Modelling Digital Health Data: the ExaMode Ontology for Histopathology

Dennis Dosso [a], Manfredo Atzori [b,c], Svetla Boytcheva [d,h], Francesco Ciompi [e], Giorgio Maria Di Nunzio [a], Filippo Fraggetta [f], Fabio Giachelle [a], Stefano Marchesin [a], Niccolò Marini [b,g], Henning Müller [b], Todor Primov [d] and Gianmaria Silvello [a]

[a] *Department of Information Engineering, University of Padua, Italy*
E-mails: dennis.dosso@unipd.it, dinunzio@dei.unipd.it, giachelle@dei.unipd.it, stefano.marchesin@unipd.it, gianmaria.silvello@unipd.it
[b] *University of Applied Sciences Western Switzerland, HES-SO Valais, Switzerland*
E-mails: manfredo.atzori@hevs.ch, henning.mueller@hevs.ch, niccolo.marini@hevs.ch
[c] *Department of Neuroscience, University of Padua, Italy*
[d] *Sirma AI, Bulgaria*
E-mails: todor.primov@ontotext.com, svetla.boytcheva@ontotext.com
[e] *Radboud University Medical Center, The Netherlands*
E-mail: francesco.ciompi@radboudumc.nl
[f] *Azienda Ospedaliera Cannizzaro, Italy*
E-mail: filippo.fraggetta@aoec.it
[g] *Department of Computer Science, University of Geneva, Switzerland*
[h] *Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria*

Abstract. Histopathology is the gold standard for cancer diagnostics and its digital health counterpart – Computational Pathology – is gaining traction in the clinical practice for unlocking innovative approaches for patient care. In this context, data processing and learning are key aspects for the advancement of the field and ontologies are needed to model the domain of interest, standardize terminology and to make methods and services interoperable and reusable.

This paper presents the ExaMode ontology defining the classes and relationships concerning diagnosing four largely diffused and studied histopathology diseases: colon cancer, lung cancer, uterine cervix cancer, and celiac disease.

The ontology holistically models the classes and relations concerning these diseases and the medical context in which they are diagnosed. The concepts are divided into semantic areas about different aspects of the diseases and the diagnosis process: the patient and the clinical trial to which s/he might participate; the outcome of the diagnosis; the anatomical location, i.e., where the disease was found or from where tissues were taken; the procedures used to take the samples; the tests performed on the patients; and possible annotations containing further information about the disease. The ontological modeling was based on real-world anonymized clinical reports provided by two hospitals in Italy and The Netherlands.

Overall, the ExaMode ontology has been employed to develop automatic methods to extract pathological concepts from medical reports, which are used to annotate medical images associated with the records themselves. These are then used to train prediction algorithms aimed to improve clinical decision systems. Moreover, the presented ontology is currently used to improve systems for clinical support and triaging.

Keywords: Ontology, Linked data, Histopathology, Digital pathology

1. Introduction

In the past 20 years, we have witnessed significant volume and complexity increment in data production within the life science domain [1]. The progress in the high-throughput experimental techniques and the development of new diagnostic methods, therapies, and medications have become a precious information flow used to support medical applications [2].

Computational pathology is an emerging domain centered on computer-assisted diagnosis tools to automatically analyze histopathology data in the form of images and text.

This field revolves around *digital pathology*, a process where specialized hardware is used to generate substantial high-resolution digital images – i.e., Whole Slide Imaging (WSI) – of histological sections.

These images are visualized/analyzed on-screen, annotated by expert pathologists, processed by image analysis tools [3], and used to train Machine Learning (ML) models such as Convolutional Neural Networks (CNN) – the state-of-the-art method for medical image analysis [4]. CNNs, in particular, are used in the diagnostic process to make more precise assessments of findings [5, 6].

While, on the one hand, the number of hospitals using WSI is increasing, pursuing the collection of thousands of images and related diagnoses [7], on the other hand, analyzing the slides still requires up to one hour per image [8] and is subjected to low inter-pathologist agreement [9, 10].

The best performing CNN and other Deep Learning (DL) algorithms for image analysis require large volumes of high-quality annotated images to be successfully employed in clinical practice. The few large image datasets publicly available usually have some drawbacks preventing their use to train CNN, since annotations can be few, sparse, unbalanced (often including few normal tissue cases) [4] and highly variable depending on the pathologist that performed them.

One recently explored solution gaining traction in the field is to employ the diagnostic text (containing diagnoses and other expert observations) often associated with the WSI slides to extract so-called weak labels to automatically annotate the WSI [11].

Although structured reporting is being increasingly applied, much diagnosis information is still represented as *free text* and is used to record the health progression of patients [12]. The information contained in this free text is also invaluable for the clinical process. In this case, the process of knowledge extraction is often manual and thus highly time-consuming due to the high volumes of data, noise, and the lack of a standard shared terminology and structure between institutions [13, 14].

Natural Language Processing (NLP) methods are being developed to enable the efficient automatic processing of thousands of medical reports and the extraction of crucial information from text [13, 15, 16], which, in turn, can be helpful to support clinical tasks such as medical image analysis based on WSI [17]. Nonetheless, NLP is also shifting from using expertly curated rules to ML approaches, with the advantage of learning from data and generalizing to previously undiscovered patterns [18]. Therefore, NLP applications, as well as CNNs, require extensive training datasets to achieve a high degree of accuracy and robustness. In this case, it is necessary to deal with the *high data variability* in the text produced in clinical practice [19] and with the traditional wealth of terminologies in the medical domain [20].

In this context, great help comes from ontologies that, thanks to their formalism, are well-suited to represent a domain in a form that computers can handle [20–22]. Today, they are used in many applications relying on domain-specific terms, including NLP [1, 20] and deep learning [5]. Ontologies are a key means for annotating text to establish links between the concepts expressed in it, the class defined in the ontology, and the information associated with it. In this way, ontologies help reach a shared model of the reality of interest, thus overcoming data heterogeneity and integration problems. Much effort has been made to produce standard ontologies in the medical and biological domains [20, 23]. These fields have been benefiting greatly from their use also due to their impressive and rapidly growing number of terms, concepts, and definitions that traditionally characterize them [20].

Among the different disciplines in the life science field, *Histopathology* (and Digital Pathology) can significantly benefit from the use of ontologies to standardize the employed terminology and concepts and helping with (semi-)automating the knowledge extraction process [5]. While ontologies play an important role in the cancer domain [4, 22], as of now, there are no well-known and specific ontologies in the histopathology area [4]. Some example of ontologies in the area is the one proposed in [21], which represents concepts pertaining hematologic

malignancies; the Quantitative Histopathology Image Ontology (QHIO), which covers terms representing different types of histopathological images, imaging processes and techniques, and computational algorithms [24]; and the SNOMED CT (Systematic NOmedclature in MEDicine, Clinical Terminology) ontology [25], that was used in [4] as support for the annotation labeling process on two hundred WSIs for the colon and skin cancer cases.

This paper proposes a new ontology to organically define the classes and relationships concerning the diagnosis of four histopathology diseases largely diffused and studied: colon cancer, uterine cervix cancer, lung cancer, and celiac disease. The ontology holistically models the classes and relations concerning these diseases and the medical context in which they are diagnosed. The concepts are divided into semantic areas about different aspects of the diseases and the diagnosis process: the patient and the clinical trial s/he might participate; the outcome of the diagnosis; the anatomical location, i.e., where the disease was found or from where tissues were taken; the procedures used to take the samples; the tests performed on the patients; and possible annotations containing further information about the disease.

The ontology was developed in the context of the *ExaMode project*[1], co-financed by the European Commission[2]. The ontological modeling proceeded bottom-up starting from the anonymized clinical reports about the four considered diseases provided by Azienda Ospedaliera per l'Emergenza Cannizzaro (AOEC) in Italy and the Radboud University Medical Center (Radboud UMC) in The Netherlands. We analyzed these records and worked together with the pathologists and physicians, following shared co-design principles, to accurately identify the classes and relations to include in the ontology. We maximized the re-use of concepts defined in already available and well-known ontologies and vocabularies, thus limiting the creation of new classes and relations to a minimum.

Overall, the ExaMode project aims to allow weakly supervised knowledge discovery of multimodal heterogeneous data, limiting human interaction. important pathological concepts are extracted from medical reports, used to weakly annotate WSI associated with the records themselves, used to to train prediction algorithms such as CNN which finally translate to healthcare decision-making applications. Moreover, Chepygina et al., in their review [17] discuss different ML methods to learn with less or other types of supervision, such as weakly labeled images. The ExaMode ontology is inspired by these approaches and provides a new tool for accurately modeling the four diseases considered.

The paper aims to describe the proposed ontology, its creation, its main features, and how it can be used. The rest of the paper is organized as follows: Section 2 presents the related work; Section 3 presents the motivations of the ExaMode project and the background provided by the four considered diseases; Section 4 presents an overview of the ontology; Section 5 describes application scenarios where the ontology is currently employed; finally, in Section 6 we present some final remarks and future work.

2. Related Work

2.1. Medical terminologies and ontologies

In the domain of medicine there can be found different resources such as controlled vocabularies, taxonomies, thesauri, and ontologies [23]. According to the BioPortal[3] - one of the most comprehensive repositories with biomedical ontologies - there are more than 900 ontologies in the domain.

Among the most famous resources in the medical field belonging to one of these categories, we count:

– *ICD - the International Classification of Diseases.* Started in 1880, the ICD was the only medical terminology resource for a long period of time [26], and was maintained by the World Health Organization (WHO) since 1946. There are available official translations of ICD on the majority of the European languages of the most widely used version ICD, 10th revision. It has now reached its 11th revision (starting from January 2022). Its scope is always extending, and it includes injuries, external causes of health problems, signs and symptoms.

[1] https://www.examode.eu/
[2] https://cordis.europa.eu/project/id/825292
[3] https://bioportal.bioontology.org/

It worth to mention also one special version of ICD with special focus to oncology diseases - ICD-O, 3rd revision. [4]

- *Systematized Nomenclature of Medicine - Clinical Terms.* SNOMED-CT, often only SNOMED [25], is a widely-used clinical terminology, consisting on more than 311k concepts and more than 1 million relationships between them, created to cover the whole patient record. SNOMED is a hierarchy of concepts that includes high level categories such as: procedures, pharmaceutical/biological products, specimen, event, environmental/geographical location, body structure, organism, and others.
- *MeSH - Medical Subject Headings.*[5] A vocabulary thesaurus edited and maintained by the US National Library of Medicine, used for indexing, cataloging, and searching of biomedical and health-related information. It includes the subject headings appearing in the MEDLINE/PubMED database, the U.S. National Library of Medicine (NLM) catalog, and other NLM databases.
- *UMLS - Unified Medical Language System.*[6] It is a metathesaurus initiated by the US by the National Library of Medicine (NLM) in 1986. It is a controlled and hierarchically-organized vocabulary that integrates and distributes key terminology, classifications, coding standards, and associated resources. Its aim is to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.
- *OBO - the Open Biomedical Ontologies.*[7] Open Biomedical Ontologies (OBO) was created in 2003 and it has been evolving constantly as a library of online, public-domain biomedical ontologies. It is an expanding family of ontologies designed to be interoperable and logically well formed and to incorporate accurate representations of biological reality [27].
- *NCI resources.* The National Cancer Institute (NCI) has developed the *NCI Metathesaurus*, containing most terminology used by the NCI for clinical care, translational and basic research, together with public information and administrative activities [28]. It is an extensive biomedical terminology database with 1.4M concepts mapped to 3.6M terms with 17M relationships.The *NCI Thesaurus* is a public domain description logic-based terminology that supports rich semantic interrelationships between the nodes of its taxonomies, evolved from the Metathesaurus. More than an ontology, it is a nomenclature with ontological features, providing definitions, synonyms, and relevant information of cancers and related diseases.
- The Gene Ontology (GO) ontology[8] [29] is the result of a major community-based bioinformatics project to provide structured and controlled vocabularies and classifications that cover several domain of molecular and cellular biology

2.2. Cancer-related ontologies

Ontologies are an important component of numerous biomedical tools and applications within the cancer domain [21]. Ontologies support text mining applications, clinical decision support systems, the analyzing of adverse events, and the targeting of cancer drugs [30–32].

One of the most successful ontology in cancer research is the GO [29] thanks to its widespread use. Other ontologies have seen employed in cancer research, such as the Foundational Model of Anatomy used to annotate biomarkers for brain tumors [33]. The TMN-O ontology represents the staging systems [34]; the ENT COBRA ontology deals with cancer treatments and brachytherapy [35]; while the Profile Ontology for Adolescent and Young Adult Cancer Survivors deals with after-care treatment plans that enhance patient engagement [36]. The NCI Thesaurus is one of the most widely used resources within the field of cancer: it covers 11K terms in 36K concepts in the cancer research domain, and arose from a need to integrate varied data systems through a unified coding system [21].

[4] https://apps.who.int/iris/bitstream/handle/10665/96612/9789241548496_eng.pdf
[5] https://www.nlm.nih.gov/mesh/meshhome.html
[6] https://www.nlm.nih.gov/research/umls/index.html
[7] http://www.obofoundry.org/
[8] http://www.geneontology.org/

2.3. Ontology repositories in the life science domain

The use of standard identifiers for classes and relations is a key component to ensure data integration across multiple disconnected databases, files or web sites [1], and there are today sufficiently stable ontologies to permit routine re-use of classes from multiple ontologies [37]. The ability to use parts of ontologies to generate new ones specifically tailored for some new application or context, while at the same time maintaining interoperability with other data sets, is now essential [1]. One of the first biological ontology to be developed, GO [29], had exactly the aim to help connect the many datasets emerging in the early 2000s, and in 2010 the 33% of ontologies had at least half of their concepts mapped to concepts in other ontologies [38].

An ontology-based annotation associates an entity to an ontology class. This annotated entity can then be represented by an identifier in a database and be combined with different types of data, such as provenance information, used for many tasks such as quality control. Data annotation and integration are often performed together, and in particular for complex multimodal datasets single ontologies are no more sufficient to perform the annotation [1]. When it becomes necessary to use more than a single ontology for the annotation process, ontology repositories can aid in finding ontologies suitable for annotating data within a domain. Among the main resources in the life science domain, to find the needed annotations for the entities of the ExaMode ontology we primarily referred to *OntoBee*[9] [39], an ontology repository in which ontologies are presented as Linked Data. It provides information about the classes and relations used by the OBO project [40].

BioPortal[10] is the largest repository for ontologies in biology and medicine, containing more than 400 ontologies and also accessible through a SPARQL endpoint; the Ontology Lookup Service[11], consisting in ontologies represented in the OBO Flatfile Format; and the OBO Library[12], composed of a set of ontologies developed accordingly to a set of agreed specifications.

3. Motivation and background

It is widely known that ever-increasing volumes of diverse data from distributed hystopathologic sources are continuously produced. Healthcare data stand out in size – production was expected to be over 2K exabytes in 2020; in heterogeneity – due to the different types of media and acquisition methods involved; in the included knowledge – for example, the diagnosis produced for the patients; and, finally, for its commercial value.

Among the current challenges in medical imaging, and more specifically in digital pathology, there is the management of the highly heterogeneous data available to train robust deep learning models and the need to overcome the lack of locally-annotated (or strongly-annotated) datasets [41].

The lack of labeled data, which are expensive and time-consuming to produce [42], precludes models from extracting knowledge and value from them, and it persists despite the increasing amount of publicly available datasets, such as The Cancer Genome Atlas (TCGA) [43] or The Cancer Imaging Archive (TCIA) [44]. While deep CNN models are currently the backbone of the state-of-the-art methods to analyze WSI [45], large locally-annotated datasets usually allow to train models that can reach the highest performance and that can generalize better to unseen new data [42]. Thus, one of the main objectives of the ExaMode project is to provide automatic and semi-automatic methods to improve the efficiency and the effectiveness of the diagnoses in the pathology domain to reduce the pathologists' workload. Ontologies allow us to formalize and exploit the extensive, unstructured, multimodal, and often distributed data contained in clinical records [1], making it possible to capture and represent information and knowledge related to diseases, diagnostic procedures, drugs, and other aspects of the medical domain in a machine-interpretable way.

ExaMode focuses on *histopathological diagnosis* of tissues to detect cancer-related diseases. Taking into account the future cancer incidence and mortality burden worldwide, which is predicted to be increasing by 63% from 2018

[9] http://www.ontobee.org/
[10] http://bioportal.bioontology.org/
[11] http://www.ebi.ac.uk/ontology-lookup/
[12] http://obofoundry.org

until 2040 [46], it has been decided to focus on four high-incidence diseases: (i) Colon cancer; (ii) Cervix cancer; (iii) Lung cancer; and (iv) the Celiac disease.

3.1. Colon Cancer

The estimated number of colon cancer incidence from 2018 to 2040 is going to increase by up 75%, for both sexes and all ages [47]. The American Cancer Society (ACS) recommends regular screening for colon cancer for people over 45 years [48]. The screening can be done either with a stool-based molecular tests or with a visual exams, so the at this stage the screening process does not include histopathological examination. The majority of colorectal cancers derives from precursor lesions which can be identified using endoscopic procedure (colonoscopy), leading to excision of these lesions, known as *polyps* [49]. Good endoscopic practice, together with an accurate histopathological diagnosis, decreases the incidence of colorectal cancer. There are different precursor lesions with different diagnostic and prognostic significance [50].

The main task for a pathologist is to detect cancerous polyps (e.g., for population screening) and to identify the answers to these main diagnostic criteria:

1. Presence or absence of polyp;
2. Presence of adenoma-serrated polyps/malignant polyps;
3. The number of polyps observed (for each type);
4. Presence or absence of dysplasia;
5. Presence of high or low grade dysplasia.

Moreover, in the microscopic analysis of colon excisional biopsy sample, there is a minimum of data that also need to be provided by the pathologist:

1. *Type* of the polyp. It is important to distinguish between the two main types of polyps: adenoma-serrated polyps and malignant polyps. From the diagnostic perspective, it can be also useful to know the *number* of each type of polyps.
2. The *grade* of the dysplasia, if is present (low, medium or high grade).
3. In case of malignant polyps (considered as cancer, e.g. Colon Adenocarcinoma or Metastatic Adenocarcinoma), several critical *histological features* need to be assessed, which include tumor type, histological tumor grade, lymphovascular invasion and margin involvement.

This information is a prognostic factor leading to the decision about the patient's management. For example, polyps with a negative polypectomy margin, low grade histology, and no lymphovascular invasion can be safely treated with endoscopic polypectomy, whereas positive margin, high grade (poorly differentiated) histology, and lymphovascular invasion are associated with an increased risk of adverse outcomes and surgical resection is indicated [51].

3.2. Lung Cancer

According to the International Agency for Research on Cancer, the number of Lung Cancer (LC) in both sexes and at all ages is estimated to increase by 72% from 2018 to 2040. In general, lung cancer is the second most common cancer in both men and women - about 13% of all new cancers diagnoses are lung cancers. Moreover, lung cancer is the leading cause of cancer death among both men and women (18% of all cancer deaths), being the leading cause of cancer death in men [52, 53].

The average overall survival rate for metastatic lung cancer is very low, whereas early stage has higher survival rates. The treatment of low-stage lung cancer is complete surgical resection. Instead, for metastatic lung cancer the surgical option is often impossible. An accurate diagnosis from lung biopsies targets the most correct prognostic and therapeutic management of the patients.

Moreover, a correct World Health Organization (WHO) classification is very important for metastatic tumors since there is therapeutic implication of distinguishing histological subtypes such as adenocarcinoma and squamous cell carcinoma. The identification of new therapeutic targets over the past decade resulted in an urgent need for a

classification system for both non-resection specimens (particularly small biopsies) and cytology samples. For this reason an accurate and specific pathology report is important to establish diagnosis and patient's treatment.

Starting from the analysis of lung biopsies, microscopic analysis section of the clinical report on lung cancer biopsy sample must provide the following information, with prognostic and predictive implications:

1. Histologic type
2. Histologic grade
3. Spread Through Air Spaces (STAS) – information about the presence of micropapillary clusters, solid nests or single cells of tumor extending beyond the edge of the tumor into the air spaces of the surrounding lung parenchyma
4. Visceral pleura invasion
5. Direct invasion of adjacent structures
6. Margins – information about involvement of the tissue margins, indicating a negative outcome
7. Lymphovascular invasion - provides information about vascular/lymphatic vessel invasion
8. Pathologic stage classification – based on the classification system proposed by the WHO
9. Extranodal extension – indicates presence of metastasis.

3.3. Uterin Cervix Cancer

Cervical cancer is the fourth most common cancer in women, and the eight most commonly occurring cancer overall[13]. Approximatively 570K cases of cervical cancer and 311K deaths from the disease occurred in 2018 [54].

The cervical biopsy (colposcopy) is a procedure made when previous tests provide evidence of precancerous/abnormal or neoplastic lesions in the uterine cervix. The cervical tissue removed has to be analyzed by an expert pathologist to identify if the tumor lesions are present or not. If present, the pathology report provides not only the diagnostic information, but it works also as a prognostic tool for the patient's treatment. Colposcopy with directed biopsy is currently one of the "gold standard" practices for the diagnosis of cervical pre-cancer [55].

Thus, the aim of the histopathologist is to recognize and identify these precursor lesions well known as Cervical Intraepithelial Neoplasia (CIN), which displays proliferation of atypical basaloid cells [56]. Based on proliferation spread, the WHO classification categorizes this dysplasia into three grades:

1. CIN 1 (Mild Dysplasia)
2. CIN 2 (Moderate Dysplasia)
3. CIN 3 (Severe Dysplasia or Carcinoma in Situ)

CIN1 corresponds to Low-Grade Squamous Intraepithelial Lesion (LSIL), whereas CIN2-3 correspond to High-Grade Squamous Intraepithelial Lesion (HSIL). A strong association between these precursor lesions and HPV infection has been observed, where LSIL is strongly associated with low intermediate risk HPV, and HSIL is associated to high risk HPV [57]. Therefore, the first feature that has to be identified and reported is the presence and the grade of dysplasia with possible HPV association. In the presence of cervical carcinoma main microscopic features and measurements of uterine cervix colposcopy biopsy are identified, and they should be provided in the pathology report:

1. Histologic Type
2. Histologic Grade
3. Stromal Invasion – provides information about cancer invasion into stromal tissue
4. Margins - indicate a negative outcome
5. Lymphovascular Invasion – provides information about vascular/lymphatic vessel invasion.

Also, the immunohistochemistry (p16 and Ki-67 staining) assists in the histological differential diagnosis of precursors to reactive and metaplastic epithelium. For invasive cervical carcinoma, stage is the strongest prognostic factor [56].

[13] https://www.who.int/health-topics/cervical-cancer#tab=tab_1

3.4. Celiac Disease

Celiac Disease (CD) is one of the most common diseases, resulting from both environmental (gluten) and genetic factors [58]. It is now recognized as a global disease affecting about 0.7% of the world's population [59]. Because of these reasons, CD was chosen as non-cancerous disease to be included in the ExaMode priority list. CD is an immune-mediated disease with chronic outcome and genetic predisposition to an intolerance to gluten and its proteins. This intolerance leads to abnormal immune response, followed by a chronic inflammation and alteration of the small intestinal mucosa. The diagnosis of this pathology is based on the description of the histopathological alterations of small intestine (after duodenal biopsy) by expert pathologists [60].

Microscopic analysis of small colon biopsy sample for celiac disease provides information about:

1. Orientation of biopsy — indicates biopsy position on cellulose acetate filter and is very important for the diagnostic criteria;
2. Normal intestinal mucosa description – includes information about: villi, enterocytes, intra-epithelial lymphocytic infiltrate and Glandular crypts. The absence or alteration of these structures must be reported;
3. Pathological intestinal mucosa – including features which have to be reported and well described, with particular attention to increased intraepithelial T lymphocytes, decreased enterocyte height, crypt hyperplasia and villous atrophy;
4. Pathologic Stage Classification – based on the classification system proposed by Marsh-Oberhuber and Corazza-Villanacci, in presence of intestinal mucosa alterations previously described.

4. The ExaMode Ontology: an overview

We represent the ontology as a graph where nodes are classes and edges are typed relationships amongst the classes. Classes (nodes) represent real-world objects such as a person, a project, a tissue or an anatomical part. Relationships (edges) describe how the classes interact one with each other. We limited the creation of new classes by maximizing the re-use of existing ontologies. The ExaMode ontology is composed of 124 classes, 89 named individuals, 22 object properties, 13 data properties, and 27 annotation properties. The prefixes used in the ontologies are reported in Table 1. Amongst the most used external ontologies, we count the Mondo Disease Ontology[14], a semi-automatically constructed ontology that merges in multiple disease resources, aiming to harmonize disease definitions across the world; Uberon, an integrated cross-species ontology covering anatomical structures in animals species, with a focus on vertebrates[15] [61]; and the NCI Thesaurus OBO Edition (NCIt)[16], a reference terminology that includes broad coverage of the cancer domain, including cancer related diseases and cell abnormalities. In many cases, when possible, we also annotated each class with its corresponding URL in the Unified Medical Language System (UMLS) metathesaurus[17], a large biomedical thesaurus that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems [62]. The URLs of the ontology are secure and permanent by using the re-direction service provided by the W3 Permanent Identifier Community Group[18]. The service works as a switchboard connecting requests for information with the true location of the information on the Web. It can be, therefore, reconfigured to point to a new location if the old location stops working.

The ExaMode Ontology is publicly available in several serialization formats along with a Web-based documentation at https://w3id.org/examode/ontology.

The ExaMode ontology is organized in five main semantic areas:

1. **ExaMode cases.** This area, represented in Figure 1, contains classes that are common to all clinical records for all four pathological cases (three cancer types and the celiac disease) considered in ExaMode. These

[14] http://purl.obolibrary.org/obo/mondo.owl, https://github.com/monarch-initiative/mondo
[15] https://uberon.github.io/about.html
[16] https://www.ebi.ac.uk/ols/ontologies/ncit
[17] https://www.nlm.nih.gov/research/umls/index.html
[18] https://w3id.org/

Table 1
Prefixes used in the ExaMode Ontology.

Prefix	URL
bto:	http://purl.obolibrary.org/obo/BTO_
cl:	http://purl.obolibrary.org/obo/CL_
dc:	http://purl.org/dc/elements/1.1/
doid:	https://www.ebi.ac.uk/ols/ontologies/doid/
exa:	https://w3id.org/examode/ontology/
fabio:	http://purl.org.spar/fabio/
fma:	http://purl.obolibrary.org/obo/FMA_
foaf:	http://xmlns.com/foaf/0.1/
go:	http://purl.obolibrary.org/obo/GO_
hp:	https://hpo.jax.org/app/browse/term/HP:
mondo:	http://purl.obolibrary.org/obo/MONDO_
mp:	http://purl.obolibrary.org/obo/MP_
ncit:	http://purl.obolibrary.org/obo/NCIT_
oae:	http://purl.obolibrary.org/obo/OAE_
oba:	http://purl.obolibrary.org/obo/OBA_
oboinowl:	http://www.geneontology.org/formats/oboInOwl#
omit:	http://purl.obolibrary.org/obo/OMIT_
owl:	http://www.w3.org/2002/07/owl#
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
skos:	http://www.w3.org/2004/02/skos/core#
symp:	http://purl.obolibrary.org/obo/SYMP_
uberon:	http://purl.obolibrary.org/obo/UBERON_
umls:	http://linkedlifedata.com/resource/umls/id/
uniprot:	https://www.uniprot.org/uniprot/
vt:	http://purl.obolibrary.org/obo/VT_
xml:	http://www.w3.org/XML/1998/namespace
xsd:	http://www.w3.org/2001/XMLSchema#

classes describe the patients with the basic information usually available for public display as gender, age, and the disease. Each use case is classified on the basis of the disease it analyses: colon cancer, lung cancer, cervical carcinoma and celiac disease, represented through four named individuals. Each of the two organizations collaborating in the project, the Cannizzaro Hospital (AOEC) in Catania (Italy), and the Radboud UMC in Nijmegen (The Netherlands) provided datasets of clinical case records. Each dataset contains several `exa:ClinicalCaseReport` instances. The "Clinical Case Report" class has four subclasses: one for each disease. In the example of Figure 1 it is reported the subclass `exa:ColonClinicalCaseReport`, representing the set of all instances of clinical case reports about a case of colon carcinoma.

2. **Diagnosis.** This area contains the classes related to the diagnosis which are extracted from the medical reports. The main class is `exa:Outcome`, specialized into three main sub-classes: `exa:NegativeResult`, `exa:PositiveOutcome` and `exa:InconclusiveOutcome`. "Positive outcome" has many subclasses, one for each type of issues detected in the diagnosis depending on the disease at hand. This is a taxonomy of the possible diagnoses that can be found in the reports. For each clinical case report only one positive outcome subclass can be instantiated at each time. A graphical representation of this part of the ontology, pertaining lung cancer, is given in Figure 2.
3. **Annotation.** It may be the case that a clinical case report contains further information about the status of the disease beyond the positive outcome diagnosis itself, e.g., the presence of necrosis in a case of lung carcinoma. This area was designed to describe this type of information. It is composed by named individuals and

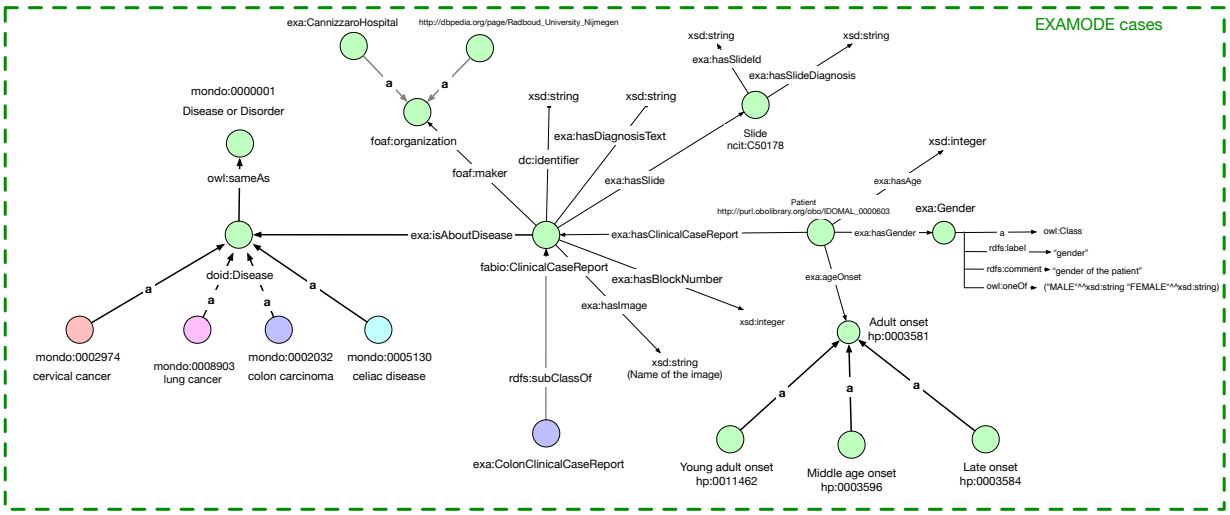

Fig. 1. The four main pathological diseases on which the ExaMode ontology is focusing.

classes that are instances/subclasses of the `exa:Annotation` superclass. Figure 2 shows how a case of lung carcinoma may be annotated with the presence of necrosis and/or metastasis.

4. **Anatomical location/Topography.** This area contains the classes modeling the taxonomy of the possible *anatomical locations* where the "disease" can be located. This information is extracted from the medical reports. These classes also describe where the tissue was withdrawn for the analyses via the `hasTopography` property. Figure 4 presents the part of the ontology composed, among other things, of the Anatomical location area pertaining cervix cancer.
5. **Procedure.** This area contains the classes that model the procedure adopted to obtain the material where the analysis was performed. The main class is `ncit:C25218`, "Intervention or Procedure". Figure 4 presents the procedure area for the cervix cancer.
6. **Test.** It may be the case that the patient undergoes some test to ascertain the presence or characteristic of their disease. The classes in this area describe the possible tests.

4.1. ExaMode cases

`exa:ClinicalCaseReport` is the central class of this part of the ontology. All the modeled medical reports are instances of subclasses of this one, such as the `exa:ColonClinicalCaseReport` example in Figure 1. Each clinical case is associated to one of the four diseases, that in the ontology are represented with four named individuals, instances of `doid:Disease`, a class representing a general disease, through the `exa:isAboutDisease` relationship. These four named individuals are `mondo:0002974` (cervical cancer), `mondo:0008903` (lung cancer), `mondo:0002032` (colon carcinoma), and `mondo:0005130` (coeliac disease).

Figure 1, in particular, represents the subclass `exa:ColonClinicalCaseReport`, associated to the named individual `mondo:0002032`, representing the colon cancer. Each instance of Clinical Case Report presents a unique identifier (through the `dc:identifier` data property) and a the diagnosis as it is found in the report (through the `exa:hasDiagnosisText` property). Each medical report may also be related to an WSI file. In the ontology this WSI slide is represented with an instance of `ncit:C50178` (Slide). The block number of the report, identified by the `exa:hasBlockNumber`, refers to internal ids related to the reports or the images.

All medical reports are associated with a patient (normally anonymized), and a single patient can have more than one associated medical report. There is minimal patient information represented in the ontology, namely the age at the time of the report, the gender, and an age onset to classify the patients into three categories, represented in the ontology as three named individuals: young adult (`hp:0011462`), middle age (`hp:0003596`), and late (`hp:0003584`).

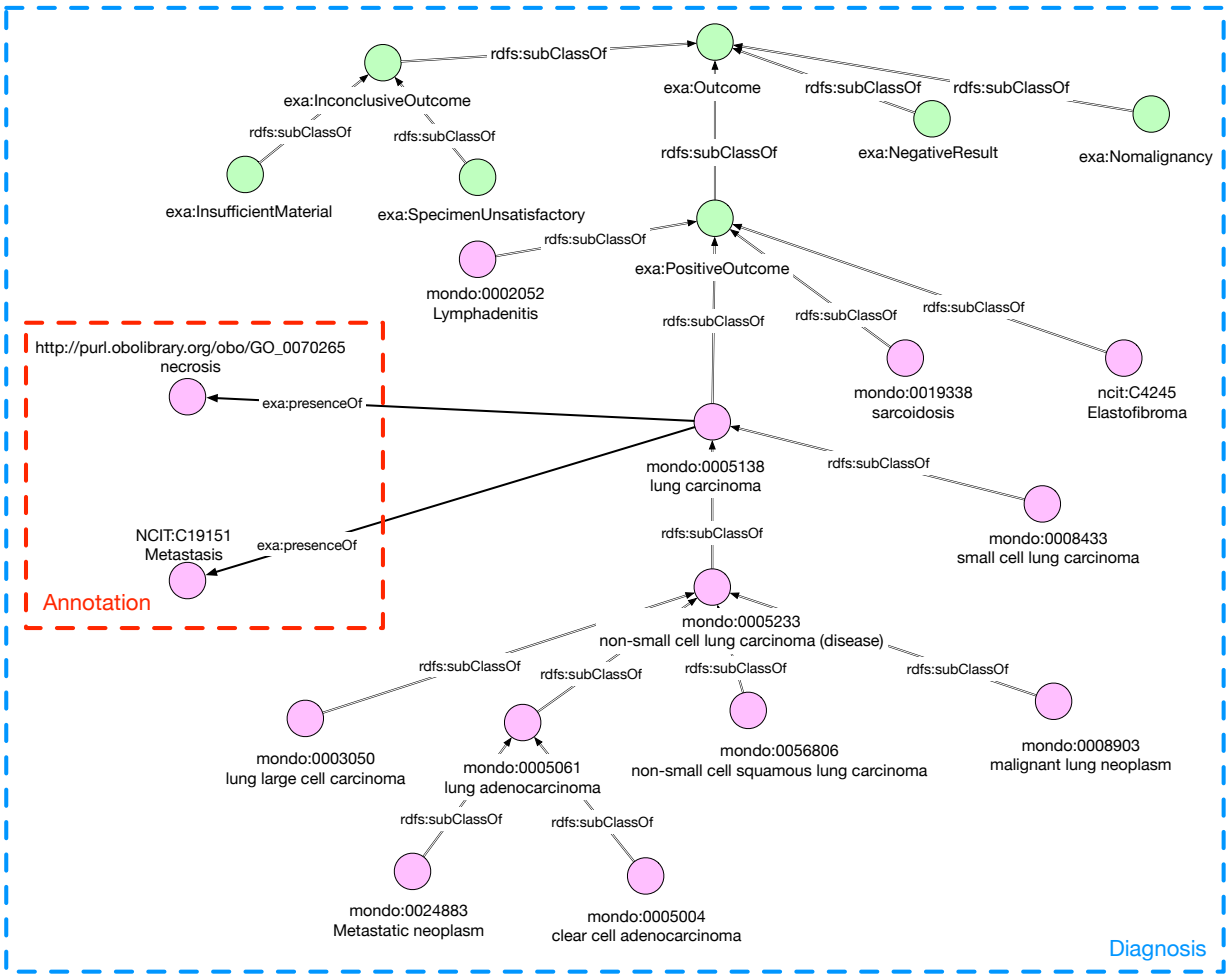

Fig. 2. Graphical representation of the Diagnosis and Annotation areas of the Ontology pertaining Lung Cancer.

All medical reports are also associated with the Organization that produced it (`foaf:Organization`, connected through the `foaf:organization` property). Currently in the ontology two organizations are represented as named individuals: AOEC and Radboud UMC.

4.2. Diagnosis

Each of the four diseases presents a different set of classes in this semantic area, organized in different taxonomies deriving from the nature of the disease.

Some of these classes at the top of the taxonomies, however, describe the general outcome of a diagnosis, and are common to all cases. The root of each taxonomy is the general `exa:Outcome` class. This, in turn, can be specialized in a positive outcome (`exa:PositiveOutcome`), negative outcome (`exa:NegativeOutcome`), an outcome where no evidence of malignancy was found (`exa:NoMalignancy`), or be an inconclusive outcome (`exa:InconclusiveOutcome`). This last case can be further specialized in two cases: one due to the presence of insufficient material to reach a conclusion (`exa:InsufficientMaterial`) or the other due to an unsatisfactory specimen for diagnosis (`exa:SpecimenUnsatisfactory`).

Colon Cancer. In this case the Positive Outcome is further specialized in many subclasses: Metastatic Adenocarcinoma (`ncit:C4124`), Colon Adenocarcinoma (`mondo:0002032`), Colitis (`mondo:0005292`), Ul-

cer (ncit:C3426), Granuloma (oae: 0001850) and the more general Polyp of Colon (mondo:0021400). The polyp of colon class has, in turn, different subclasses: Colon Hyperplastic Polyp (ncit:C4930), Colon Inflammatory Polyp (mondo:0006152), and the Adenoma (mondo:0006498). Adenoma, in turn, can be of different types, each represented with a subclass: Colon Tubular Adenoma (ncit:C7041), Serrated Adenoma (ncit:C38458), Colon Villous Adenoma (mondo:0021271), and Colon Tubulovillous Adenoma (ncit:C5496).

Lung Cancer. Figure 2 presents the part of the ontology pertaining the diagnosis of the lung cancer. In this case, positive outcome presents as subclasses: lymphadenitis (mondo:0002052), lung carcinoma (mondo:0005138), sarcoidosis (mondo:0019338), and elastofibroma (ncit:C4245).

Depending on the type of cells being found, the lung carcinoma can then be sub-classed in small cell lung carcinoma (mondo:0008433) and non-small cell lung carcinoma (mondo:0005233). This last class is further sub-classed in lung large cell carcinoma (mondo:0003050), non-small squamous lung carcinoma (mondo:0056806), malignant lung neoplasm (mondo:0008903), and lung adenocarcinoma (mondo:0005061), which can manifest itself as a metastatic neoplasm (mondo:0024883) or clear cell adenocarcinoma (mondo:0005004).

Uterine Cervix Cancer. The case of the cervix cancer presents one of the most articulated taxonomy of possible diagnosis. A positive outcome can be classified as: cervical polyp (mondo:0000751), focal acute inflammation (ncit:C82967), cervicitis (mondo:0002345) and its subclass chronic cervicitis (ncit:C27057), cervical intraepithelial neoplasia (commonly referred as CIN, mondo:0022394), cervical carcinoma (mondo:0005131), condyloma (ncit:C2960), metaplasia (ncit:C3236) and its subclass squamous metaplasia (ncit:C3237), atrophic vulva (mondo: 0001932), acanthosis (ncit:C35265), dyskeratosis (ncit:C62570), and hyperkeratosis (ncit:C35541).

The class CIN is further sub-classed in low grade cervical squamous intraepithelial neoplasia (CIN I) (ncit:C40196), and cervical intraepithelial neoplasia grade 2/3 (CIN 2/3) (mondo:0006137). This last has two other sub-classes: cervical squamous intraepithelial neoplasia 2 (CIN II) (ncit:C40198) and squamous carcinoma in situ (CIN III) (mondo:0004693).

The class cervical carcinoma has, as sub-classes, the following: cervical squamous cell carcinoma (mondo:0006143), the uterine cervix carcinoma in situ (mondo:0042487), and cervical adenocarcinoma (mondo:0005153), with its sub-class cervical adenocarcinoma in situ (ncit:C45420).

Coealiac Disease. For the diagnosis of the coeliac disease, the outcome is fairly simpler that the one of the other three diseases: if the patient is positive, they can either have the coeliac disease (exa:PositiveToCoealiacDisease) or the Duodenitis (mondo:0004627), an inflammation which does not necessarily imply coeliac disease.

4.3. Annotation

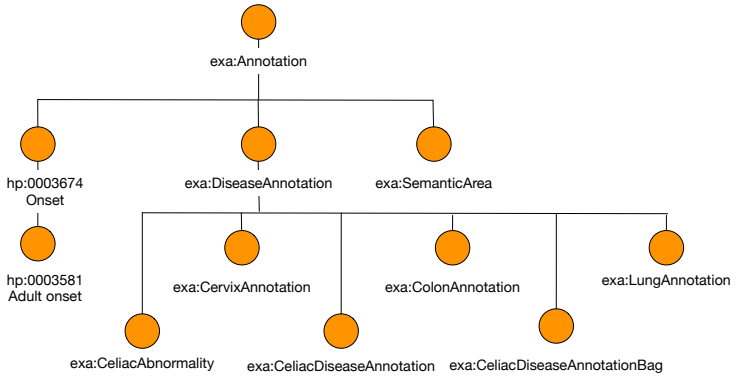

Fig. 3. Taxonomic hierarchy of the classes used to represent annotations for the clinical cases. Each line represents a relationship rdfs:subClassOf.

It is often the case that the clinical case reports are annotated with additional information about the presence or absence of some characteristic of the disease, or describing the evolution of the disease itself. To represent this type of information we defined a new class, called `exa:Annotation`, and different sub-classes depending on the necessary information to be modeled. Figure 3 reports the hierarchy of classes that we defined in this way.

The Onset (`hp:0003674`) class is used to represent age groups. As of now we defined the subclass Adult Onset (`hp:0003581`) with three related named entities: Late onset (`hp:0003584`, 60+ years), Middle age onset (`hp:0003596`, 40-60 years), and Young adult onset (`hp:0011462`, 16-40 years). Patients are connected to one onset through the `exa:hasAgeOnset` property.

The class `exa:SemanticArea` is used to annotate the different classes of the ontology with a corresponding area. This class contains named individuals that represent the areas described in this section, namely: General Entity (`exa:General`), to represent the classes belonging to the ExaMode disease cases areas; Diagnosis (`ncit:C15220`); Anatomical Location (`ncit:C13717`), Procedure (`exa:Procedure`); and Test `exa:Test`. All of these are named individuals. The classes of the ontology are connected to their semantic area through the `exa:hasSemanticArea` property.

Finally, class `exa:DiseaseAnnotation` contains other subclasses, each of one is used to represent types of information pertaining to the different diseases, as described in the next section.

Colon cancer. In the case of colon cancer an outcome instance of type "polyp of colon", or one of its subclasses, may highlight the presence of a form of dysplasia. If this is the case, the instance may be annotated, through the object property `exa:hasDysplasia`, to a named individual representing a type of dysplasia. These named individuals are all instances of the class `exa:ColonAnnotation`, and they are colon dysplasia (`ncit:C4847`) and its subclasses, namely mild colon dysplasia (`ncit:C4848`), moderate colon dysplasia (`ncit:C4849`), and severe colon dysplasia (`exa:SevereColonDysplasia`).

Uterine Cervix Cancer. In the case of the cervix cancer, a positive outcome may be accompanied by the presence of the Human Papilloma Virus (HPV) (`ncit:C14226`), and of koilocytotic squamous cells (`ncit:C36808`). Both of these in the ontology are named individuals of type cervix annotation (`exa:CervixAnnotation`). They may be respectively connected to an instance of positive outcome through the properties `exa:detectedHumanPapillomaVirus`, and `exa:koylociteDetected`.

Lung cancer. In the case of a diagnosis of type "lung carcinoma", as can be seen in Figure 2, the outcome may be annotated, through the `exa:presenceOf` object property, with the named individuals necrosis (`go:0070265`), and metastasis (`ncit:C19151`), both of type lung annotation (`exa:LungAnnotation`).

Celiac Disease. The diagnosis of a patient tested for the presence of the coeliac disease, irrespectively of the nature of the outcome itself, may present different types of additional information. Because of this, the annotations are connected to the clinical case report instance, and not the instance of the outcome.

Different sub-classes of `exa:DiseaseAnnotation` are used to represent these annotations:

- `exa:CeliacDiseaseAnnotation` The named individuals granulocyte (`cl:0000094`) and lymphocyte (`cl:0000542`) are instances of this class. Instances of celiac clinical case report are connected to these individuals by means of the `exa:presenceOf` object property.
- `exa:CeliacAbnormality` This class is used to describe the presence of abnormalities in a report. The named individuals that are instances of this class are: Brunner's Gland hyperplasia (`ncit:C135565`), hyperemia (`symp:000299`), edema `omit:0005738`, and intestinal fibrosis (`mp:0011748`). Report instances are connected to these individuals using the `exa:presenceOfCeliacAbnormality` object property.
- `exa:CeliacDiseaseAnnotationBag` A type of RDF bag. In this case, for each clinical trial, one or more instances of this class can be created and associated with the `exa:hasCeliacAnnotation` property to a report. Each of these may contain a combination of three values belonging to datatypes that are subclasses of `xsd:string`. These are:

  1. Intraepithelial lymphocyte amount (`oba:0004525`). The instances are associated to a value belonging to this datatype with the `exa:hasIEL` data property.

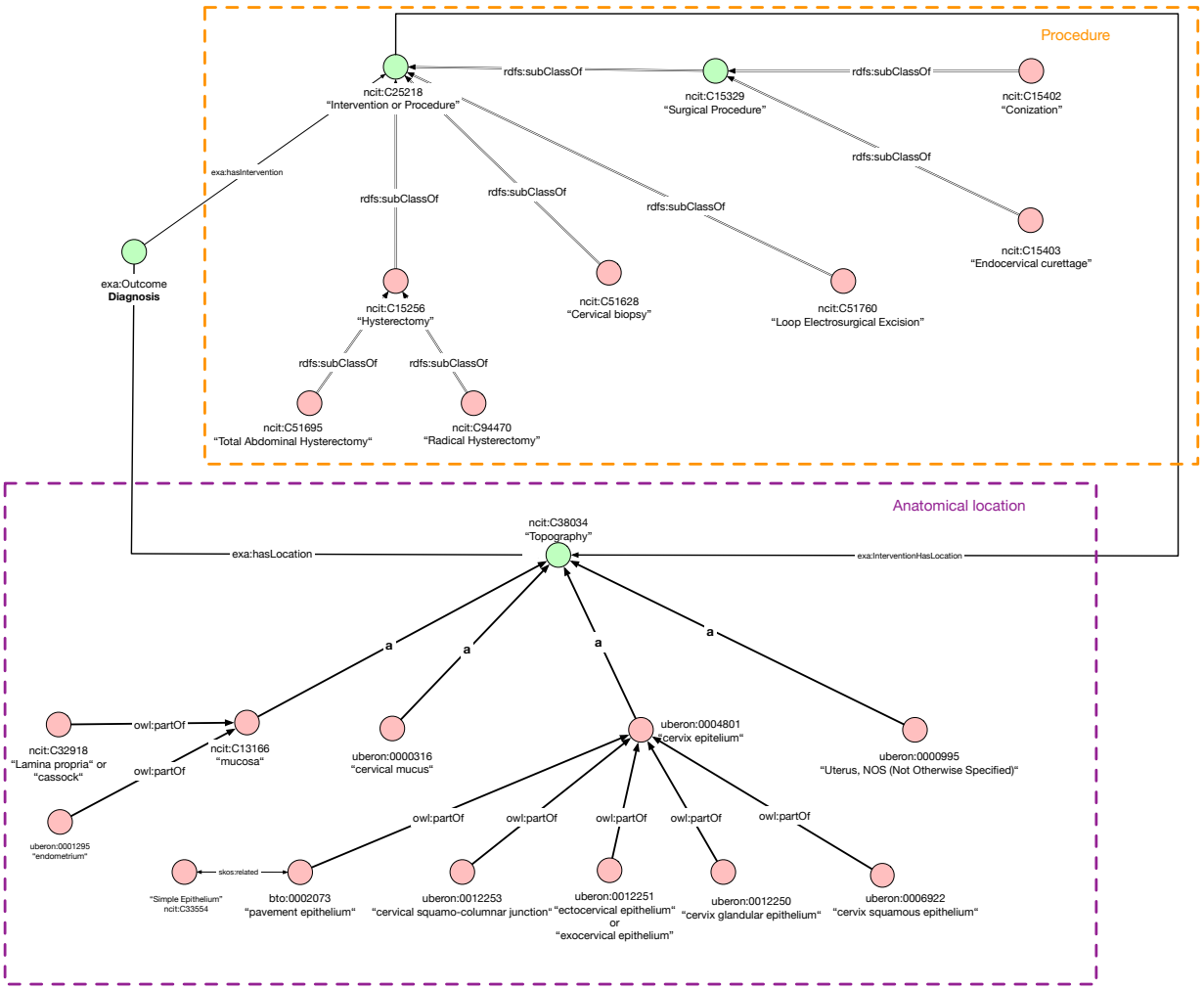

Fig. 4. Graphical representation of the Procedure and Anatomical Location semantic areas for the Uterine Cervix Cancer.

2. Villi to crypt of Lieberkuhn Ratio (`exa:villusCryptRatio`). The instances are associated to a value belonging to this datatype with the `exa:hasVillusCryptRatio` data property.
3. Duodenum villi length (`vt:0010209`). The instances are associated to a value belonging to this datatype with the `exa:villiStatus` data property.

4.4. Anatomical Location/Topography

In the ExaMode ontology the "Anatomical Location" area contains entities that represent different parts of the human body. These entities accomplish two goals: (i) indicate where the disease of the patient is located; and, (ii) where the samples to study the disease were taken. The main class of this semantic area is "Topography" (`ncit:C38034`), an entity that represents the "description of an anatomical region or of a body part", as defined by the NCI.

In the ExaMode ontology we do not instantiate a new location each time a record is registered in the database. Thus, we decided to use named individuals, not subclasses, for this class. The different named individuals are all of type Topography, and are also organized in a hierarchy through the use of the `owl:partOf` property.

4.4.1. Colon Cancer

The colon is a part of the human anatomy that can be divided in many sub-parts. In the ontology we represented the abdomen (uberon:0000916), the ileum (uberon:0002166), and the colorectum (uberon:0012652). In turn, the colorectum can be divided in rectum, Not Otherwise Specified (NOS) (uberon:0001052) and the colon, NOS (uberon:0001155). The NOS classes are usually adopted when the physician was not able to be more precise in the diagnosis. Generally, the colon is divided in different parts and we represented: the caecum (uberon:0001153), the transverse colon (uberon:0001157), the descending colon (uberon:0001158), the rectal mucous membrane (uberon: 0003346), the ascending colon (uberon:0001156), which can be assimilated to the right colon (uberon:0008972), the sigmoid colon (uberon:000159), the left colon (uberon:0008971), and the rectosigmoid junction (uberon:0036214).

4.4.2. Lung Cancer

Lung cancer presents the largest set of anatomical locations in the ontology due to the complexity of the human lung, which is composed of different parts. In the ontology we represented whole lung (uberon:0002048), the bronchus (uberon: 0002185), the mediastinum (uberon:0003728), the thoracic lymph nodes (uberon:-0007644), the pleura (uberon:0000977), and the pulmonary lymph node (uberon:0035764). The bronchus parts of the bronchus are the main bronchus (ncit:C12284), and the lobar bronchus (ncit:C32998). Part of the main bronchus are, in turn, the left main bronchus (ncit:C32968) and the right main bronchus (ncit:C33486). Parts of the lobar bronchus are, instead, the left inferior lobar bronchus (fma:7432), the right inferior lobar bronchus (fma:7404), the left superior lobar bronchus (fma:7423), and the right superior lobar bronchus (fma:7397).

4.4.3. Uterine Cervix cancer

We can see the classes of the Anatomical Location semantic area of the uterine cervix cancer in Figure 4.

These are the mucosa (ncit:C13166), the cervical mucus (uberon:0000316), the cervix epitelium (uveron:0004801), and, finally, the Uterus, NOS (uberon:0000995).

The mucosa presents two other parts: the lamina propria (ncit:C32918), and the endometrium (uberon:-0001295). The cervix epitelium presents many different parts: the pavement epithelium (bto:0002073), the cervical squamo-columnar junction (uberon:0012253), the ectovervical epithelium (uberon:0012251), the cervix glandular epithelium (uberon:0012250), and the cervix squamous epithelium (uberon:0006922).

4.4.4. Celiac disease

The celiac disease is located in the area of the Duodenum (uberon:0002114). Parts of it are: the duodenal mucosa (uberon:0000320), the crypt of Lieberkuhn of the duodenum (uberon:0013482), the intestinal villus of the duodenum (uberon:0008342), the duodenal ampulla (uberon:0013644), the duodenal epithelium (uberon:0008346), the duodenum lamina propria (uberon:0015834), the duodenal gland (uberon:0001212).

Moreover, there are possible operations aimed at diagnosing the disease that may also be performed in other areas of the human anatomy where it is possible to find symptoms connected to the presence of the disease. These are the pyloric antrum (uberon:0001165) and the Greater curvature of the stomach (uberon:0001164).

4.5. Procedure

The Procedure semantic area includes entities that describe the operations performed to collect the tissues used to diagnose the disease. Depending on the type of disease and its location in the human anatomy, different procedure may be used. The main class of this area is "Intervention or Procedure" (ncit:C25218), which is a general entity describing an activity intended to alter the course of a disease. This class is connected with the property exa:hasTopography to the topography class. We also frequently use its subclass "Surgical Procedure" (ncit:C15329), a general entity the represents a procedure to remove part of the body to find out whether a disease is present.

For the cases considered by the ExaMode ontology, one of the most important subclass of surgical procedure is Biopsy (ncit:C15189), i.e., the removal of tissue specimens or fluid from a living body performed to establish a diagnosis, which is used between all four cases, although in different form and subclasses.

4.5.1. Colon cancer

In the case of colon cancer the subclasses of surgical procedure are: resection (ncit:C158758), anastomosic (ncit:C15609), hemiectomy (ncit:C51580), and the endoscopic biopsy (ncit:C15389). Subclass of this last one is the biopsy of colon (ncit:C51678), that in turn is further sub-classed in"polypectomy" (ncit:C25349).

4.5.2. Lung cancer

In this case the surgical procedure presents as unique subclass the general biopsy (ncit:C15189). Subclasses of this in the specific case of the lung cancer are the the bronchial biopsy (ncit:C51782) and the lung biopsy (ncit:C51748).

4.5.3. Uterine Cervix Cancer

For the cervix cancer there are different types of procedures, as can be seen in Figure 4. Among the direct subclass of intervention or procedure, we count: hysterectomy (ncit:C15256) and its subclasses, total abdominal hysterectomy (ncit:C51695) and radical hysterectomy (ncit:C94470); the cervical biopsy (ncit:C51628); the Loop electrosurgical excision (ncit:C51760), Among the sub-classes of surgical procedure we see the conization (ncit:C15402), and the endocervical curettage (ncit:C15403).

4.5.4. Celiac disease

In this case, as for the lung cancer, the unique subclass of surgical procedure is the biopsy, further subclasses in: biopsy of duodenum (ncit:C51683), biopsy of the greater curvature (exa:GreaterCurvatureBiopsy), and biopsy of the pyloric antrum (exa:AntrumPyloriBiopsy).

4.6. Test

It is possible that the doctors producing the diagnosis of a disease also decided to perform some tests on the patient to better identify its characteristics. Connected to the outcome class through the exa:hasText property, there is the "Test" (ncit:C47891) entity, a general class describing a procedure for critical evaluation. Depending on the disease, different tests may be performed on the patient.

4.6.1. Colon cancer

The only immunohistochemical test that is considered in the ExaMode ontology for the colon cancer is the immunoprecipitation (ncit:C16724), returning a boolean value (true or false).

4.6.2. Lung cancer

When a patient is afflicted by lung cancer, it is possible to perform an immunohistochemical test (ncit:C51944), or, more specifically, one of its many subclass entities, as represented in Figure 5. As can be seen, each of these classes is associated with a boolean or float value, the outcome of the test itself.

4.6.3. Uterine Cervix cancer

No test is considered in this case.

4.6.4. Celiac disease

A test called "Cluster of differentiation 3" (exa:CD3) can be issued.

5. Applications

5.1. Entity Linking.

One of the most relevant tasks that can be performed with an ontology is Entity Linking (EL) [63]. EL is the task of assigning unique meanings to entities mentioned within text. In a nutshell, the aim of EL is to determine if a given (extracted) entity refers to a specific concept or object within a reference ontology. To this end, we developed the

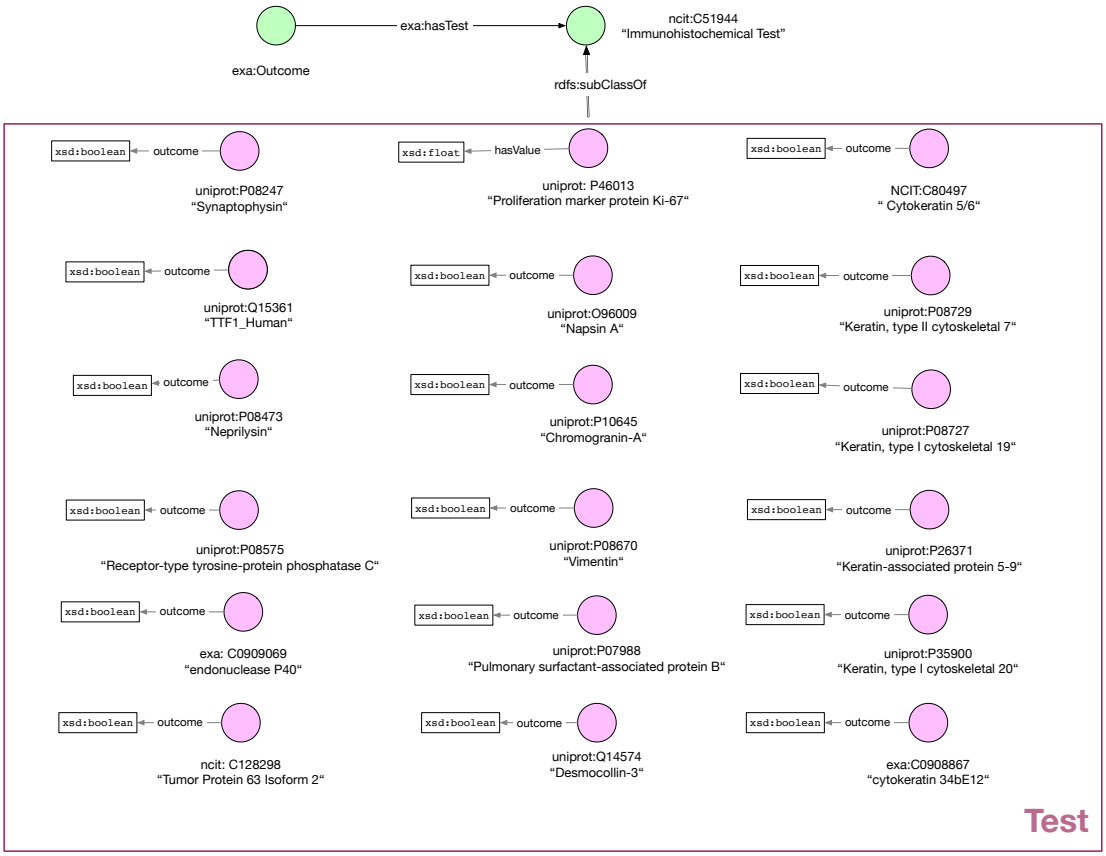

Fig. 5. Representation of the Test semantic area pertaining the lung cancer case, including the possible immunohistochemical tests that can be issued to a patient. All the classes are in a relationship `rdfs:subClassOf` with the classe `ncit:C51944`.

Semantic Knowledge Extractor Tool (SKET),[19] an unsupervised hybrid knowledge extraction system that combines a rule-based expert system with pre-trained machine learning models to extract relevant concepts from pathology reports.

SKET combines pre-trained Named Entity Recognition (NER) models with unsupervised Entity Linking (EL) methods to extract relevant entities from pathology reports and link them to the ExaMode ontology. The use of pre-trained NER models and unsupervised EL methods makes SKET suitable for weak supervision tasks [64]. Therefore, SKET can be used in the clinical workflow without the expensive and time-consuming cost of human annotations.

We adopted SKET to extract concepts from pathology reports coming from the clinical workflow of AOEC and Radboud UMC. We considered reports from three use cases modeled by the ExaMode ontology, i.e., Colon Cancer, Uterine Cervix Cancer, and Lung Cancer. We report the statistics of the extraction process for each use case in Table 2. The extracted concepts can then be used for different applications, such as weak annotations to train models for computer aided diagnosis (or decision support) for digital pathology [17]. This demonstrates the importance of developing medical ontologies, which empower many different computer-aided diagnosis tools [65, 66].

[19] SKET source code is publicly available at https://github.com/ExaNLP/sket/

Table 2

SKET extraction statistics. Columns represent, from left to right, the considered hospital, the considered use case, the total number of reports, the total number of concepts, the maximum number of concepts per report, the minimum number of concepts per report, and the mean number of concepts per report. The "–" symbol represents the lack of pathology reports for the considered hospital and use case.

Hospital	Use case	N. of rep.s	N. of conc.s	Max conc.s/rep.	Min conc.s/rep.	Mean conc.s/rep.
AOEC	Colon	2,069	7,979	10	0	3.86
	Uterine Cervix	2,518	10,308	15	0	4.09
	Lung	2,077	6,211	7	0	2.99
Radboud UMC	Colon	6,532	7,629	11	0	1.17
	Uterine Cervix	4,711	8,346	10	0	1.77
	Lung	–	–	–	–	–

5.2. Decision support systems and triaging

The designed ExaMode ontology, alongside with other domain ontologies (ICD-10, ICD-O, UMLS, SNOMED, Human Disease Ontology[20], PathLex[21], Mondo[22], ProstateCancer[23]) was integrated with patient data into the "HistoGrapher demonstrator"[24] for decision support systems and triaging.

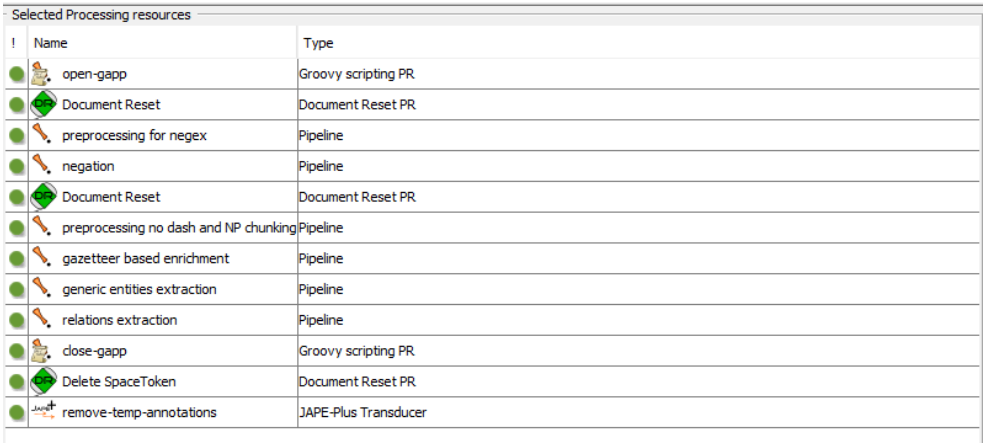

Fig. 6. NER pipeline consist of multiple text pre- and post-processing resources (PRs) and ontology based semantic annotation PR "gazetteer-based-enrichment", which normalize the identified phrases to ontology terms

HistoGrapher is a software platform for building holistic solution integrating multimodal histopathology data. This solution supports pathologists in making more informed decisions based on a larger amount of data (judging from similarity to other cases in their clinical practice or the identified likelihood in scientific literature). Prioritization of cases is considered, so that the cases with higher severity and likelihood of specific diseases will be presented for confirmation first, while the cases with smaller likelihood - later when there is enough time.

The data model of the HistoGrapher platform is based on the defined ExaMode ontology, which is used for information extraction and semantic data normalization of medical synopsis texts provided by the laboratory information management system (LIMS). The platform applies machine translation in order to translate the source text data, provided in Italian and Dutch, into corresponding English representations. All translated text fields from the synopsis record are processed with a Natural Language Processing (NLP) pipeline (Fig. 6). The pipeline is based

[20] https://disease-ontology.org/
[21] https://bioportal.bioontology.org/ontologies/PATHLEX
[22] https://mondo.monarchinitiative.org/
[23] https://bioportal.bioontology.org/ontologies/PCAO
[24] http://examode.ontotext.com/

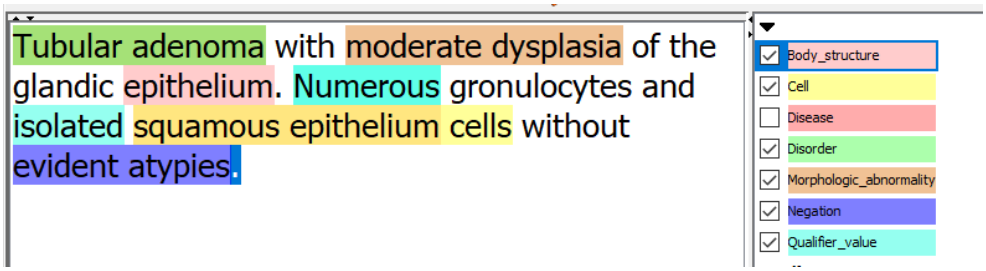

Fig. 7. Semantically annotated synopsis report with highlighter terms from various annotation classes (in different colors)

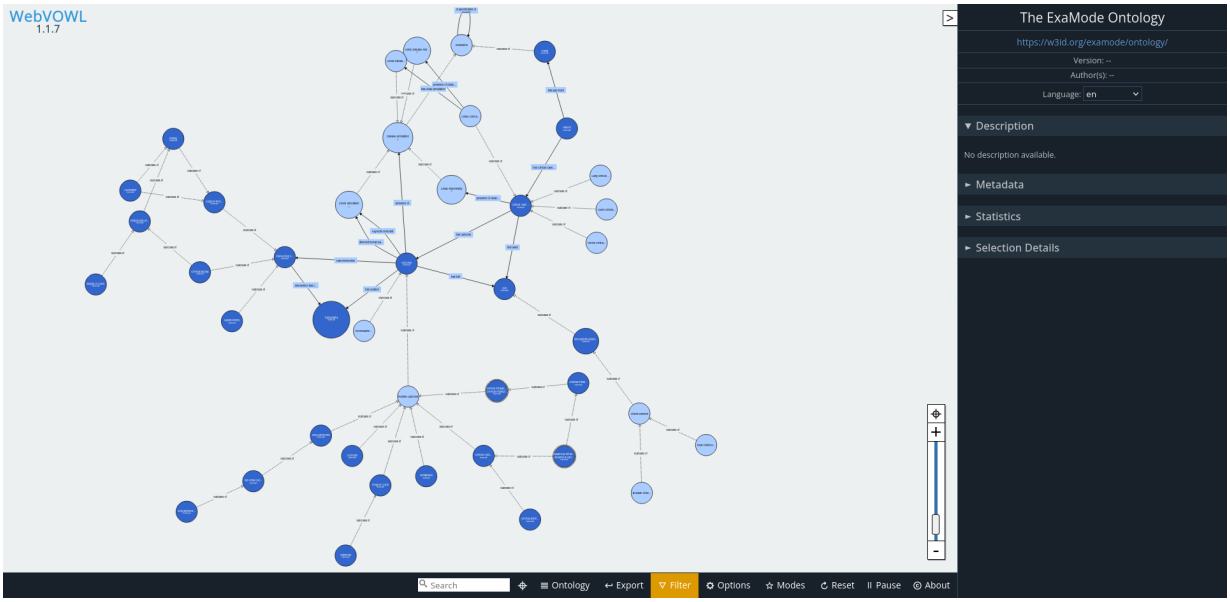

Fig. 8. ExaMode ontology visualized using WebVOWL.

on the ExaMode ontology and all created semantic annotations in-lined in the source text are referred to concepts from the ontology or to other biomedical terminologies, which are part of the platform semantic solution (fig. 7). The output of the NLP pipeline, in the form of RDF triples, is imported in GraphDB[25], which is the semantic RDF triples store used to build the knowledge graph behind the Histographer platform. As the NLP pipeline output is fully harmonized with the ExaMode ontology, the extracted data can be further explored in the context of the ontological model and the terminologies included in the knowledge graph of the system. The HistoGrapher platform provides fully configurable semantic search and data exploration dashboards that are adapted to the specific data schema and ontologies used. The system provides capabilities for intuitive faceted filtering for relevant clinical case records and both semantic (concept) and free-text search. The focused data exploration dashboards provide different views over the data in the knowledge graph. An important dashboard component is the case report similarity widget used to retrieve similar case reports. The similarity of case reports is based on graph embedding configured per certain knowledge graph class and it's `DataType` and `ObjectType` properties.

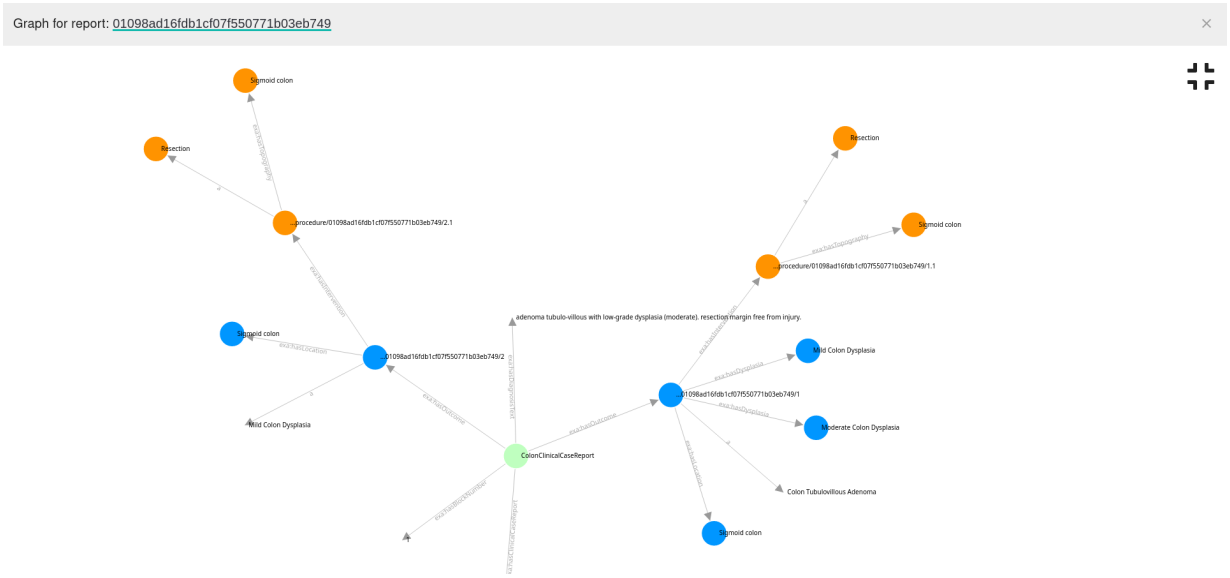

Fig. 9. ExaNet visualization of a clinical report (full screen mode).

5.3. *Visual Exploration*

The visualization of ontologies is an important task for assessing ontologies, enabling users to explore, verify, and understand them and their underlying structures [67–70].

To this aim, we used WebVOWL [68] to render a dynamic visualization of the ExaMode ontology, as shown in Figure 8. The WebVOWL interface allows the users to interactively explore the ontology, filter nodes (classes) and search for properties of interest. Since the ontology visualization obtained with WebVOWL focuses primarily on the Terminological Box (TBox) (definition of classes and properties), we developed a visual application - i.e. ExaNet - that focuses instead on the Assertional Box (ABox) (individuals and instance data). ExaNet is provided as a web-based application[26] allowing the users (e.g., experts and pathologists) to explore the ExaMode clinical reports linked data, using an interactive graph visualization tool. Figure 9 shows the ExaNet visualization layer for a clinical report related to colon carcinoma.

To enhance readability and data comprehension, ExaNet replaces the classes' IRI with the corresponding literals. For instance, in Figure 9 the IRI http://purl.obolibrary.org/obo/NCIT_C4848, describing a report's outcome, is replaced with the matching literal *Mild Colon Dysplasia*. ExaNet has been implemented using the JavaScript library D3[27] and its force-directed graph layout, enabling users to explore graph connections by taking advantage of pan and zoom functionalities. In addition, ExaNet allows the users to visualize an interactive JSON serialization of each clinical report and to possibly download it.

6. Final Remarks and Future Work

We described the ExaMode ontology designed to represent the diagnostic aspects of histopathology diseases by focusing on four use cases of high societal and clinical impact. The ontology development followed co-design principles working iteratively by involving experts and pathologists in the process. We relied on shared community best practices by employing standard languages like OWL, existing medical vocabularies and ontologies as UBERON and NCIT, and URL consistent naming convention.

[25] https://www.ontotext.com/products/graphdb/

[26] ExaNet is publicly available at http://w3id.org/exanet, using *demo* as username and password credentials.

[27] https://d3js.org/

The ontology is divided into six main "semantic areas" covering the different aspects of medical reports in histopathology. The general ExaMode disease cases area contains entities describing the patient and the clinical case report. The diagnosis area contains entities describing the outcome of the disease, distinguishing between an inconclusive outcome, a negative result, the absence of malignancy, or a positive outcome. In the case of a positive outcome, and depending on the disease, a taxonomy of the ontological classes covering the possible diagnosis was developed. The patient's clinical report may contain additional information regarding the nature or the state of the disease. We modeled these aspects in the Annotation semantic area. The Anatomical Location semantic area contains the named individuals identifying the parts of the human body where the disease is found or the samples were taken. The entities belonging to the *Procedure* area contain the classes describing the surgical procedures. Finally, the *Test* semantic area includes the classes describing the tests that may be done to the patient.

Currently, the ontology supports several tasks as automatic entity-linking from medical reports, and the production of "weak" annotations for WSI used to train prediction algorithms. Moreover, the ontology is employed as a key component in a decision support system that supports pathologists in making more informed decisions based on the large amount of data available. We showed how the ontology empowers visualization tools such as ExaNet, allowing users to explore, verify and understand the graph of the ontology and the ones built from the entities extracted from the reports. Indeed, the ExaMode ontology is at the basis of a graph interconnecting many reports from two hospitals (AOEC and Radboud UMC), aiming to discover relations between reports and helping the medical doctors to find new knowledge. We employed the ontology also in the NanoWeb application that relates gene-disease associations RDF triples with the facts extracted from ExaMode medical records [71].

Future work will focus on multilingual terminology in the medical domain. Indeed, it has been underlined that communication in the healthcare domain is characterized by a rigid and closed nomenclature which, in many cases, produces an opaque and complex to understand lexicon [72]. The medical language presents many inconsistencies, such as semantic ambiguity of the scientific terms, redundancy in the formulation of compounds, and other etymological inconsistencies. As a result, non-expert users may have difficulties finding complete and accurate information on health issues or understanding diagnostic reports. Moreover, although the biomedical domain offers many linguistic resources for NLP, including terminologies, corpora, and ontologies, most of these resources are prominently available for the English language. Thus, the access to terminological resources in other languages is not straightforward for non-native speakers. So, we aim to make the ExaMode ontology fully multilingual by providing verified translations of the medical terms contained also complementing those available in UMLS that are already included in the ExaMode ontology. We are working on Italian and Dutch, which are the medical reports available to the project. Moreover, we are developing XML terminological forms for mapping the terms in French and, for Italian, in popular language to enhance comprehension for the general public.

References

[1] R. Hoehndorf, P.N. Schofield and G.V. Gkoutos, The role of ontologies in biological and biomedical research: a functional perspective, *Briefings in Bioinformatics* **16**(6) (2015), 1069–1080. doi:10.1093/bib/bbv011.

[2] B.M. Konopka, Biomedical ontologies—A review, *Biocybernetics and Biomedical Engineering* **35**(2) (2015), 75–86.

[3] A.J. Evans, M.E. Salama, W.H. Henricks and L. Pantanowitz, Implementation of whole slide imaging for clinical purposes: issues to consider from the perspective of early adopters, *Archives of pathology & laboratory medicine* **141**(7) (2017), 944–959.

[4] K. Lindman, J.F. Rose, M. Lindvall, C. Lundström and D. Treanor, Annotations, ontologies, and whole slide images–Development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue, *Journal of pathology informatics* **10** (2019).

[5] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken and C.I. Sánchez, A survey on deep learning in medical image analysis, *Medical image analysis* **42** (2017), 60–88.

[6] J. Griffin and D. Treanor, Digital pathology in clinical use: where are we now and what is holding us back?, *Histopathology* **70**(1) (2017), 134–145.

[7] M.G. Hanna, V.E. Reuter, J. Samboy, C. England, L. Corsale, S.W. Fine, N.P. Agaram, E. Stamelos, Y. Yagi, M. Hameed et al., Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings, *Archives of pathology & laboratory medicine* **143**(12) (2019), 1545–1555.

[8] E.A. Krupinski, A.R. Graham and R.S. Weinstein, Characterizing the development of visual search expertise in pathology residents viewing whole slide images, *Human pathology* **44**(3) (2013), 357–364.

[9] P. Vennalaganti, V. Kanakadandi, J.R. Goldblum, S.C. Mathur, D.T. Patil, G.J. Offerhaus, S.L. Meijer, M. Vieth, R.D. Odze, S. Shreyas et al., Discordance among pathologists in the United States and Europe in diagnosis of low-grade dysplasia for patients with Barrett's esophagus, *Gastroenterology* **152**(3) (2017), 564–570.
[10] M. Costantini, S. Sciallero, A. Giannini, B. Gatteschi, P. Rinaldi, G. Lanzanova, L. Bonelli, T. Casetti, E. Bertinelli, O. Giuliani et al., Inter-observer agreement in the histologic diagnosis of colorectal polyps: the experience of the multicenter adenoma colorectal study (SMAC), *Journal of clinical epidemiology* **56**(3) (2003), 209–214.
[11] G. Campanella, M.G. Hanna and L. Geneslaw, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, *Nature Medicine* **25** (2019), 1301—1309.
[12] K. Jensen, C. Soguero-Ruiz, K.O. Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, S.O. Skrovseth and K.M. Augestad, Analysis of free text in electronic health records for identification of cancer patient trajectories, *Scientific reports* **7**(1) (2017), 1–12.
[13] T. Davenport and R. Kalakota, The potential for artificial intelligence in healthcare, *Future healthcare journal* **6**(2) (2019), 94.
[14] A. Dhrangadhariya, S. Otálora, M. Atzori and H. Müller, Classification of Noisy Free-Text Prostate Cancer Pathology Reports Using Natural Language Processing, in: *International Conference on Pattern Recognition*, Springer, 2021, pp. 154–166.
[15] G. Burger, A. Abu-Hanna, N. de Keizer and R. Cornet, Natural language processing in pathology: a scoping review, *Journal of clinical pathology* **69**(11) (2016), 949–955.
[16] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn et al., Clinical information extraction applications: a literature review, *Journal of biomedical informatics* **77** (2018), 34–49.
[17] V. Cheplygina, M. de Bruijne and J.P.W. Pluim, Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, *Medical Image Anal.* **54** (2019), 280–296.
[18] L. Chiticariu, Y. Li and F. Reiss, Rule-based information extraction is dead! long live rule-based information extraction systems!, in: *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 827–832.
[19] A. Madabhushi and G. Lee, Image analysis and machine learning in digital pathology: Challenges and opportunities, *Medical image analysis* **33** (2016), 170–175.
[20] F. Freitas, S. Schulz and E. Moraes, Survey of current terminologies and ontologies in biology and medicine, *RECIIS-Electronic Journal in Communication, Information and Innovation in Health* **3**(1) (2009), 7–18.
[21] L.M. Serra, W.D. Duncan and A.D. Diehl, An ontology for representing hematologic malignancies: the cancer cell ontology, *BMC Bioinform.* **20-S**(5) (2019), 231–236. doi:10.1186/s12859-019-2722-8.
[22] J.A. Turner, J.L. Mejino, J.F. Brinkley, L.T. Detwiler, H.J. Lee, M.E. Martone and D.L. Rubin, Application of neuroanatomical ontologies for neuroimaging data annotation, *Frontiers in neuroinformatics* **4** (2010), 10.
[23] M. Ivanovic and Z. Budimac, An overview of ontologies and data resources in medical domains, *Expert Syst. Appl.* **41**(11) (2014), 5158–5166. doi:10.1016/j.eswa.2014.02.045.
[24] M.N. Gurcan, J. Tomaszewski, J.A. Overton, S. Doyle, A. Ruttenberg and B. Smith, Developing the quantitative histopathology image ontology (QHIO): a case study using the hot spot detection problem, *Journal of biomedical informatics* **66** (2017), 129–135.
[25] M. García-Rojo, C. Daniel and A. Laurinavicius, SNOMED CT in pathology, *Perspectives on Digital Pathology* (2012), 123–140.
[26] K.W. Fung, J. Xu and O. Bodenreider, The new International Classification of Diseases 11th edition: a comparative analysis with ICD-10 and ICD-10-CM, *J. Am. Medical Informatics Assoc.* **27**(5) (2020), 738–746. doi:10.1093/jamia/ocaa030.
[27] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature biotechnology* **25**(11) (2007), 1251–1255.
[28] J. Golbeck, G. Fragoso, F.W. Hartel, J.A. Hendler, J. Oberthaler and B. Parsia, The National Cancer Institute's Thésaurus and Ontology, *J. Web Semant.* **1**(1) (2003), 75–80. doi:10.1016/j.websem.2003.07.007.
[29] Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Res.* **32**(Database–Issue) (2004), 258–261. doi:10.1093/nar/gkh036.
[30] I. Spasic, J. Livsey, J.A. Keane and G. Nenadic, Text mining of cancer-related information: Review of current status and future directions, *Int. J. Medical Informatics* **83**(9) (2014), 605–623. doi:10.1016/j.ijmedinf.2014.06.009.
[31] Y. Shen, J. Colloc, A. Jacquet-Andrieu and K. Lei, Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system, *J. Biomed. Informatics* **56** (2015), 307–317. doi:10.1016/j.jbi.2015.06.012.
[32] K. Regan, S. Raje, C. Saravanamuthu and P.R.O. Payne, Conceptual Knowledge Discovery in Databases for Drug Combinations Predictions in Malignant Melanoma, in: *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics*, Studies in Health Technology and Informatics, Vol. 216, IOS Press, 2015, pp. 663–667. doi:10.3233/978-1-61499-564-7-663.
[33] T. Wu, L.M. Schriml, Q. Chen, M. Colbert, D.J. Crichton, R.P. Finney, Y. Hu, W.A. Kibbe, H. Kincaid, D.M. Meerzaman, E. Mitraka, Y. Pan, K. Smith, S. Srivastava, S. Ward, C. Yan and R. Mazumder, Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis, *Database J. Biol. Databases Curation* **2015** (2015). doi:10.1093/database/bav032.
[34] M. Boeker, F. França, P. Bronsert and S. Schulz, TNM-O: ontology support for staging of malignant tumours, *J. Biomed. Semant.* **7** (2016), 64:1–64:11. doi:10.1186/s13326-016-0106-9.
[35] L. Tagliaferri, G. Kovács, R. Autorino, A. Budrukkar, J.L. Guinot, G. Hildebrand, B. Johansson, R.M. Monge, J.E. Meyer, P. Niehoff et al., ENT COBRA (Consortium for Brachytherapy Data Analysis): interdisciplinary standardized data collection system for head and neck patients treated with interventional radiotherapy (brachytherapy), *Journal of Contemporary Brachytherapy* **8**(4) (2016), 336.
[36] S. Myneni, M. Amith, Y. Geng and C. Tao, Towards an Ontology-driven Framework to Enable Development of Personalized mHealth Solutions for Cancer Survivors' Engagement in Healthy Living, in: *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics*, I.N. Sarkar, A. Georgiou and P.M. de Azevedo Marques, eds, Studies in Health Technology and Informatics, Vol. 216, IOS Press, 2015, pp. 113–117. doi:10.3233/978-1-61499-564-7-113.

[37] J. Huang, J. Dang, G.M. Borchert, K. Eilbeck, H. Zhang, M. Xiong, W. Jiang, H. Wu, J.A. Blake, D.A. Natale et al., OMIT: dynamic, semi-automated ontology development for the microRNA domain, *PLoS One* **9**(7) (2014), e100855.
[38] C. Jonquet, M.A. Musen and N.H. Shah, Building a biomedical ontology recommender web service, *J. Biomed. Semant.* **1**(S–1) (2010), S1. http://www.jbiomedsem.com/content/1/S1/S1.
[39] Z. Xiang, C. Mungall, A. Ruttenberg and Y. He, OntoBee: A Linked Data Server and Browser for Ontology Terms, in: *Proceedings of the 2nd International Conference on Biomedical Ontology*, CEUR Workshop Proceedings, Vol. 833, CEUR-WS.org, 2011. http://ceur-ws.org/Vol-833/paper48.pdf.
[40] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature biotechnology* **25**(11) (2007), 1251–1255.
[41] V. Cheplygina, M. de Bruijne and J.P.W. Pluim, Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, *Medical Image Analysis* **54** (2019), 280–296. doi:https://doi.org/10.1016/j.media.2019.03.009. https://www.sciencedirect.com/science/article/pii/S1361841518307588.
[42] D. Komura and S. Ishikawa, Machine learning methods for histopathological image analysis, *CoRR* **abs/1709.00786** (2017). http://arxiv.org/abs/1709.00786.
[43] K. Tomczak, P. Czerwińska and M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Contemporary oncology* **19**(1A) (2015), A68.
[44] F.W. Prior, K.W. Clark, P.K. Commean, J.B. Freymann, C.C. Jaffe, J.S. Kirby, S.M. Moore, K.E. Smith, L. Tarbox, B.A. Vendt and G. Marquez, TCIA: An information resource to enable open science, in: *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2013, pp. 1282–1285. doi:10.1109/EMBC.2013.6609742.
[45] O.A.J. del Toro, S. Otálora, M. Atzori and H. Müller, Deep Multimodal Case-Based Retrieval for Large Histopathology Datasets, in: *Patch-Based Techniques in Medical Imaging - Third International Workshop, Patch-MI 2017, Held in Conjunction with MICCAI 2017*, Vol. 10530, Springer, 2017, pp. 149–157. doi:10.1007/978-3-319-67434-6_17.
[46] International Agency for Research on Cancer [Internet]. Cancer Tomorrow: A tool that predicts the future cancer incidence and mortality burden worldwide from the current estimates in 2020 up until 2040, [cited 2021 July 13] Available from: https://gco.iarc.fr/tomorrow/en..
[47] Y. Xi and P. Xu, Global colorectal cancer burden in 2020 and projections to 2040, *Translational Oncology* **14**(10) (2021), 101174.
[48] A.M. Wolf, E.T. Fontham, T.R. Church, C.R. Flowers, C.E. Guerra, S.J. LaMonte, R. Etzioni, M.T. McKenna, K.C. Oeffinger, Y.-C.T. Shih et al., Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society, *CA: a cancer journal for clinicians* **68**(4) (2018), 250–281.
[49] M. Fleming, S. Ravula, S.F. Tatishchev and H.L. Wang, Colorectal carcinoma: Pathologic aspects, *Journal of gastrointestinal oncology* **3**(3) (2012), 153.
[50] A.G. Zauber, S.J. Winawer, M.J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B.F. Hankey, W. Shi, J.H. Bond, M. Schapiro, J.F. Panish, E.T. Stewart and W.J. D., Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths, *N Engl J Med* **366** (2012), 687–696.
[51] J.M. Butte, P. Tang, M. Gonen, J. Shia, M. Schattner, G.M. Nash, L.K.F. Temple and M.R. Weiser, Rate of residual disease after complete endoscopic resection of malignant colonic polyp, *Diseases of the colon & rectum* **55**(2) (2012), 122–127.
[52] World Health Organization, International Agency for Research on Cancer, Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020, press release no. 292, 2020.
[53] W.D. Travis, E. Brambilla, M. Noguchi, A.G. Nicholson, K.R. Geisinger, Y. Yatabe, D.G. Beer, C.A. Powell, G.J. Riely, P.E. Van Schil et al., International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma, *Journal of thoracic oncology* **6**(2) (2011), 244–285.
[54] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay and F. Bray, Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis, *The Lancet Global Health* **8**(2) (2020), e191–e203.
[55] M.K. Zarchi, F. Binesh, Z. Kazemi, S. Teimoori, H.R. Soltani and Z. Chiti, Value of colposcopy in the early diagnosis of cervical cancer in patients with abnormal pap smears at Shahid Sadoughi hospital, Yazd, *Asian Pacific Journal of Cancer Prevention* **12**(12) (2011), 3439–3441.
[56] S. Lax, Histopathology of cervical precursor lesions and cancers, *Acta Dermatoven APA* **20**(3) (2011), 125–133.
[57] D.R. Lowy and J.T. Schiller, Reducing HPV-associated cancer globally, *Cancer prevention research* **5**(1) (2012), 18–23.
[58] N. Gujral, H. Freeman and A. Thomson, Celiac disease: Prevalence, diagnosis, pathogenesis and treatment, *World journal of gastroenterology: WJG* **18** (2012), 6036–59. doi:10.3748/wjg.v18.i42.6036.
[59] J.R. Glissen Brown and S. P., Celiac disease, *Paediatr Int Child Health* **39** (2019), 23–31. doi:10.1080/20469047.2018.1504431.
[60] V. Villanacci, P. Ceppa, E. https://www.overleaf.com/project/617c1835b0e93b5f11026f98Tavani, C. Vindigni and U. Volta, On behalf of the "Gruppo Italiano Patologi Apparato Digerente (GIPAD)" and of the "Società Italiana di Anatomia Patologica e Citopatologia Diagnostica"/International Academy of Pathology, Italian division (SIAPEC/IAP) Coeliac disease: the histology report, *Digestive and Liver Disease* **43** (2011), S385–S395. doi:10.1016/S1590-8658(11)60594-X.
[61] C.J. Mungall, C. Torniai, G.V. Gkoutos, S.E. Lewis and M.A. Haendel, Uberon, an integrative multi-species anatomy ontology, *Genome biology* **13**(1) (2012), 1–20.
[62] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research* **32**(suppl_1) (2004), D267–D270.
[63] W. Shen, J. Wang and J. Han, Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions, *IEEE Trans. Knowl. Data Eng.* **27**(2) (2015), 443–460.

[64] P. Courtiol, E.W. Tramel, M. Sanselme and G. Wainrib, Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach, *CoRR* **abs/1802.02212** (2018).
[65] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn and H. Liu, Clinical information extraction applications: A literature review, *J. Biomed. Informatics* **77** (2018), 34–49.
[66] G. Burger, A. Abu-Hanna, N. de Keizer and R. Cornet, Natural Language Processing in Pathology: a Scoping Review, *Journal of Clinical Pathology* **69**(11) (2016), 949–955.
[67] S. Lohmann, S. Negru, F. Haag and T. Ertl, Visualizing ontologies with VOWL, *Semantic Web* **7**(4) (2016), 399–419.
[68] S. Lohmann, V. Link, E. Marbach and S. Negru, WebVOWL: Web-based visualization of ontologies, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2014, pp. 154–158.
[69] S. Lohmann, S. Negru, F. Haag and T. Ertl, VOWL 2: User-oriented visualization of ontologies, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2014, pp. 266–281.
[70] M. Lanzenberger, J. Sampson and M. Rester, Visualization in ontology tools, in: *2009 International Conference on Complex, Intelligent and Software Intensive Systems*, IEEE, 2009, pp. 705–711.
[71] F. Giachelle, D. Dosso and G. Silvello, Search, access, and explore life science nanopublications on the Web, *PeerJ Comput. Sci.* **7** (2021), e335. doi:10.7717/peerj-cs.335.
[72] F. Vezzani, G.M.D. Nunzio and G. Henrot, TriMED: A Multilingual Terminological Database, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, European Language Resources Association (ELRA), 2018. http://www.lrec-conf.org/proceedings/lrec2018/summaries/715.html.
[73] J.M. Buckley, S.B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A.K. Belli, E.M. Kim, J.E. Garber, B.L. Smith, M.A. Gadd, M.C. Specht, C.A. Roche, T.M. Gudewicz and K.S. Hughes, The Feasibility of Using Natural Language Processing to Extract Clinical Information from Breast Pahology Reports, *J. Pathol Inform* **3**(1) (2012), 23.
[74] mbf Bioscience [Internet]. What is Whole Slide Imaging?, [cited 2021 July 13] Available from : https://www.mbfbioscience.com/whole-slide-imaging..
[75] International Agency for Research on Cancer [Internet]. Cancer, [updated 2021 March 3, cited 2021 July 13] Available from https://www.who.int/news-room/fact-sheets/detail/cancer.
[76] N. Sioutos, S. de Coronado, M.W. Haber, F.W. Hartel, W. Shaiu and L.W. Wright, NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information, *Journal of biomedical informatics* **40**(1) (2007), 30–43.
[77] M.A. Musen, N.F. Noy, N.H. Shah, P.L. Whetzel, C.G. Chute, M.D. Storey and B. Smith, The National Center for Biomedical Ontology, *J. Am. Medical Informatics Assoc.* **19**(2) (2012), 190–195. doi:10.1136/amiajnl-2011-000523.
[78] M. Girdea, S. Dumitriu, M. Fiume, S. Bowdin, K.M. Boycott, S. Chénier, D. Chitayat, H. Faghfoury, M.S. Meyn, P.N. Ray et al., PhenoTips: Patient phenotyping software for clinical and research use, *Human mutation* **34**(8) (2013), 1057–1065.
[79] J.C. Denny, Chapter 13: Mining Electronic Health Records in the Genomics Era, *PLoS Comput. Biol.* **8**(12) (2012). doi:10.1371/journal.pcbi.1002823.
[80] L. Lan, N. Djuric, Y. Guo and S. Vucetic, MS-kNN: protein function prediction by integrating multiple data sources, *BMC Bioinform.* **14**(S–3) (2013), S8. doi:10.1186/1471-2105-14-S3-S8.
[81] N. Noy, T. Tudorache, C. Nyulas and M. Musen, The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies, in: *AMIA Annual Symposium Proceedings*, Vol. 2010, American Medical Informatics Association, 2010, p. 552.
[82] N. Marini, S. Otálora, H. Müller and M. Atzori, Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification, *Medical image analysis* **73** (2021), 102165.
[83] N. Farahani, A.V. Parwani and L. Pantanowitz, Whole slide imaging in pathology: advantages, limitations, and emerging perspectives, *Pathology and Laboratory Medicine International* **7** (2015), 23–33.
[84] M. van Rijthoven, M. Balkenhol, M. Atzori, P. Bult, J. van der Laak and F. Ciompi, Few-shot weakly supervised detection and retrieval in histopathology whole-slide images, in: *Medical Imaging 2021: Digital Pathology*, Vol. 11603, International Society for Optics and Photonics, 2021, p. 116030N.
[85] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig et al., Gene ontology: tool for the unification of biology, *Nature genetics* **25**(1) (2000), 25–29.
[86] B.C. Grau, I. Horrocks, B. Motik, B. Parsia, P.F. Patel-Schneider and U. Sattler, OWL 2: The next step for OWL, *J. Web Semant.* **6**(4) (2008), 309–322. doi:10.1016/j.websem.2008.05.001.
[87] B. Chandrasekaran, J.R. Josephson and V.R. Benjamins, What are ontologies, and why do we need them?, *IEEE Intelligent Systems and their Applications* **14**(1) (1999), 20–26. doi:10.1109/5254.747902.
[88] T.R. Gruber, Toward principles for the design of ontologies used for knowledge sharing, *International Journal of Human-Computer Studies* **43**(5) (1995), 907–928. doi:https://doi.org/10.1006/ijhc.1995.1081. https://www.sciencedirect.com/science/article/pii/S1071581985710816.
[89] J. Frankovich, C.A. Longhurst and S.M. Sutherland, Evidence-based medicine in the EMR era, *N Engl J Med* **365**(19) (2011), 1758–1759.
[90] S.A. Murray, M. Kendall, K. Boyd and A. Sheikh, Illness trajectories and palliative care, *Bmj* **330**(7498) (2005), 1007–1011.
[91] C. Van Walraven, C. Bennett, A. Jennings, P.C. Austin and A.J. Forster, Proportion of hospital readmissions deemed avoidable: a systematic review, *Cmaj* **183**(7) (2011), E391–E402.
[92] S. Sell, Readmission within 30 days cost the NHS £1.6bn a year, 2010, online, available at https://www.gponline.com/readmissions-within-30-days-cost-nhs-16bn-year/article/1011933, retrieved October 2021.

[93] F. Aeffner, K. Wilson, N.T. Martin, J.C. Black, C.L.L. Hendriks, B. Bolon, D.G. Rudmann, R. Gianani, S.R. Koegler, J. Krueger et al., The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth, *Archives of pathology & laboratory medicine* **141**(9) (2017), 1267–1275.

[94] A. Kadadi, R. Agrawal, C. Nyamful and R. Atiq, Challenges of data integration and interoperability in big data, in: *2014 IEEE international conference on big data (big data)*, IEEE, 2014, pp. 38–40.

[95] S.W. Jahn, M. Plass and F. Moinfar, Digital Pathology: Advantages, Limitations and Emerging Perspectives, *Journal of Clinical Medicine* **9**(11) (2020), 3697.