AIDA: combining text mining and e-science for bioinformatics

Marco Roos[1,a], Scott Marshall[1,b], Sophia Katrenko[1], Edgar Meij[1], Willem van Hage[2], Frans Verster[1], Pieter Adriaans[1]

Adaptive Information Disclosure: 1) Institute of Informatics, Faculty of Science, University of Amsterdam, The Netherlands;
2) TNO Industrie & Techniek / Vrije Universiteit Amsterdam, The Netherlands

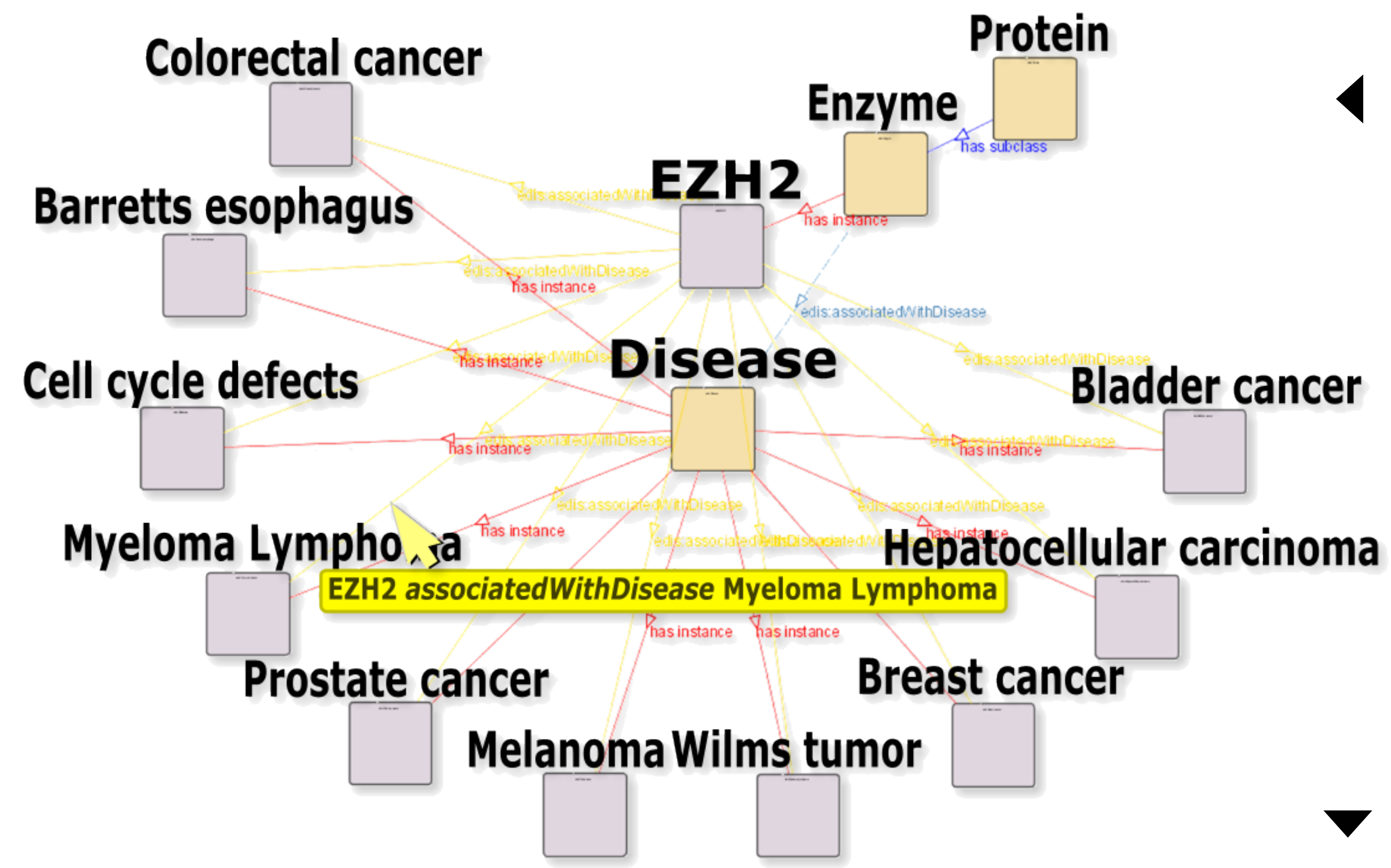
a) Project or Area Liaison (PAL) for OMII-UK (e.g. Taverna) b) Participating in W3C Semantic Web Health Care and Life Sciences Interest Group (http://www.w3.org/2001/sw/hcls/)

Correspondence: marshall@science.uva.nl, Website: http://adaptivedisclosure.org

Introduction

Our objective is to provide heuristic support for hypothesis construction from literature

We start with a seed of knowledge, a 'proto-ontology', that we want to extend. For example, information from a review article about histones and disease.



Snapshot of Jambalaya plug-in for Protégé/OWL showing relationships between Enhancer of Zeste Homologue 2 and various diseases.

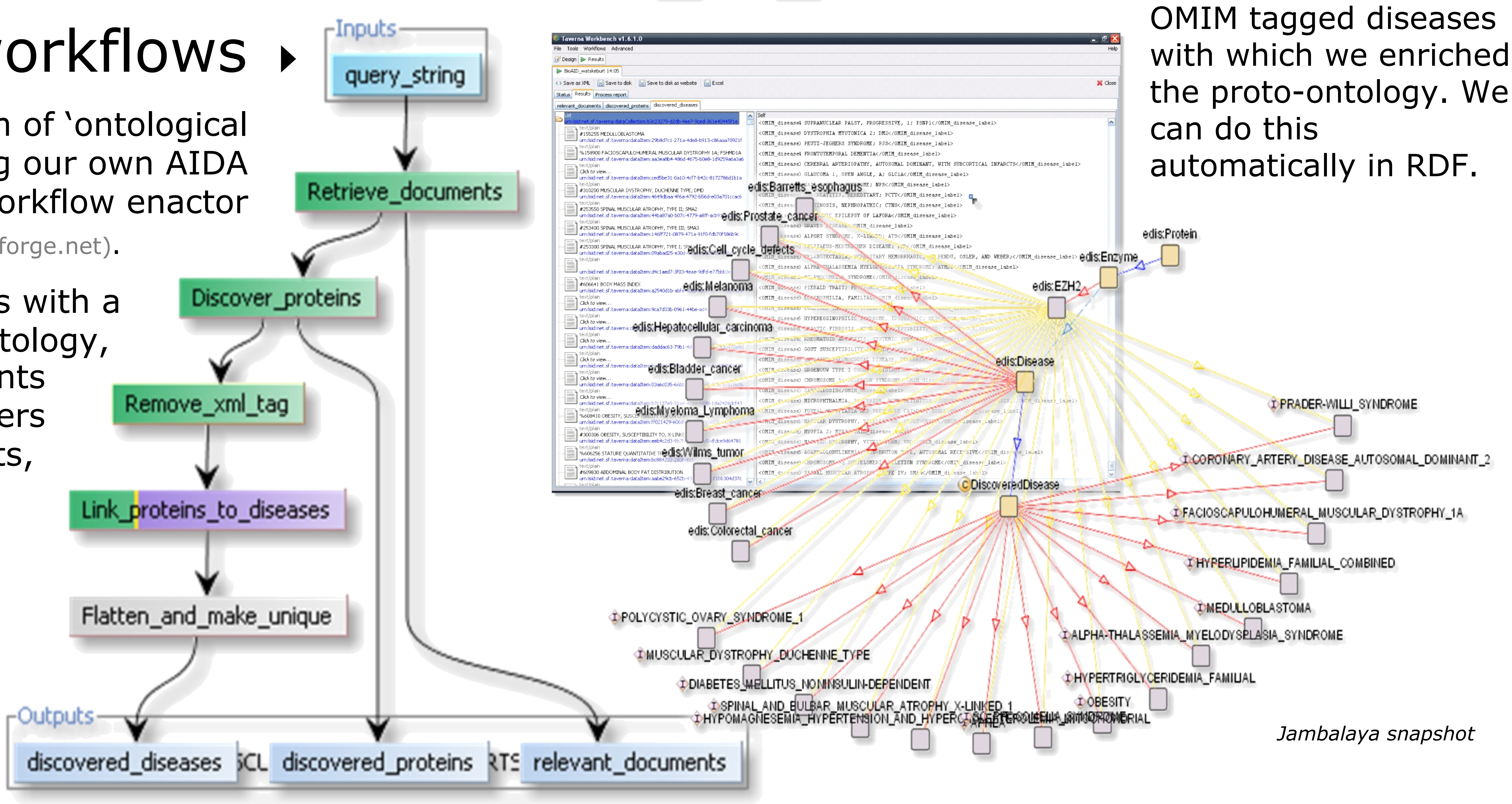
From table by Moss TJ, Wallrath LL, Mutation Research, Vol. 618, pp. 163-174.

Text mining workflows

We implemented a form of 'ontological enrichment' by connecting our own AIDA services and others in the workflow enactor tool Taverna (http://taverna.sourceforge.net).

The workflow here starts with a query from the proto-ontology, retrieves relevant documents from medline, discovers proteins in the abstracts, and then links these proteins to diseases using the OMIM database

OMIM service from the Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics (see http://xml.nig.ac.jp)


Inputs
query_string
Retrieve_documents
Discover_proteins
Remove_xml_tag
Link_proteins_to_diseases
Flatten_and_make_unique
Outputs
discovered_diseases
discovered_proteins
relevant_documents
edis:Protein
edis:Enzyme
edis:EZH2
edis:Disease
edis:Prostate_cancer
edis:Barretts_esophagus
edis:Cell_cycle_defects
edis:Melanoma
edis:Hepatocellular_carcinoma
edis:Bladder_cancer
edis:Myeloma_Lymphoma
edis:Wilms_tumor
edis:Breast_cancer
edis:Colorectal_cancer
DiscoveredDisease
PRADER-WILLI_SYNDROME
CORONARY_ARTERY_DISEASE_AUTOSOMAL_DOMINANT_2
FACIOSCAPULOHUMERAL_MUSCULAR_DYSTROPHY_1A
HYPERLIPIDEMIA_FAMILIAL_COMBINED
MEDULLOBLASTOMA
ALPHA-THALASSEMIA_MYELODYSPLASIA_SYNDROME
HYPERTRIGLYCERIDEMIA_FAMILIAL
OBESITY
POLYCYSTIC_OVARY_SYNDROME_1
MUSCULAR_DYSTROPHY_DUCHENNE_TYPE
DIABETES_MELLITUS_NONINSULIN-DEPENDENT
SPINAL_AND_BULBAR_MUSCULAR_ATROPHY_X-LINKED_1
HYPOMAGNESEMIA_HYPERTENSION_AND_HYPERCHOLESTEROLEMIA_MITOCHONDRIAL

Jambalaya snapshot

Results

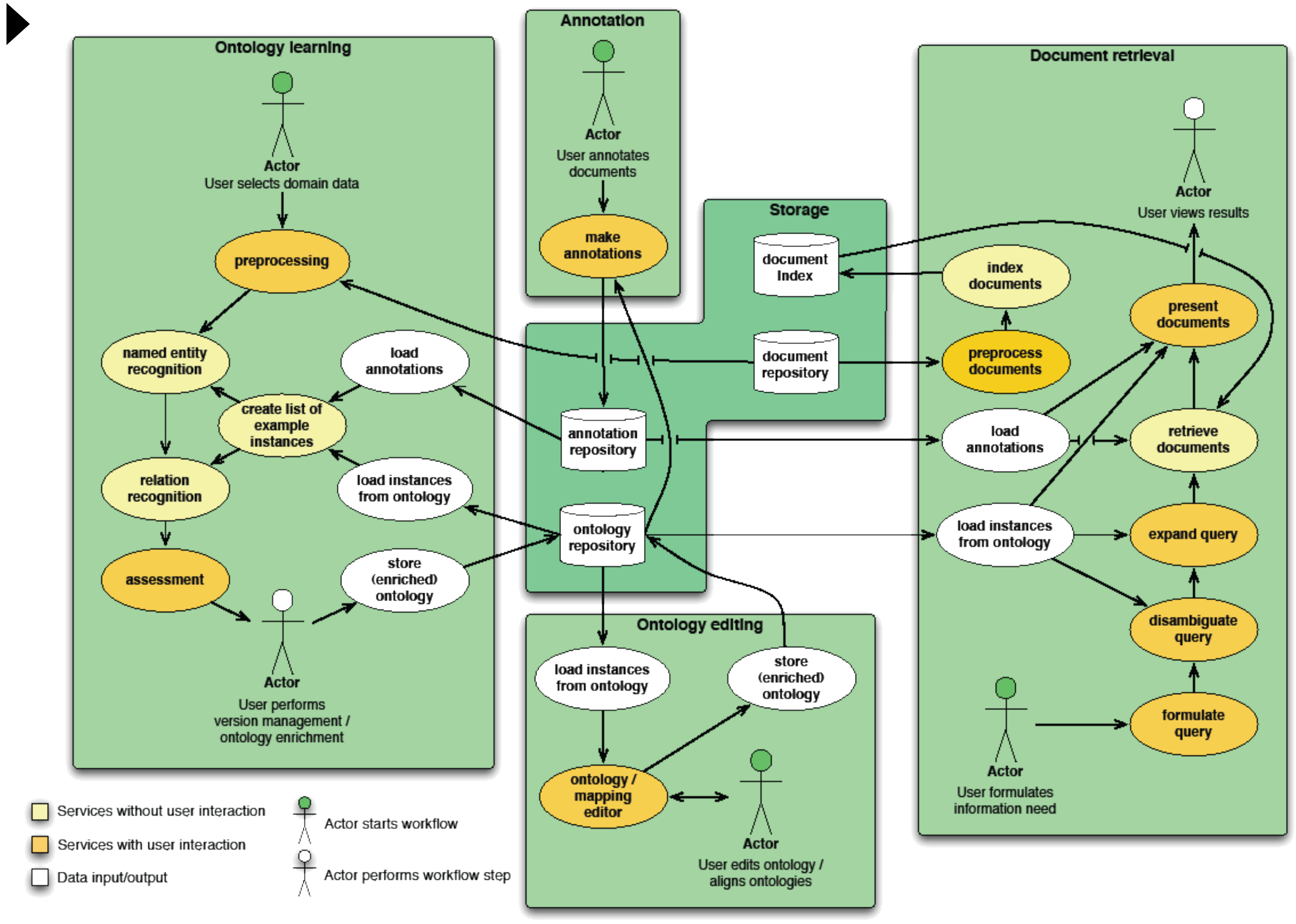
The workflow produces OMIM tagged diseases with which we enriched the proto-ontology. We can do this automatically in RDF.

The AIDA toolbox

An alternative to building special purpose bioinformatics tools is to use *web services* within a general computing environment such as the workflow tool *Taverna*.

Our workflows apply this 'e-science' approach to text mining. The components come from our 'AIDA* toolbox', which contains services that can be flexibly combined for a variety of applications, including various forms of text mining.

*Adaptive Information Disclosure Application



Discussion

Our example text mining workflow supports hypothesis construction by enriching an ontology with hypothetically related elements that were discovered from literature. We can explore various text mining strategies by adapting workflows and by adding new services from the AIDA toolbox, such as machine learning services to make discoveries in terms of one's own ontology, and semantic web services to steer data handling in terms of ontologies. The proportion of true positives returned (recall) can be improved, for instance, by incorporating synonym services. The AIDA toolbox is of general use. For instance it is also used in food informatics applications.

Acknowledgements and availability

We thank Taverna users and developers for technical assistance with Taverna.

Services, workflows, ontologies, and ontology-based data are available at http://ws.adaptivedisclosure.org , http://adaptivedisclosure.org , and http://rdf.adaptivedisclosure.org.

Our workflows are available from myexperiment.org

This work was carried out in the context of the Virtual Laboratory for e-Science program (VL-e) and the BioRange program. These programs are supported by BSIK grants from the Dutch Ministry of Education, Culture and Science (OC&W).