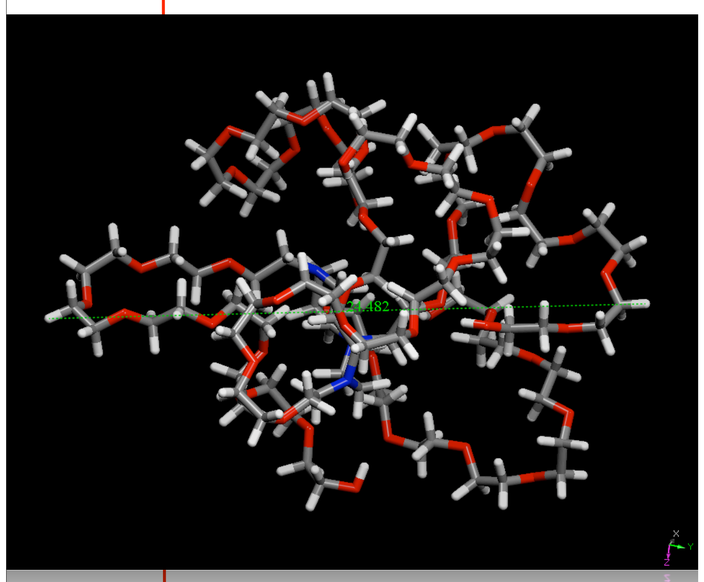
UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Science Informatics

Engineering Polymer Informatics

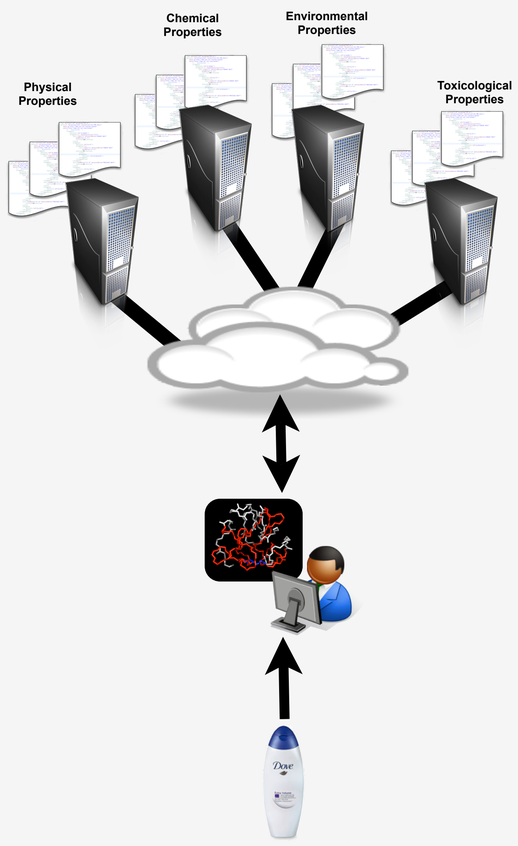
Nico Adams, Jen Ryder, Nicholas England, David Jessop, Peter Corbett, Peter Murray-Rust



Our mission is to develop an informatics toolbox, which will take into account the special computational needs of polymers and will make compromises in terms of data management and property calculation obsolete. This will be done by

- **creating a polymer information model, which captures polymer related information in a granular way, using modern semantic web technologies.**
- **developing new polymer descriptors**
- **developing new methods/models for the calculation of polymer properties**
- **developing text mining tools for polymer-related information**
- **development of selection rules for polymer benefits**

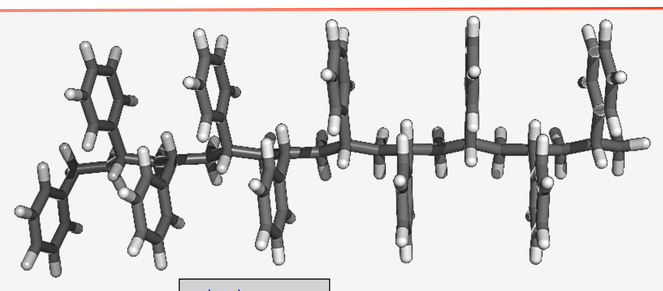
The Vision: Computer Aided Design of Polymers



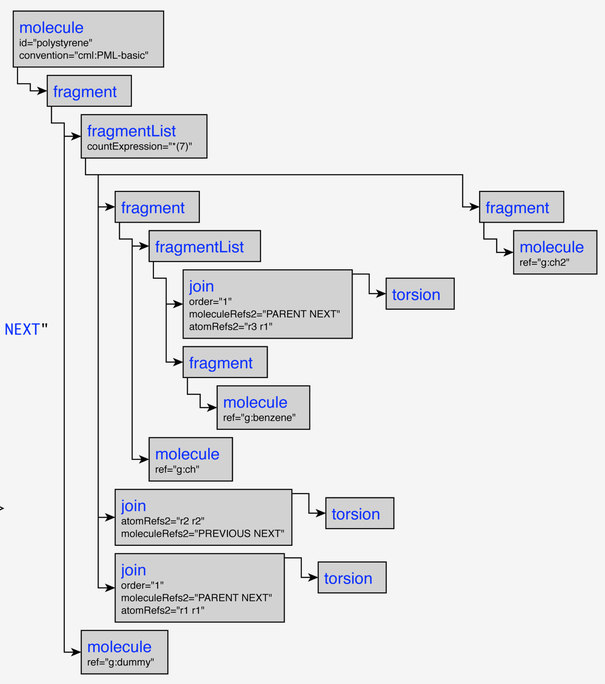
Our vision for polymer informatics is analogous to Berners-Lee's vision of the semantic web.[1] In an ideal scenario, a researcher would develop a property/requirements profile of polymer, taking into account physicochemical properties, environmental properties, toxicology etc. The researcher subsequently instructs an agent (a computer program) to try and find a polymer, which either completely or approximately fulfills the required criteria. The agent accomplishes this either via simple lookup by retrieving and combining relevant information from different sources, such as proprietary and public databases and information on the web, or by using the retrieved information and a set of rules to suggest a lead structure for a hitherto unknown polymeric entity.

Polymer Markup Language

- built on Chemical Markup Language
- extensible
- build polymers from small chemical fragments
- cope with different levels of certainty
- rich annotation capabilities
- cooperative with ThermoML, AniML



```
<?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>
```

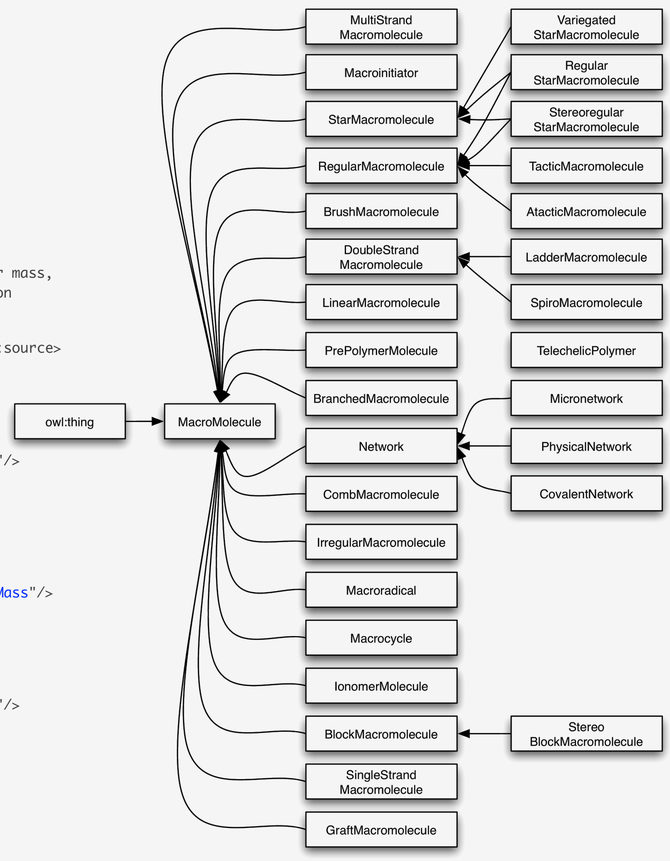


Polymer Markup Language (PML) represents a completely new approach to the representation of polymers. It is semantically completely explicit and allows polymers to be represented at various levels of certainty in a consistent manner. As an example, it is possible to represent an ill-defined system such 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 or measures of reactivity, which can be used to, for example, model competing reactive centres. Moreover, it also allows phenomena such as the law of mass action as well as molecular weight and other distributions to be taken into account when constructing a polymer. All of this represents a significant advance in comparison with other known polymer representation systems.

The Polymer Ontology

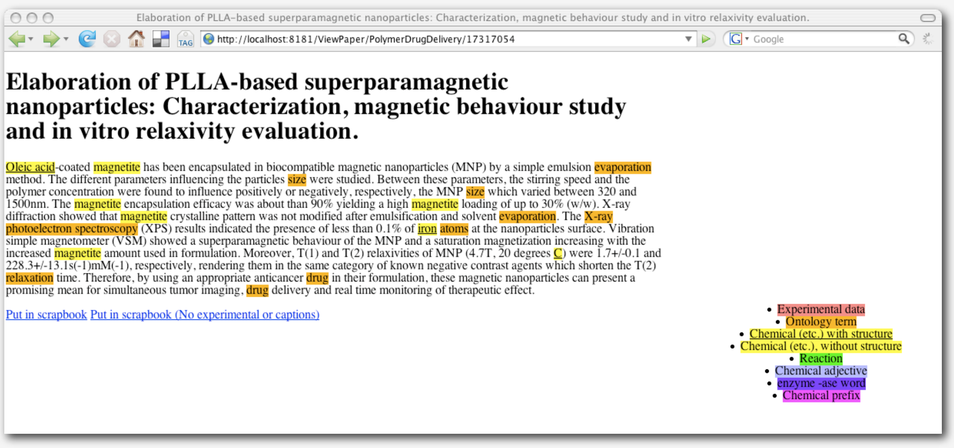
- an explicit specification of a shared conceptualization
- developed in Web Ontology Language (OWL)
- facilitates knowledge sharing
- facilitates knowledge re-use
- separation of declarative and procedural knowledge

```
<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>
```



For polymer informatics, ontologies have several uses. Firstly, an ontology serves to share a common understanding of the information structure of a domain between people and software agents. In the above scenario, a software agent is 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 by making 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.

Mining of Polymer Text - the OSCAR 3 system



Elaboration of PLLA-based superparamagnetic nanoparticles: Characterization, magnetic behaviour study and in vitro relaxivity evaluation.
Oleic acid-coated magnetite has been encapsulated in biocompatible magnetic nanoparticles (MNP) by a simple emulsion evaporation method.

```
<?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><Abstract><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>
```

OSCAR 3[2] is part of the SciBorg system[3] 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 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 above was generated automatically by OSCAR 3.

Contact Address
Dr Nico Adams
Unilever Centre for Molecular Science Informatics
University Chemical Laboratory
University of Cambridge
Lensfield Road
Cambridge CB2 1EW, United Kingdom

References
[1] T. Berners-Lee *et al.*, Scientific American (2001)
[2] P. Corbett *et al.*, Comp. Life Sci. II, Lecture Notes in Computer Science, 4216, 107 (2006)
[3] A. Copestake *et al.*, Abstracts UK eScience All Hands Meeting (2006)

Acknowledgements

We wish to thank Unilever PLC for financial support of this work.
Furthermore, we wish to acknowledge Dr Jerry Winter for many helpful discussions and suggestions.