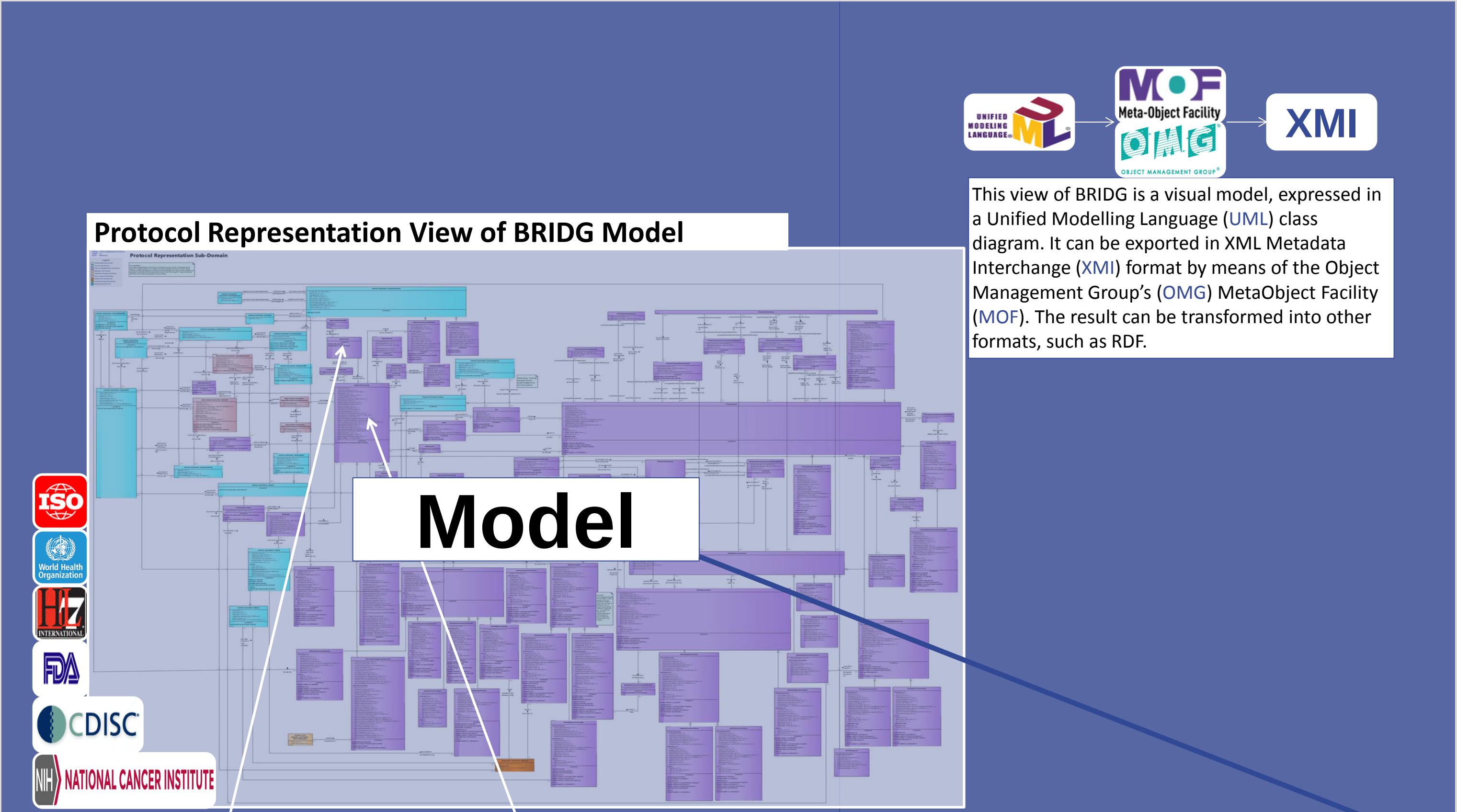


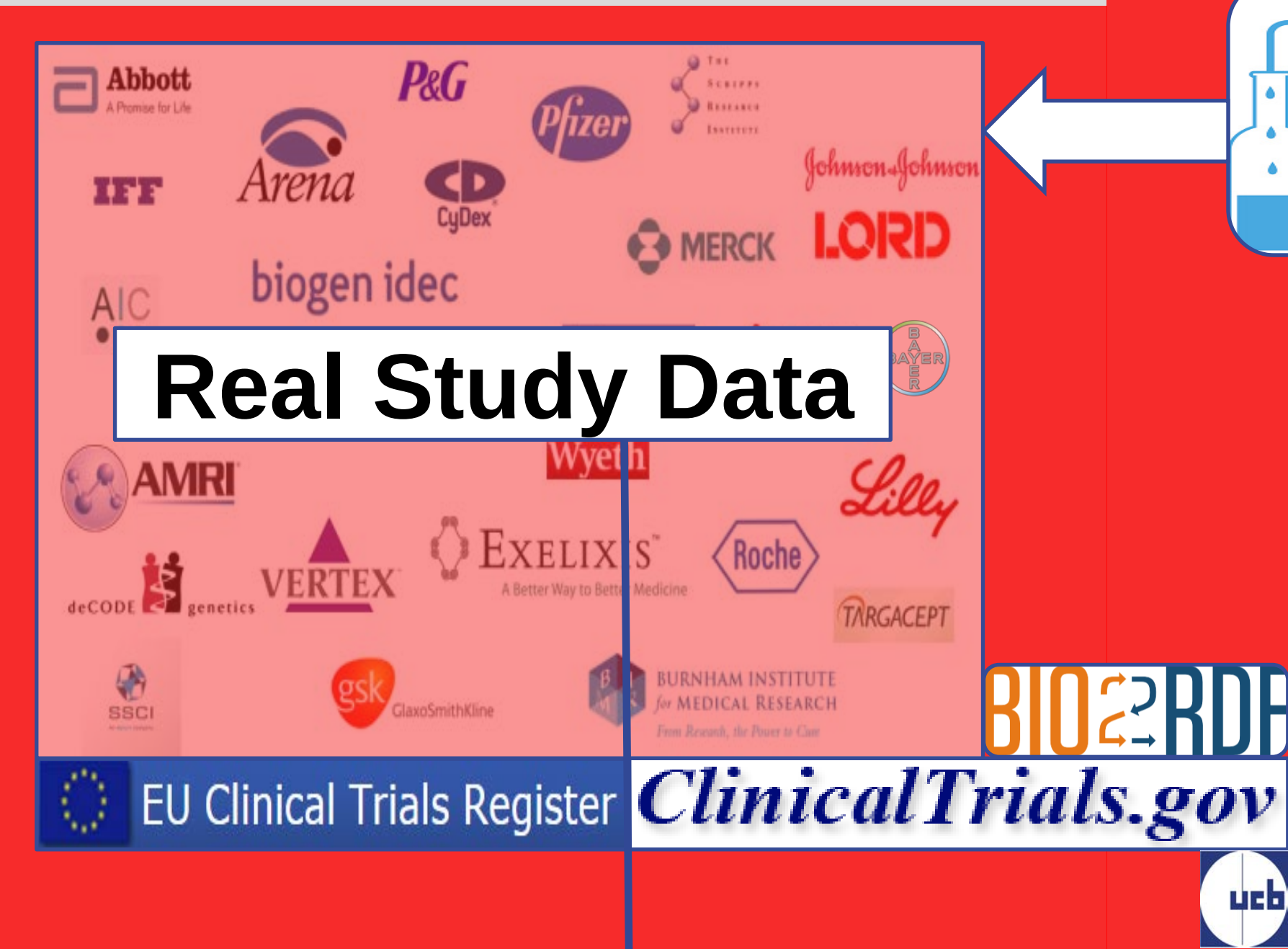
Constructing Interoperable Study Documents From A Semantic Technology-based Repository

Colin de Klerk
UCB BioSciences GmbH
Monheim, Germany



This view of BRIDG is a visual model, expressed in a Unified Modelling Language (UML) class diagram. It can be exported in XML Metadata Interchange (XMI) format by means of the Object Management Group's (OMG) MetaObject Facility (MOF). The result can be transformed into other formats, such as RDF.

This poster shows how new, structured and semantic-aware clinical documents can be generated from a repository connected to external standards and containing enterprise-level concepts and extensions.



Unstructured data from real studies might not be in a form that can be readily processed and would then need to be processed using text mining and natural language processing (NLP) techniques.

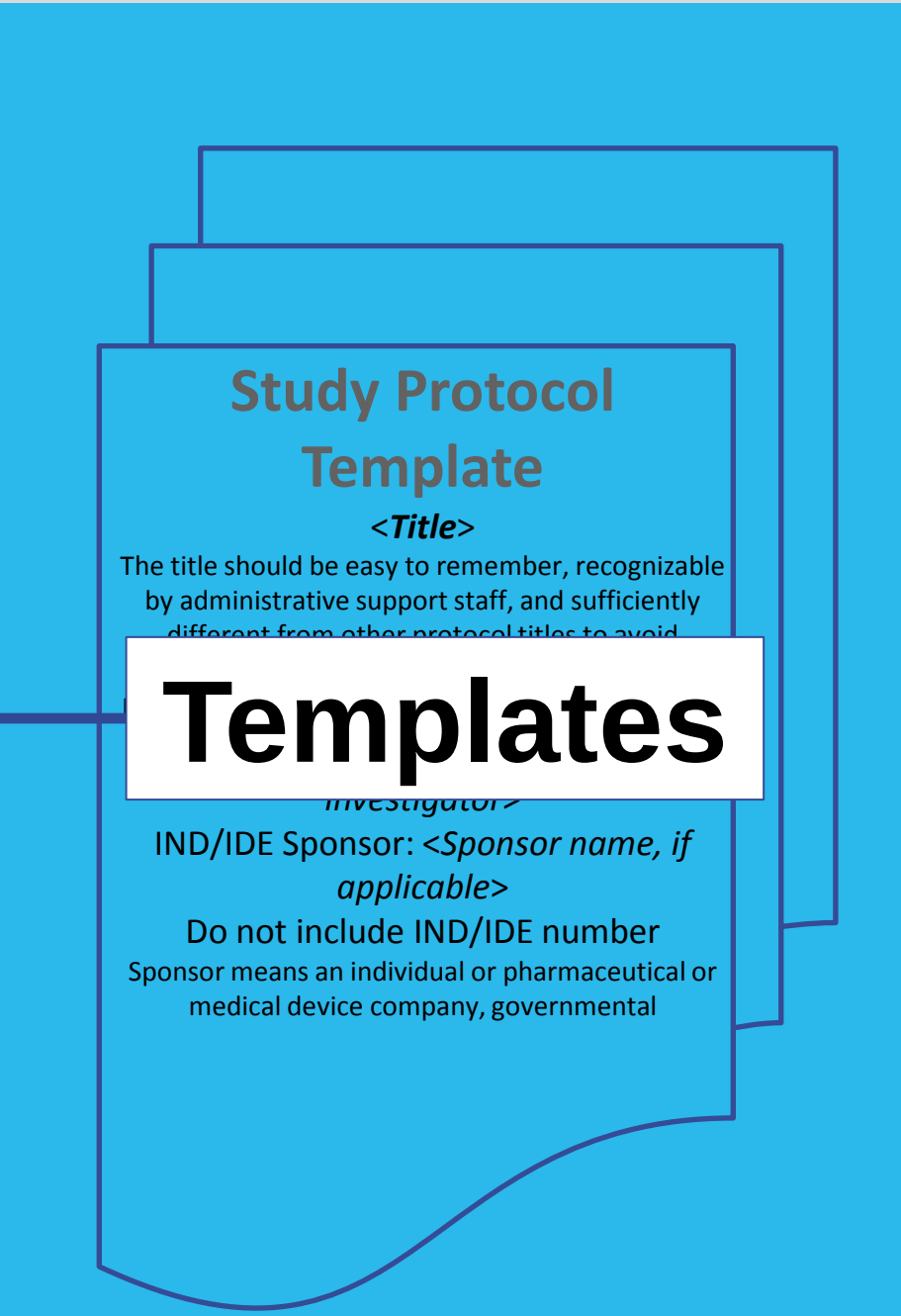
Real world study data might even need to be converted to text first using optical character recognition (OCR).



Consider the study with the reference number NCT01550003 and with UCB as the sponsor. This study is also known by the acronym PASCAL.

We can use Bio2RDF to extract publicly-available data about this UCB study from clinicaltrials.gov using <http://bio2rdf.org/describe?uri=http://bio2rdf.org/clinicaltrials:NCT01550003> to obtain this fragment of the RDF data returned by the query:

```
...<<omitted>>...
@prefix ns8: <http://identifiers.org/clinicaltrials/> .
ns1:NCT01550003 <http://bio2rdf.org/bio2rdf_vocabulary:x-identifiers.org>
ns8:NCT01550003 ;
ns2:acronym "PASCAL"^^xsd:string ;
...<<omitted>>...
```



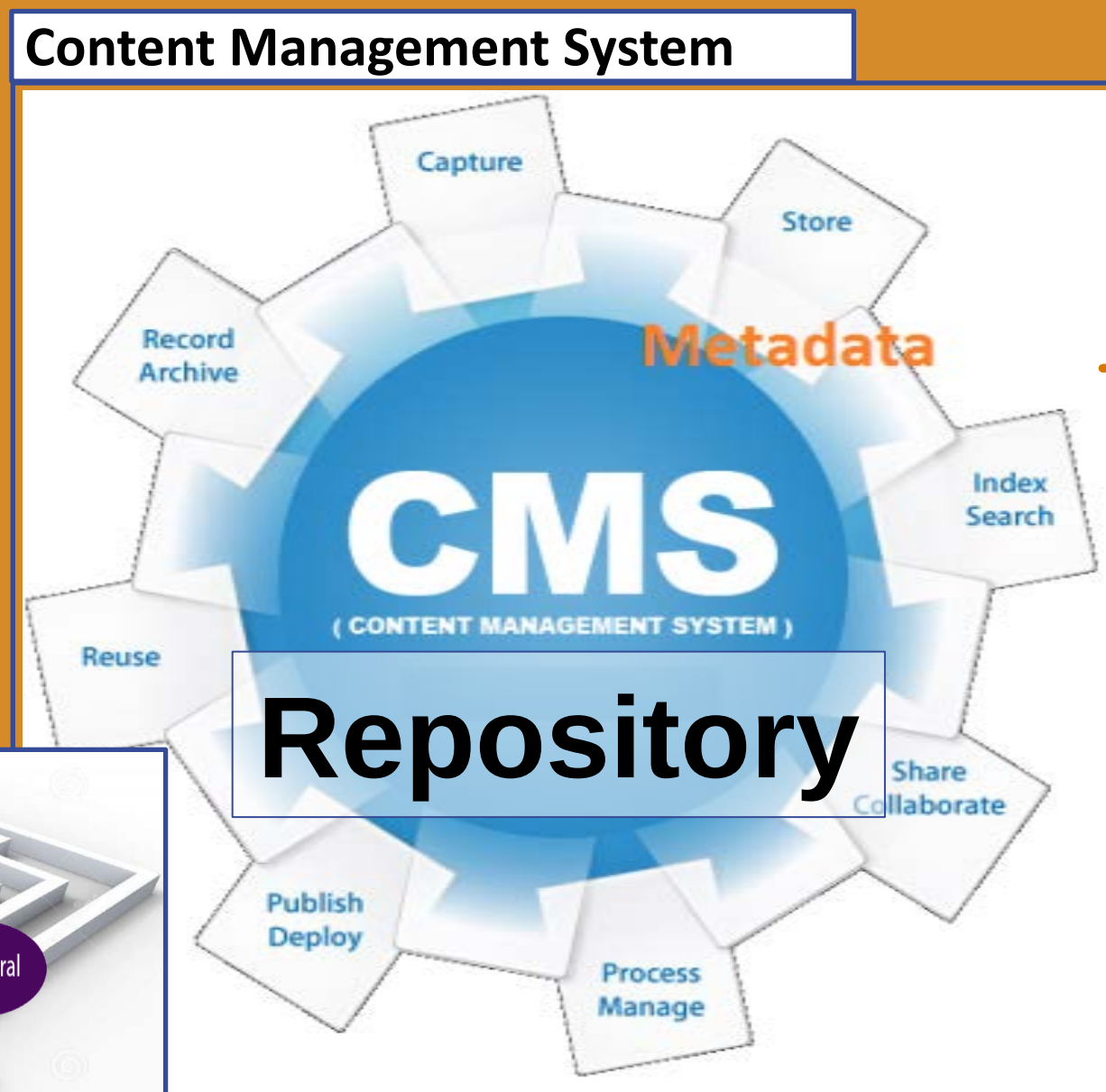
studyProtocol

- + acronym: ST[0..1]
- + mandatoryIndicator: BL [0..1]
- ...

The model captures best practises across the industry, derived from the cumulative experience of major industry players (see logos). It represents the typical business processes that agents perform in their roles in the clinical study, as well as material and data flows.

UML classes relating to the **study protocol** are shown in this view. Most of the important attributes are associated with the **study protocol version** class. This magnified view shows just a few with their associated data types.

So in the general case, the model has a placeholder for the attribute „acronym“ – which is a short-hand way to refer to a given study.



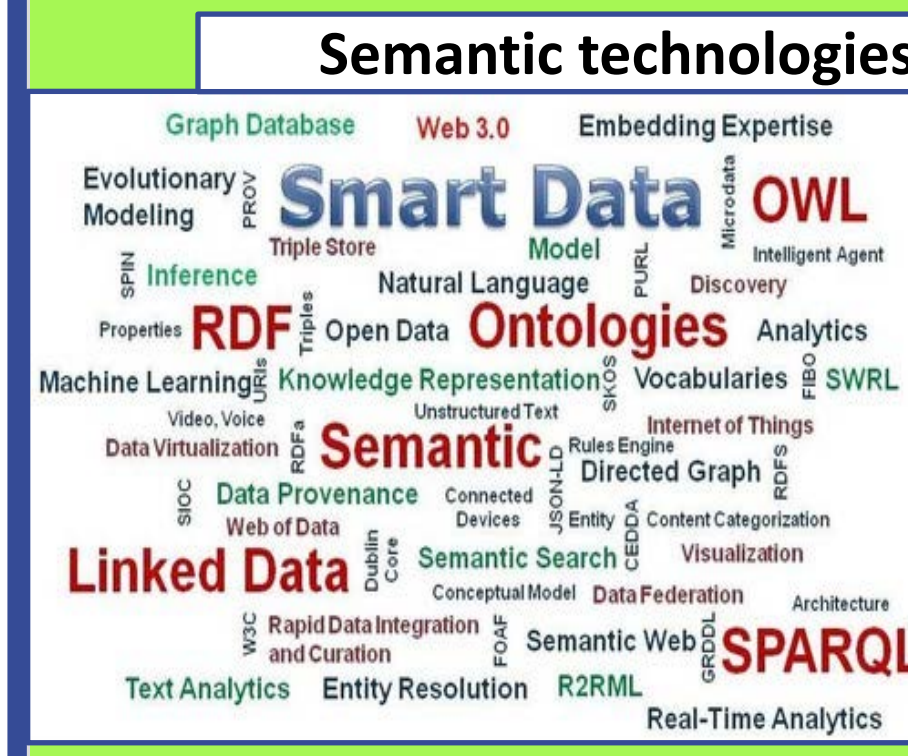
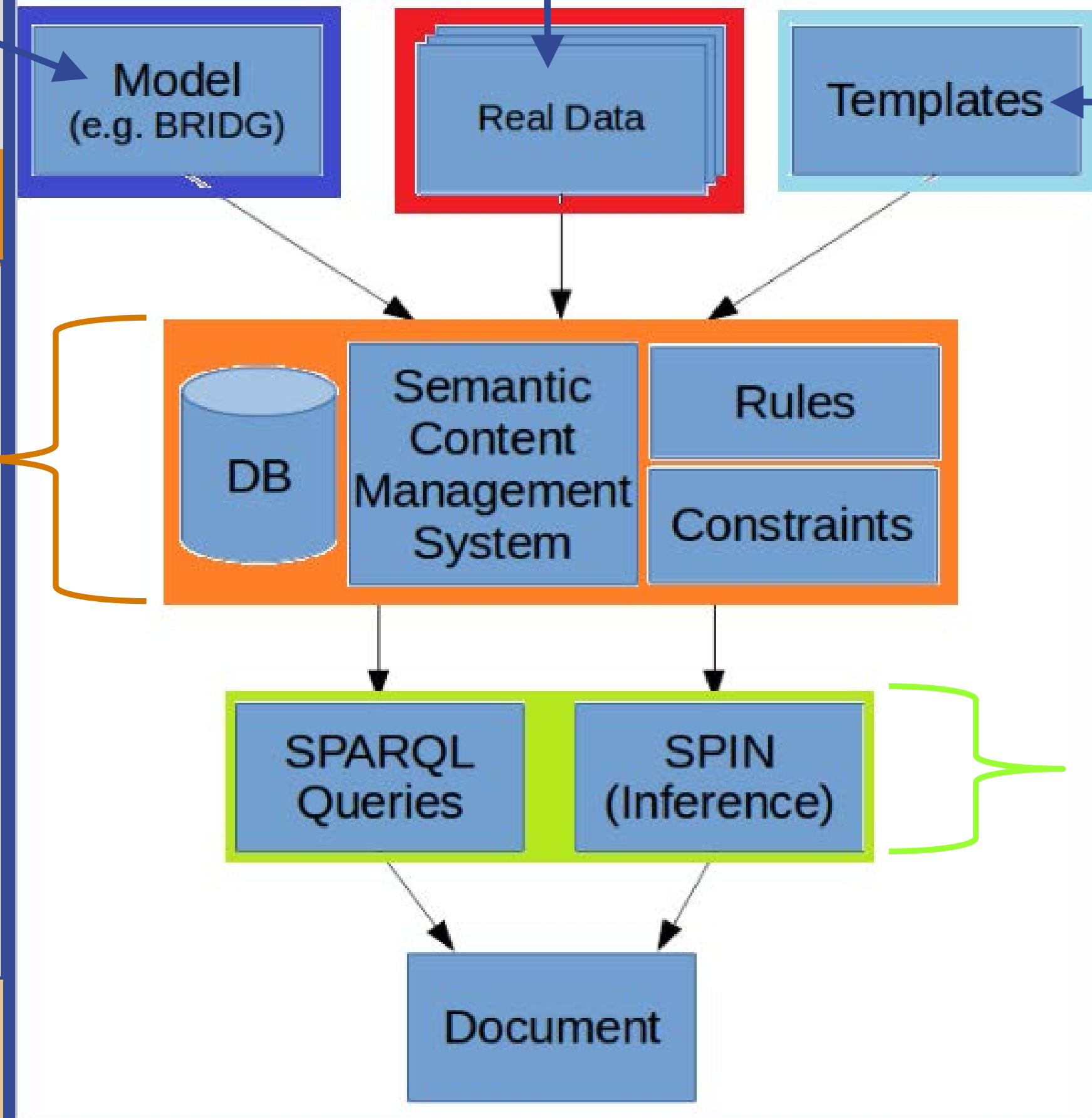
A **Content Management System** is typically used to store and index documents that are used in business processes. The addition of appropriate **metadata** is indispensable to enable meaningful retrieval of data and documents later. However, it is not always easy to find the right keywords, especially if the document is long and contains many themes. Even if the metadata can be envisaged as a cloud of terms where the size of the text indicates the relative importance of a keyword, it is not easy to get a computer to understand the significance of a term.

A CMS should also promote the re-use of its stored content and thus drive the goal of **standardisation**.

Many CMS treat their content documents like a **black box**. The cannot look inside the content; the only identification to inform the system and its users about what a document is all about are the associated **metadata** – and these often need to be added by hand.

So in the example above, such metadata could take the form of keywords such as **acronym** or **PASCAL**. But it would not be obvious that PASCAL is the value of a field called acronym in a class called studyProtocolVersion.

To really add value so that the CMS can also help to **generate** new content, the system needs to „know“ something about the content – it needs to recognize the relationship: `studyProtocolVersion.acronym = "PASCAL"`.



The system can perform **logical inference** to deduce facts from other facts that have already been specified (e.g. by using rules specified using SPIN).

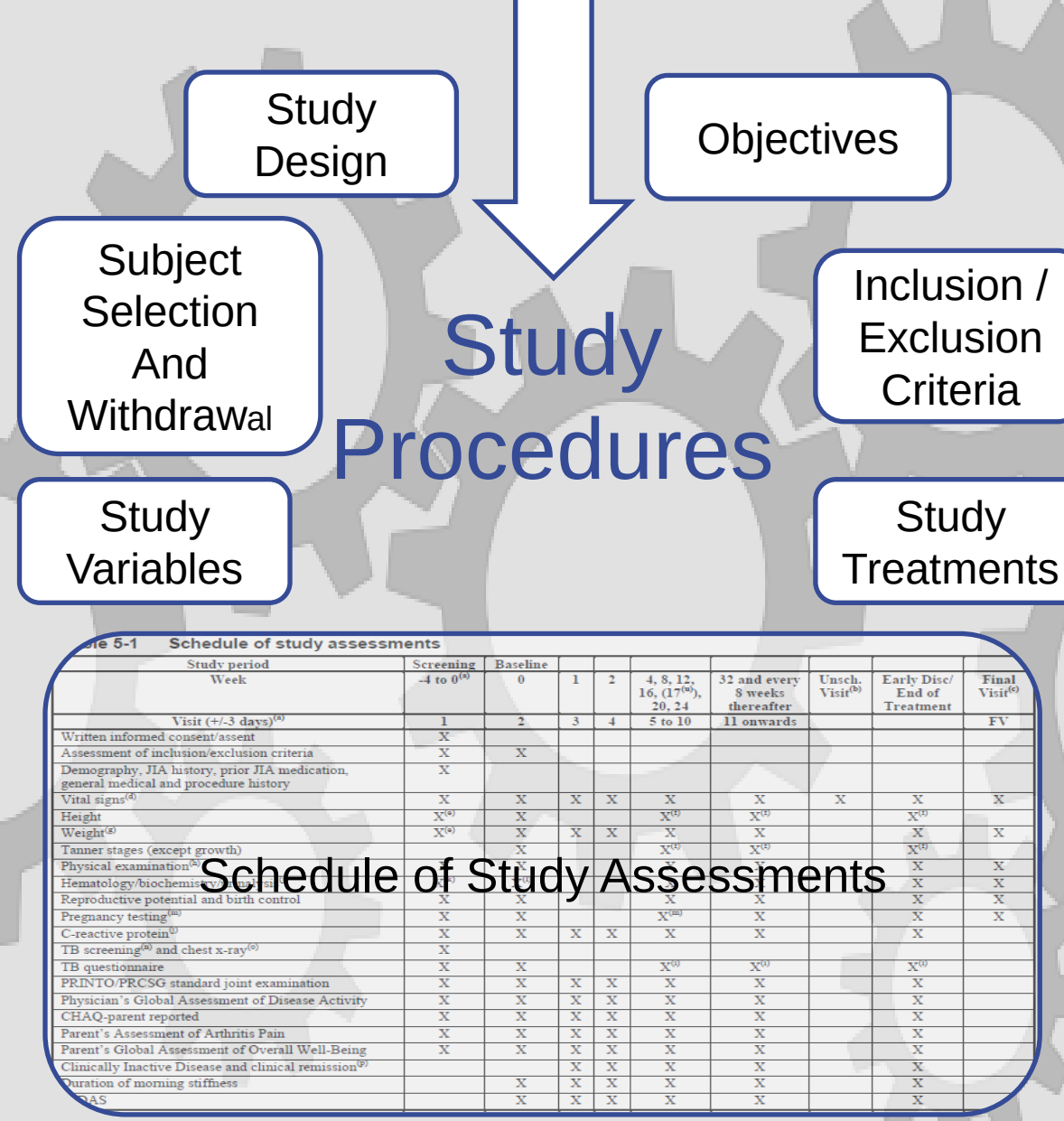
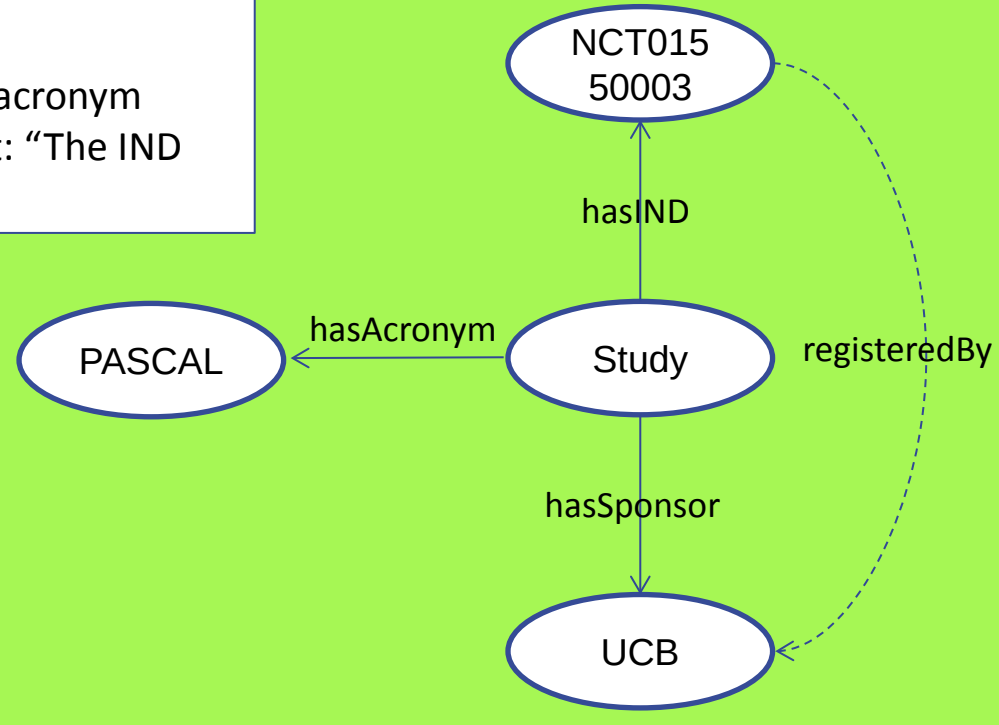
So if the system "knows" that the study with IND NCT01550003 is known by the acronym PASCAL, and that the study has UCB as a sponsor, it could then infer the new fact: "The IND NCT01550003 was registered by UCB" without this having to be stated explicitly.

TopBraid Composer
Standard Edition

PREFIX model:<http://semweb.ucb.com/CdK_Protocol#>
PREFIX uml:<<http://schema.omg.org/spec/UML/2.1.2#>>

```
SELECT ?cls1bl ?attr1bl
WHERE {
  ?cls rdf:type uml:Class .
  ?attr rdf:type uml:Property .
  ?attr model:attributeOf ?cls .
  ?cls rdfs:label ?cls1bl .
  ?attr rdfs:label ?attr1bl ;
}
ORDER BY ?cls1bl ?attr1bl
```

Inferences	References	Problems
cls1bl	attr1bl	
Activity	comment	
Activity	identifier	
Activity	reasonCode	
AdverseEvent	categoryCode	
AdverseEvent	endRelativeToReferenceCode	
AdverseEvent	expectedIndicator	
AdverseEvent	gradeCode	
AdverseEvent	highlightedIndicator	
AdverseEvent	hospitalizationRequiredIndicator	
AdverseEvent	locationDescription	



The final product of the process is a new study protocol document that conforms to the **template** and with a large part of the contents already pre-filled based on the model and on real study data.

