sci., Part A: Polym. Chem.* **2004**, *42*, 1830.

[11] R. Hoogenboom, M. W. M. Fijten, C. H. Abeln, U. S. Schubert *PMSE Preprints* **2004**, *90*, 342.

[12] M. A. R. Meier, R. Hoogenboom, U. S. Schubert *Macromol. Eng.* **2007**, *3*, 1967.

[13] N. Adams, M. Moneke, S. A. Gulmus, D. Chenouf, M. Rehahn, U. S. Schubert *Mater. Res. Soc. Symp. Proc.* **2006**, *894*, 171.

[14] N. Adams, B. J. De Gans, D. Kozodaev, C. Sanchez, C. W. M. Bastiaansen, D. J. Broer, U. S. Schubert *J. Comb. Chem.* **2006**, *8*, 184.

[15] N. Adams, U. S. Schubert *QSAR & Comb. Sci.* **2005**, *24*, 58.

[16] N. Adams, U. S. Schubert *Macromol. Rapid Commun.* **2004**, *25*, 48.

[17] T. Berners-Lee, J. Hendler, O. Lassila In *Scientific American*, 2001.

[18] *http://www.w3.org/TR/REC-xml/.*

[19] *http://www.w3.org/TR/xmlschema-0/.*

[20] *http://www.w3.org/TR/rdf-primer/.*

[21] *http://www.w3.org/TR/rdf-schema/.*

[22] *http://www.w3.org/TR/owl-features/.*

[23] D. Weininger *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

[24] S. R. Heller, S. E. Stein, D. V. Tchekhovskoi *Abstracts of Papers, 230th ACS National Meeting, Washington, DC, United States, Aug. 28-Sept. 1, 2005* **2005**, CINF.

[25] P. Murray-Rust, H. S. Rzepa *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757.

[26] G. V. Gkoutos, P. Murray-Rust, H. S. Rzepa, M. Wright *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124.

[27] P. Murray-Rust, H. S. Rzepa *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1113.

[28] P. Murray-Rust, H. Rzepa *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928.

[29] H. G. Elias *"An introduction to polymer science."* Wiley VCH: Weinheim, 1997.

[30] H. Ma, G. Melillo, L. Oliva, T. P. Spaniol, U. Englert, J. Okuda *Dalton Trans.* **2005**, 721.

[31] C. A. Service *"Chemical Abstracts Index Guide 1997"*: Columbus, 1997.

[32] E. S. Wilks *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 171.

[33] *Compendium of macromolecular nomenclature (The Purple Book)*; W. V. Metanomski, Ed.; Blackwell Scientific Publications: Oxford, 1991.

[34] E. S. Wilks *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 193.

[35] E. S. Wilks *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 209.

[36] E. S. Wilks *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 224.

[37] M. D. Rieder *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 458.

[38] M. Herz *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 469.

[39] W. V. Metanomski *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 59.

[40] L. Lamport *"LATEX: A document preparation system: user's guide and reference manual"*; Addison Wesley, 1994.

[41] *http://www.microsoft.com.*

[42] *http://www.nationalarchives.gov.uk/news/stories/164.htm?homepage=news.*

[43] P. Murray-Rust, H. Rzepa *J. Digital. Inf.* **2004**, *5*.

[44] Copyright, Designs and Patents Act, 48, 1988

[45] *http://www.soros.org/openaccess/read.shtml.*

[46] *http://creativecommons.org.*

[47] *http://sciencecommons.org.*

[48] *http://pubchem.ncbi.nlm.nih.gov/.*

[49] D. W. van Krevelen *"Properties of Polymers"*; 3rd ed.; Elsevier: Amsterdam, 2003.

[50] *http://www.polymersdatabase.com/.*

[51] *http://polymer.nims.go.jp/polyinfo_top_eng.htm.*

[52] *Polymer Handbook*; 3rd ed.; J. Brandrup, E. H. Immergut, Eds.; John Wiley: New York, 1989.

[53] *http://www.w3.org/TR/html401/.*

[54] R. Studer, V. R. Benjamins, D. Fensel *Data Knowl. Eng.* **1998**, *25*, 161.

[55] G. L. Holliday, P. Murray-Rust, H. S. Rzepa *J. Chem. Inf. Model.* **2006**, *46*, 145.

[56] P. Murray-Rust, H. S. Rzepa, M. J. Williamson, E. L. Willighagen *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 462.

[57] T. Davies, P. Lampen, M. Fiege, T. Richter, T. Froehlich *Spectroscopy Europe* **2003**, *15*, 25.

[58] M. Frenkel, R. D. Chiroco, V. Diky, Q. Dong, K. N. Marsh, J. H. Dymond, W. A. Wakeham, S. E. Stein, E. Koenigsberger, A. R. H. Goodwin *Pure Appl. Chem,* **2006**, *78*, 541.

[59] A. B. Carrell, L. Shimoni, C. J. Carrell, C. W. Bock, P. Murray-Rust, J. P. Glusker *Receptor* **1993**, *3*, 57.

[60] H. J. Baek, J. H. Han, B. I. Min; *Polyester compositions containing organic particles for films*; 99-46628 2001038589; **2001**

[61] P. Sankar, G. Aghila *J. Chem. Inf. Model.* **2007**, *47*, 1747.

[62] M. Hamoudeh, A. A. Faraj, E. Canet-Soulas, F. Bessueille, D. Leonard, H. Fessi *Int. J. Pharm.* **2007**, *338*, 248.

[63] J. Bicerano *"Prediction of Polymer Properties"*; 3rd rev edn. ed.; Marcel Dekker Ltd: New York, 2002.

[64] Y. Zhang, P. Murray-Rust, M. T. Dove, R. C. Glen, H. S. Rzepa, J. A. Townsend, S. Tyrell, J. Wakelin, E. L. Willighagen Proceedings of the UK e-Science All Hands Meeting, 2004; p 930.

[65] J. E. Gordon, J. C. Brockwell *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 117.

[66] J. E. Gordon *J. Chem. Inf. Comput. Sci.* **1984**, *23*, 81.

[67] J. E. Gordon *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 100.

[68] P. E. van der Vet *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 564.

[69] C. Brooksbank, G. Cameron, J. Thornton *Nucleic Acids Res.* **2005**, *33*, D46.

[70] P. de Matos, M. Ennis, M. Zbinden, A. D. McNaught, R. Alcantara, M. Darsow, M. Guedj, M. Ashburner, D. K. In *Nucleic Acids Res.*, 2006; Vol. Database Summary Paper 646.

[71] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori *Nucleic Acids Res.* **2004**, *32*, D277.

[72] A. Fleischmann, M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K. B. Axelsen, A. Bairoch, D. Schomburg, K. F. Tipton, R. Apweiler *Nucleic Acids Res.* **2004**, *32*, D434.

[73] *http://obofoundry.org/cgi-bin/detail.cgi?id=rex.*

[74] *http://obofoundry.org/cgi-bin/detail.cgi?id=fix.*

[75] H. J. Feldman, M. Dumontier, S. Lng, N. Haider, C. W. V. Hogue *FEBS Letters* **2005**, *579*, 4685.

[76] K. R. Taylor, J. W. Essex, J. G. Frey, H. R. Mills, G. Hughes, E. J. Zaluska *J. Web Semant.* **2006**, *4*, 84.

[77] J. G. Frey, G. V. Hughes, H. R. Mills, M. C. Schraefel, G. M. Smith, D. de Roure All Hands Meeting, Nottingham, 2003.

[78] J. G. Frey, D. de Roure, M. C. Schraefel, H. R. Mills, H. Fu, S. Peppe, G. Hughes, G. Smith, T. R. Payne, 2003.

[79] P. Sankar, G. Aghila *J. Chem. Inf. Model.* **2006**, *46*, 2355.

[80] A. D. Jenkins, P. Kratochvil, R. F. T. Stepto, U. W. Suter *Pure Appl. Chem,* **1996**, *68*, 2287.

[81] *http://bioie.ldc.upenn.edu/.*

[82] *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA*

[83] *http://www-tsujii.is.s.u-tokyo.ac.jp/medie/.*

[84] *http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed.*

[85] *http://www.textpresso.org/.*

[86] G. G. Chowdhury, M. F. Lynch *Journal Chem. Inf. Comp. Sci.* **1992**, *32*, 463.

[87] G. G. Chowdhury, M. F. Lynch *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 468.

[88] C. S. Ai, P. E. Blower, R. H. Ledwith *Journal of Chemical Information and Computer Sciences* **1990**, *30*, 163.

[89] C. S. Ai, P. E. Blower, R. H. Ledwith *Abstracts of Papers of the American Chemical Society* **1989**, *197*, 18.

[90] E. M. Zamora, P. E. Blower *J. Chem. Inf. Comp. Sci.* **1984**, *24*, 176.

[91] E. M. Zamora, P. E. Blower *J. Chem. Inf. Comp. Sci.* **1984**, *24*, 181.

[92] P. E. Blower, H. W. Whitlock *Abstracts of Papers of the American Chemical Society* **1976**, 22.

[93] P. Corbett, P. Murray-Rust *Computational Life Sciences II, Lecture Notes in Computer Science* **2006**, *4216*, 107.

[94] A. Copestake, P. Corbett, P. Murray-Rust, C. J. Rupp, A. Siddharthan, S. Teufel, B. Waldron *submitted to the UK eScience All Hands Meeting 2006* **2006**.

[95] S. Wu, T. Hayakawa, R. Kikuchi, S. J. Grunzinger, M. Kakimoto, H. Oikawa *Macromolecules* **2007**.

[96] The Wiley Database of Polymer Properties, Wiley, *http://www3.interscience.wiley.com/cgi-bin/mrwhome/104554802/HOME*

[97] Polymers - A Property Database, Taylor & Francis, *http://www.polymersdatabase.com*

Figure Captions

Figure 1: The semantic layer cake.

Figure 2: Repeat unit structure of poly(butadiene).

Figure 3: Repeat unit structure of poly(ethylene terephthalate) (PET).

Figure 4: Multiple possible repeat unit definitions for poly(butadiene).

Figure 5: Loss of information from a pdf document after conversion to plain text. The boxes indicate loss of bond multiplicity information, loss of special characters and loss of superscript/subscript information. (reproduced with permission from reference [95]).

Figure 6: Layered technologies for polymer informatics.

Figure 7: A simple RDF triple.

Figure 8: Simple CML and .mol documents describing the 2D structure of styrene.

Figure 9: A simple PML document describing a poly(styrene) oligomer.

Figure 10: Screenshots of the Cambridge Polymer Builder (A) before and (B) after building a model of a macromolecule (The builder is available at http://wwmm-svc.ch.cam.ac.uk/polydemo).

Figure 11: Graphical representation of the partial class hierarchy of the Cambridge polymer ontology.

Figure 12: Ontological description of the concept "MacroMolecule" in the OWL ontology language.

Figure 13: Polymers and polymer-related terms in a polymer paper marked up as a result of natural language processing.

Table 1

Table 1: Major sources of polymer information and accessibility notes.

Information Source (Publisher)	**Access notes**
Polymer Handbook[52] (Wiley)	Non-digital, contents copyrighted and all rights reserved by Wiley, commercial, no semantics.
The Wiley Database of Polymer Properties[96] (Wiley)	Digital, subscription basis, log-in required, contents copyrighted and all rights reserved by Wiley, commercial, no semantics. Derivative of Polymer Handbook.
Polymers – A Property Database[97] (Taylor & Francis)	Digital, subscription basis, log-in required, contents copyrighted and all rights reserved by Taylor and Francis, commercial, no semantics.
PoLyInfo Database[51] (National Institute for Materials Science, Japan)	Digital, log-in required, contents copyrighted and all rights reserved by NIMS, non-commercial, free to view, no semantics.

Table 2

Table 2: An abstract (ref. [62]) prior to markup (**A**) and marked up in SciXML (**B**).

(A) Elaboration of PLLA-based superparamagnetic nanoparticles: Characterization, magnetic behaviour study and in vitro relaxivity evaluation

Abstract. Oleic acid-coated magnetite has been encapsulated in biocompatible magnetic nanoparticles (MNP) by a simple emulsion evaporation method.

(B)

```
<?xml version="1.0" encoding="UTF-8"?>
<PAPER><TITLE>Elaboration of PLLA-based superparamagnetic
nanoparticles: Characterization, magnetic behaviour study
and in vitro relaxivity evaluation.</TITLE>[31]<ne
surface="Oleic acid" type="CM" provenance="unknown"
SMILES="CCCCCCCC\C=C/CCCCCCCC(O)=O"
InChI="InChI=1/C18H34O2/c1-2-3-4-5-6-7-8-9-10-11-12-13-
14-15-16-17-18(19)20/h9-10H,2-8,11-
17H2,1H3,(H,19,20)/b10-9-" cmlRef="cml1"
ontIDs="CHEBI:16196">Oleic acid</ne>-coated <ne
surface="magnetite" type="CM" provenance="nGramScore"
weight="0.09220993385201925">magnetite</ne> has been
encapsulated in biocompatible magnetic nanoparticles
(MNP) by a simple emulsion <ne surface="evaporation"
type="ONT" provenance="oscarLexicon"
ontIDs="REX:0000178">evaporation</ne>
method…..</ABSTRACT>
```

Figure 1

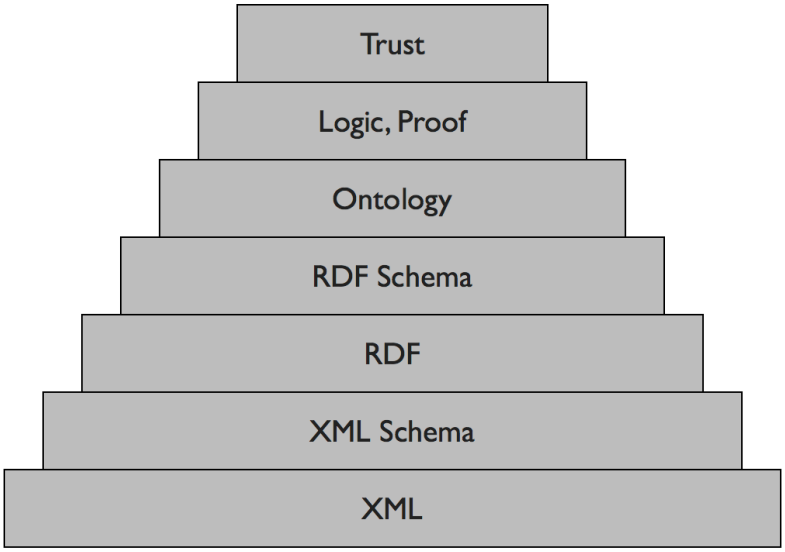


Figure 2

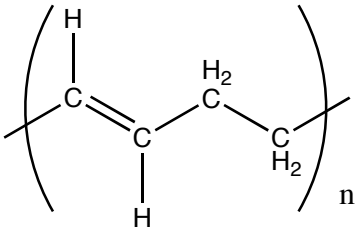


Figure 3

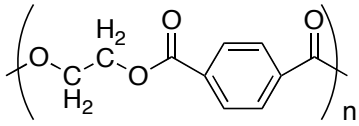
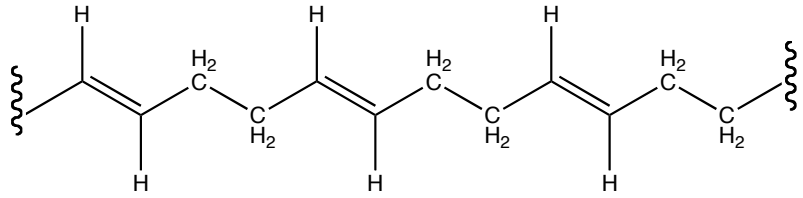


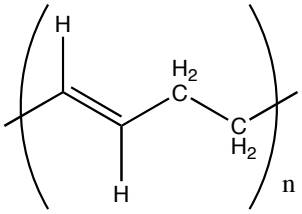
Figure 4

Poly(1,3-butadiene)

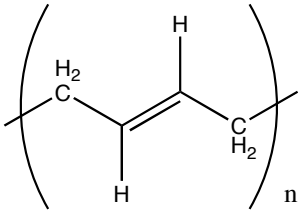


Possible Repeat Unit Definitions

(A)



(B)



(C)

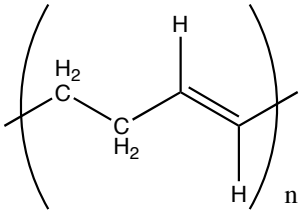


Figure 5

PDF

Synthesis of POSS—PAAs 7. In a typical experiment, **4** (0.50 g, 0.27 mol) was dissolved in 10 mL DMAc in a 25 mL three-necked, to which, BPDA (**d**, 0.08 g, 0.27 mol) was added while stirring. The suspension was stirred for 24 h at room temperature to yield a viscous solution, which was then poured into methanol. The precipitate was filtered off, washed with water, and dried under vacuum at 40 °C. The inherent viscosity of the resulting POSS−PAA **7d** is 0.38 dL/g, measured at a concentration of 0.5 g/dL in DMAc at 30 °C. IR (KBr): ν = 3445 (N−H), 1767, 1709 (amide C═O), 1634 (C═O, carboxylic acid), 1264 (Si−Me), 1234 (Si−Ph), 1132, 1083 (Si−O−Si) cm^{-1}. ^{1}H NMR (300 MHz, DMSO-d_6): 10.55 (2H, br), 8.12−7.95 (8H, m), 7.79 (2H, br), 7.46 (2H, br), 3.40(4H, br), 3.25(2H, br), 2.86(2H, br), 1.74 (4H, br), 1.51 (4H, br), 0.83 (2H, br), 0.32 (6H, br) ppm. ^{13}C NMR (75 MHz, DMSO-d_6) : 177.6, 177.5, 167.7, 167.2, 157.8, 156.4, 151.6, 136.4, 133.7, 131.5, 130.9, 130.1, 128.6, 127.1, 121.7, 120.5, 119.5, 118.1, 50.9, 48.6, 40.7, 40.1, 39.3, 26.3, 24.6, −1.7 ppm. ^{29}Si NMR (60 MHz, DMSO-d_6) : −21.6, −77.7, 78.3, −79.2 ppm.

After Conversion from PDF

Synthesis of POSS)PAAs 7. In a typical experiment, 4 (0.50 g, 0.27 mol) was dissolved in 10 mL DMAc in a 25 mL three-necked, to which, BPDA (d, 0.08 g, 0.27 mol) was added while stirring. The suspension was stirred for 24 h at room temperature to yield a viscous solution, which was then poured into methanol. The precipitate was filtered off, washed with water, and dried under vacuum at 40 °C. The inherent viscosity of the resulting POSS-PAA 7d is 0.38 dL/g, measured at a concentration of 0.5 g/dL in DMAc at 30 °C. IR (KBr): •)3445 (N-H), 1767, 1709 (amide CdO), 1634 (CdO, carboxylic acid), 1264 (Si-Me), 1234 (Si-Ph), 1132, 1083 (Si-O-Si) cm-1. 1H NMR (300 MHz, DMSO-d6): 10.55 (2H, br), 8.12-7.95 (8H, m), 7.79 (2H, br), 7.46 (2H, br), 3.40 (4H, br), 3.25(2H, br), 2.86(2H, br), 1.74 (4H, br), 1.51 (4H, br), 0.83 (2H, br), 0.32 (6H, br) ppm. 13C NMR (75 MHz, DMSO-d6) : 177.6, 177.5, 167.7, 167.2, 157.8, 156.4, 151.6, 136.4, 133.7, 131.5, 130.9, 130.1, 128.6, 127.1, 121.7, 120.5, 119.5, 118.1, 50.9, 48.6, 40.7, 40.1, 39.3, 26.3, 24.6, -1.7 ppm. 29Si NMR (60 MHz, DMSO-d6): -21.6, -77.7, 78.3, -79.2 ppm.

Figure 6

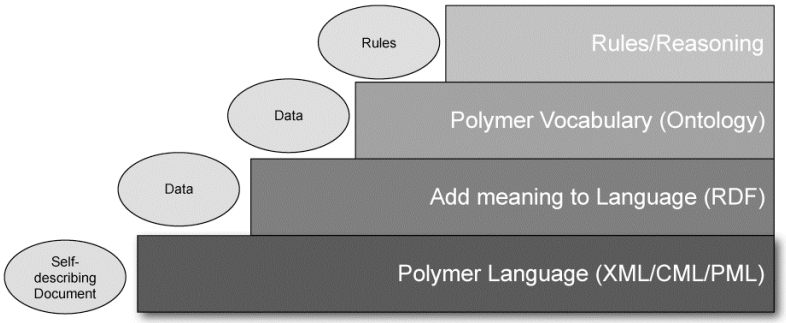

Rules
Rules/Reasoning
Data
Polymer Vocabulary (Ontology)
Data
Add meaning to Language (RDF)
Self-describing Document
Polymer Language (XML/CML/PML)

Figure 7

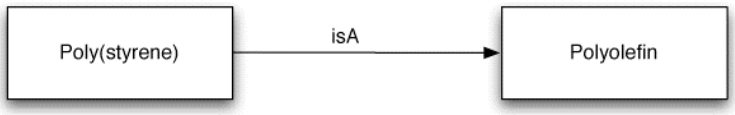
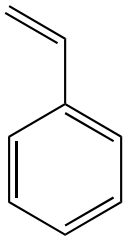


Figure 8



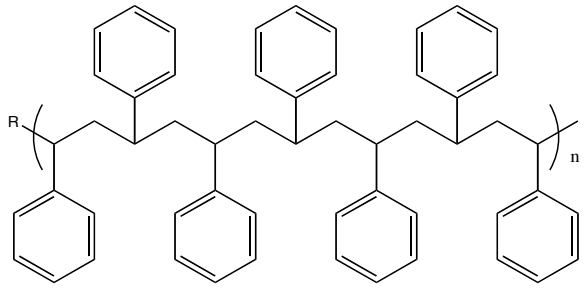
CML Connection Table Styrene

```
<?xml version="1.0"?>
<molecule xmlns="http://www.xml-
cml.org/schema/cml2/core">
 <atomArray>
  <atom id="a1" elementType="H"/>
  <atom id="a2" elementType="C"/>
  <atom id="a3" elementType="C"/>
  <atom id="a4" elementType="C"/>
  <atom id="a5" elementType="H"/>
  <atom id="a6" elementType="C"/>
  <atom id="a7" elementType="H"/>
  <atom id="a8" elementType="H"/>
  <atom id="a9" elementType="C"/>
  <atom id="a10" elementType="H"/>
  <atom id="a11" elementType="C"/>
  <atom id="a12" elementType="H"/>
  <atom id="a13" elementType="C"/>
  <atom id="a14" elementType="H"/>
  <atom id="a15" elementType="C"/>
  <atom id="a16" elementType="H"/>
 </atomArray>
 <bondArray>
  <bond atomRefs2="a1 a2" order="1"/>
  <bond atomRefs2="a2 a3" order="2"/>
  <bond atomRefs2="a3 a4" order="1"/>
  <bond atomRefs2="a4 a5" order="1"/>
  <bond atomRefs2="a4 a6" order="2"/>
  <bond atomRefs2="a6 a7" order="1"/>
  <bond atomRefs2="a6 a8" order="1"/>
  <bond atomRefs2="a3 a9" order="1"/>
  <bond atomRefs2="a9 a10" order="1"/>
  <bond atomRefs2="a9 a11" order="2"/>
  <bond atomRefs2="a11 a12" order="1"/>
  <bond atomRefs2="a11 a13" order="1"/>
  <bond atomRefs2="a13 a14" order="1"/>
  <bond atomRefs2="a13 a15" order="2"/>
  <bond atomRefs2="a2 a15" order="1"/>
  <bond atomRefs2="a15 a16" order="1"/>
 </bondArray>
</molecule>
```

Molfile Connection Table Styrene

```
  8  8  0  0  0  0  0  0  0  0999 V2000
   -0.7145   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -1.0312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -1.4438    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -1.0312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    1.0312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    1.4438    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  2  3  1  0
  3  4  2  0
  4  5  1  0
  5  6  2  0
  6  1  1  0
  6  7  1  0
  7  8  2  0
M  END
```

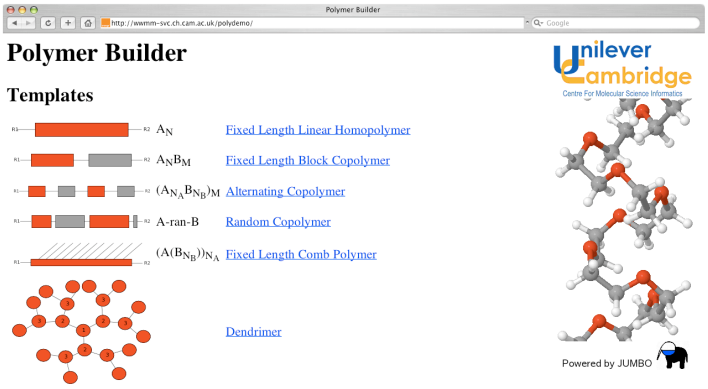
Figure 9




```
<?xml version="1.0" encoding="UTF-8"?>
<molecule id="polystyrene" convention="cml:PML-basic"
      xmlns:g="http://www.xml-cml.org/mols/geom1"
      xmlns="http://www.xml-cml.org/schema">
            <!--  polystyrene -->
            <fragment>
                  <molecule ref="g:dummy"/>
                  <fragmentList countExpression="*(7)">
                        <join order="1" moleculeRefs2="PARENT NEXT"
                              atomRefs2="r1 r1">
                              <torsion>180</torsion>
                        </join>
                        <fragment>
                              <molecule ref="g:ch"/>
                              <fragmentList>
                                    <join order="1" moleculeRefs2="PARENT NEXT"
                                          atomRefs2="r3 r1">
                              <torsion>90</torsion>
                                    </join>
                                    <fragment>
                                          <molecule ref="g:benzene"/>
                                    </fragment>
                              </fragmentList>
                        </fragment>
                        <join atomRefs2="r2 r2" moleculeRefs2="PREVIOUS NEXT">
                              <torsion>60</torsion>
                        </join>
                        <fragment>
                              <molecule ref="g:ch2"/>
                        </fragment>
                  </fragmentList>
            </fragment>
</molecule>
```

Figure 10

(A)


Polymer Builder
http://wwmm-svc.ch.cam.ac.uk/polydemo/
Polymer Builder
Templates
A_N Fixed Length Linear Homopolymer
A_NB_M Fixed Length Block Copolymer
$(A_{N_A}B_{N_B})_M$ Alternating Copolymer
A-ran-B Random Copolymer
$(A(B_{N_B}))_{N_A}$ Fixed Length Comb Polymer
Dendrimer
Unilever Cambridge
Centre For Molecular Science Informatics
Powered by JUMBO

(B)

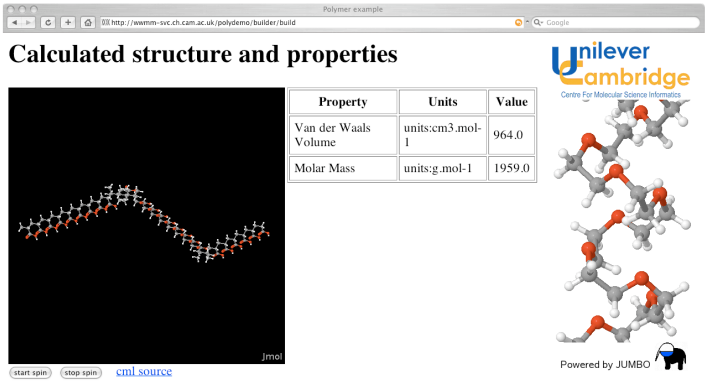

Polymer example
http://wwmm-svc.ch.cam.ac.uk/polydemo/builder/build
Calculated structure and properties
Property
Units
Value
Van der Waals Volume
units:cm3.mol-1
964.0
Molar Mass
units:g.mol-1
1959.0
Jmol
start spin
stop spin
cml source
Build another polymer
Unilever Cambridge
Centre For Molecular Science Informatics
Powered by JUMBO

Figure 11

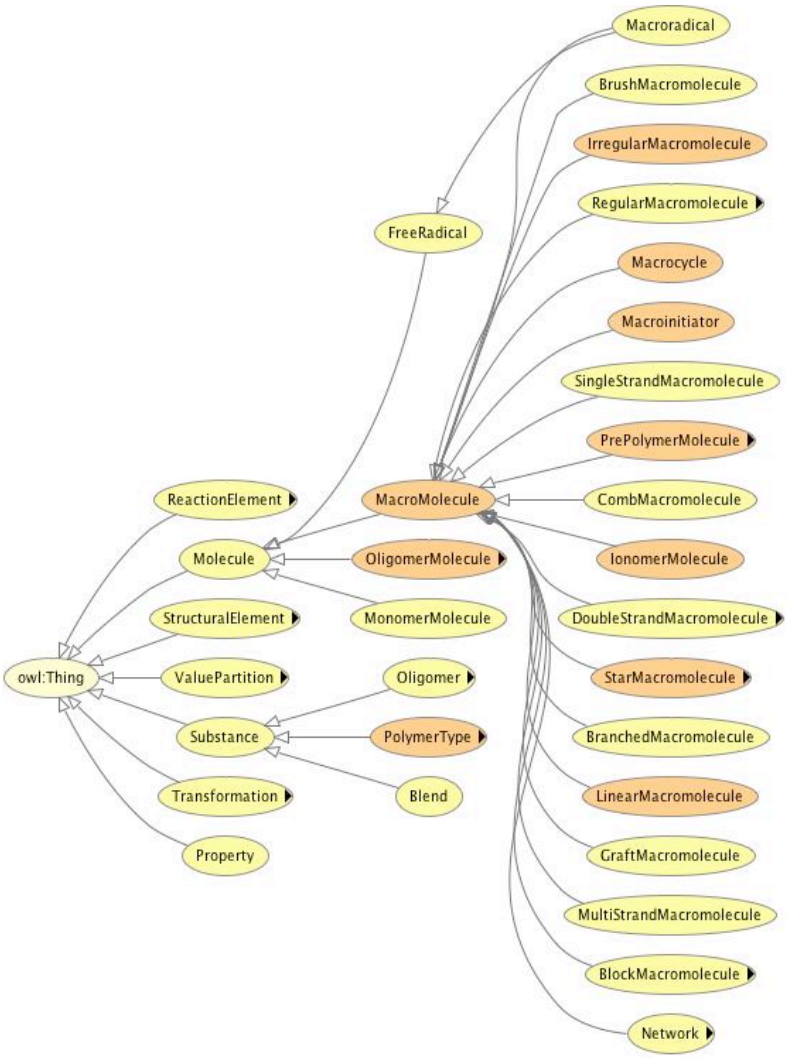

Macroradical
BrushMacromolecule
IrregularMacromolecule
RegularMacromolecule
FreeRadical
Macrocycle
Macroinitiator
SingleStrandMacromolecule
PrePolymerMolecule
ReactionElement
MacroMolecule
CombMacromolecule
Molecule
OligomerMolecule
IonomerMolecule
StructuralElement
MonomerMolecule
DoubleStrandMacromolecule
owl:Thing
ValuePartition
Oligomer
StarMacromolecule
Substance
PolymerType
BranchedMacromolecule
Transformation
Blend
LinearMacromolecule
GraftMacromolecule
Property
MultiStrandMacromolecule
BlockMacromolecule
Network

Figure 12

```
<owl:Class rdf:about="#MacroMolecule">
    <dc:creator xml:lang="en">Nico Adams</dc:creator>
    <dc:description xml:lang="en">A molecule of high relative molecular mass, the
structure of which
    essentially comprises the multiple repetition of units derived, actually
    or conceptually, from molecules of low relative molecular mass.</dc:description>
    <dc:source xml:lang="en">http://goldbook.iupac.org/M03667.html</dc:source>
    <owl:equivalentClass>
      <owl:Class>
        <owl:intersectionOf rdf:parseType="Collection">
          <owl:Restriction>
            <owl:onProperty>
              <owl:TransitiveProperty rdf:about="#hasStructuralElement"/>
            </owl:onProperty>
            <owl:someValuesFrom rdf:resource="#Endgroup"/>
          </owl:Restriction>
          <owl:Class rdf:about="#Molecule"/>
          <owl:Restriction>
            <owl:onProperty>
              <owl:FunctionalProperty rdf:about="#hasRelativeMolecularMass"/>
            </owl:onProperty>
            <owl:someValuesFrom rdf:resource="#High"/>
          </owl:Restriction>
          <owl:Restriction>
            <owl:onProperty>
              <owl:TransitiveProperty rdf:about="#hasStructuralElement"/>
            </owl:onProperty>
            <owl:someValuesFrom rdf:resource="#Chain"/>
          </owl:Restriction>
        </owl:intersectionOf>
      </owl:Class>
    </owl:equivalentClass>
  </owl:Class>
```

Figure 13

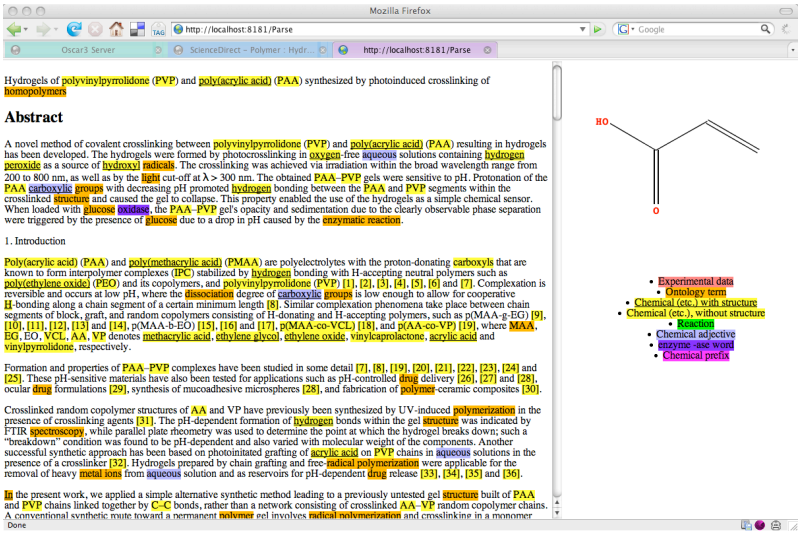

Mozilla Firefox
http://localhost:8181/Parse
Oscar3 Server
ScienceDirect – Polymer : Hydr...
http://localhost:8181/Parse
Hydrogels of polyvinylpyrrolidone (PVP) and poly(acrylic acid) (PAA) synthesized by photoinduced crosslinking of homopolymers
Abstract
A novel method of covalent crosslinking between polyvinylpyrrolidone (PVP) and poly(acrylic acid) (PAA) resulting in hydrogels has been developed. The hydrogels were formed by photocrosslinking in oxygen-free aqueous solutions containing hydrogen peroxide as a source of hydroxyl radicals. The crosslinking was achieved via irradiation within the broad wavelength range from 200 to 800 nm, as well as by the light cut-off at λ > 300 nm. The obtained PAA–PVP gels were sensitive to pH. Protonation of the PAA carboxylic groups with decreasing pH promoted hydrogen bonding between the PAA and PVP segments within the crosslinked structure and caused the gel to collapse. This property enabled the use of the hydrogels as a simple chemical sensor. When loaded with glucose oxidase, the PAA–PVP gel's opacity and sedimentation due to the clearly observable phase separation were triggered by the presence of glucose due to a drop in pH caused by the enzymatic reaction.
1. Introduction
Poly(acrylic acid) (PAA) and poly(methacrylic acid) (PMAA) are polyelectrolytes with the proton-donating carboxyls that are known to form interpolymer complexes (IPC) stabilized by hydrogen bonding with H-accepting neutral polymers such as poly(ethylene oxide) (PEO) and its copolymers, and polyvinylpyrrolidone (PVP) [1], [2], [3], [4], [5], [6] and [7]. Complexation is reversible and occurs at low pH, where the dissociation degree of carboxylic groups is low enough to allow for cooperative H-bonding along a chain segment of a certain minimum length [8]. Similar complexation phenomena take place between chain segments of block, graft, and random copolymers consisting of H-donating and H-accepting polymers, such as p(MAA-g-EG) [9], [10], [11], [12], [13] and [14], p(MAA-b-EO) [15], [16] and [17], p(MAA-co-VCL) [18], and p(AA-co-VP) [19], where MAA, EG, EO, VCL, AA, VP denotes methacrylic acid, ethylene glycol, ethylene oxide, vinylcaprolactone, acrylic acid and vinylpyrrolidone, respectively.
Formation and properties of PAA–PVP complexes have been studied in some detail [7], [8], [19], [20], [21], [22], [23], [24] and [25]. These pH-sensitive materials have also been tested for applications such as pH-controlled drug delivery [26], [27] and [28], ocular drug formulations [29], synthesis of mucoadhesive microspheres [28], and fabrication of polymer-ceramic composites [30].
Crosslinked random copolymer structures of AA and VP have previously been synthesized by UV-induced polymerization in the presence of crosslinking agents [31]. The pH-dependent formation of hydrogen bonds within the gel structure was indicated by FTIR spectroscopy, while parallel plate rheometry was used to determine the point at which the hydrogel breaks down; such a "breakdown" condition was found to be pH-dependent and also varied with molecular weight of the components. Another successful synthetic approach has been based on photoinitated grafting of acrylic acid on PVP chains in aqueous solutions in the presence of a crosslinker [32]. Hydrogels prepared by chain grafting and free-radical polymerization were applicable for the removal of heavy metal ions from aqueous solution and as reservoirs for pH-dependent drug release [33], [34], [35] and [36].
In the present work, we applied a simple alternative synthetic method leading to a previously untested gel structure built of PAA and PVP chains linked together by C–C bonds, rather than a network consisting of crosslinked AA–VP random copolymer chains. A conventional synthetic route toward a permanent polymer gel involves radical polymerization and crosslinking in a monomer
HO
O
Experimental data
Ontology term
Chemical (etc.) with structure
Chemical (etc.), without structure
Reaction
Chemical adjective
enzyme -ase word
Chemical prefix
Done

Table of Contents

The development of modern polymer informatics is an essential prerequisite for the success of increasingly data-driven polymer science. The paper discusses some of the challenges which need to be overcome in order to successfully establish this discipline and demonstrates some first technical solutions.

Table of Content Graphics