Ontologies, Data Curation and Text Mining

Thérèse Vachon, Global Head of Text Mining Services

Novartis Institutes for Biomedical Research, NX

Information Retrieval and Text Mining for Biology

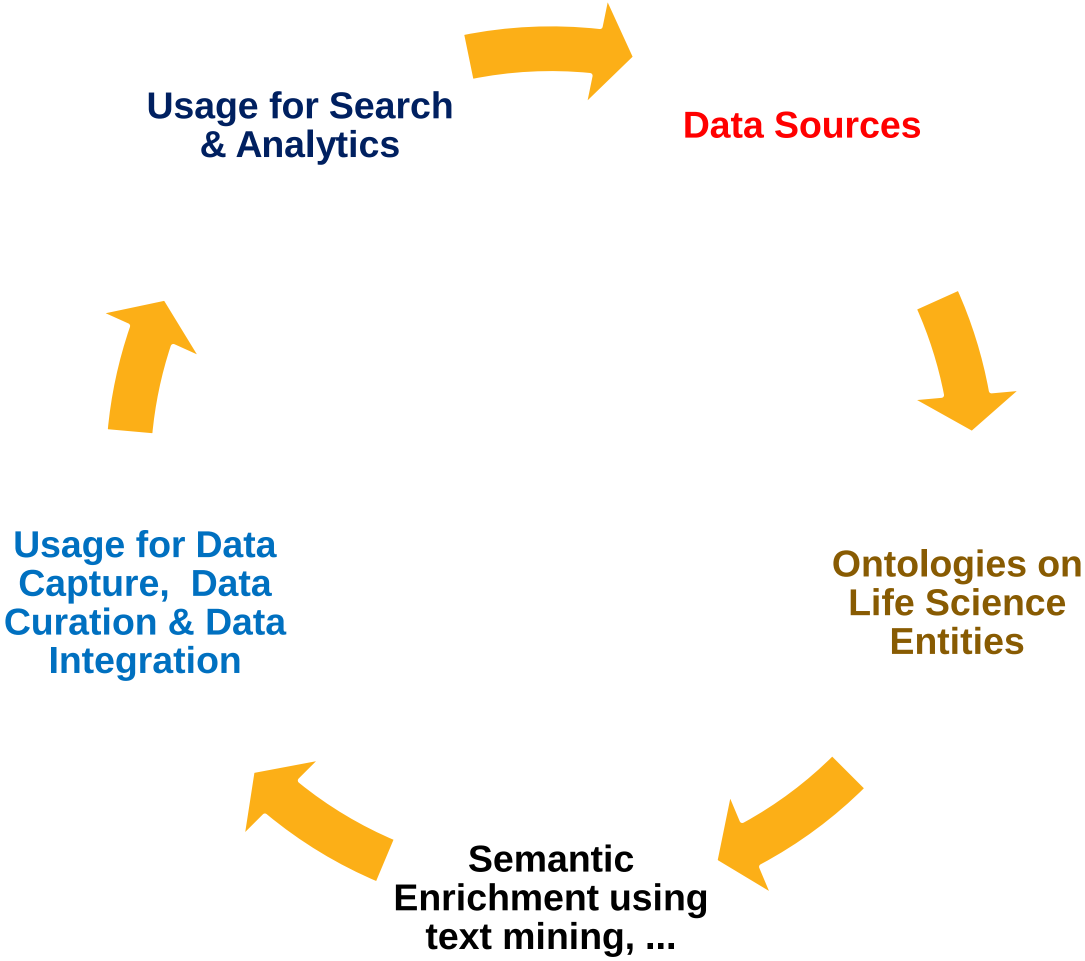
Geneva, 05.06.2015

NOVARTIS

Main activities in Text Mining Services

- NIBR Ontologies
- Scientific Data Curation / Data Integrity & Consistency
- Scientific Data Integration
- RDF Graph DBs
- Federated Queries
- Search
- Text Mining
- Patent Mining

Agenda



Data Sources

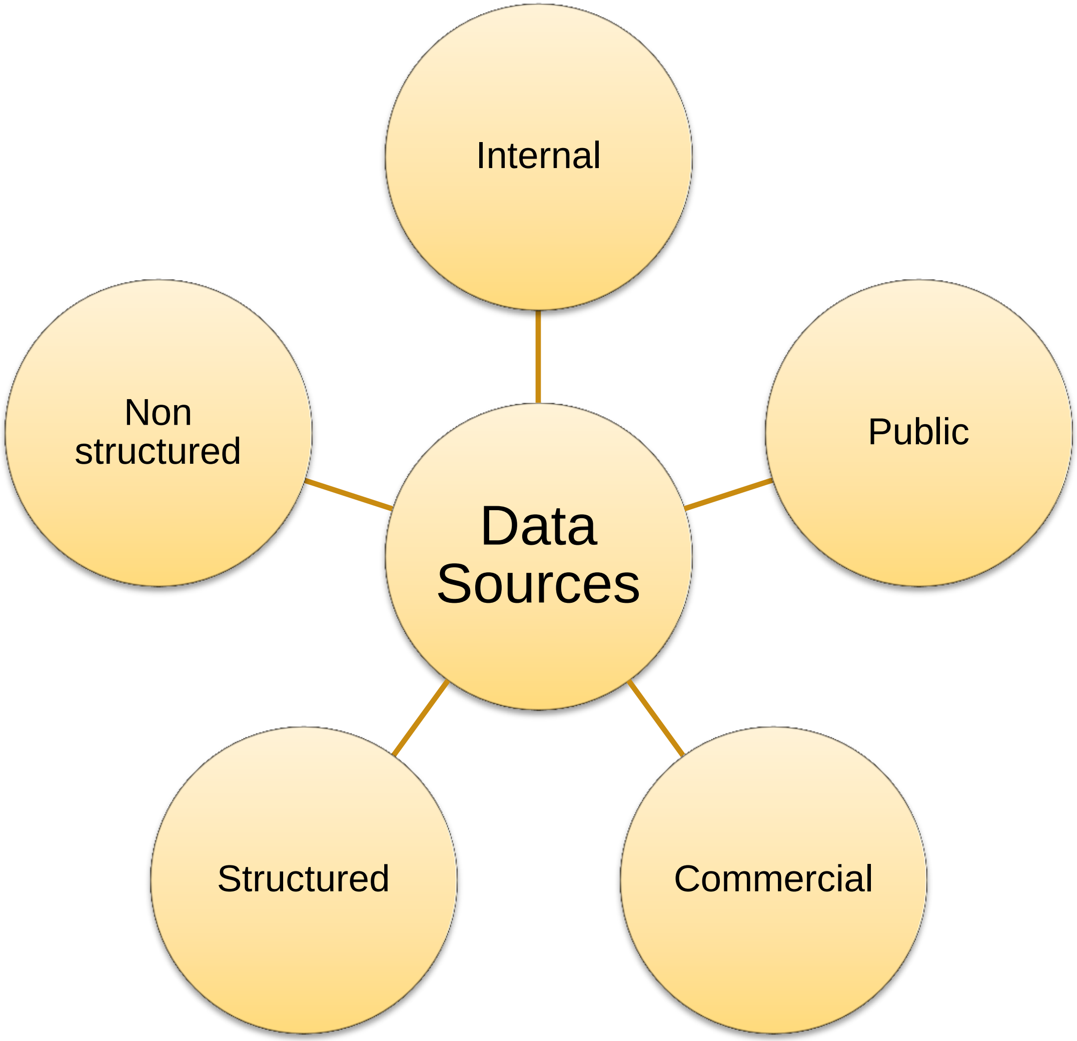
Ontologies on Life Science Entities

Semantic Enrichment using text mining, ...

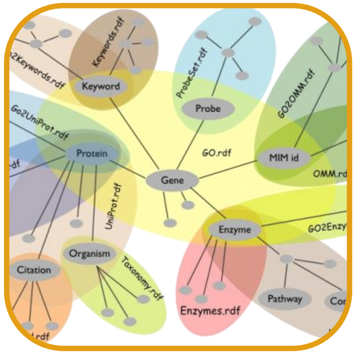
Usage for Data Capture, Data Curation & Data Integration

Usage for Search & Analytics

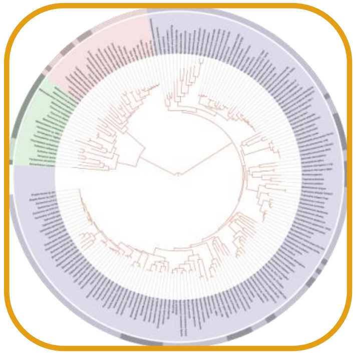
Data Sources



Central Repository for NIBR Ontologies



It aims at providing a uniform vocabulary within and across data repositories and ontology services to informatics teams for usage within NIBR applications.



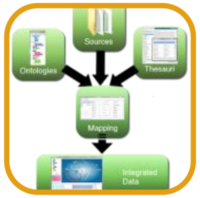
Why is Ontology good for science?

- Standardized vocabulary with definitions and synonyms for unified database annotations across different databases
- Hierarchical organization for aggregation and multi-level comparison of results
- Community adoption for easy comparison of results to other project results worldwide
- Explicit relationships and underlying logical definitions for automated reasoning
- Explicit bridging relationships between different ontologies for exploring underlying mechanisms

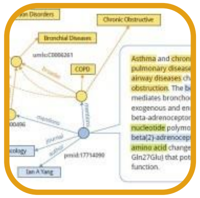
NIBR Ontologies, domain of applications

It can be used for multiple purposes e.g.

Registration systems (metadata capture), data curation



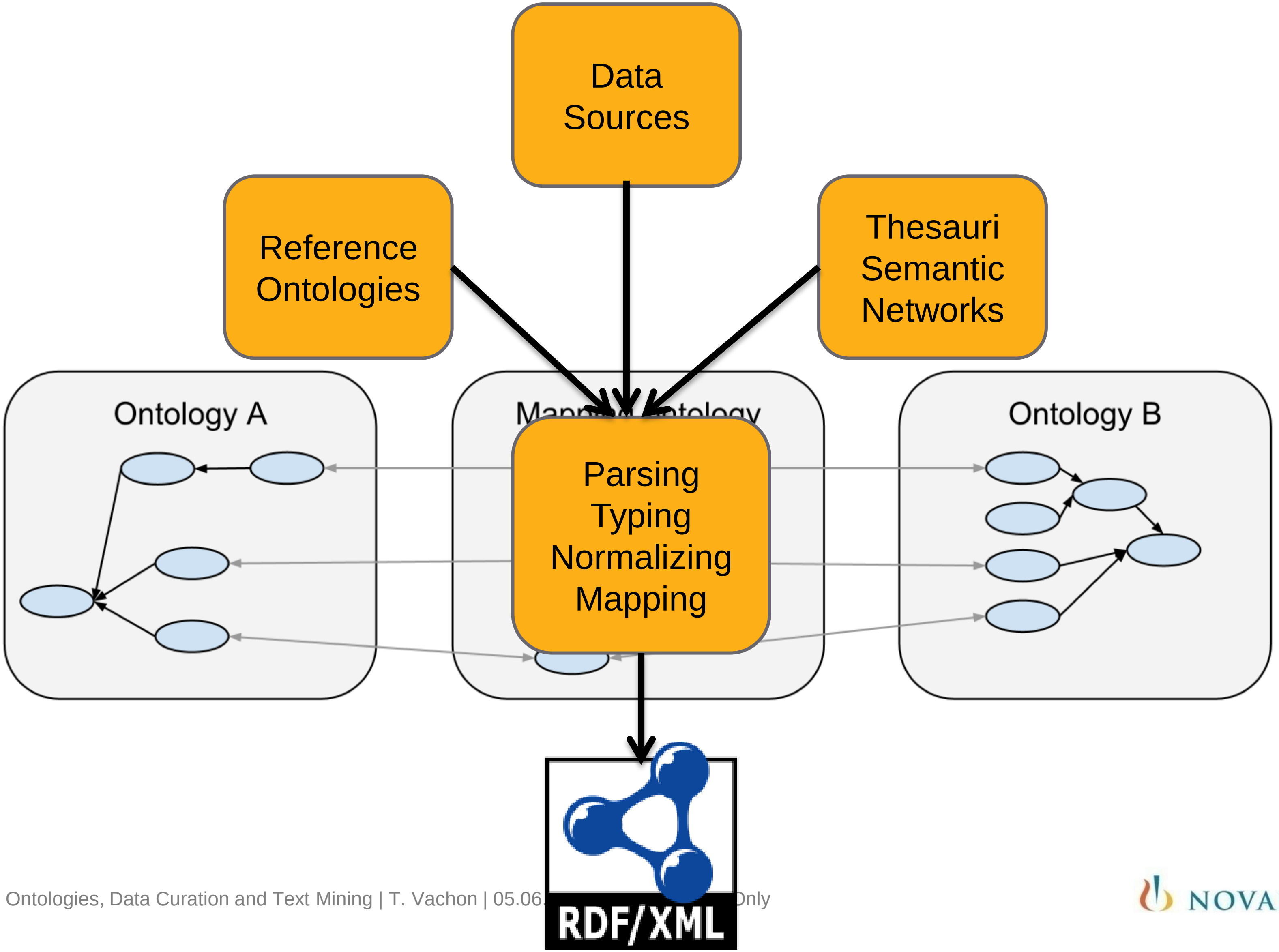
Data mapping, bridging ontologies, navigation between concepts and referential data



Text and data mining

Contribution to scientific queries, semantic data federation

NIBR Ontologies, Content Generation



Data Sources

Reference Ontologies

Thesauri Semantic Networks

Ontology A

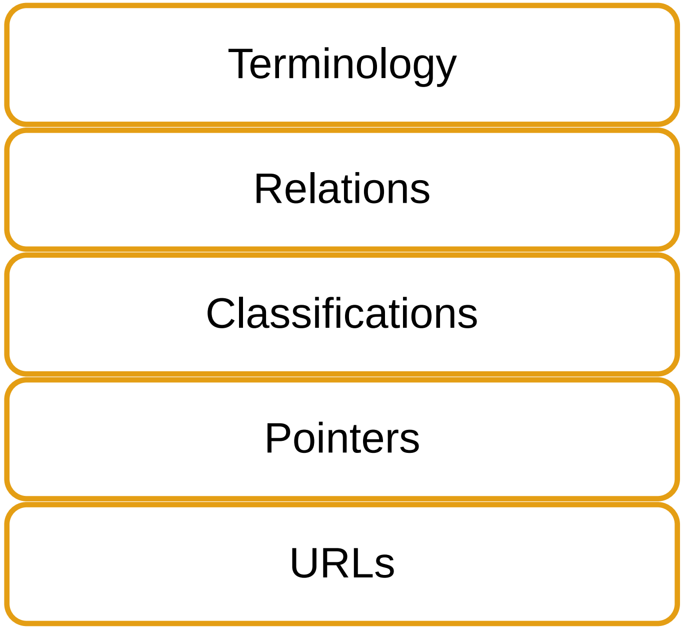
Mapp ntology

Ontology B

Parsing Typing Normalizing Mapping

RDF/XML

NIBR Ontologies, Organization by concept type



Terminology

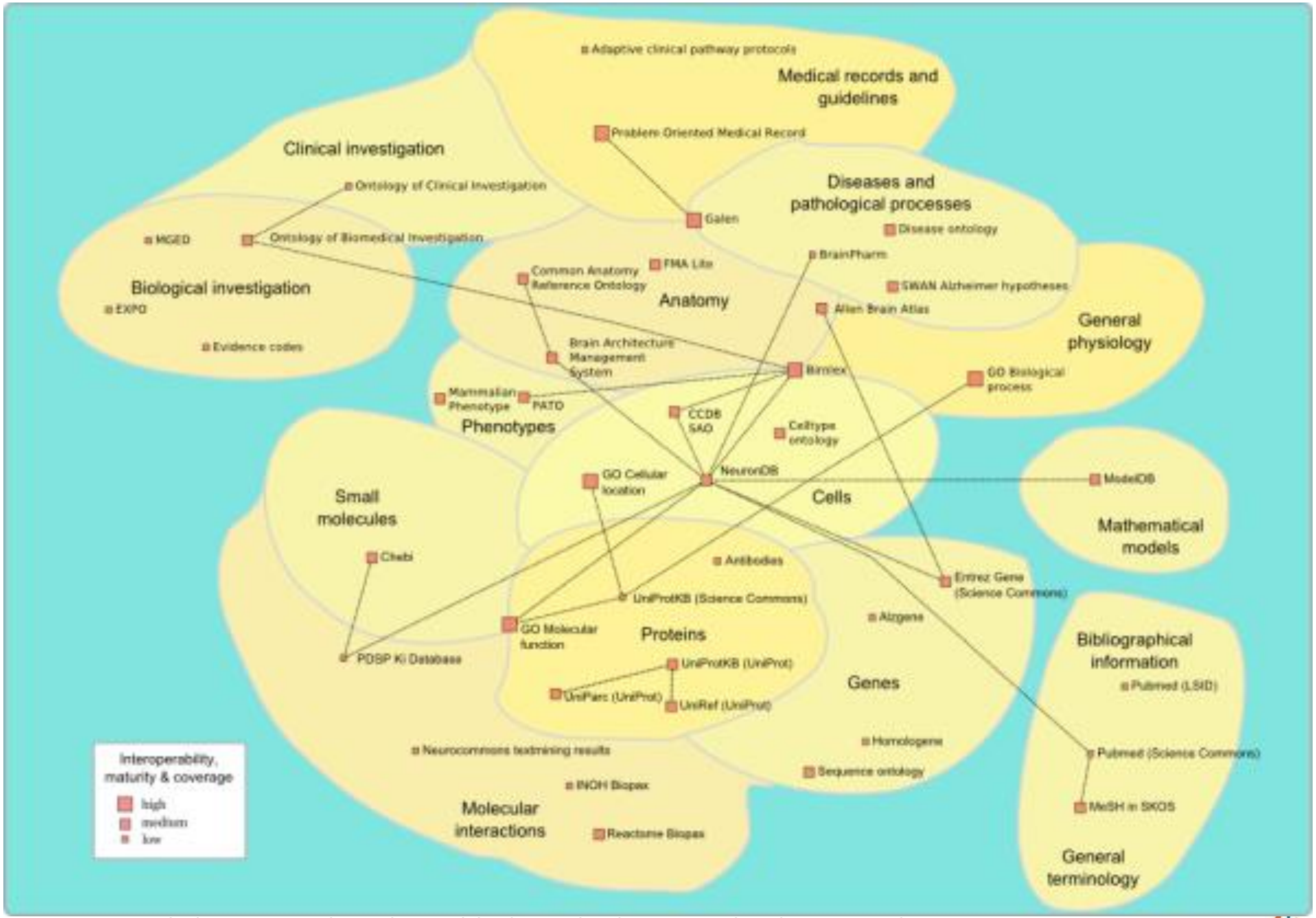
Relations

Classifications

Pointers

URLs

NIBR Ontologies, Concept Types



Genes
Proteins
Pathways
Species
Target Classes

Cell lines
Cell types
Anatomy

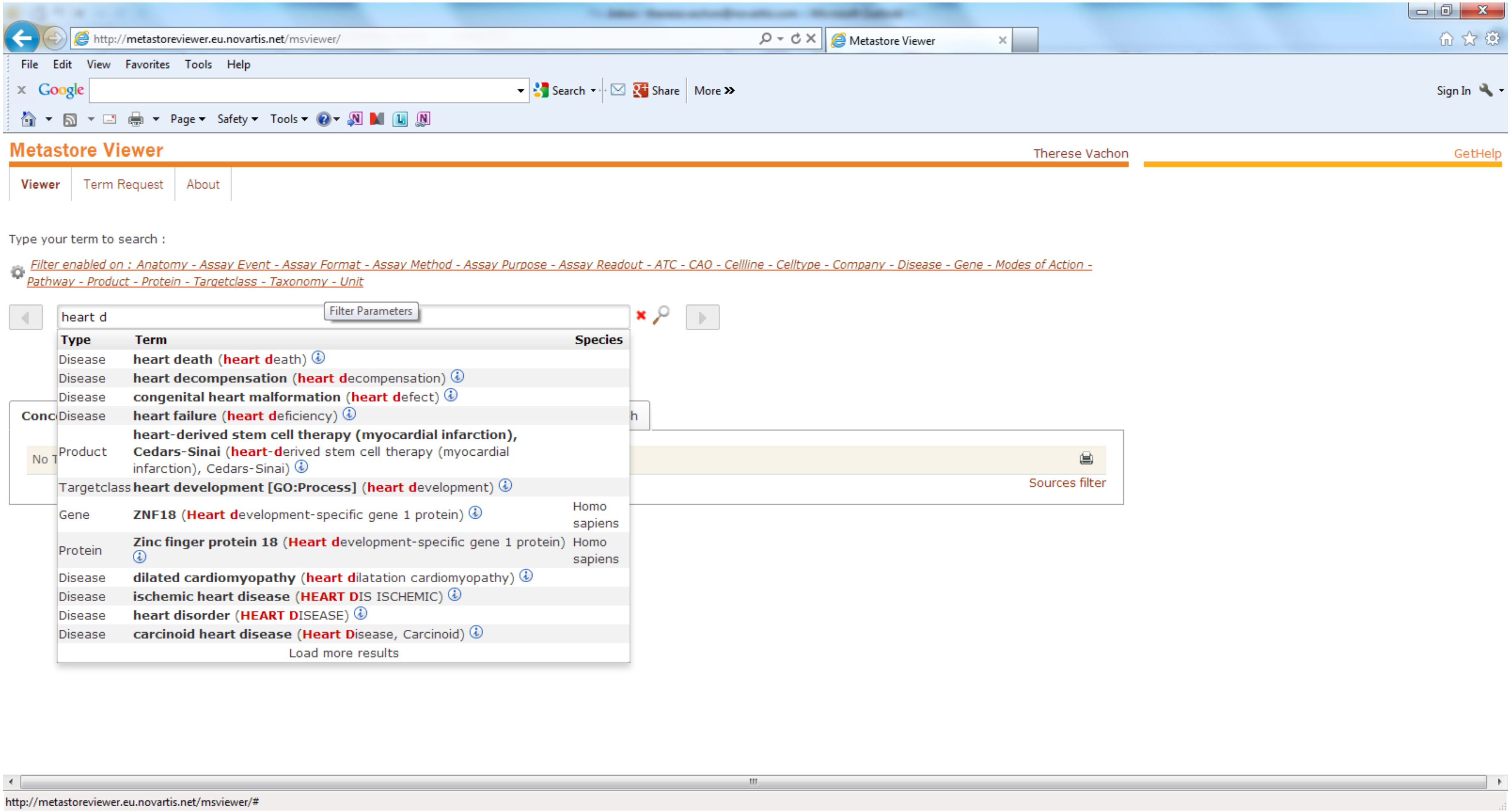
Diseases

Compounds,
Products

Companies
Public
Institutions
Countries

ADCs
Instruments
Skills

Auto-Suggest



Metastore Viewer

Therese Vachon

GetHelp

Type your term to search :

Filter enabled on : Anatomy - Assay Event - Assay Format - Assay Method - Assay Purpose - Assay Readout - ATC - CAO - Cellline - Celltype - Company - Disease - Gene - Modes of Action - Pathway - Product - Protein - Targetclass - Taxonomy - Unit

heart d

Type	Term	Species
Disease	**heart death** (**heart d**eath)	
Disease	**heart decompensation** (**heart d**ecompensation)	
Disease	**congenital heart malformation** (**heart d**efect)	
Disease	**heart failure** (**heart d**eficiency)	
Product	**heart-derived stem cell therapy (myocardial infarction), Cedars-Sinai** (**heart-d**erived stem cell therapy (myocardial infarction), Cedars-Sinai)	
Targetclass	**heart development [GO:Process]** (**heart d**evelopment)	
Gene	**ZNF18** (**Heart d**evelopment-specific gene 1 protein)	Homo sapiens
Protein	**Zinc finger protein 18** (**Heart d**evelopment-specific gene 1 protein)	Homo sapiens
Disease	**dilated cardiomyopathy** (**heart d**ilatation cardiomyopathy)	
Disease	**ischemic heart disease** (**HEART D**IS ISCHEMIC)	
Disease	**heart disorder** (**HEART D**ISEASE)	
Disease	**carcinoid heart disease** (**Heart D**isease, Carcinoid)	

Load more results

Conc

No T

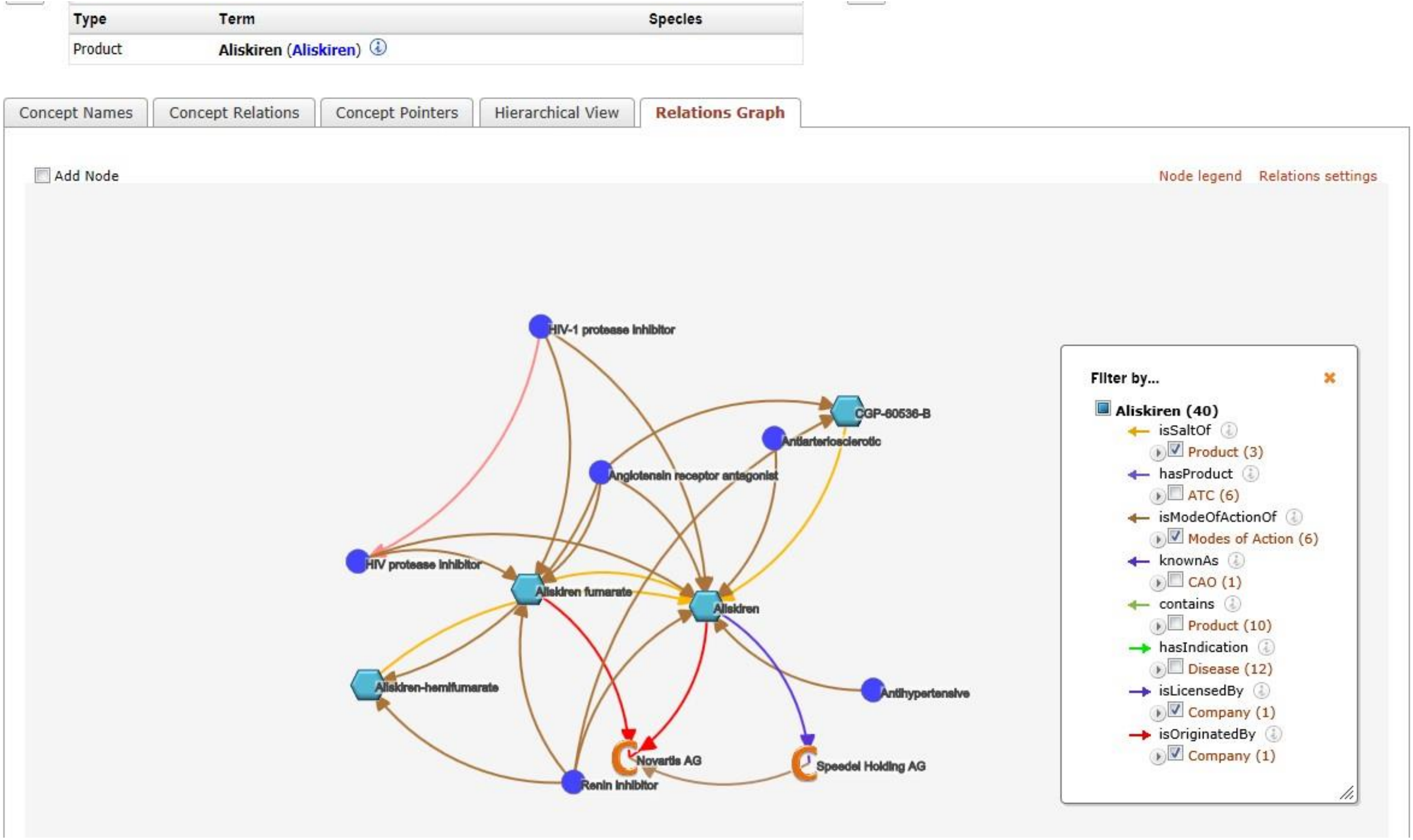
h

Sources filter

http://metastoreviewer.eu.novartis.net/msviewer/#

NOVARTIS

Bridging Ontologies



Type	Term	Species
Product	Aliskiren (Aliskiren) ⓘ	

Concept Names | Concept Relations | Concept Pointers | Hierarchical View | Relations Graph

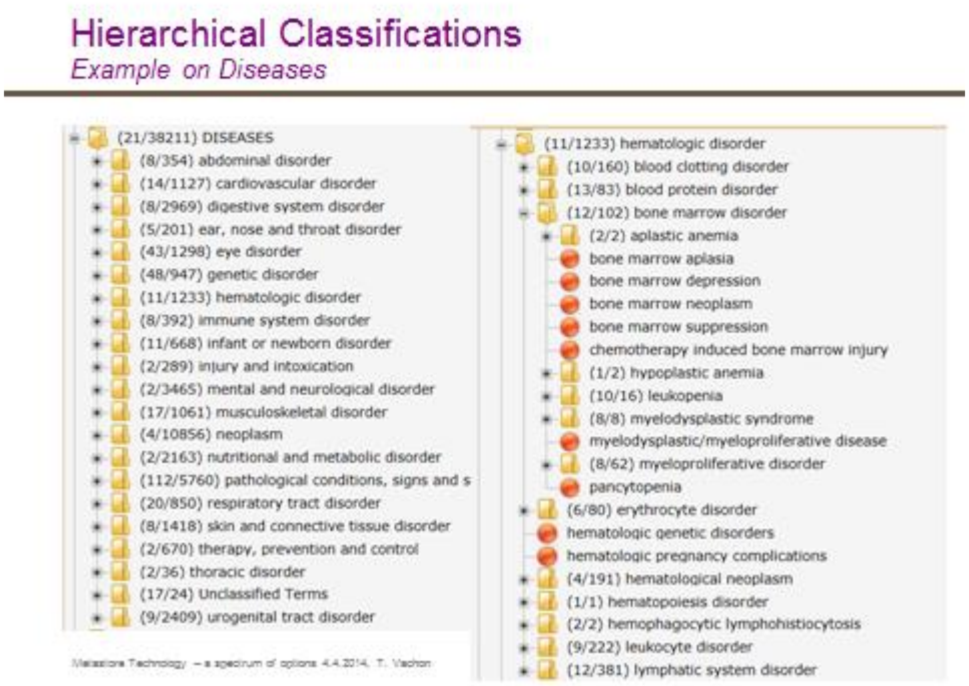
□ Add Node

Node legend   Relations settings

Filter by... ✖

- Aliskiren (40)
  - ← isSaltOf ⓘ
    - ☑ Product (3)
  - ← hasProduct ⓘ
    - □ ATC (6)
  - ← isModeOfActionOf ⓘ
    - ☑ Modes of Action (6)
  - ← knownAs ⓘ
    - □ CAO (1)
  - ← contains ⓘ
    - □ Product (10)
  - → hasIndication ⓘ
    - □ Disease (12)
  - → isLicensedBy ⓘ
    - ☑ Company (1)
  - → isOriginatedBy ⓘ
    - ☑ Company (1)

Classifications



Hierarchical Classifications

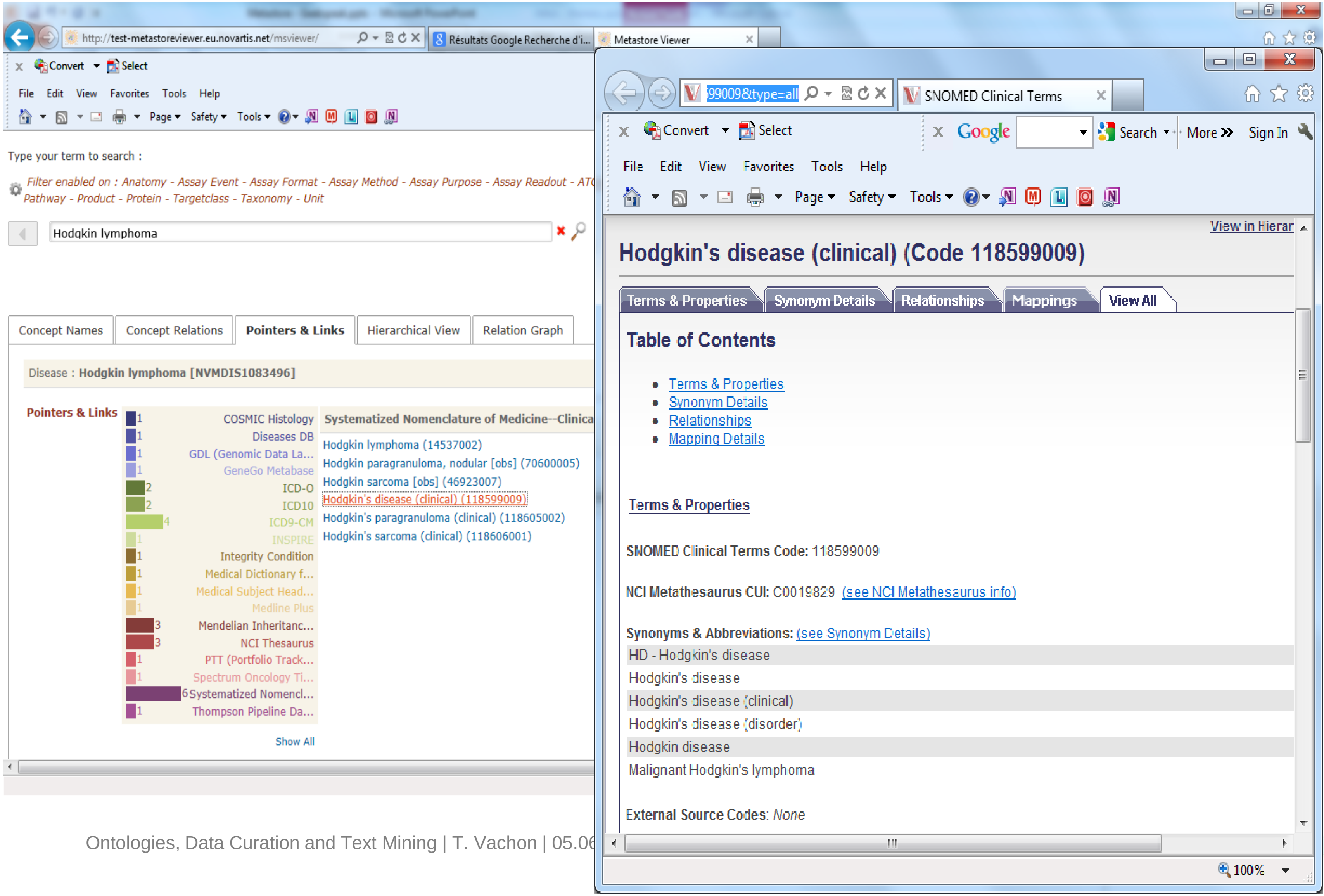
Example on Diseases

- (21/38211) DISEASES
  - (8/354) abdominal disorder
  - (14/1127) cardiovascular disorder
  - (8/2969) digestive system disorder
  - (5/201) ear, nose and throat disorder
  - (43/1298) eye disorder
  - (48/947) genetic disorder
  - (11/1233) hematologic disorder
  - (8/392) immune system disorder
  - (11/668) infant or newborn disorder
  - (2/289) injury and intoxication
  - (2/3465) mental and neurological disorder
  - (17/1061) musculoskeletal disorder
  - (4/10856) neoplasm
  - (2/2163) nutritional and metabolic disorder
  - (112/5760) pathological conditions, signs and s
  - (20/850) respiratory tract disorder
  - (8/1418) skin and connective tissue disorder
  - (2/670) therapy, prevention and control
  - (2/36) thoracic disorder
  - (17/24) Unclassified Terms
  - (9/2409) urogenital tract disorder

Metastore Technology – a spectrum of options 4.4.2014. T. Vachon

- (11/1233) hematologic disorder
  - (10/160) blood clotting disorder
  - (13/83) blood protein disorder
  - (12/102) bone marrow disorder
    - (2/2) aplastic anemia
    - bone marrow aplasia
    - bone marrow depression
    - bone marrow neoplasm
    - bone marrow suppression
    - chemotherapy induced bone marrow injury
    - (1/2) hypoplastic anemia
    - (10/16) leukopenia
    - (8/8) myelodysplastic syndrome
    - myelodysplastic/myeloproliferative disease
    - (8/62) myeloproliferative disorder
    - pancytopenia
  - (6/80) erythrocyte disorder
  - hematologic genetic disorders
  - hematologic pregnancy complications
  - (4/191) hematological neoplasm
  - (1/1) hematopoiesis disorder
  - (2/2) hemophagocytic lymphohistiocytosis
  - (9/222) leukocyte disorder
  - (12/381) lymphatic system disorder

Navigation



http://test-metastoreviewer.eu.novartis.net/msviewer/

Résultats Google Recherche d'i...
Metastore Viewer
Convert
Select
File Edit View Favorites Tools Help
Page Safety Tools

Type your term to search :

Filter enabled on : Anatomy - Assay Event - Assay Format - Assay Method - Assay Purpose - Assay Readout - AT(
Pathway - Product - Protein - Targetclass - Taxonomy - Unit

Hodgkin lymphoma

Disease : **Hodgkin lymphoma [NVMDIS1083496]**

Pointers & Links

1	COSMIC Histology
1	Diseases DB
1	GDL (Genomic Data La...
1	GeneGo Metabase
2	ICD-O
2	ICD10
4	ICD9-CM
1	INSPIRE
1	Integrity Condition
1	Medical Dictionary f...
1	Medical Subject Head...
1	Medline Plus
3	Mendelian Inheritanc...
3	NCI Thesaurus
1	PTT (Portfolio Track...
1	Spectrum Oncology Ti...
6	Systematized Nomencl...
1	Thompson Pipeline Da...

Show All

Systematized Nomenclature of Medicine--Clinica

- Hodgkin lymphoma (14537002)
- Hodgkin paragranuloma, nodular [obs] (70600005)
- Hodgkin sarcoma [obs] (46923007)
- Hodgkin's disease (clinical) (118599009)
- Hodgkin's paragranuloma (clinical) (118605002)
- Hodgkin's sarcoma (clinical) (118606001)

Hodgkin's disease (clinical) (Code 118599009)

Table of Contents

- Terms & Properties
- Synonym Details
- Relationships
- Mapping Details

Terms & Properties

SNOMED Clinical Terms Code: 118599009

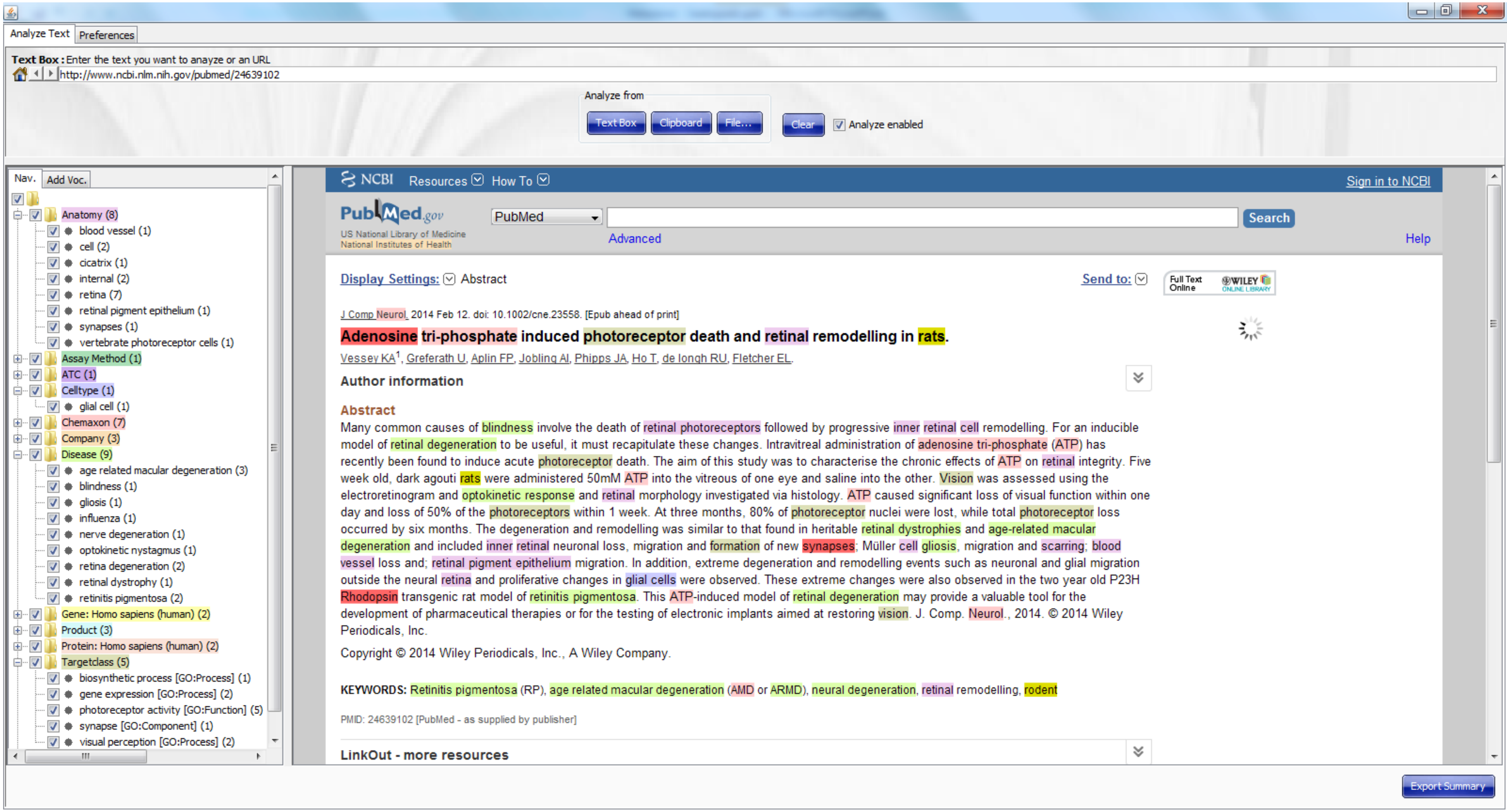
NCI Metathesaurus CUI: C0019829 (see NCI Metathesaurus info)

Synonyms & Abbreviations: (see Synonym Details)

HD - Hodgkin's disease
Hodgkin's disease
Hodgkin's disease (clinical)
Hodgkin's disease (disorder)
Hodgkin disease
Malignant Hodgkin's lymphoma

External Source Codes: *None*

Annotation using text mining and NIBR Ontologies



Text Box : Enter the text you want to anayze or an URL

http://www.ncbi.nlm.nih.gov/pubmed/24639102

☑ Analyze enabled

- ☑ Anatomy (8)
  - ☑ ● blood vessel (1)
  - ☑ ● cell (2)
  - ☑ ● cicatrix (1)
  - ☑ ● internal (2)
  - ☑ ● retina (7)
  - ☑ ● retinal pigment epithelium (1)
  - ☑ ● synapses (1)
  - ☑ ● vertebrate photoreceptor cells (1)
- ☑ Assay Method (1)
- ☑ ATC (1)
- ☑ Celltype (1)
  - ☑ ● glial cell (1)
- ☑ Chemaxon (7)
- ☑ Company (3)
- ☑ Disease (9)
  - ☑ ● age related macular degeneration (3)
  - ☑ ● blindness (1)
  - ☑ ● gliosis (1)
  - ☑ ● influenza (1)
  - ☑ ● nerve degeneration (1)
  - ☑ ● optokinetic nystagmus (1)
  - ☑ ● retina degeneration (2)
  - ☑ ● retinal dystrophy (1)
  - ☑ ● retinitis pigmentosa (2)
- ☑ Gene: Homo sapiens (human) (2)
- ☑ Product (3)
- ☑ Protein: Homo sapiens (human) (2)
- ☑ Targetclass (5)
  - ☑ ● biosynthetic process [GO:Process] (1)
  - ☑ ● gene expression [GO:Process] (2)
  - ☑ ● photoreceptor activity [GO:Function] (5)
  - ☑ ● synapse [GO:Component] (1)
  - ☑ ● visual perception [GO:Process] (2)

PubMed

Display Settings: Abstract

J Comp Neurol. 2014 Feb 12. doi: 10.1002/cne.23558. [Epub ahead of print]

Adenosine tri-phosphate induced photoreceptor death and retinal remodelling in rats.

Vessey KA[1], Greferath U, Aplin FP, Jobling AI, Phipps JA, Ho T, de Iongh RU, Fletcher EL.

Author information

Abstract

Many common causes of blindness involve the death of retinal photoreceptors followed by progressive inner retinal cell remodelling. For an inducible model of retinal degeneration to be useful, it must recapitulate these changes. Intravitreal administration of adenosine tri-phosphate (ATP) has recently been found to induce acute photoreceptor death. The aim of this study was to characterise the chronic effects of ATP on retinal integrity. Five week old, dark agouti rats were administered 50mM ATP into the vitreous of one eye and saline into the other. Vision was assessed using the electroretinogram and optokinetic response and retinal morphology investigated via histology. ATP caused significant loss of visual function within one day and loss of 50% of the photoreceptors within 1 week. At three months, 80% of photoreceptor nuclei were lost, while total photoreceptor loss occurred by six months. The degeneration and remodelling was similar to that found in heritable retinal dystrophies and age-related macular degeneration and included inner retinal neuronal loss, migration and formation of new synapses; Müller cell gliosis, migration and scarring; blood vessel loss and; retinal pigment epithelium migration. In addition, extreme degeneration and remodelling events such as neuronal and glial migration outside the neural retina and proliferative changes in glial cells were observed. These extreme changes were also observed in the two year old P23H Rhodopsin transgenic rat model of retinitis pigmentosa. This ATP-induced model of retinal degeneration may provide a valuable tool for the development of pharmaceutical therapies or for the testing of electronic implants aimed at restoring vision. J. Comp. Neurol., 2014. © 2014 Wiley Periodicals, Inc.

Copyright © 2014 Wiley Periodicals, Inc., A Wiley Company.

KEYWORDS: Retinitis pigmentosa (RP), age related macular degeneration (AMD or ARMD), neural degeneration, retinal remodelling, rodent

PMID: 24639102 [PubMed - as supplied by publisher]

LinkOut - more resources

NOVARTIS

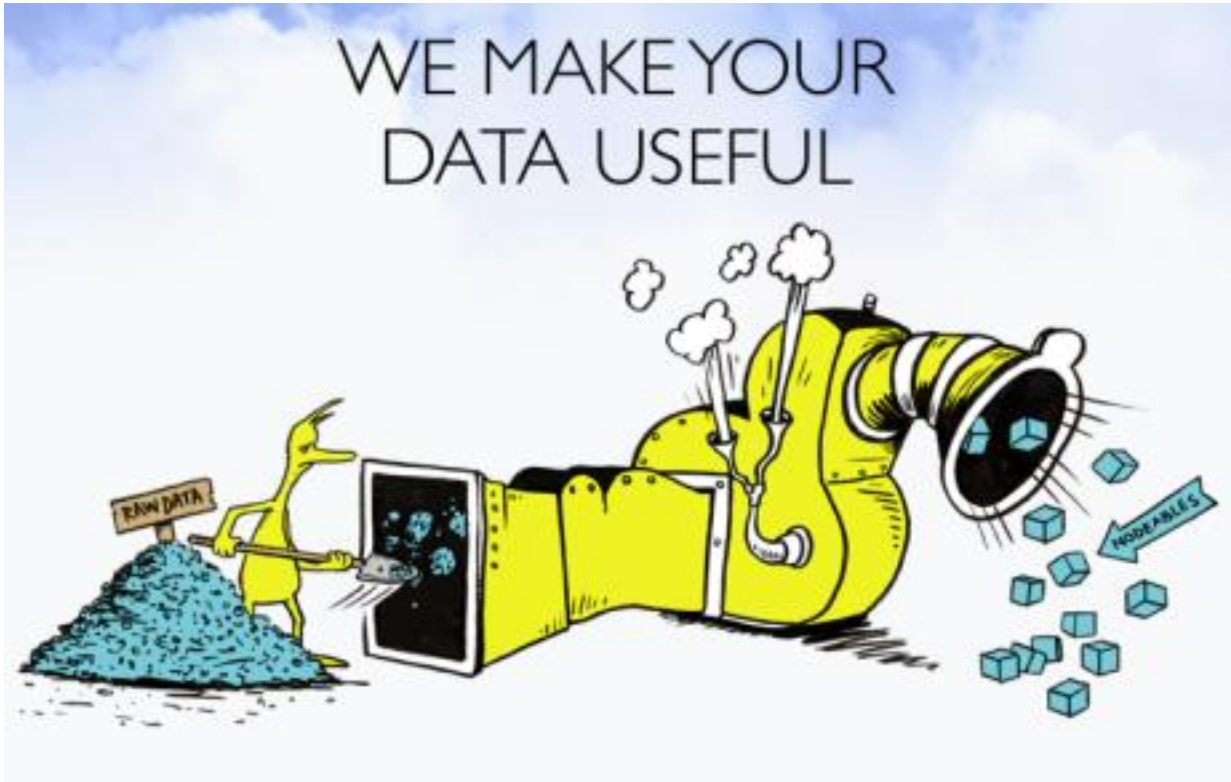
Text extractors

- Lexical extraction
- Pattern extraction
  - E.g. IDs, Patent numbers, Compound numbers
- Chemical Entity Extraction
  - Trivial names
  - IUPAC
  - Smiles
  - Inchi
  - Images
- Bulk Terminology Extractor (see examples)

More sophisticated text mining

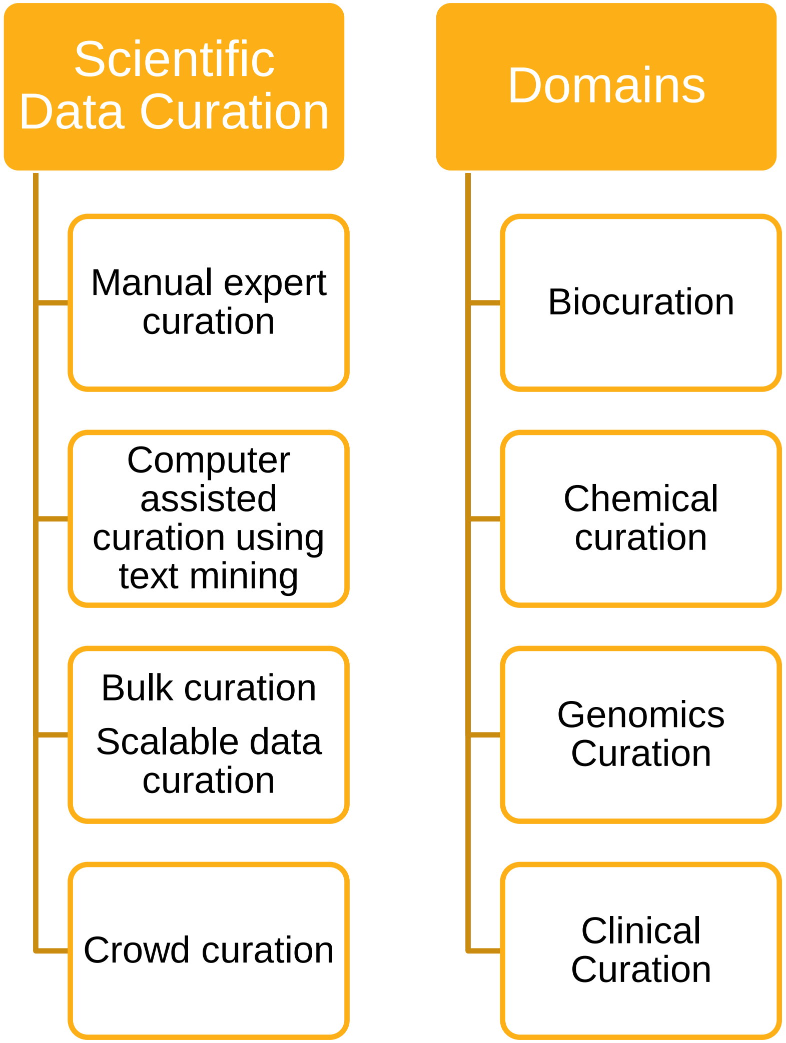
- Extraction of entities relationships from literature
- Extraction of skills and disambiguation of authors from literature
- Text mining projects specific for a team

Data Curation



Scientific Data Curation

Curation types, Domains



Scientific Data Curation

- Manual expert curation
- Computer assisted curation using text mining
- Bulk curation Scalable data curation
- Crowd curation

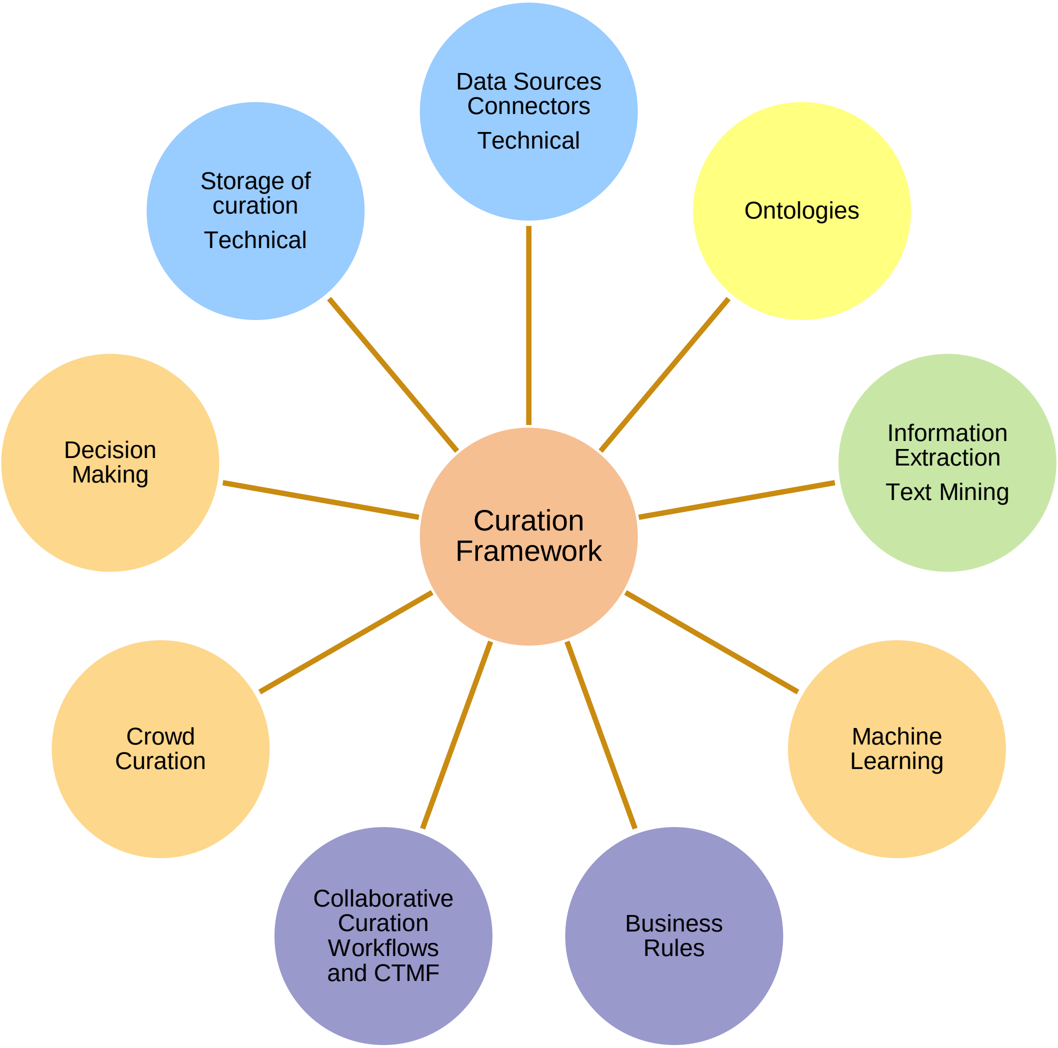
Domains

- Biocuration
- Chemical curation
- Genomics Curation
- Clinical Curation

Expectations about scientific data curation in NIBR

- Implementing rules and best practices for data and metadata capture
- Helping to register / publish new data in the existing systems
- Adding / correcting metadata using authoritative vocabularies
- Identifying missing terms / terminologies
- Correcting faulty data in the source systems
- QA/QC: Assessing the data integrity in different systems & implementing corrective measures
- Making sure that the source data is consumable by the systems
- Curating different entities (assays, targets, compounds, genomics data, etc)
- Curating internal / external data sources (literature, patents, etc)

Curation Framework



CTMF: an introduction

What is the CTMF ?

- **NIBR Ontologies (Metastore)** provide an integrated terminology and synonyms resolution service.
- The **C**ollaborative **T**erminology **M**anagement **F**ramework enables a distributed creation of the Metastore content.
- In its **first version**, the CTMF provides the following functionalities:
  - Users can request a new term to be added
  - The CTMF supports content owners and users in the clarification and resolution process.
  - Applications properly linked to the CTMF can make use of "temporary ID" before the term resolution process is completed.

CTMF: an introduction

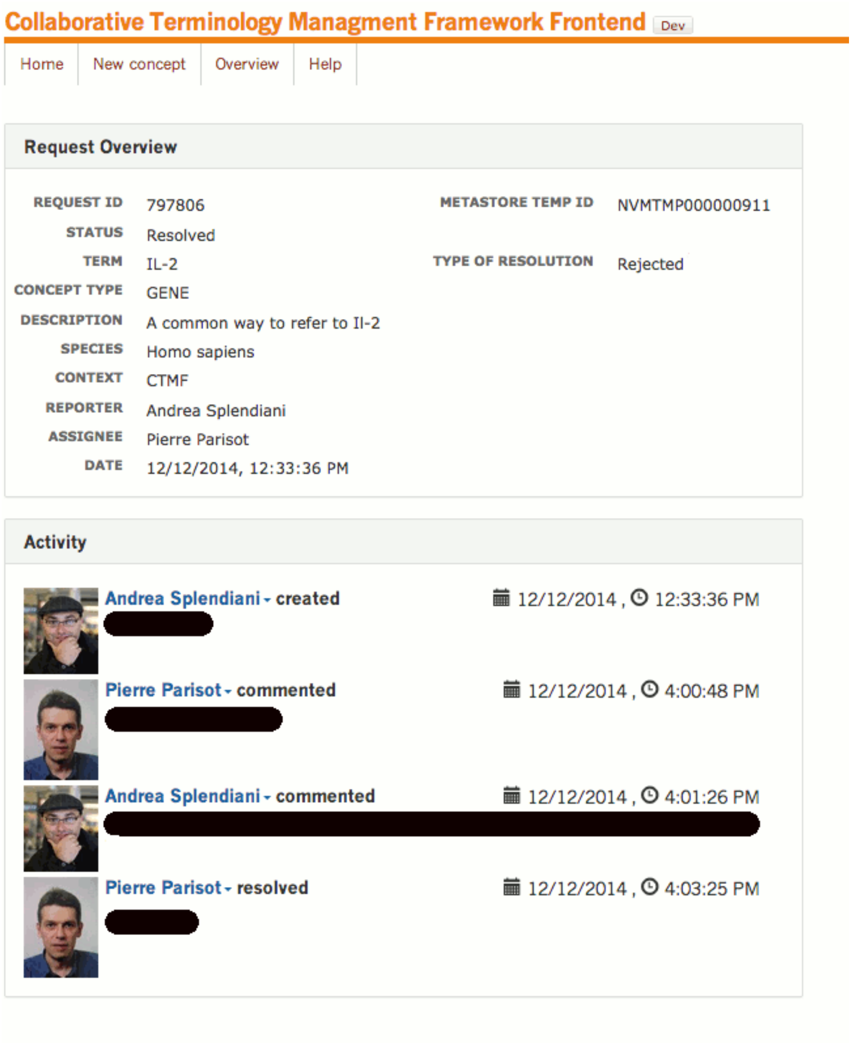
Request for new terms

- A user can request a new term via the **CTMF application**, or in any application that properly includes the **CTMF widget**.

- **The CTMF is not a registration systems**: no automatic bulk submissions of terms are possible, as each term is validated by content owners.

CTMF: an introduction

Clarification and resolution process



Request Overview

REQUEST ID	797806	METASTORE TEMP ID	NVMTMP000000911
STATUS	Resolved		
TERM	IL-2	TYPE OF RESOLUTION	Rejected
CONCEPT TYPE	GENE		
DESCRIPTION	A common way to refer to Il-2		
SPECIES	Homo sapiens		
CONTEXT	CTMF		
REPORTER	Andrea Splendiani		
ASSIGNEE	Pierre Parisot		
DATE	12/12/2014, 12:33:36 PM		

Activity

Andrea Splendiani - created — 12/12/2014 , 12:33:36 PM
[redacted]

Pierre Parisot - commented — 12/12/2014 , 4:00:48 PM
[redacted]

Andrea Splendiani - commented — 12/12/2014 , 4:01:26 PM
[redacted]

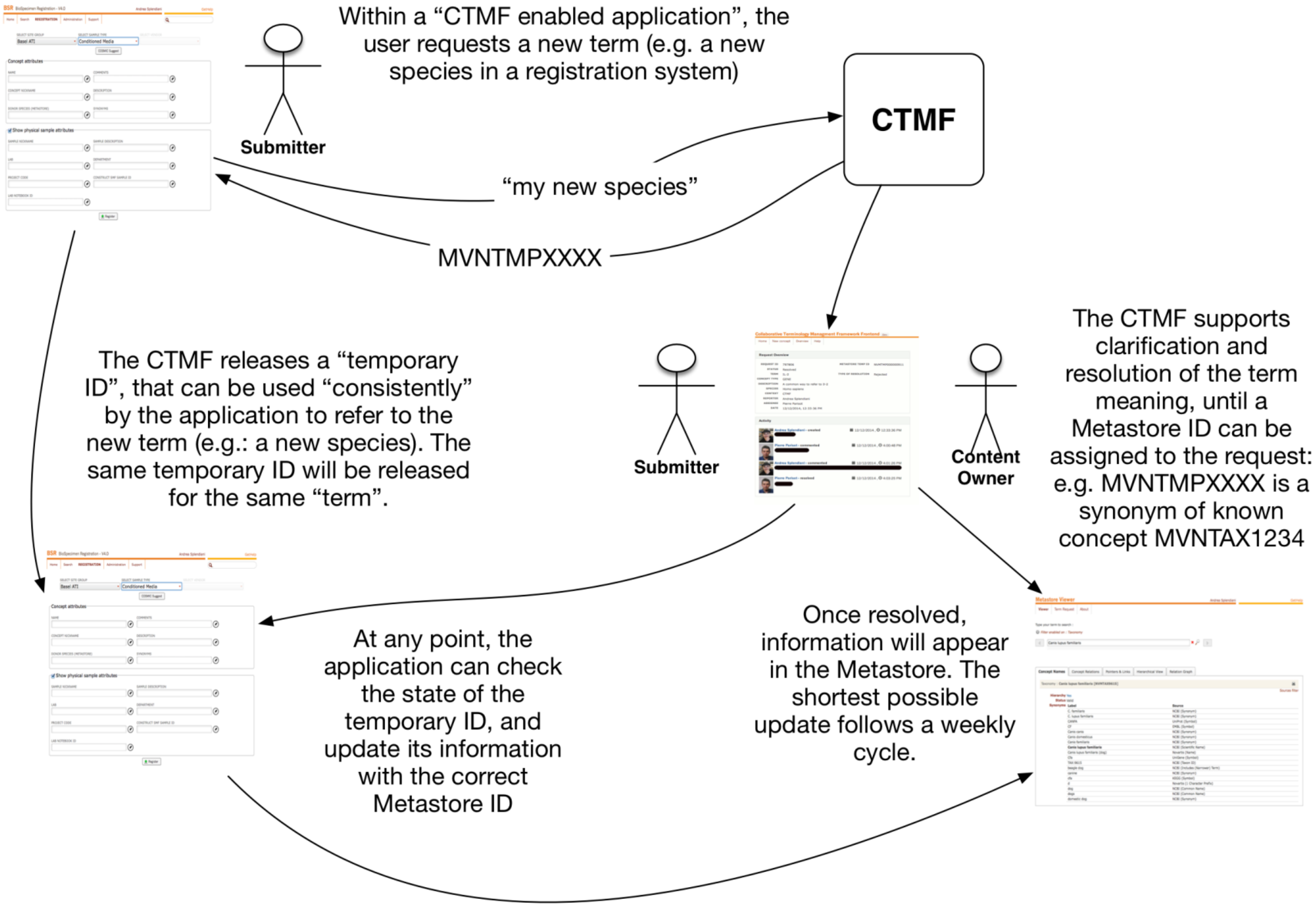
Pierre Parisot - resolved — 12/12/2014 , 4:03:25 PM
[redacted]

- The CTMF presents a “term status page” where discussion on the meaning of the term can be recorded.
- In general, content owners can “resolve” a new term as relative to a new concept, as a synonym of an existing concept, or as an error.

NOVARTIS

CTMF: an introduction

Use of temporary IDs by applications



Within a "CTMF enabled application", the user requests a new term (e.g. a new species in a registration system)

The CTMF releases a "temporary ID", that can be used "consistently" by the application to refer to the new term (e.g.: a new species). The same temporary ID will be released for the same "term".

The CTMF supports clarification and resolution of the term meaning, until a Metastore ID can be assigned to the request: e.g. MVNTMPXXXX is a synonym of known concept MVNTAX1234

At any point, the application can check the state of the temporary ID, and update its information with the correct Metastore ID

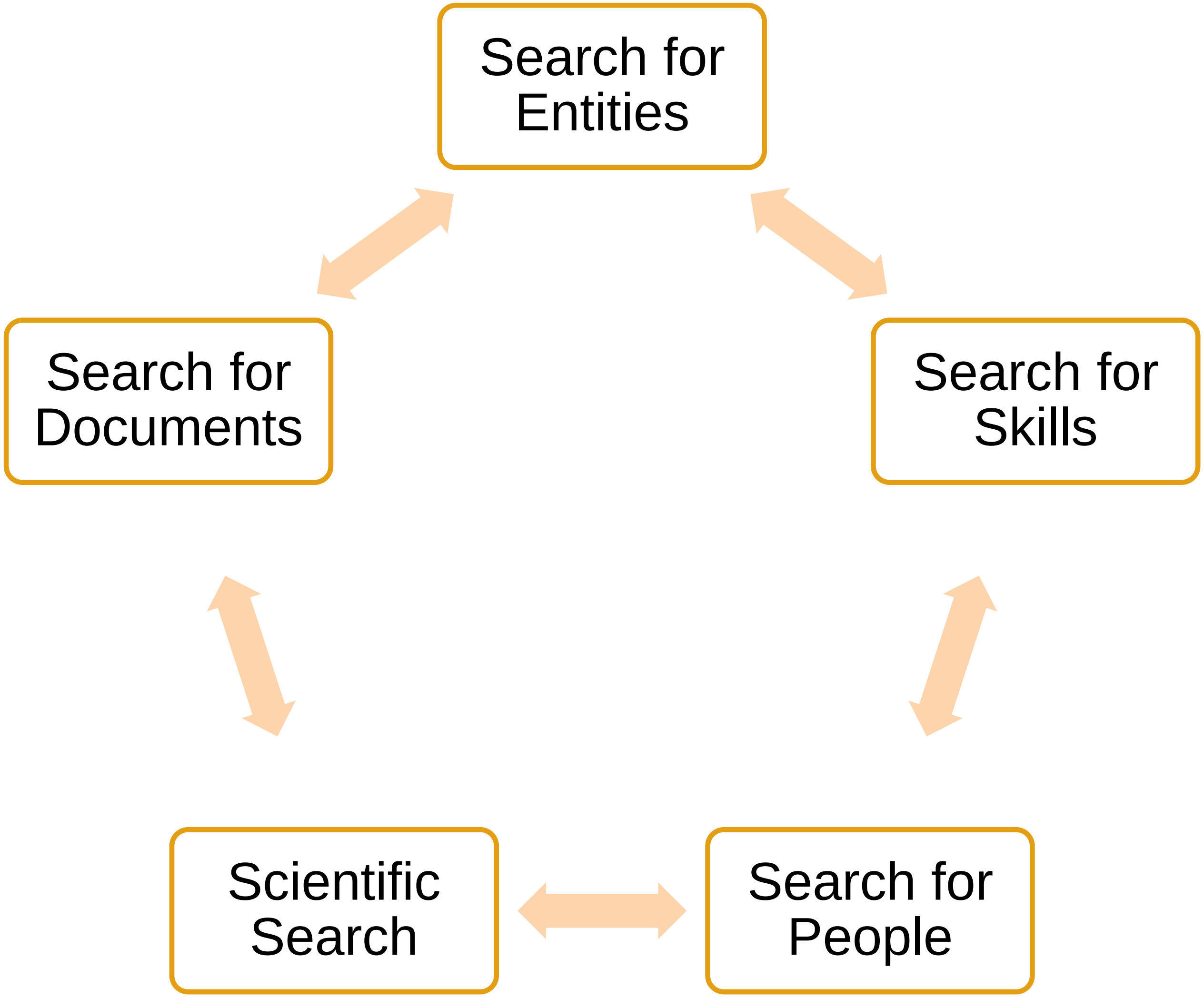
Once resolved, information will appear in the Metastore. The shortest possible update follows a weekly cycle.

CTMF: an introduction

The CTMF in perspective

- The CTMF has deep implications in the way “concepts” and terminologies are used and produced. As such it is a system that needs to evolve with its user base.
- For this reason, we have released a first core-set of functionalities.

Search


Search for Entities

Search for Skills

Search for People

Scientific Search

Search for Documents

Search for entities

- Genes, proteins, target classes, pathways
- Cell lines, cell types, diseases, anatomy
- Compounds, ADC
- Projects...

Search for Skills

- Disciplines, Lab skills, Instruments, Methods, Technologies, Assays, ...
- Tools/applications

Search for People

- People name, Department, Sub-department, Location

Scientific Search

- Search for
  - assays and assay results
  - samples, studies, experiments (ELNs)
  - pre-clinical and clinical data, safety data
  - Studies
  - Clinical trials
  - Project information, project decisions
  - Complex queries

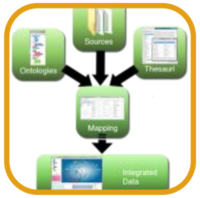
Search for documents ...

- Search for documents in SharePoint/Intranet, Document Databases
- Search for Publications, Patents
- Search for Applications, Tools
- Search for training courses
- ...

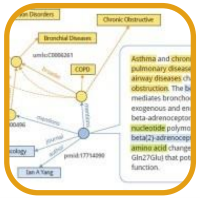
Contribution to scientific queries

NIBR Ontologies can be used for multiple purposes e.g.

Registration systems (metadata capture), data curation



Data mapping, bridging ontologies, navigation between concepts and referential data



Text and data mining

Contribution to scientific queries, semantic data federation

Querying Metastore and other data sources

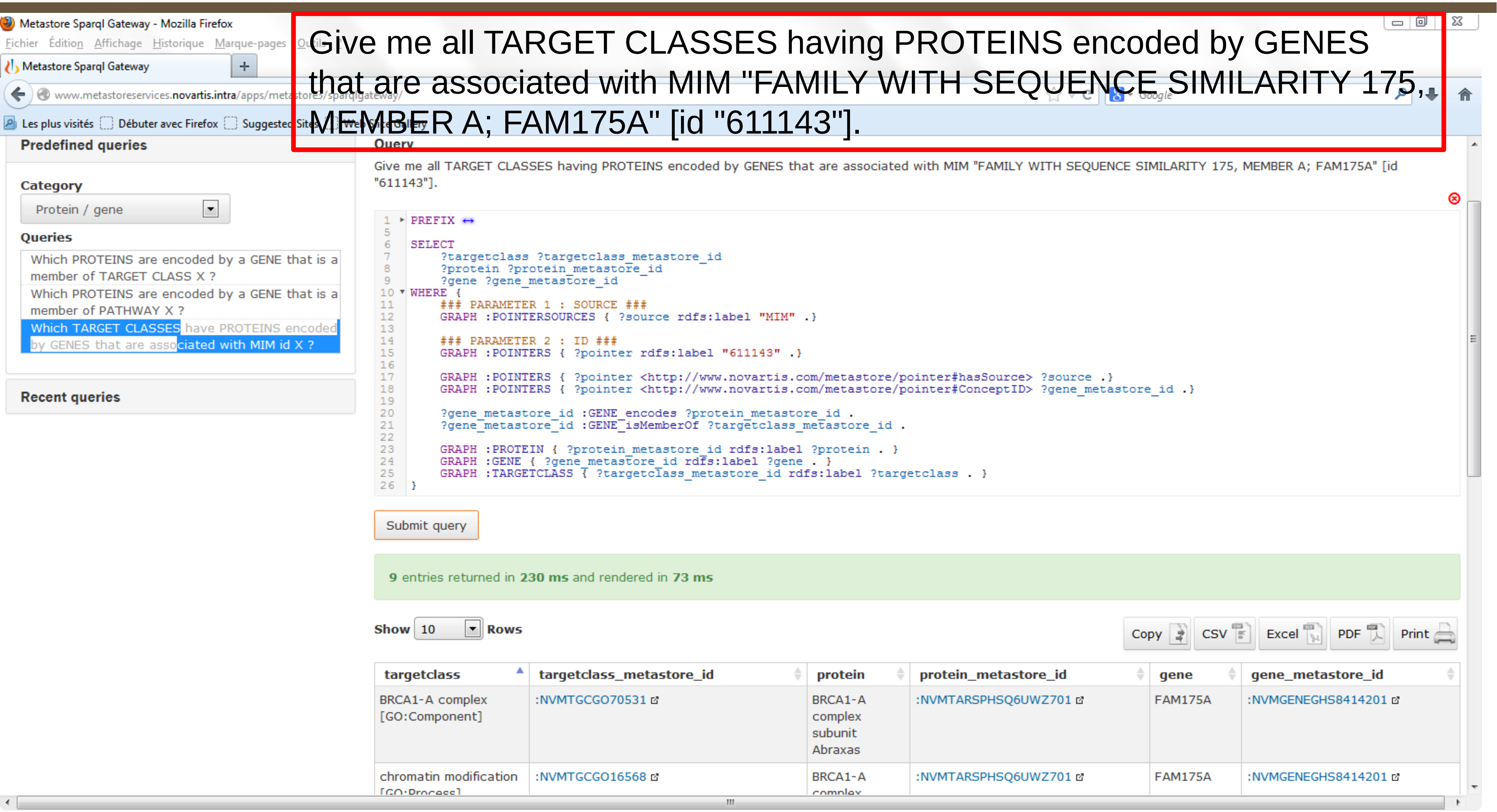
SPARQL queries

Querying Materialized Views

Federated SPARQL queries with other triplestores

Combining queries in e.g. a datasource, a DWH with querying Metastore materialized views

SPARQL Query



Give me all TARGET CLASSES having PROTEINS encoded by GENES that are associated with MIM "FAMILY WITH SEQUENCE SIMILARITY 175, MEMBER A; FAM175A" [id "611143"].

Predefined queries

Category: Protein / gene

Queries

- Which PROTEINS are encoded by a GENE that is a member of TARGET CLASS X ?
- Which PROTEINS are encoded by a GENE that is a member of PATHWAY X ?
- Which TARGET CLASSES have PROTEINS encoded by GENES that are associated with MIM id X ?

Recent queries

Query

Give me all TARGET CLASSES having PROTEINS encoded by GENES that are associated with MIM "FAMILY WITH SEQUENCE SIMILARITY 175, MEMBER A; FAM175A" [id "611143"].

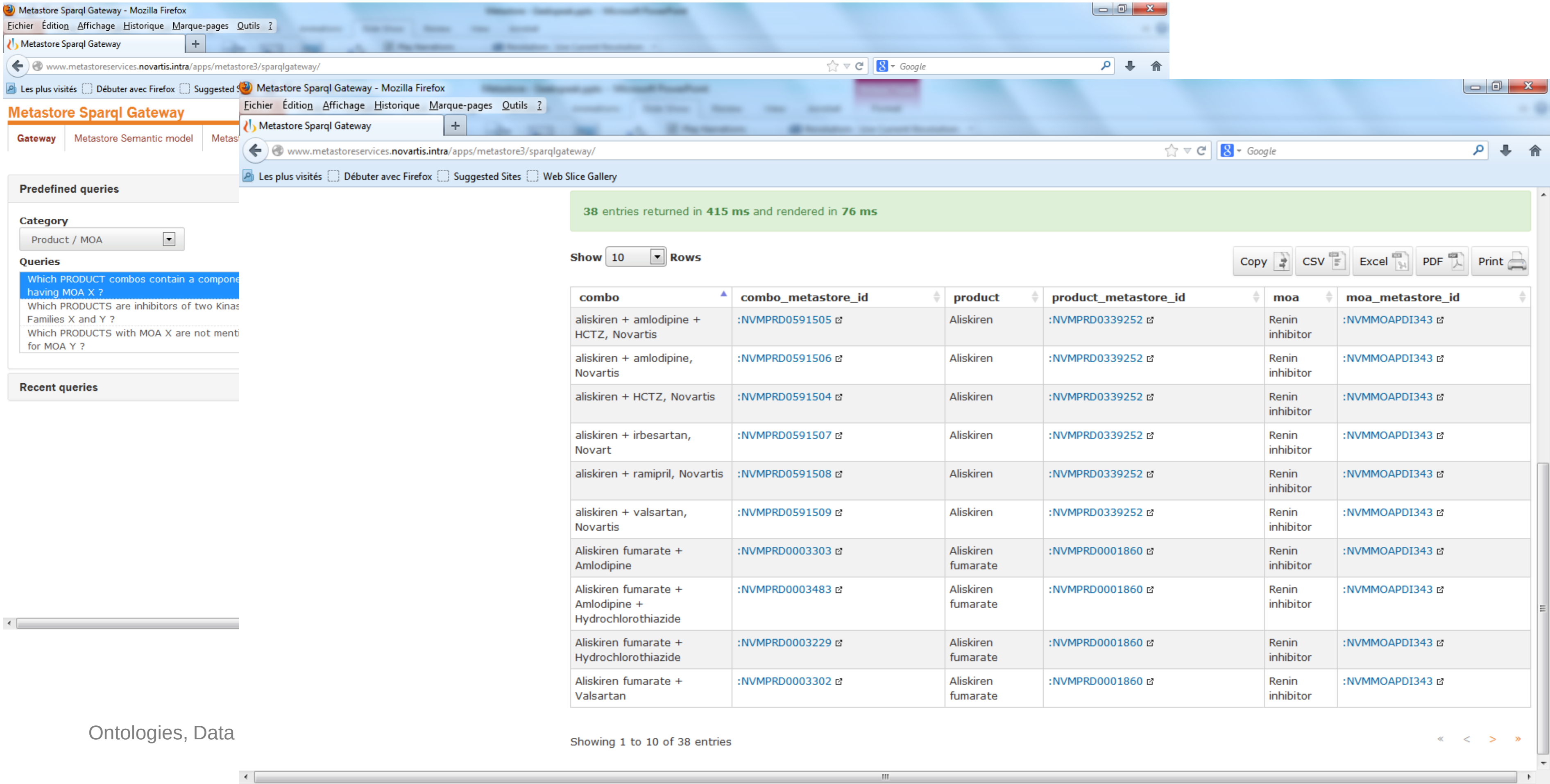
```
PREFIX ↔

SELECT
    ?targetclass ?targetclass_metastore_id
    ?protein ?protein_metastore_id
    ?gene ?gene_metastore_id
WHERE {
    ### PARAMETER 1 : SOURCE ###
    GRAPH :POINTERSOURCES { ?source rdfs:label "MIM" .}

    ### PARAMETER 2 : ID ###
    GRAPH :POINTERS { ?pointer rdfs:label "611143" .}

    GRAPH :POINTERS { ?pointer <http://www.novartis.com/metastore/pointer#hasSource> ?source .}
    GRAPH :POINTERS { ?pointer <http://www.novartis.com/metastore/pointer#ConceptID> ?gene_metastore_id .}

    ?gene_metastore_id :GENE_encodes ?protein_metastore_id .
    ?gene_metastore_id :GENE_isMemberOf ?targetclass_metastore_id .

    GRAPH :PROTEIN { ?protein_metastore_id rdfs:label ?protein . }
    GRAPH :GENE { ?gene_metastore_id rdfs:label ?gene . }
    GRAPH :TARGETCLASS { ?targetclass_metastore_id rdfs:label ?targetclass . }
}
```

9 entries returned in **230 ms** and rendered in **73 ms**

targetclass	targetclass_metastore_id	protein	protein_metastore_id	gene	gene_metastore_id
BRCA1-A complex [GO:Component]	:NVMTGCGO70531	BRCA1-A complex subunit Abraxas	:NVMTARSPHSQ6UWZ701	FAM175A	:NVMGENEGHS8414201
chromatin modification [GO:Process]	:NVMTGCGO16568	BRCA1-A complex	:NVMTARSPHSQ6UWZ701	FAM175A	:NVMGENEGHS8414201

SPARQL Query

- Give me all PRODUCT combos for component(s) acting as "Renin inhibitor" (MOA) and its subclasses.



Metastore Sparql Gateway - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?
Metastore Sparql Gateway
www.metastoreservices.novartis.intra/apps/metastore3/sparqlgateway/
Google
Les plus visités
Débuter avec Firefox
Suggested

Metastore Sparql Gateway

Gateway
Metastore Semantic model
Metas

Predefined queries

Category

Product / MOA

Queries

Which PRODUCT combos contain a compone having MOA X ?

Which PRODUCTS are inhibitors of two Kinas Families X and Y ?

Which PRODUCTS with MOA X are not menti for MOA Y ?

Recent queries

Metastore Sparql Gateway - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?
Metastore Sparql Gateway
www.metastoreservices.novartis.intra/apps/metastore3/sparqlgateway/
Google
Les plus visités
Débuter avec Firefox
Suggested Sites
Web Slice Gallery

38 entries returned in **415 ms** and rendered in **76 ms**

Show 10 **Rows**

combo	combo_metastore_id	product	product_metastore_id	moa	moa_metastore_id
aliskiren + amlodipine + HCTZ, Novartis	:NVMPRD0591505	Aliskiren	:NVMPRD0339252	Renin inhibitor	:NVMMOAPDI343
aliskiren + amlodipine, Novartis	:NVMPRD0591506	Aliskiren	:NVMPRD0339252	Renin inhibitor	:NVMMOAPDI343
aliskiren + HCTZ, Novartis	:NVMPRD0591504	Aliskiren	:NVMPRD0339252	Renin inhibitor	:NVMMOAPDI343
aliskiren + irbesartan, Novart	:NVMPRD0591507	Aliskiren	:NVMPRD0339252	Renin inhibitor	:NVMMOAPDI343
aliskiren + ramipril, Novartis	:NVMPRD0591508	Aliskiren	:NVMPRD0339252	Renin inhibitor	:NVMMOAPDI343
aliskiren + valsartan, Novartis	:NVMPRD0591509	Aliskiren	:NVMPRD0339252	Renin inhibitor	:NVMMOAPDI343
Aliskiren fumarate + Amlodipine	:NVMPRD0003303	Aliskiren fumarate	:NVMPRD0001860	Renin inhibitor	:NVMMOAPDI343
Aliskiren fumarate + Amlodipine + Hydrochlorothiazide	:NVMPRD0003483	Aliskiren fumarate	:NVMPRD0001860	Renin inhibitor	:NVMMOAPDI343
Aliskiren fumarate + Hydrochlorothiazide	:NVMPRD0003229	Aliskiren fumarate	:NVMPRD0001860	Renin inhibitor	:NVMMOAPDI343
Aliskiren fumarate + Valsartan	:NVMPRD0003302	Aliskiren fumarate	:NVMPRD0001860	Renin inhibitor	:NVMMOAPDI343

Showing 1 to 10 of 38 entries

Acknowledgements

- Pierre Parisot
- Andrea Splendiani
- Katia Vella
- Daniel Cronenberger
- TMS, Data Curation team