Named Entity Resolution in OSCAR4

Daniel Lowe

Unilever Centre for Molecular Science Informatics

dl387@cam.ac.uk

13th April 2011, OSCAR4 Launch

Contents

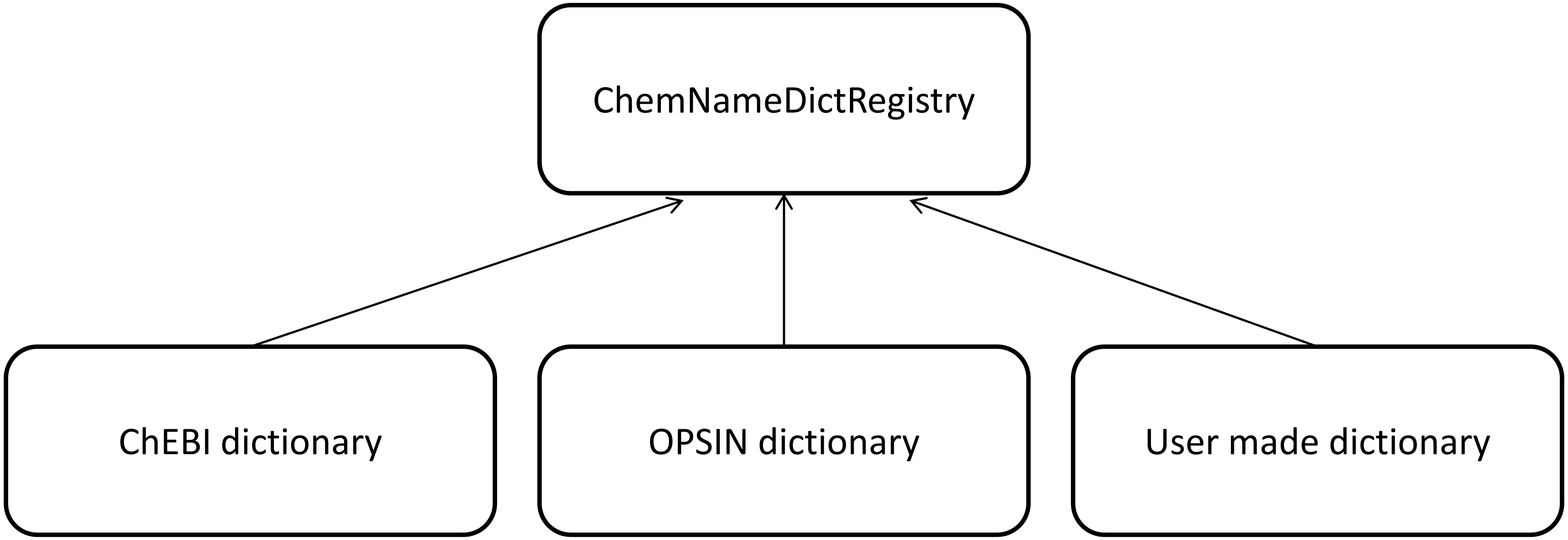
- Named entity resolution architecture
- Systematic name resolution (OPSIN)
- Extending entity resolution
- Conclusions

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

Dictionary Architecture

- Dictionaries are registered with a dictionary registry which then provides a central point of access
- To be registerable a class need only implement the ChemNameDict interface



Using dictionaries for Name to Structure

- Dictionaries that contain name to structure mappings implement the appropriate interface for that type of chemical identifier.

- IInChIProvider

  InChI=1/C6H12/c1-2-4-6-5-3-1/h1-6H2

- ISMILESProvider

  C1CCCCC1

- ICMLProvider

  <cml>…</cml>

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

Implementing a provider

```
public interface IInChIProvider {
        /**
         * Returns a set containing all of the known InChIs for
         * the given query name.
         */
        public Set<String> getInchis(String queryName);
}
```

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

Resolving entities against the dictionaries

ChemNameDictRegistry:

- resolveNamedEntity(NamedEntity ne)
- getInchis(String name)
- …

OSCAR API:

- findAndResolveNamedEntities(String input)
- findResolvableEntities(String input)

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

Ontology Term Resolution

- ChEBI ontology, REX (physico-chemical process ontology), FIX (biophysical chemistry ontology)
- Ontology terms are stored as a multimap between terms and Ids.
- Ontology terms are found and resolved during the process of finding named entities or can be resolved using an OntologyTerms object's methods.

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

Extending Ontology Support

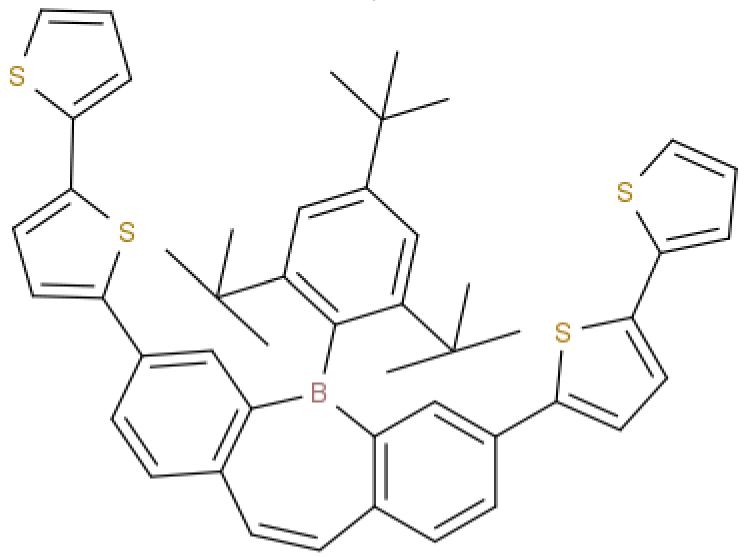
- A custom OntologyTerms can be created from a multimap of terms and ids.

```
OntologyTerms myCustomOntologyTerms = new
OntologyTerms(myTermIdMappings);
Oscar oscar = new Oscar();
oscar.setOntologyTerms(myCustomOntologyTerms);
```

OPSIN

(Open Parser for Systematic IUPAC Nomenclature)

3,7-Di([2,2′-bithiophen]-5-yl)-5-(2,4,6-tri-tert-butylphenyl)dibenzo[b,f]borepin



Centre For Molecular Science Informatics

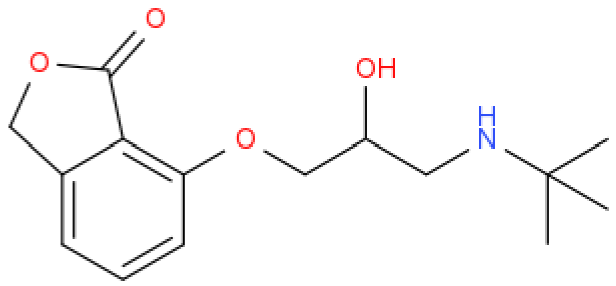
Why is dictionary lookup insufficient?

- Infinite number of chemical compounds

- Many different ways of naming the same compound even if names are computer generated

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE



Name on Wikipedia and in dictionaries:
7-[3-(tert-butylamino)-2-hydroxy-propoxy]-3H-isobenzofuran-1-one

Other suitable names generated by structure to name programs:

7-(3-(tert-butylamino)-2-hydroxypropoxy)isobenzofuran-1(3H)-one

7-[3-(tert-butylamino)-2-hydroxypropoxy]-1,3-dihydro-2-benzofuran-1-one

7-[3-(tert-butylamino)-2-hydroxypropoxy]-2-benzofuran-1(3H)-one

Why is dictionary lookup insufficient?

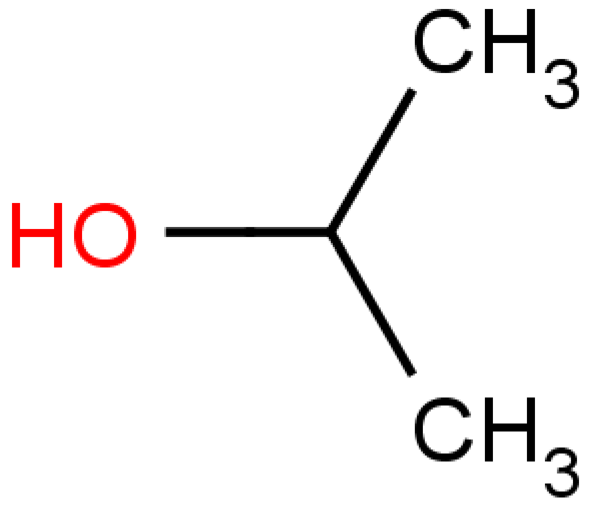
- Infinite number of chemical compounds
- Many different ways of naming the same compound even if names are computer generated
- Names for the same compound can have very little in common if different nomenclature is employed

Different naming styles

- isopropanol
- Isopropyl alcohol
- 2-Propanol
- Propan-2-ol
- 2-Hydroxypropane
- sec-Propyl alcohol
- 1-Methylethanol
- i-Propanol
- i-Propyl alcohol
- n-Propan-2-ol
- sec-Propanol
- 2-Propyl alcohol
- 1-Methylethyl alcohol
- Iso-propyl alcohol
- iso-propylalcohol
- Dimethylcarbinol

$\longrightarrow$

InChI=1/C3H8O/c1-3(2)4/h3-4H,1-2H3



Centre For Molecular Science Informatics

Key Facts

- Outputs to CML, InChI or SMILES
- Provides high precision conversions
- Offers comparable recall to commercial offerings on systematic names

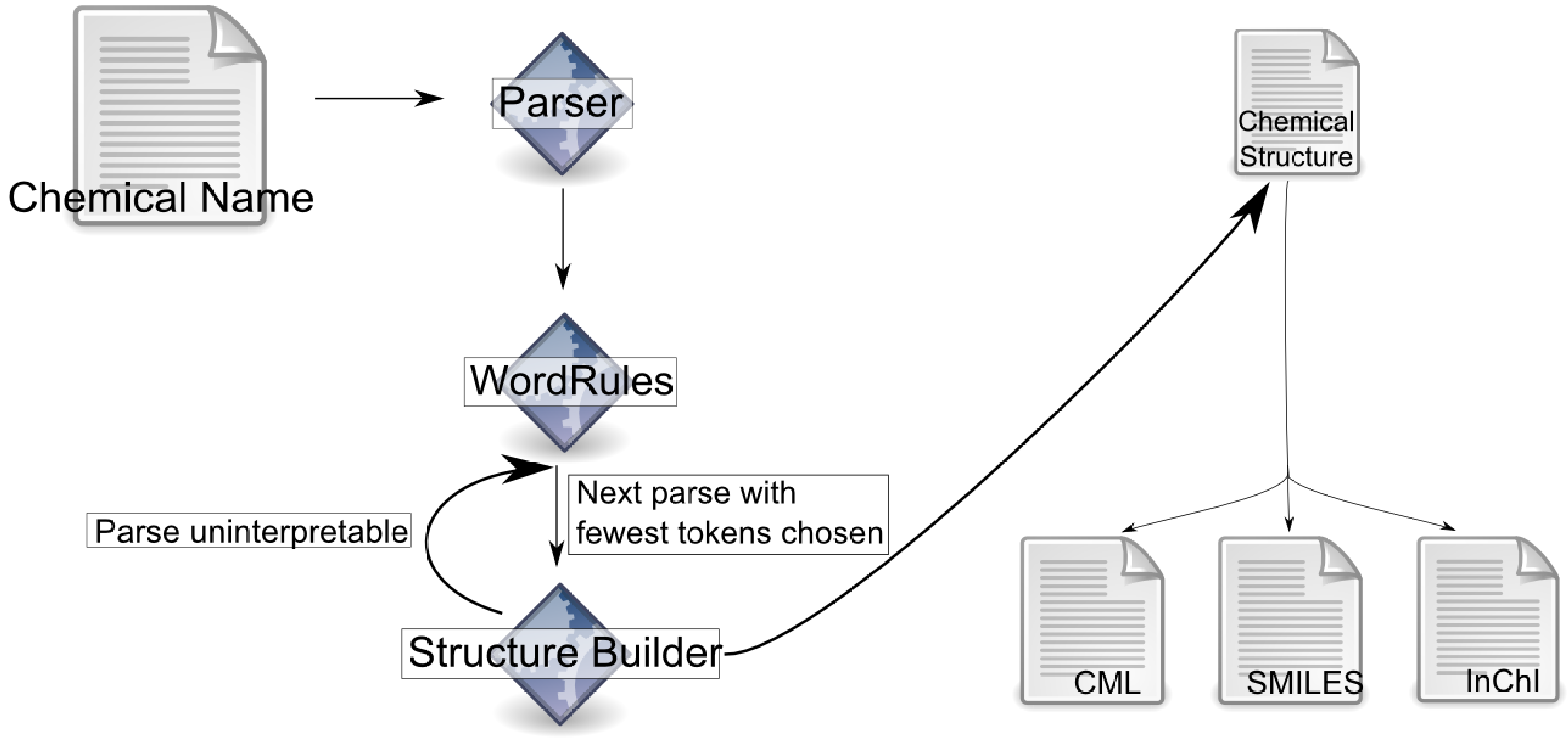
Chemical Name to Structure: OPSIN, an Open Source Solution
Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, Robert C. Glen
Journal of Chemical Information and Modeling **2011** *51* (3), 739-753

Most Read Paper in JCIM in the previous month!

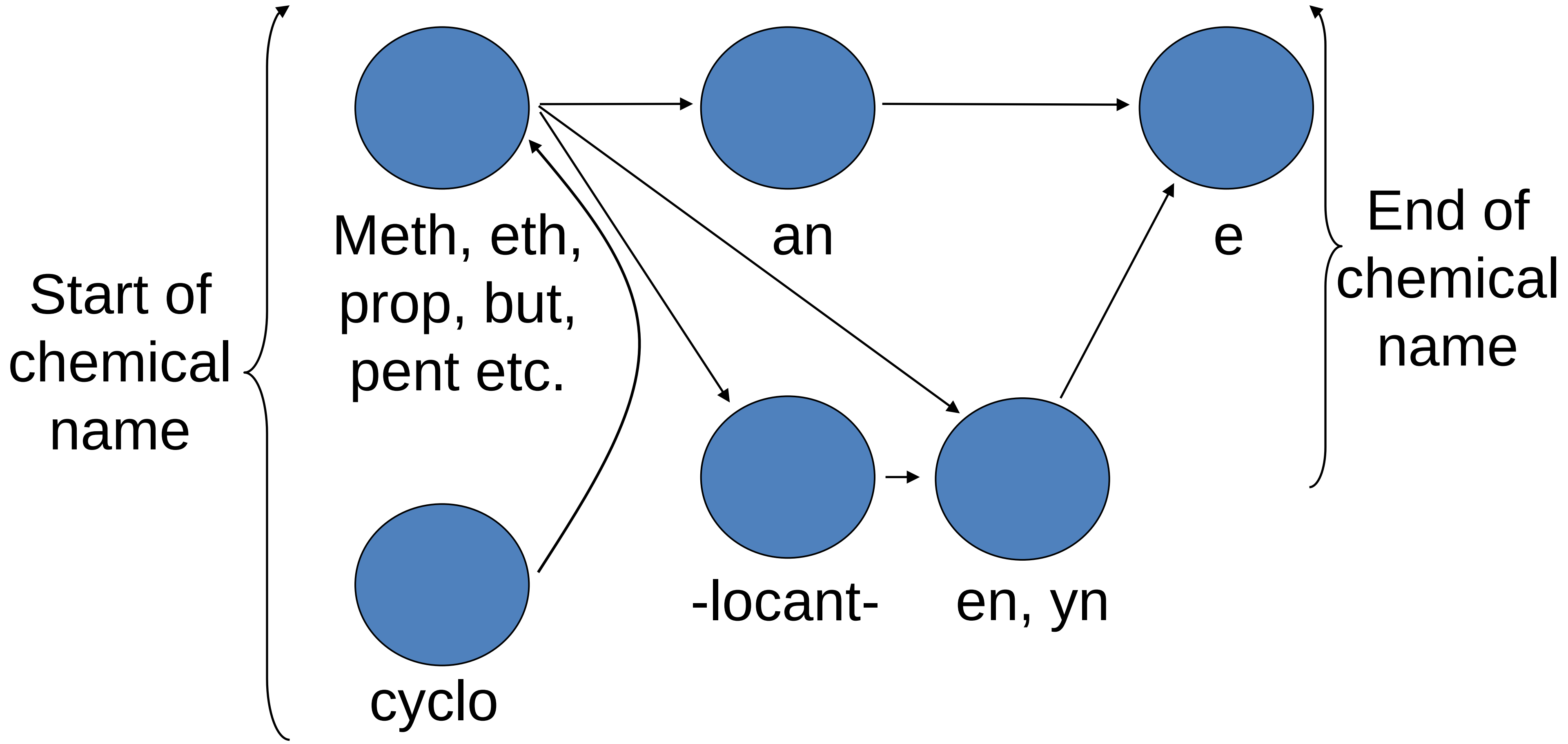
Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

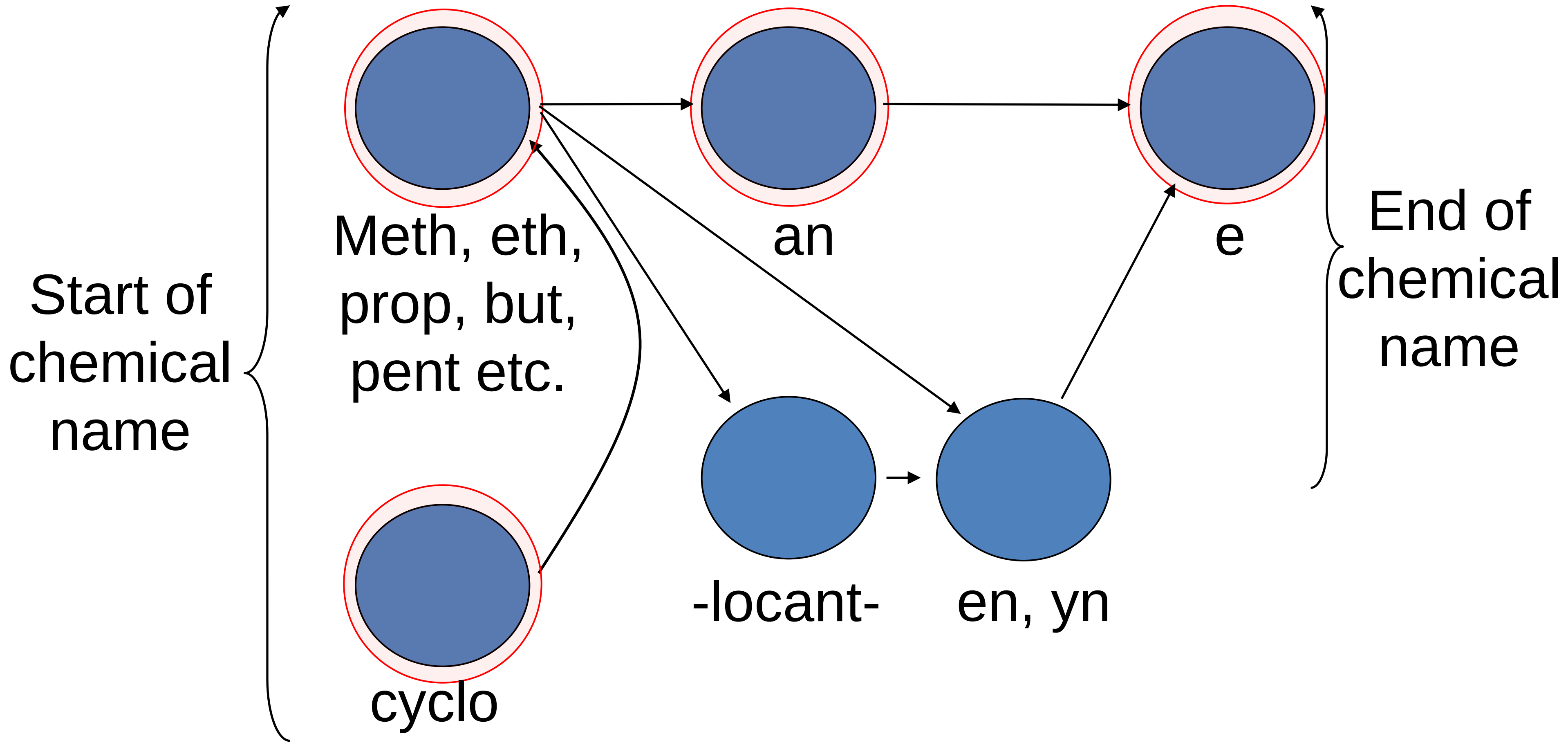
Schematic of program



A regular grammar for hydrocarbons



e.g. cyclopropane



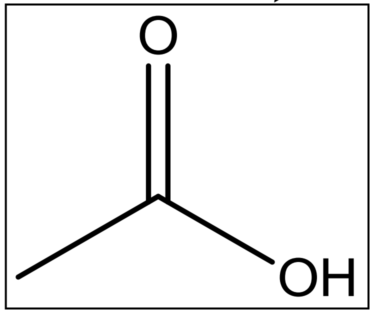
- OPSIN's grammar describes a finite state machine with 10091 states
- Around 3500 discrete morphemes form the program's vocabulary
- These are grouped into 117 morpheme classes e.g. multiplier (mono, di, tri, tetra, penta…), acidStem (acet, oxal, succin…)

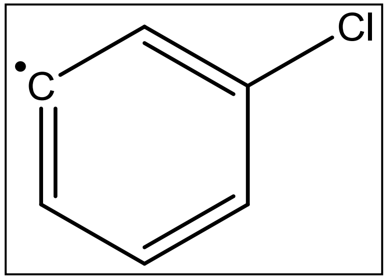
Example

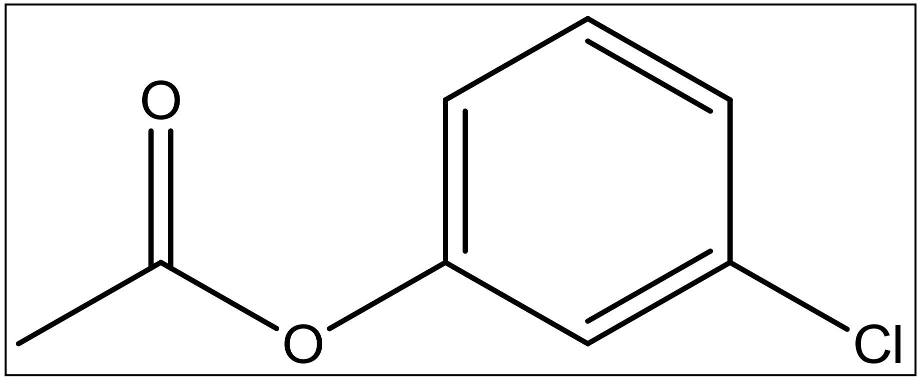
ethanoic acid m-chlorophenyl ester

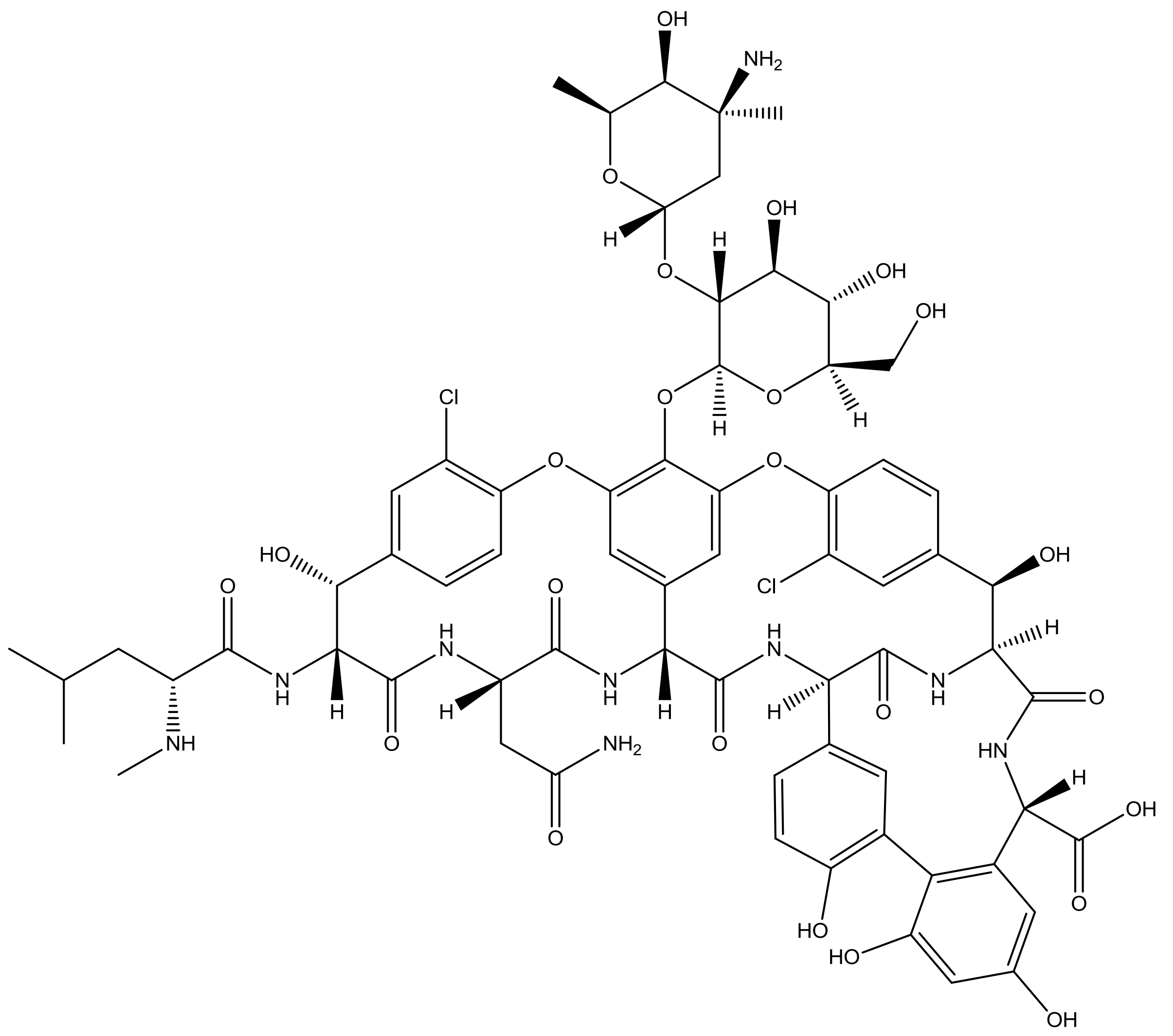
full substituent functionalTerm

WordRule="functionalClassEster"









(1S,2R,18R,19R,22S,25R,28R,40S)-48-{[(2S,3R,4S,5S,6R)-3-{[(2S,4S,5S,6S)-4-amino-5-hydroxy-4,6-dimethyloxan-2-yl]oxy}-4,5-dihydroxy-6-(hydroxymethyl)oxan-2-yl]oxy}-22-(carbamoylmethyl)-5,15-dichloro-2,18,32,35,37-pentahydroxy-19-[(2R)-4-methyl-2-(methylamino)pentanamido]-20,23,26,42,44-pentaoxo-7,13-dioxa-21,24,27,41,43-pentaazaoctacyclo[26.14.2.2^{3,6}.2^{14,17}.1^{8,12}.1^{29,33}.0^{10,25}.0^{34,39}]pentaconta-3,5,8(48),9,11,14,16,29(45),30,32,34,36,38,46,49-pentadecaene-40-carboxylic acid

Extending entity resolution

Proof of concept code for retrieving InChIs from the NIH's Chemical Identifier Resolver

```
public class CIRDictionary implements IChemNameDict, IInChIProvider{
 ........................................
public Set<String> getInchis(String queryName) {
    Set<String> inchis = new HashSet<String>();
    try{
        URI uri = new URI("http", null, "cactus.nci.nih.gov", 80, "/chemical/structure/" + queryName +"/stdinchi", null, null);
        URLConnection connection = uri.toURL().openConnection();
        BufferedReader in = new BufferedReader(new InputStreamReader(connection.getInputStream()));
        StringBuilder input = new StringBuilder();
        String inputLine = null;
        while ((inputLine = in.readLine()) != null){
            input.append(inputLine);
        }
        inchis.add(input.toString());
    }
    catch (Exception e) {
        LOG.debug("Request to CIR failed", e);
    }
    return inchis;
}

public boolean hasName(String queryName) {
    return getInchis(queryName).size()>0;
}

........................................

}
```

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

```
Oscar oscar = new Oscar();

ChemNameDictRegistry myChemNameDictReg  = new ChemNameDictRegistry();

myChemNameDictReg.register(new CIRDictionary());

oscar.setDictionaryRegistry(myChemNameDictReg);

List<ResolvedNamedEntity> resolvedEntities =
oscar.findResolvableEntities("Cholecalciferol is not in ChEBI.");
```

Conclusions

- OSCAR4 provides out of the box named entity resolution through simple APIs.

- OPSIN's provides commercial level name to structure performance across a wide range of IUPAC chemical nomenclature.

- OSCAR4 contains a readily extensible system for adding custom dictionaries and ontology term/id pairs.

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

Any Questions?

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE