Relations between Ontologies and other Knowledge Structures:
Two Case Studies

CrEDIBLE Workshop
October, 16, 2012

Presented by Jean Charlet
Thanks to collaboration of Rémy Choquet and Ferdinand Dhombres

Plan

- Ontologies in DebugIT
- Ontology in OrphaOnto

- Ontologies in DebugIT

- Nota: I Will discuss about ontologies and ontologies metrics in the project and Remy Choquet will be present the (complex!)matching mechanism to query and answer the CDRs

The debugIT Project in short

- Funded by the European Community's Seventh Framework Program under grant agreement n° FP7–217139 (7M€)
- Project period: from Jan 1st, 2008 to December 31st, 2011 – extended until mid-2012.
- 14 Partners (next slide)

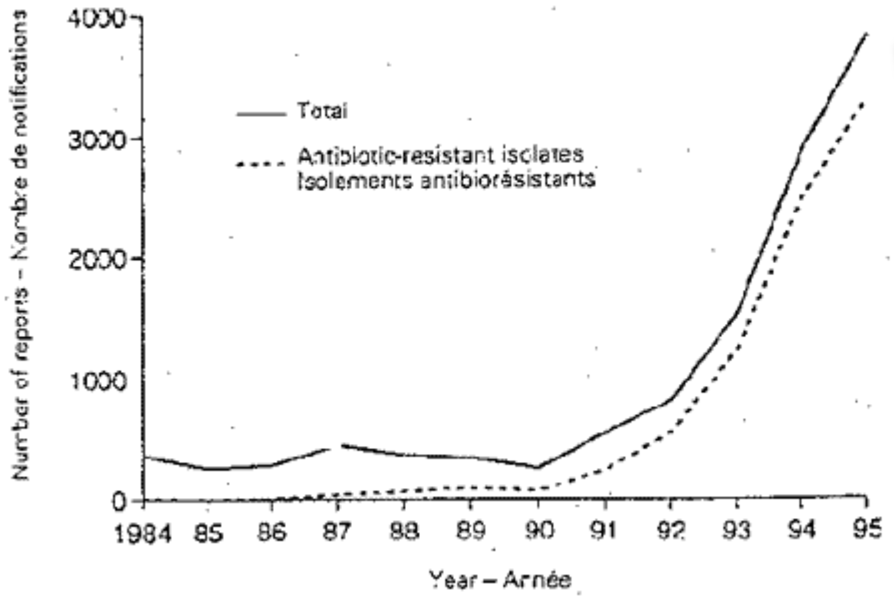
Disclaimer: this presentation reflects solely the views of the DebugIT team. The European Commission, Directorate General Information Society and Media, Brussels is not liable for any use that may be made of the information contained therein

The Partners

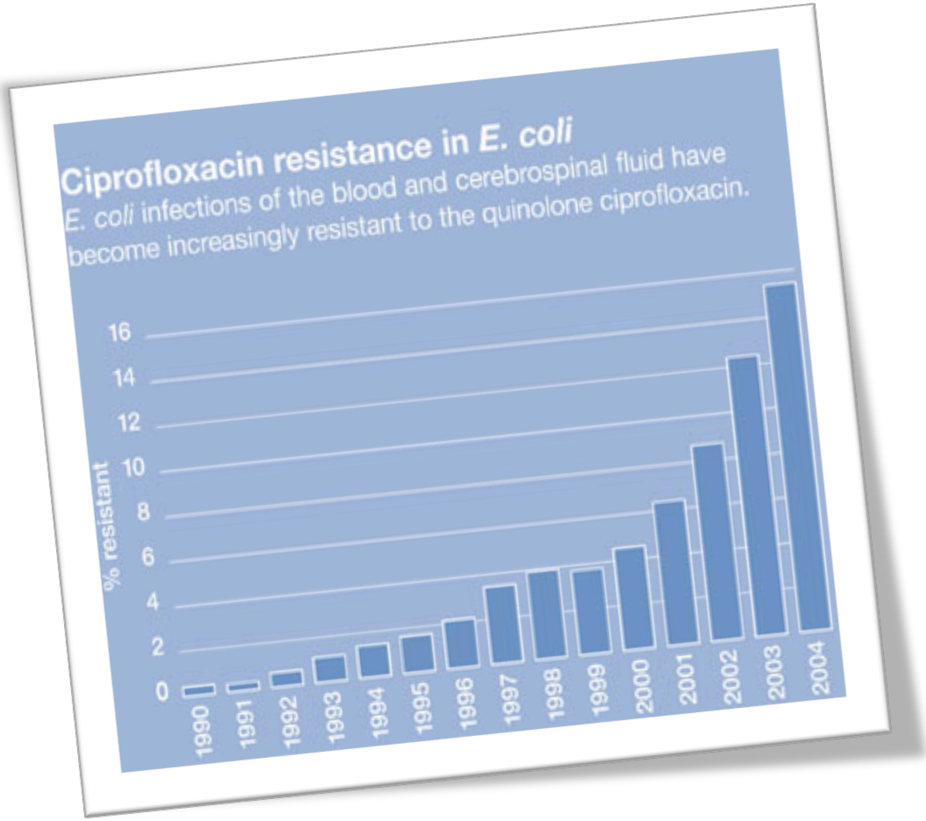
- Agfa HealthCare, Belgium (coordinator)
- empirica Gesellschaft für Kommunikations- und Technologieforschung mbH, Germany
- Gama Sofia Ltd., Bulgaria
- Institut National de la Santé et de la Recherche Médicale, France
- Internetový Pristup Ke Zdravotním Informacím Pacienta (IZIP), Czech Republic
- Linköpings Universitetet, Sweden
- Technologiko Expedeftiko Idrima Lamias, Greece
- University College London, United Kingdom
- Les Hôpitaux Universitaires de Genève, Switzerland
- Universitätsklinikum Freiburg, Germany
- Université de Genève, Switzerland
- Averbis, Freiburg, Germany
- MDA, Czech Republic
- HEG, Geneva, Switzerland

The Problem

antibiotic resistance in Salmonella typhimurium DT104, England and Wales, 1984-1995



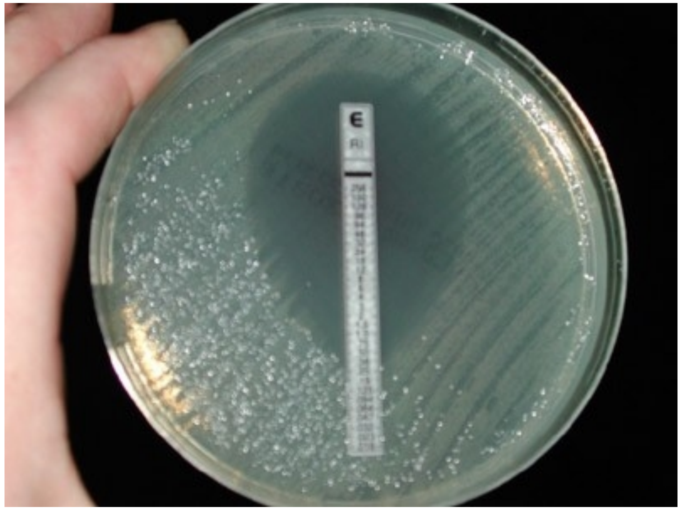
WHO Weekly Epidemiological Record, Vol 71, No 18, 1996



The debugIT Response

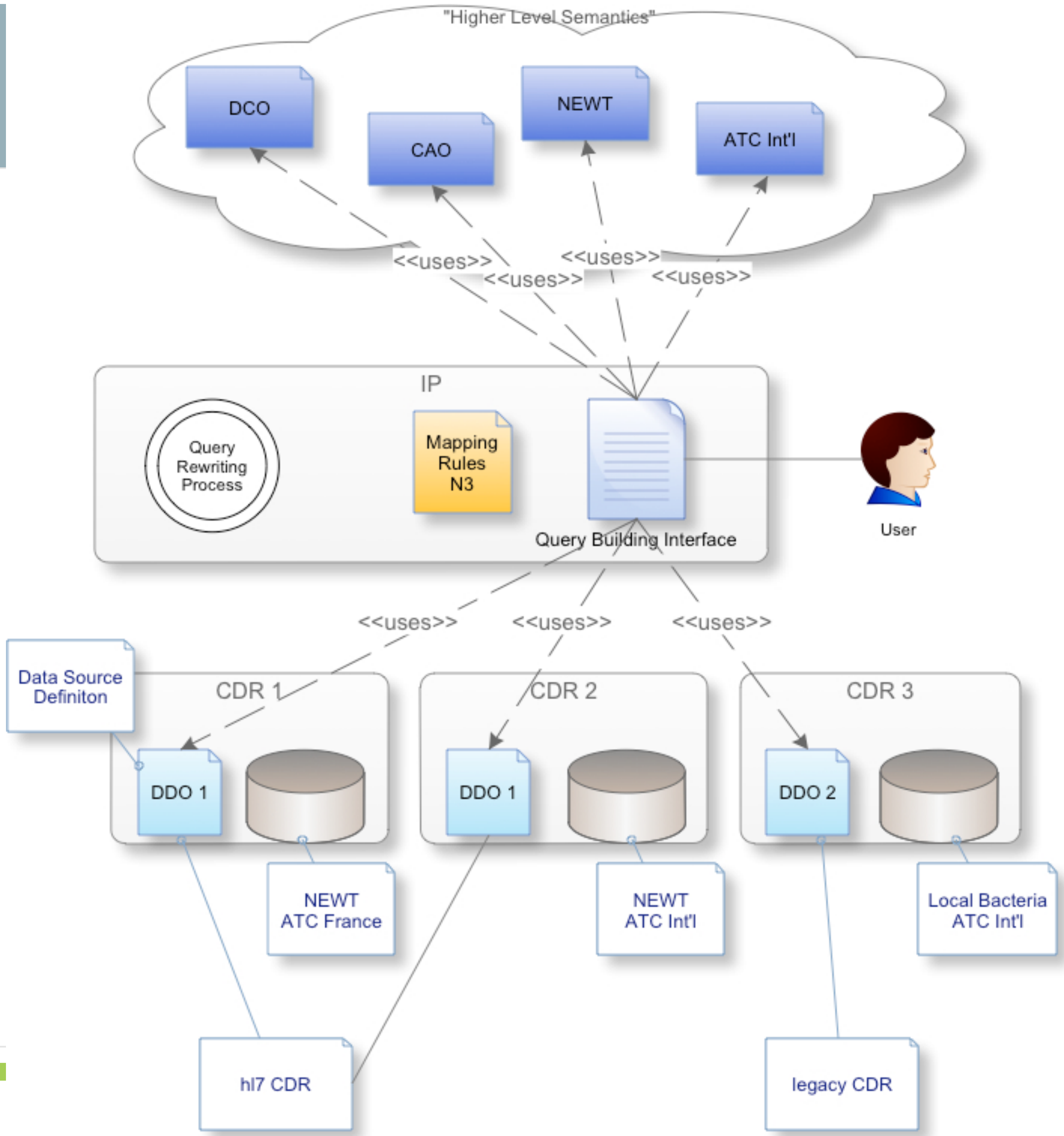
- If new antibiotics can not keep up with the bacterial resistance and a race against evolution can only be lost,
  → we need new solutions

- If this is a war and current weapons don't work anymore,
  → we need a new weapon
  → **ITbiotics** to help antibiotics



Objectives

- Objectives of Ontologies in DebugIT
  - Provide formal computer-interpretable meaning
    - exploitable by logics & rule-based reasoners
  - Enable SPARQL queries and mapping rules
    - To express research questions on different abstraction levels
    - To align CDRs via different ontology layers
    - Allow for cross-site data integration & comparison
- “Could you give me antibiograms to Escherichia coli tests where we found a resistance to Beta lactam antibiotics ?”


"Higher Level Semantics"
DCO
CAO
NEWT
ATC Int'l
<<uses>>
<<uses>>
<<uses>>
<<uses>>
IP
Query Rewriting Process
Mapping Rules N3
Query Building Interface
User
<<uses>>
<<uses>>
<<uses>>
Data Source Definiton
CDR 1
CDR 2
CDR 3
DDO 1
DDO 1
DDO 2
NEWT ATC France
NEWT ATC Int'l
Local Bacteria ATC Int'l
hl7 CDR
legacy CDR

All DebugIT ontologies

Ontology name	Acronym	Ontology Type	Content
BioTop	**BioTop**	upper ontology	Biology domain top level connecting the DCO to top level ontologies like BFO or DOLCE
DebugIT Core Ontology	**DCO**	core ontology	Core health care domain including human infectious diseases, their analysis and therapies
Medical Evidence Ontology	**MEO**	operational ontology	Types of medical evidence as described in SIGN 50 guidelines[i]
Clinical Analysis Ontology	**CAO**	operational ontology	Derived and concise formal representation of clinical domain ontologies, permitting concise expression, easy query building by a clinician, and N3 rules formulation for data mining
Analysis Ontology	**AO**	operational ontology	Derived and concise formal representation of non-clinical domain ontologies, permitting concise expression, easy query building by a clinician and data miner, and N3 rules formulation for data mining
Clinical and Biological SKOS Schemes Ontologies	**CSSO** **BSSO**	operational ontologies	Instances permitting mapping of clinical terminology/coding systems and ontologies
Quantities and Units Extension Ontologies	**QEO** **UEO**	operational ontologies	formal description of quantities and units, elaborating on work done by NASA in their SWEET ontology series[ii]
Decision Support Ontology **Document Ontology** **SPARQL Ontology** **SPARQL Analysis Ontology**	**DSO** **DO** **SO** **SAO**	operational ontologies	series of ontologies used in the document life cycle of different DebugIT services
Workflow Ontology	**WO**	operational ontology	formal description of workflow
Data Definition Ontologies for all sites	**DDO**	data definition ontology	formal representation of clinical database schemes

[i] http://www.sign.ac.uk/pdf/sign50.pdf
[ii] http://sweet.jpl.nasa.gov/2.0

Ontology Layers

7 Data Definition Ontologie (DDO) **average 40 Entities**

- **Mediation layer** closing the ‘formality’-gap
- Describing **site-specific local** CIS Data models in RDF
- For SPARQL data access to local hospital data

13 Operational ontologies (OO, e.g CAO) **average 35 Entities**

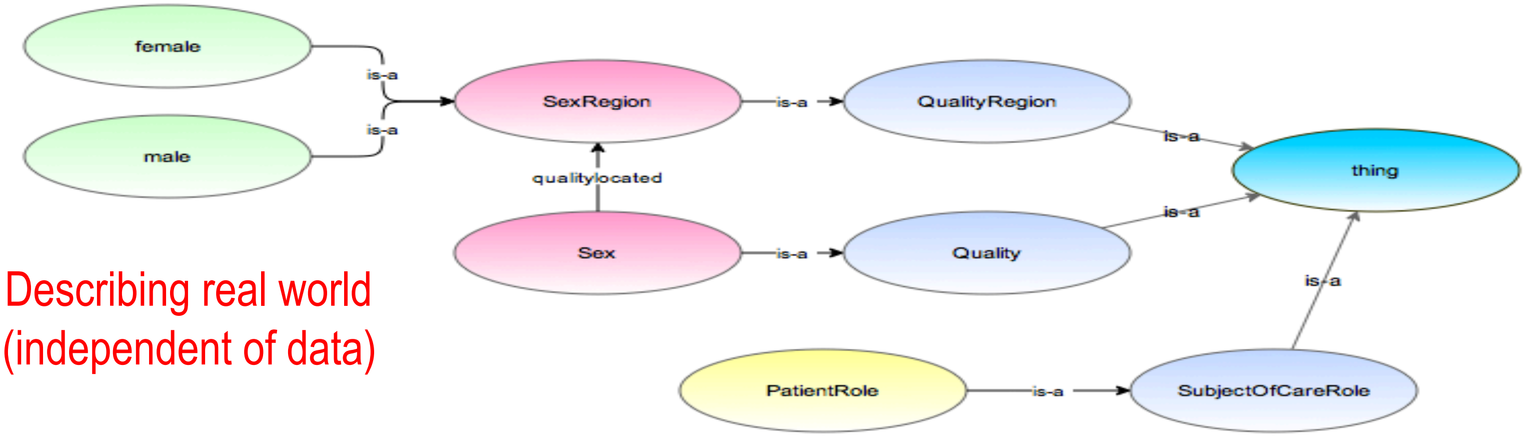
- **Mediation & Integration layer**
- Implementation, module crosstalk, data mining analysis, query building, statistics, evidences, maths, units, …
- OWL-Full →Coherent Logic reasoning (e.g. rule-based)

1 DebugIT Core Ontology (DCO) **~ 1720 Entities**

- **Integration layer**, mapped to DDOs & external Terminologies
- Rooted in Biotop upper level ontology
- **Global,** clinical domain of infectious diseases
- OWL-DL →DL & Coherent Logic reasoning

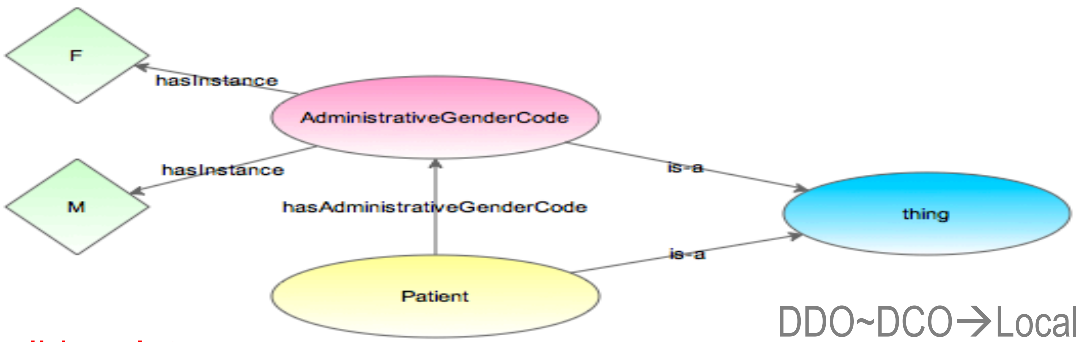
„Female patient' in different ontology layers

DCO



Describing real world (independent of data)

DDO



Describing data

DDO~DCO→Local to Global rules to later create SPARQL CONSTRUCT/WHERE clauses

E/R

Patient	
ID	FKREFADMINISTRATIVEGENDERCODE
23	F
45	M
58	M

DCO design principles

- OWL-DL
  - Reasoner for autoclassification & consistency checks during OE
  - Reasoner infers multiple parenthood
- Reusing BioTop
  - Ensure a rigid modeling view
  - Provides reuseable constraints
- Concepts harvested from
  - Hospital CDR schemata
  - Competency questions from clinical use case
    - Datadriven *bottom up*
  - Domain terminologies in use
    - Via UMLS
    - Ontology modularisation tools (A.Rector)
    - HL7 v3 based

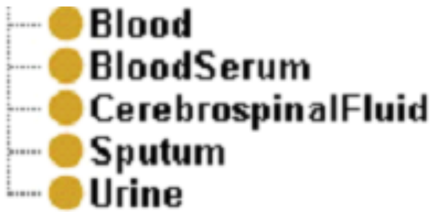
Inference of new facts
(BloodSample is a BodyLiquidSample)

Stated Facts

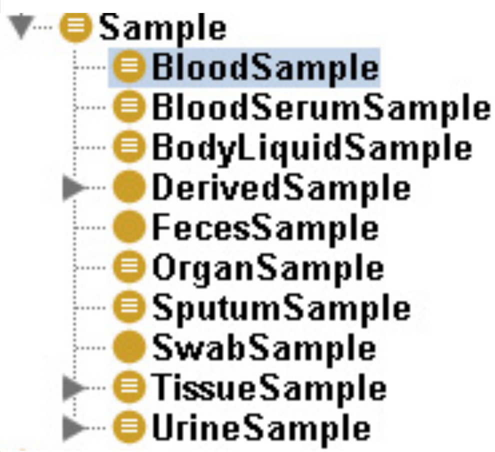
BodyLiquidSample =

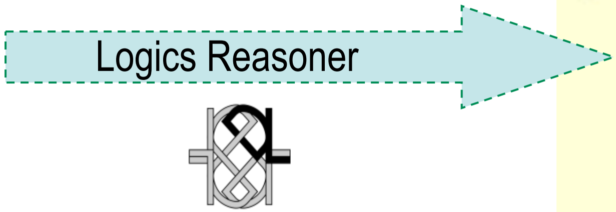
BloodSample =

BodyLiquid =

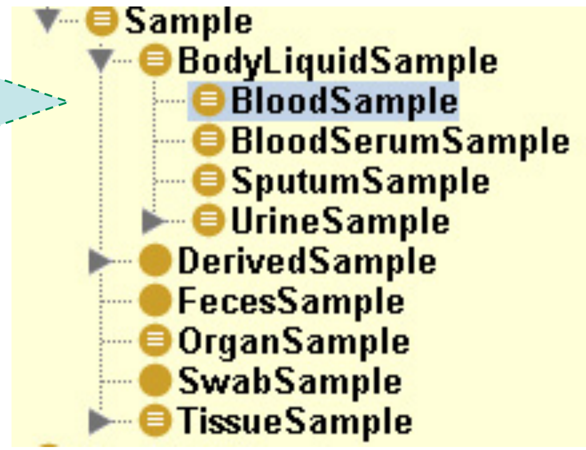


Asserted Hierarchy (flat list)





Inferred Hierarchy (more structure)



DCO/Biotop Metrics (April 2011)

Ontology elements and axioms	Count (all)	DCO	BioTop
Classes	1732	1445	375
Object Properties (relations)	85	43	76
Datatype Properties	12	12	1
Subclass Axioms	2022	1578	454
Equivalent Class Axioms	212	117	99
Disjoint Axioms	77	2	75

OO Metrics (March 2011)

Name	Classes	Instances	Properties		Total	Property axioms			
			ObjectP	DataP		Domain	Range	Inverse	Chain
CAO	61	-	10	-	71	10	10	1	-
AO	25	19	34	5	82	1	1	-	2
CSSO	-	24	-	2	26	-	-	-	2
BSSO	-	2	-	-	2	-	-	-	-
DSO	9	-	6	-	17	5	5	2	-
DO	15	1	18	7	40	25	25	2	2
SO	7	-	4	-	11	4	4	-	-
SAO	12	-	4	-	16	4	4	2	-
SATO	3	-	-	-	3	-	-	-	-
WO	12	-	-	-	12	-	-	-	-
QEO	37	3	29	2	71	31	31	8	4
UEO	-	87	-	-	87	-	-	-	-
Total	**181**	**136**	**105**	**16**	**438**	**80**	**80**	**15**	**10**

Ontology Layers in their IP query context

Analysis step	Role	Ontology
1. Clinical Question	physician	(NL)
2. Clinical Analysis Query (CAQ)	clinical researcher	OO (CAO), DCO
3. Data Set Query (DSQ) per CDR	data manager	DDO
4. Conversion of DDO to DCO (define N3 mapping rules)	data manager	DDO/DCO
5. Apply N3 rules on data sets	data miner	DCO, OO
6. Result in CAQ CONSTRUCT	data miner	OO, DCO
7. NL Answer	clinical researcher	(NL or CNL)

Two parallel data-to-instance conversion approaches

- Local/unformal to global/formal data conversion
- Two parallel bottom up formalization approaches
- Formalization approach chosen depending on CIS datatype
  - **Rule-driven Formalization**
    - For **Freetext Data**, we exploit only
      manually generated DTB to DDO (D2R)
      and local2global DDO to DCO mapping rules ⇠ implies creation of DDO2DCO mappings at development time
  - **Terminology-driven Formalization**
    - For **Terminologies/codes** (e.g. ATC, ICD-10, NewT), we also use a chain of SKOS Terminology mappings, e.g.
      - ICD10 to SNOMEDCT
      - SNOMEDCT to DCO ⇠ implies creation of DCO2SNOMEDCT mappings at development time

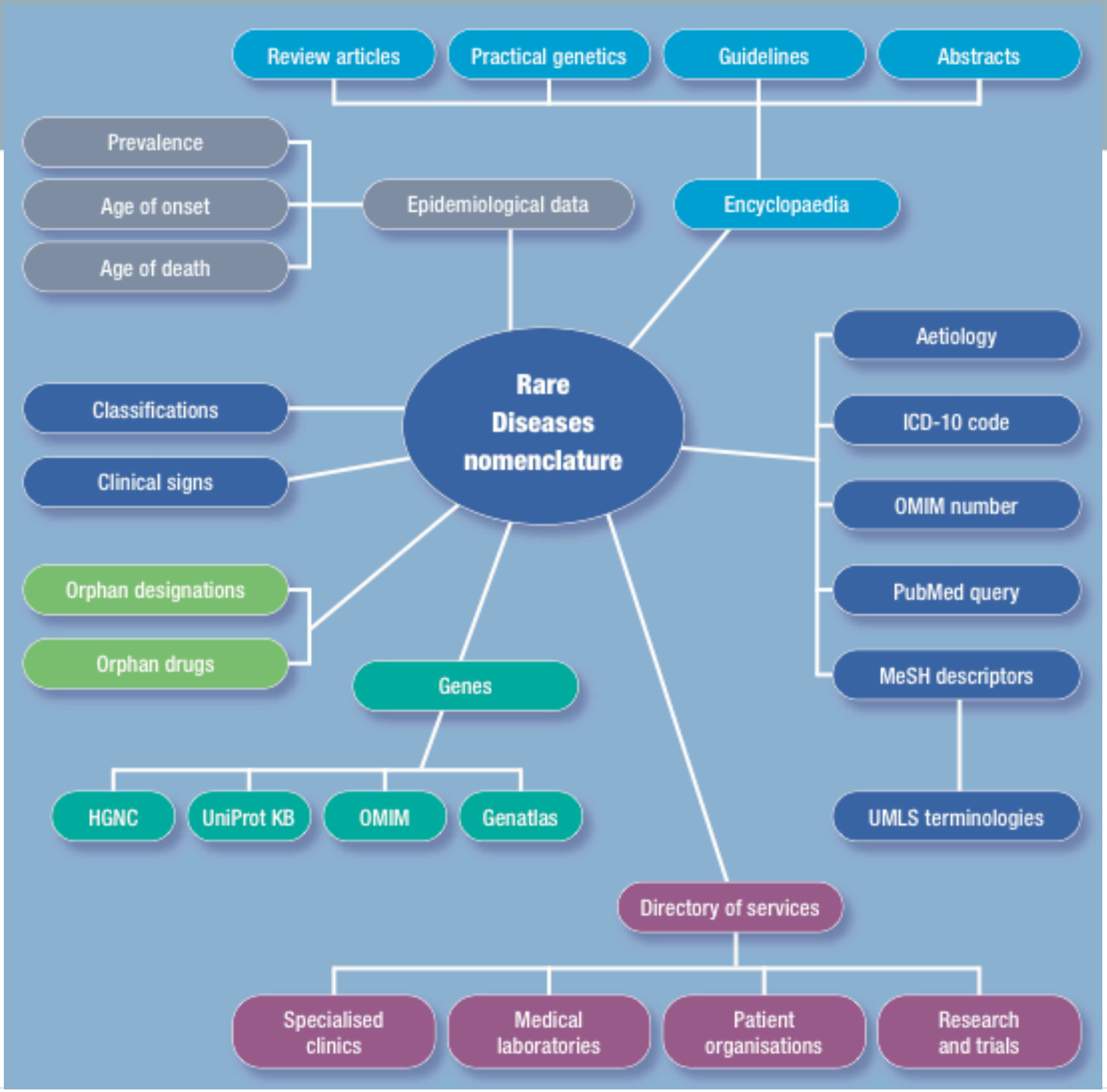
Conclusion about DebugIT ontologies…

… after presentation of

Rare diseases ontology

- Ontology in OrphaOnto

http://www.orpha.net



Review articles
Practical genetics
Guidelines
Abstracts
Prevalence
Age of onset
Age of death
Epidemiological data
Encyclopaedia
Aetiology
Classifications
Rare Diseases nomenclature
ICD-10 code
Clinical signs
OMIM number
Orphan designations
PubMed query
Orphan drugs
Genes
MeSH descriptors
HGNC
UniProt KB
OMIM
Genatlas
UMLS terminologies
Directory of services
Specialised clinics
Medical laboratories
Patient organisations
Research and trials

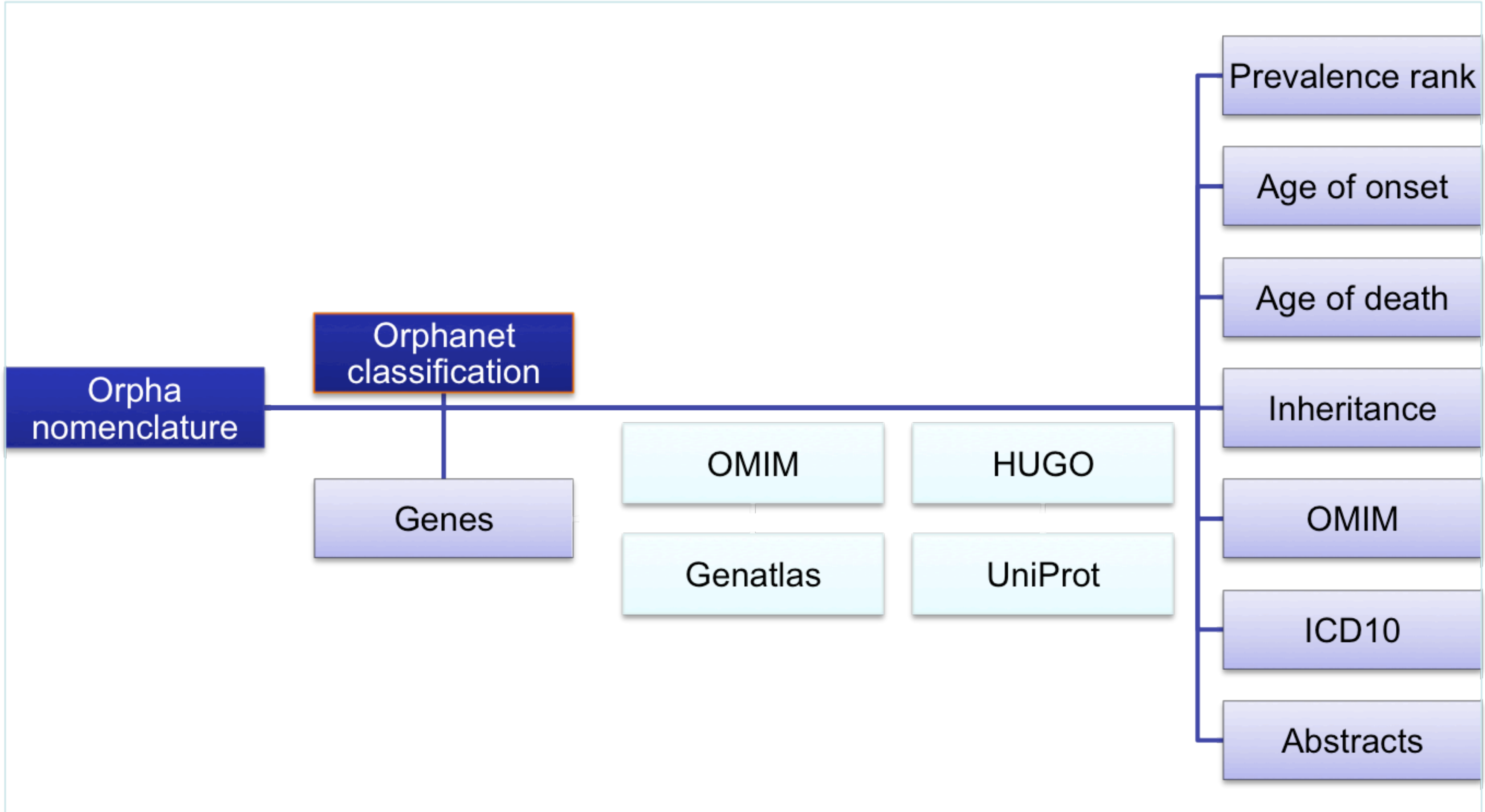
View classifications by disease or by group of diseases

- Rare metabolic disease
  - Metabolic disease involving complex molecules
    - Peroxisomal disease
      - **Adrenoleukodystrophy, X-linked**
        - Adrenoleukodystrophy, X-linked, cerebral form
        - Adrenomyeloneuropathy

- Rare neurologic disease
  - Neurometabolic disease
    - **Adrenoleukodystrophy, X-linked**
      - Adrenoleukodystrophy, X-linked, cerebral form
      - Adrenomyeloneuropathy
- Rare neurologic disease
  - Rare epilepsy
    - Metabolic diseases with epilepsy
      - Peroxisomal disease
        - **Adrenoleukodystrophy, X-linked**
          - Adrenoleukodystrophy, X-linked, cerebral form
          - Adrenomyeloneuropathy
- Rare neurologic disease
  - Leukodystrophy
    - **Adrenoleukodystrophy, X-linked**
      - Adrenoleukodystrophy, X-linked, cerebral form
      - Adrenomyeloneuropathy

- Rare endocrine disease
  - Rare adrenal disease
    - Primary adrenal insufficiency
      - Chronic primary adrenal insufficiency
        - Genetic chronic primary adrenal insufficiency
          - **Adrenoleukodystrophy, X-linked**
            - Adrenoleukodystrophy, X-linked, cerebral form

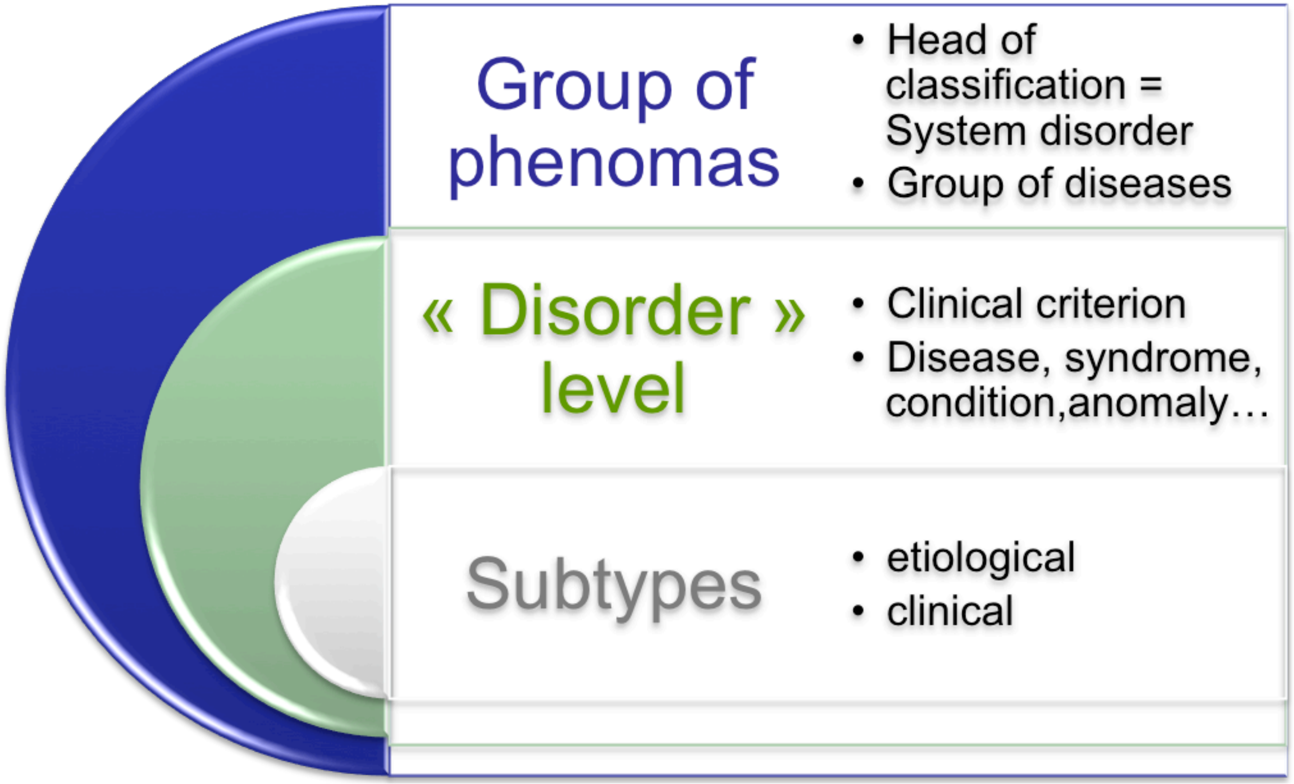
Current RD database Content



Need for improvement of RD database (1)

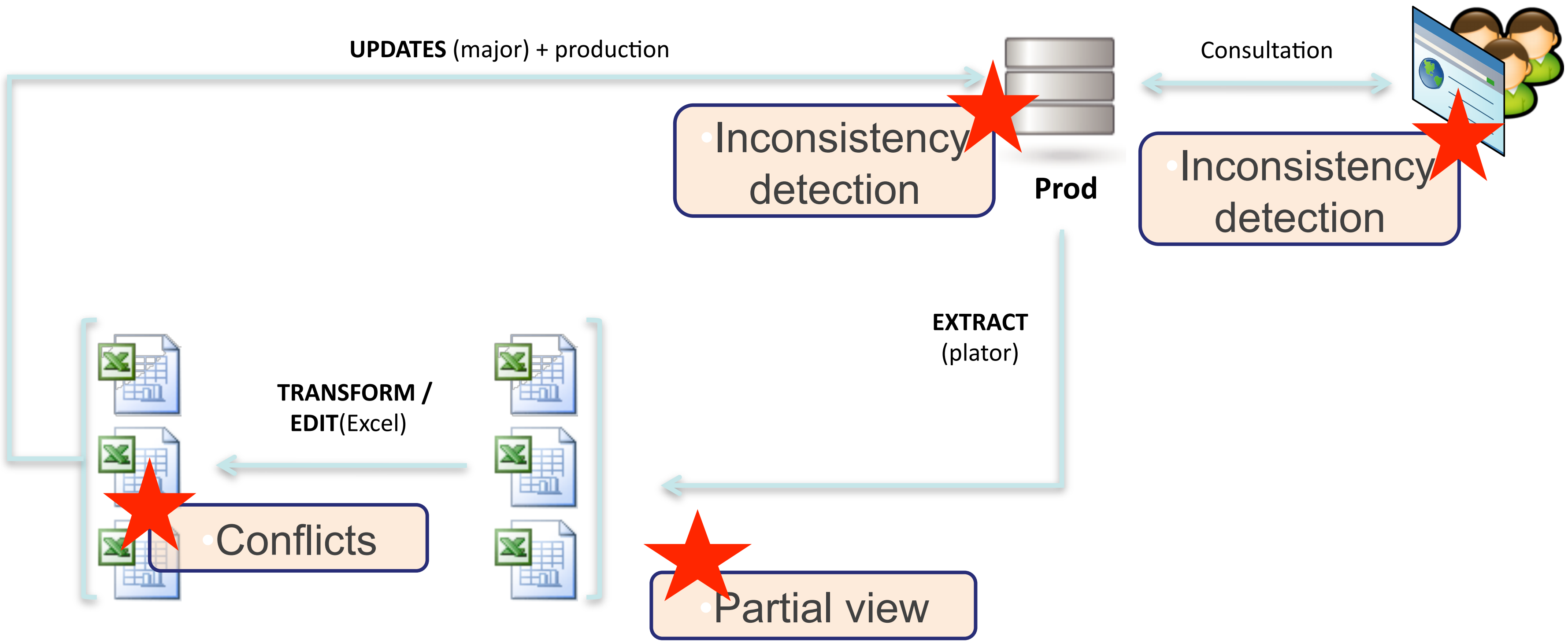
- Identifying relationships between entries in a classification
  - « IsA »
  - « IsCauseOf »
  - « Has(very frequently/frequently/ occasionnally)CinicalSign »
  - « IsAPredisposingGeneFor »
  - …

Need for improvement of RD database (2)



– To better represent the continuum of phenomas
– To better represent the relationships between phenomas and genes-to-disorders

Need for improvement of workflow



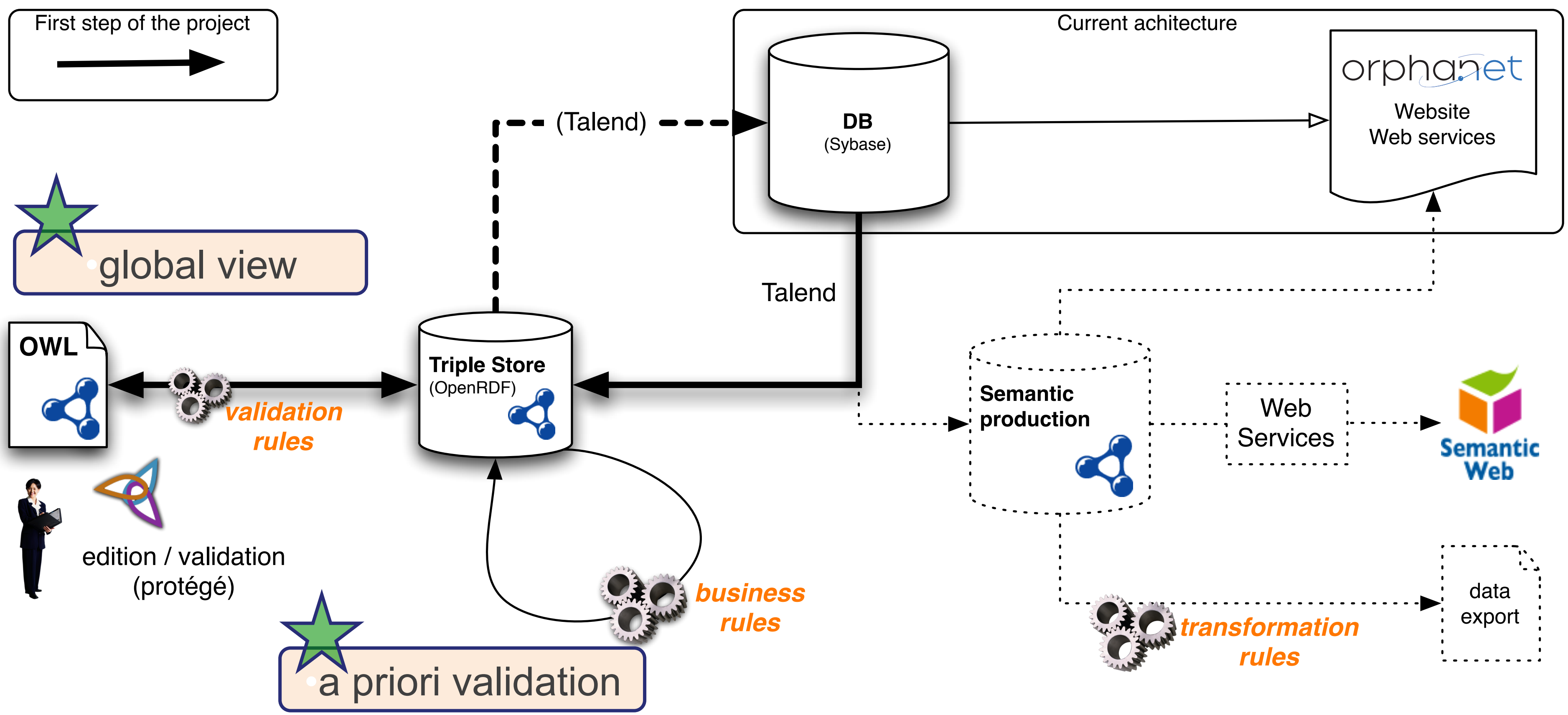
Excel is unappropriate for both knowledge vizualisation and edition
Conflicts caused by knowledge pieces spread in few files
Knowledge Model = DB schema
a posteriori validation procedure

Objectives of OrphaOnto Project

Moving to an Orphanet ontology in a two steps process

- Ontological view of the current RD database in order to:
  - improve annotations
  - allow quality control (i.e. detection of inconsistencies)
  - achieve generalization of « IsA » relationships across the classification, if possible
  - adopt Protégé as an edition tool for RD database update
- Build and edit the Orphanet ontology
  - besides the relational database but connected to it
  - the ontology will feed the database,
  - will be freely available,
  - and will allow serving new needs

Future workflow



Tools dedicated to knowledge vizualisation and edition
Knowledge model = OWL graph

Orphanet: a Classifications System

- Orphanet is a classification system
- 32 classifications of rare diseases
- Semantic representation of classifications
  - Turning hierarchical relationships into subsumption relationships seems inapropriate (inheritance issues)
  - Approach 1 : coloured graph

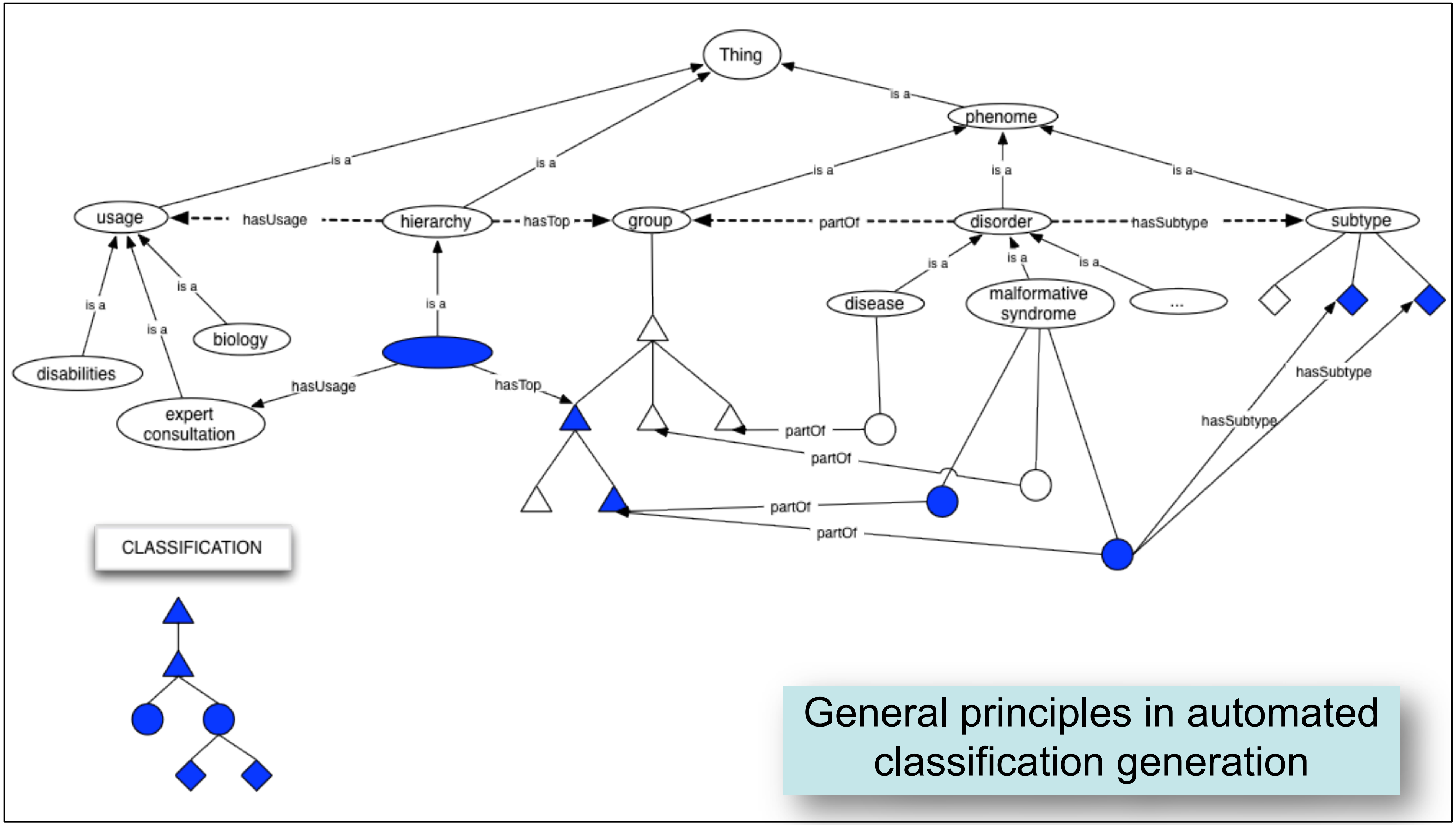
    A classification (links between diseases) is a set of “hierarchical relationships” between its elements. Then browsing diseases in a specific classification is browsing the diseases graph using ONLY one relationship
  - Approach 2 : Rule-based hierarchy generation from a core ontology

Ongoing Core-ontology

- Multidisciplinary approach to define the core-ontology of rare diseases (aka the meta-model of the domain)
  - Expert of rare diseases (MD and researchers)
  - Computer engineers
  - Knowledge engineers

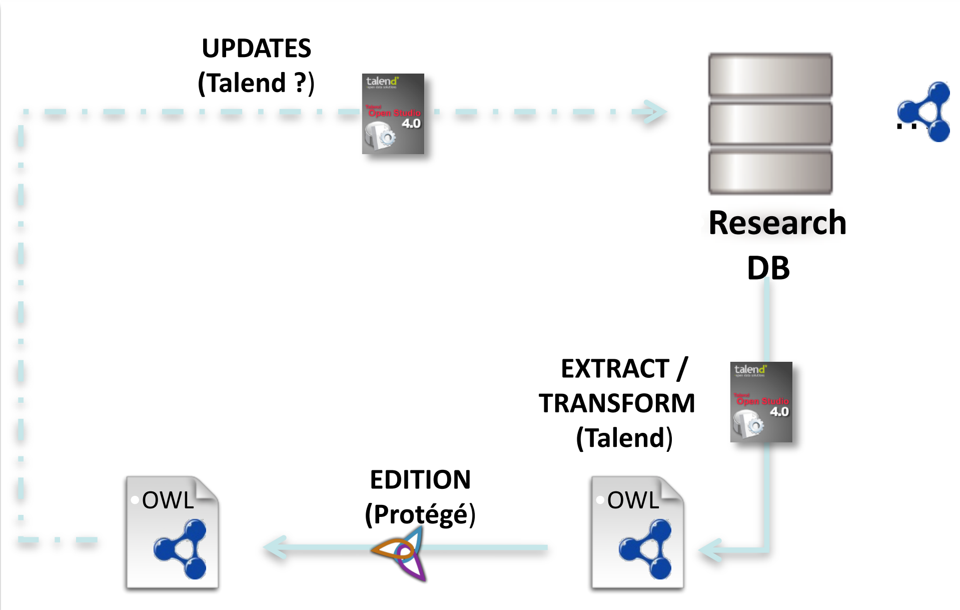
- Must comply with :
  - Expert representation of the domain
  - Appropriate rules expression support
  - Description logic

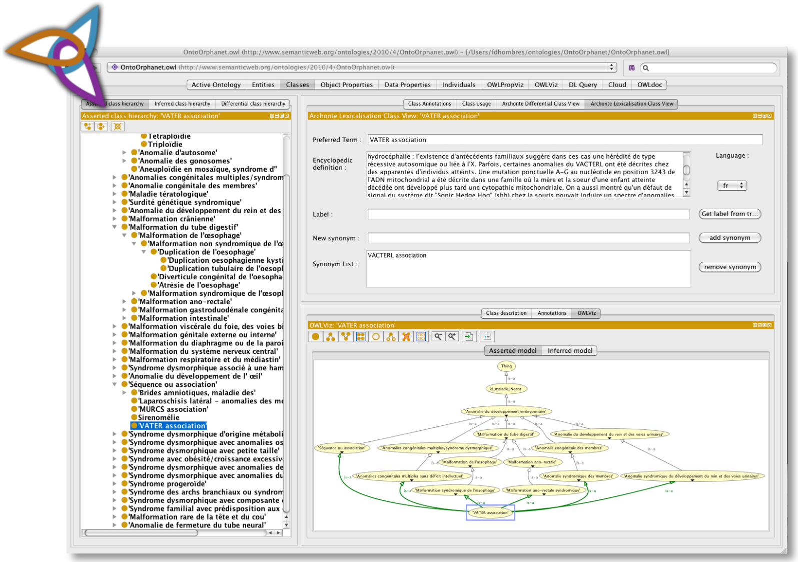
Ontological commitment :
Automated classification generation

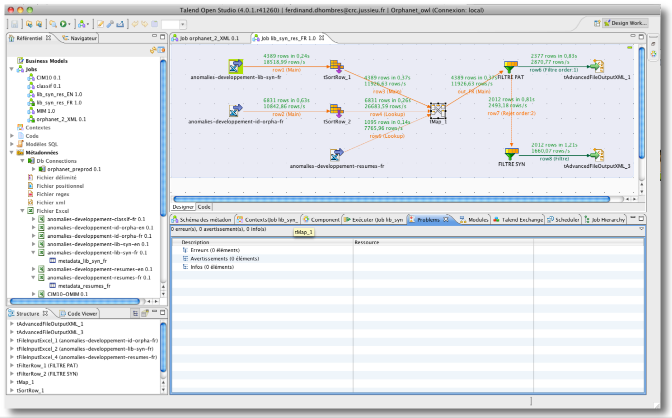


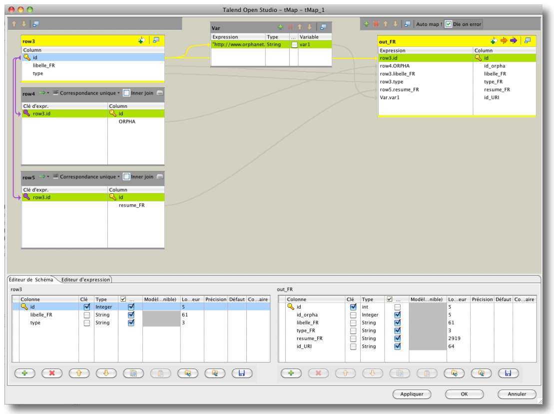
General principles in automated classification generation

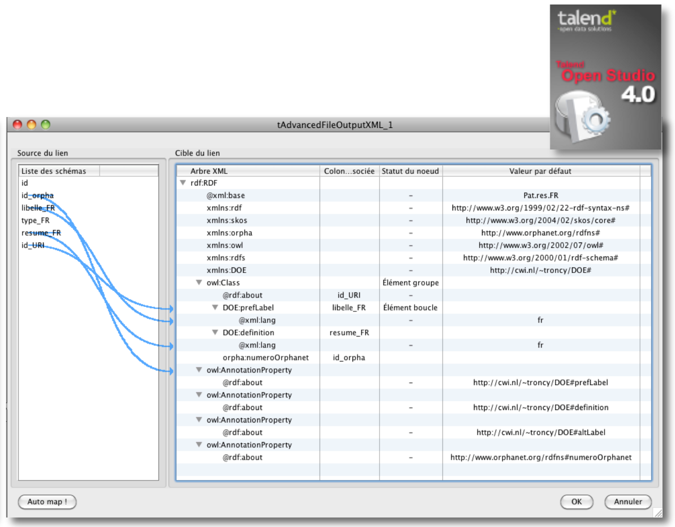
The tools : Talend and Protégé



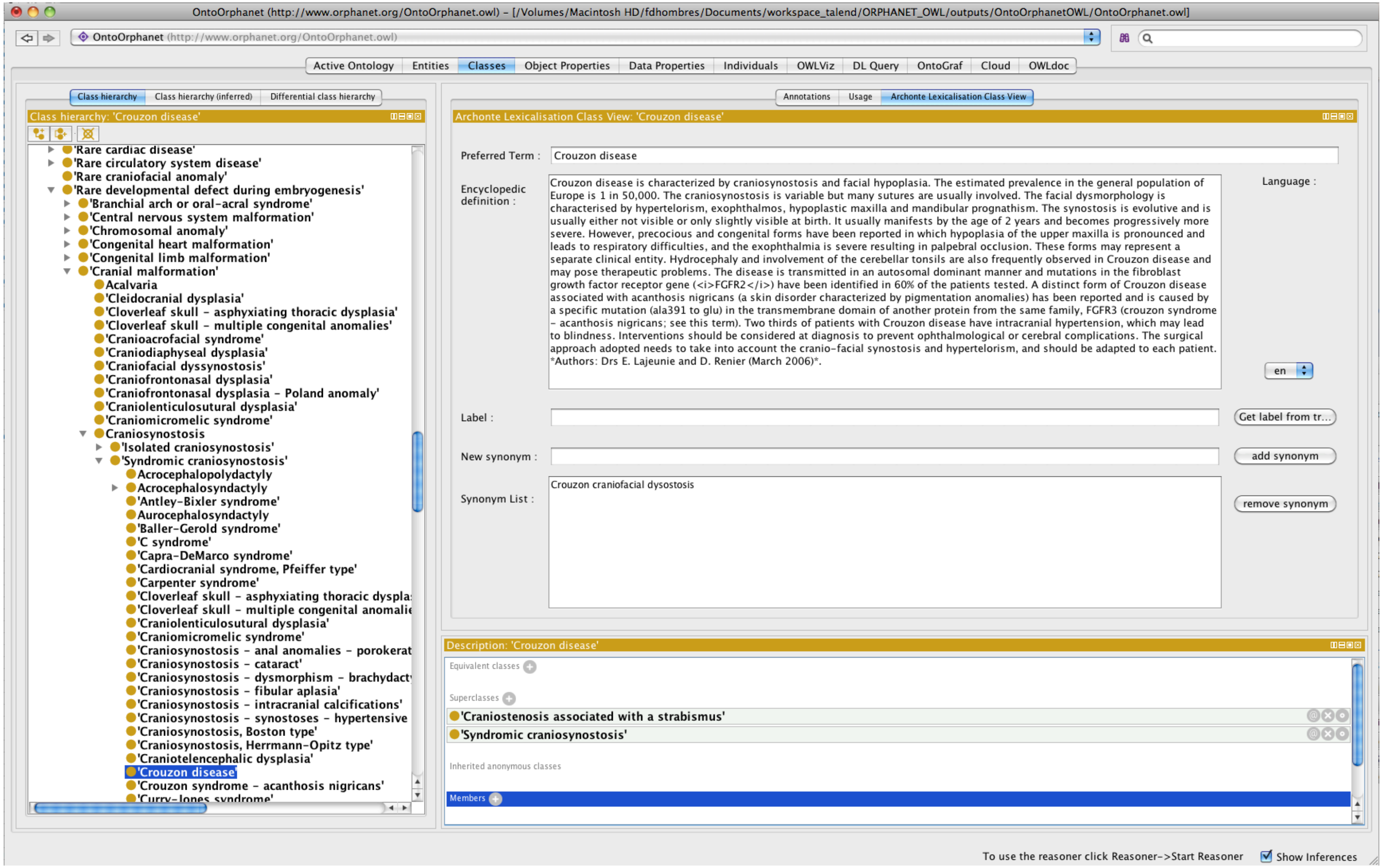




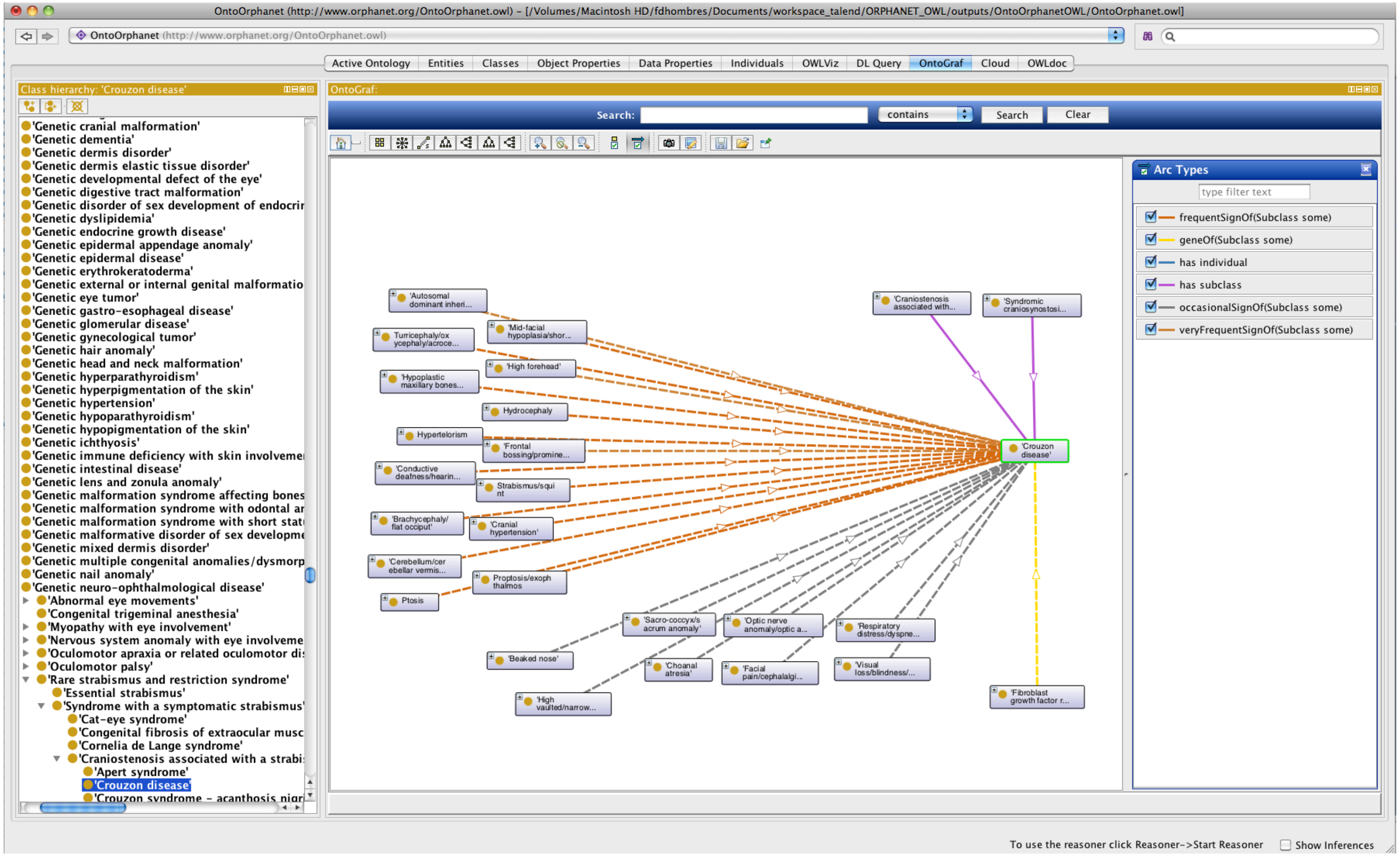




Protegé / dedicated plugin



Protegé / Vizualisation tools



Sesame OpenRDF / quality control

SPARQL QUERIES (N3)

Audit of data
Annotations monitoring
Quality control procedures
(Rule-based procedures)

0	maladie(s) sans libellé
1	mode de transmission non lié à une maladie
4	signe(s) non liés à une maladie
20	gène(s) non liés à une maladie
2 494	maladie(s) génétique(s) sans mode de transmission (sur un total de 5272 maladies génétiques rares)

		owl :Restriction	
owl :Class	11 077	orpha :occasionalSignOf	10 530
owl :ObjectProperty	10	orpha :frequentSignOf	12 384
owl :AnnotationProperty	29	orpha :veryFrequentSignOf	21 281
AnnotationAssertion	179 567	orpha :geneOf	3 819
Classes polyparentales	2 843		

	total	fr	en	de	es	it	pt
AnnotationAssertion	153 513	23 203	23 641	20 808	18 625	19 253	14 126
skos :prefLabel	42 970	7 163	7 163	7 161	7 161	7 161	7 161
skos :altLabel	27 984	5 615	5 990	4 836	4 837	4 389	2 317
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Summary

- **Use of data management environment**
  - ETL procedures with Talend
  - OWL file generated from a relational database
  - relational database updates

- **Use of ontology editor**
  - Protégé : knowledge modelling
  - Editorial work supported by home-made plug-in

- **Quality control procedures**
- **Classifications generation**
  - Sesame : triple store / SPARQL queries
  - Consistency checking / EULER

Annotations

- **Property and annotations in OWL**
  - ObjectProperty : inheritance
  - Annotations : no inheritance

- **Annotation is an appropriate choice for :**
  - Disease name (“Label”) and synonyms (“altLabel”)
  - External references (ICD-10 ref, MIM number…)
  - Epidemiological data
  - Disease definition / abstract
  - Classification belonging

Conclusion about the 2 ontologies (I/II)

- Ontologies used in production mode
  - DebugIT
    - Data integration across 6 languages over 7 EU Hospitals
    - Practical SPARQL query building
    - Result aggregation/integration in DCO
      - Enables secondary data usage over SemanticWeb
  - OrphaOnto
    - The workflow is not modified but the ontology is regularly updated and available on bioportal: http://http://bioportal.bioontology.org/ontologies/1586
- Ontology engineering & evaluation
  - DebugIT
    - DL-reasoning helps ontology engineering & evaluation
  - OrphaOnto
    - Quality control procedure by SPARQL request in a triplestore

Conclusion about the 2 ontologies (II/II)

- Model Complexity and usage
  - DebugIT
    - DDO-based data set queries and layered rule mappings from mediation to integration layer are complex

    … but approach scales over increasing number of new participants (!)
  - OrphaOnto
    - OntoOrpha is in fact a meta model from which we can produce different classifications for practitioners (genetic diseases, neurologic diseases, cardiologic diseases, …)

Next steps about the 2 ontologies

- DebugIT
  - Implement ontology release checker
    - Validate ontologies on term redundancies, required metadata, naming conventions, …
  - Generate simpler DCO subset
    - Based on checked available bindings
- OrphaOnto
  - Implement validation process in ETL program
    - Insure coherence of information stored in Orphanet DB
  - Work about signs in collaboration with HPO
  - Validate the generation model and the query against the old classifications (in the case of genetic disease classification, it's OK)

End

Thank you for your attention