chem-bla-ics

The CDK data model #1

Egon Willighagen

Published July 16, 2007

Citation
Willighagen, E. (2007, July 16). The CDK data model #1. *Chem-bla-ics*. https://doi.org/10.59350/4xfms-7nn46

Keywords
Cdk, Cheminf

Abstract
The Chemistry Development Kit has a rich set of data classes, each of which is defined by an interface. While the classes for atoms, bonds and a connectivity table are fairly straightforward, but beyond that it is sometimes not entirely clear. I will now discuss all interfaces in a series of blog items. I'll start with the IChemFile. Christoph, please correct me if I move to far away from our Notre Dame board sketch.

The Chemistry Development Kit has a rich set of data classes, each of which is defined by an interface. While the classes for atoms, bonds and a connectivity table are fairly straightforward, but beyond that it is sometimes not entirely clear. I will now discuss all interfaces in a series of blog items. I'll start with the IChemFile. Christoph, please correct me if I move to far away from our Notre Dame board sketch.

IChemFile

The IChemFile is the class to hold a chemical document, e.g. a MDL molfile or a PDB file. The idea of this class is that it can hold anything we can expect from a chemical document. But nothing beyond that either; a XHTML document with embedded CML is outside the scope of a IChemFile. You might wonder why the IChemObjectReaders not always just return a IChemFile. That would be a fair point, any many actually do, but somethings it is handier to return an IMolecule. A reader for MDL molfiles would be expected to return a IMolecule.

However, a document may contain much more, and the approach taken by the CDK is that a file contains one or more models. A MDL molfile is an example document with one model, while a MDL SD file would be a document with more than one model.

IChemSequence

However, the IChemFile can hold more than one IChemSequence. Now, I honestly cannot remember why that is; a single IChemSequence should be enough. And, I actually do not remember more than one IChemSequence being used. (Anyone?) As said, the IChemSequence contains IChemModels, and nothing more really. The interface therefore just contains the basic logic of a list. Let's move on.

IChemModel

The IChemModel is much more interesting. In the CDK a model is defined as anything that occurs in one actual volume of 3D (or 2D) space. A CIF file with a crystal structures is, therefore, one IChemModel. A supramolecular aggregation of lipids, e.g. a mono- or bilayer, would be IChemModel too. This could be a time step in a molecular dynamics run. Additionally, the IChemModel may also be a chemical reaction, possibly a multistep reaction. It could be, for example, a enzyme reaction mechanism entry from the MACiE database . These three types of content are captured in the ICrystal, IMoleculeSet, and IReactionSet.

Some Examples

A CIF file would be read as an IChemFile contains an IChemSequence with one IChemModel containing an ICrystal. An MDL molfile would be read as an IChemFile containing an IChemSquence with one IChemModel containing a IMoleculeSet with one IMolecule. And, an MDL SD file, however, would be read is an IChemFile with an IChemSequence with as many

chem-bla-ics

IChemModels as there are molecules in the SD file; and, each IChemModel would contains a IMoleculeSet with only one IMolecule. Counter-intuitively, because one may expect the SD file, which is a set of molecules, being stored in a IMoleculeSet.

Enough for tonight. More later. For the impatient, previously I wrote up a short blog about the update notification scheme in the CDK interfaces .