International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

A Semantic Framework Supporting Multilayer Networks Analysis for Rare Diseases

Nicola Capuano, University of Basilicata, Italy*
Pasquale Foggia, University of Salerno, Italy
https://orcid.org/0000-0002-7096-1902
Luca Greco, University of Salerno, Italy
Pierluigi Ritrovato, University of Salerno, Italy

ABSTRACT

Understanding the role played by genetic variations in diseases, exploring genomic variants, and discovering disease-associated loci are among the most pressing challenges of genomic medicine. A huge and ever-increasing amount of information is available to researchers to address these challenges. Unfortunately, it is stored in fragmented ontologies and databases, which use heterogeneous formats and poorly integrated schemas. To overcome these limitations, the authors propose a linked data approach, based on the formalism of multilayer networks, able to integrate and harmonize biomedical information from multiple sources into a single dense network covering different aspects on Neuroendocrine Neoplasms (NENs). The proposed integration schema consists of three interconnected layers representing, respectively, information on the disease, on the affected genes, on the related biological processes and molecular functions. An easy-to-use client-server application was also developed to browse and search for information on the model supporting multilayer network analysis.

KEYWORDS

Biomedical Ontologies, Human Genome, Linked Data, Multilayer Network Analysis, Neuroen-Docrine Neoplasms, Rare Diseases, Semantic Information Integration

1. INTRODUCTION

The last few years have marked the explosion of data in the field of biomedicine. Several key events, such as the completion of the Human Genome Project, the advent of next-generation sequencing technologies and the Internet of Things, have led to a significant increase of the volume and variety of available biomedical data including medical records, imaging data, sequencing data, sensor data, etc. (Kamdar, Fernández, Polleres, Tudorache, & Musen, 2019).

DOI: 10.4018/IJSWIS.297141 *Corresponding Author

Ontologies and open databases are widely used in biology and medicine to store this huge and ever-increasing amount of information. Unfortunately, this often results in hundreds of large, fragmented, isolated, and heterogeneous data sources, each using a different format and scheme. As a matter of fact, healthcare professionals and biomedical researchers are facing serious difficulties in finding the information they need and even in mastering the enormous amount of available data. Furthermore, it should be considered that, while some information sources are primary (i.e., they collect data directly from articles published in biomedical journals), others are the result of systematic reviews. Without a method of critical evaluation and synthesis of this information, its integration, analysis, visualization and, in other words, translation into knowledge is almost impossible.

To overcome these limitations, new tools are needed, capable of querying multiple databases behind the scenes and providing researchers with integrated biomedical information and semantically interconnected entities (Fathalla, 2018). This integration must be transparent for researchers who would no longer have to worry about finding information sources, interpreting their syntax and schemas or mapping elements to reconcile concepts, relationships, and entities (Kamdar, 2018).

The research described in this paper goes exactly in this direction, aiming at the definition and implementation of a linked data application for the analysis, aggregation and study of available data related to Neuroendocrine Neoplasms (NENs), which are relatively rare neoplasms with 6.4-times increasing age-adjusted annual incidence during the last four decades (Grigoris Effraimidis, 2021). The developed system harmonizes the way information is stored in the existing biomedical information sources, thus contributing to the interoperability between these sources and improving the work of scientists in investigating these rare diseases.

The proposed solution interconnects information sources, representing a wide spectrum of current studies and expertise, by means of a single and robust ecosystem, thus providing the researcher with a quick access point to a dense network of information. Connected information include the National Cancer Institute Thesaurus, the Mondo Disease Ontology, the MedGen database, the Disease Ontology, the Orphanet Rare Disease Ontology, the DisGeNet database and the Gene Ontology.

Given the heterogeneity of interconnected information, the multilayer network formalism (implemented with semantic web languages and technologies) has been adopted to semantically link the available data sources (Hammoud & Kramer, 2020). Such networks are made up of distinct "layers" (each grouping concepts and relations), corresponding to different "aspects" of the domain, which are in turn connected with interlayer relationships. In particular, three interconnected layers have been designed that represent, respectively, information on diseases, affected genes, and biological processes and molecular functions of such genes and related gene products.

To the best of our knowledge, this is the first example of a multilayer network based on linked data and semantic web and the first tool for analyzing, aggregating, and studying data on rare tumors. By querying the system, researchers and healthcare professionals can obtain, through a user-friendly interface, answers to scientific questions such as relations between pathologies, involved genes and their mutations.

The paper is organized as follows: in section 2 the related work on biomedical data integration is summarized and the work is contextualized in the relevant literature; in section 3 background information on NENs and related biomedical data sources is provided; in section 4 the knowledge base architecture and the related integration issues are presented; in section 5 the developed prototype is described. The last section summarizes the conclusions and outlines the ongoing work.

2. RELATED WORK

The difficulty in using biomedical data is mainly due to the great heterogeneity of the data sources: querying them and exploiting the wealth of information they contain is a complex task full of obstacles. The problems that the researchers are required to face are mainly due to syntactic and semantic conflicts between data sources (Messaoudi, Fissoune, & Hassan, 2016). Syntactic conflicts are related to the

diversity and multiplicity of models (structured, semi-structured, unstructured) and data formats. These can be represented, for example, with a relational, object-oriented, or semi-structured XML model. Semantic conflicts are due to the presence of data from multiple sources, that may lead to different interpretations depending on local contexts, causing misunderstandings.

This is a problem also felt in other areas including cultural (Capuano, Gaeta, Guarino, Miranda, & Tomasiello, 2016), formative (Capuano, Longhi, Salerno, & Toti, 2015), legal (Hasan, et al., 2021), etc. but particularly relevant in biology and medicine. To address this, researchers need to have advanced skills and knowledge to find relevant data and perform their research effectively. Such skills can sometimes range from learning multiple systems' configurations and requirements up to coding. This process can greatly increase the complexity and time of scientific research so that, in many cases, the researcher ends up simply looking at the web portals and using the available search engines (e.g., PubMed) to retrieve, as best as possible, information they need.

Being a deeply felt problem in biomedicine, several projects for the integration of biological data sources have been proposed over time. For example, the Gene Expression Data Warehouse (GEDAW) project developed an object-oriented data warehouse to store and manage relevant information on liver gene expression data and related biomedical resources (Guérin, et al., 2005). It systematically integrates gene information from a multitude of structured data sources including GenBank, BioMeKe, and an internal database with detailed experimental data on liver genes.

Bio2RDF is an open-source project aimed at transforming a vast collection of heterogeneously formatted biomedical data into linked data through semantic web technologies (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008). With Bio2RDF, public bioinformatics database documents including Kegg, PDB, MGI, HGNC etc. are made available in RDF. The third edition of Bio2RDF is made up of 11 billion triples in 35 datasets and constitutes one of the largest collections of linked data for life sciences. It also includes scripts to automatically convert data from various formats (e.g., text, XML, SQL, etc.) into RDF (Dumontier, et al., 2014).

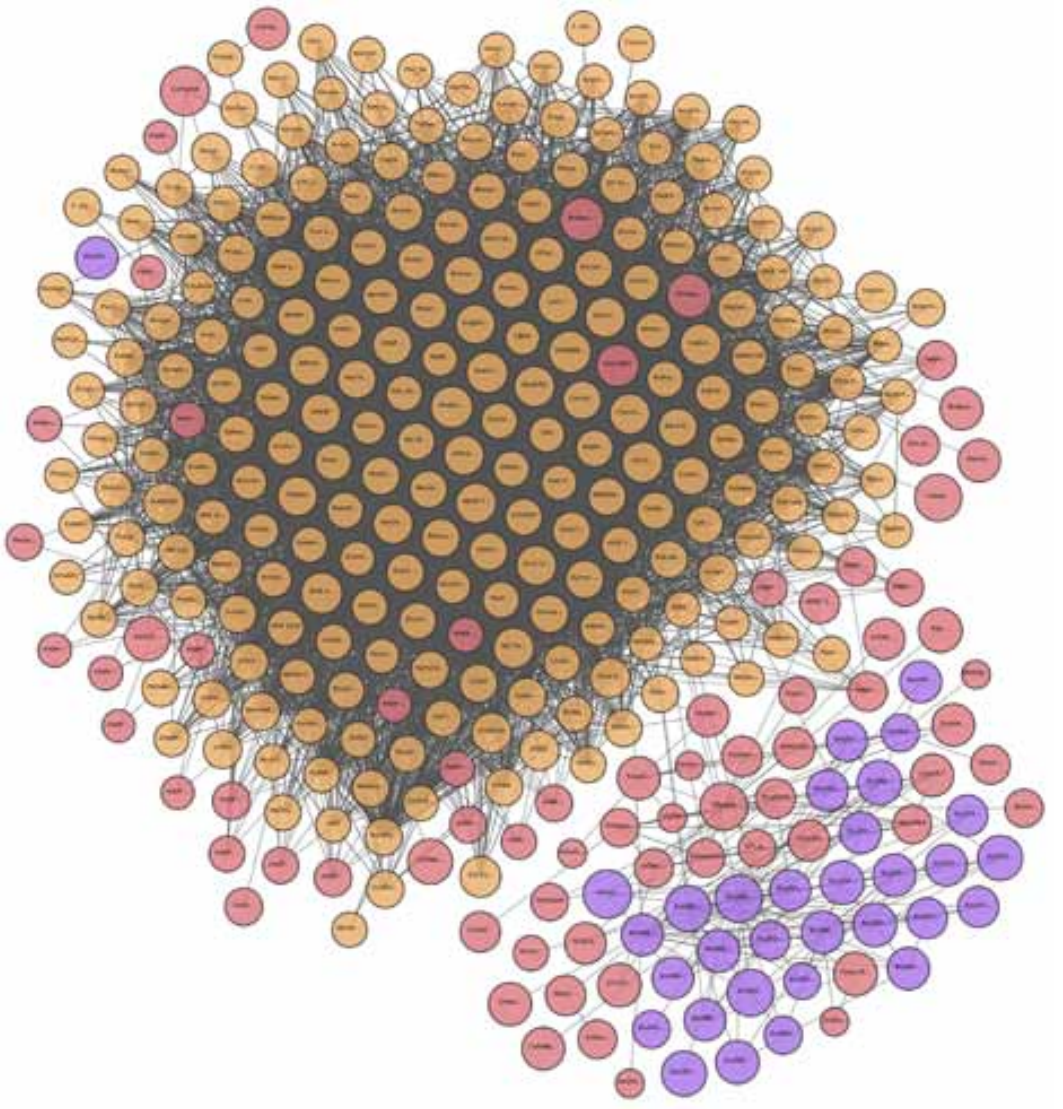
Bio2RDF, together with other biomedical ontologies (most of which are collected in the BioPortal repository[2]) constitutes the Life Sciences Linked Open Data (LSLOD), part of the wider Linked Open Data initiative (Bizer, Heath, & Berners-Lee, 2009). Figure 1 shows the LSLOD cloud where data sources are represented as circles while semantic relationships between data sources are represented with gray lines.

The Knowledge Base of Biomedicine (KaBOB) is another project aimed at integrating 18 biomedical data sources using 14 ontologies from the Open Biomedical Ontologies (OBO) initiative[3], thus facilitating the interaction of these sources with data and tools that already rely on these ontologies (Livingston, Bada, Baumgartner, & Hunter, 2015). In KaBOB, identity between data sources is maintained through the generation of a single biomedical entity for each set of equivalent data source-specific identifiers. These entities, connected with the ontology concepts, serve as building blocks for common biomedical representations, which can be simultaneously modeled and queried at multiple abstraction levels.

The Genomic and Proteomic Knowledge Base (GPKB) integrates several biomedical data sources including Entrez Gene, UniProt, IntAct, Expasy Enzyme, GO, GOA, BioCyc, Kegg, Reactome and OMIM (Masseroli, Canakoglu, & Ceri, 2016). Like other initiatives described so far, it adopts a global schema based on the abstraction and generalization of the integrated sources. It also includes a set of procedures for the integration and maintenance of data, capable of updating the knowledge base considering the evolutions of the integrated sources, ensuring coherence, and performing the semantic closure of the hierarchical relationships of the adopted ontologies. The system also provides a web interface[4] for the composition of queries on the knowledge base.

The Software for Flexible Integration of Annotation (SoFIA) is a framework for workflow-driven integration of omics information from multiple sources (Childs, Mamlouk, Brandt, Sers, & Leser, 2016). To avoid the information overload caused by similar software, returning all available information related to a given query, SoFIA applies a goal-oriented approach. It conceptualizes a

Figure 1. The Life Sciences Linked Open Data cloud[1] in May 2021



set of workflow templates that cover different integration goals. Given a specific integration task, consisting of a goal, a set of data sources and required outputs, SoFIA composes a minimal workflow that completes the task and returns only the subset of information that is really needed.

SysCancer is a research project aimed at developing an integrated system that combines all stages of cancer studies (Bensz, et al., 2016). The central data warehouse, developed as part of the project, is used to support multidimensional analysis starting from local databases after the data, intended for public access, has been gathered and integrated. A computational cluster is responsible for performing complex analysis on this data with advanced algorithms.

Based on the analysis of existing systems and in accordance with (Messaoudi, Fissoune, & Hassan, 2016), the current approaches to biomedical data integration, like those described so far, can be broadly classified into three categories: data warehouse (i.e., databases that integrate a selected set of data into a common schema); linked data integration systems (i.e., based on the adoption of

semantic web standards); workflow-based integration systems (i.e., integrating external data sources, based on a predefined pattern, to respond to a specific request).

In this work we propose a hybrid integration approach based on linked data supporting multilayer network analysis. Furthermore, like workflow-based systems, our approach is also capable of integrating external, non-semantic data sources according to specific integration patterns. Since the integration of data from these external sources is made "on the fly" on a frequently updated local copy, according to user requests, an advantage of this approach is that there is no danger of using obsolete data.

3. BACKGROUND

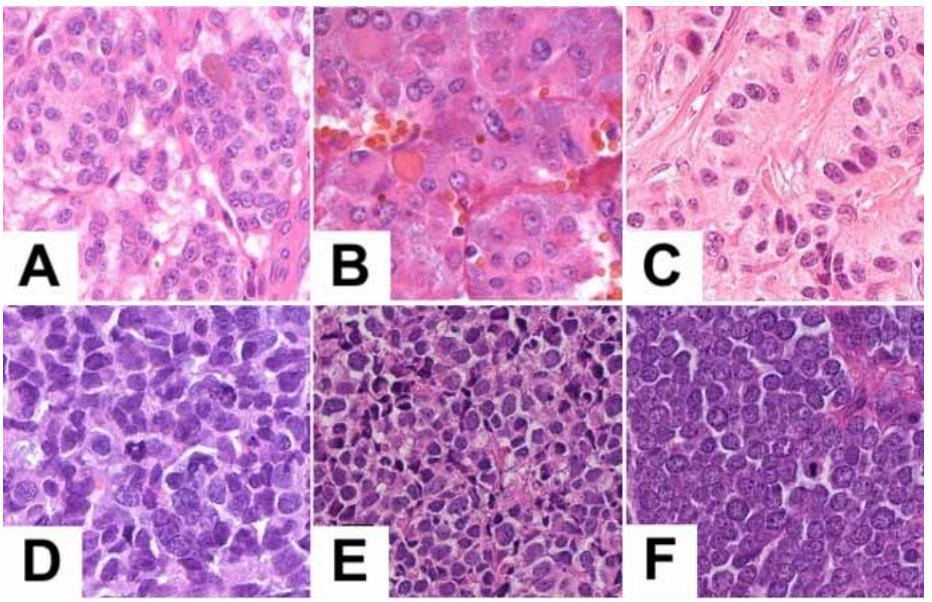
As we are interested in the definition and development of a linked data application for the analysis, aggregation and study of existing information about Neuroendocrine Neoplasms (NENs), in this section some background information about such rare diseases (section 3.1) is provided, followed by a brief overview of the main related biomedical information sources used for the semantic integration task (section 3.2).

3.1. Neuroendocrine Neoplasms

NENs can arise in most of the epithelial organs of the body, in pure endocrine organs, in nerve structures or in the diffuse neuroendocrine system. They can be organized in two different groups: poorly differentiated Neuroendocrine Carcinomas (NECs) and well-differentiated Neuroendocrine Tumors (NETs). Figure 2 (Rindi & Inzani, 2020) shows sections of NEN cancerous tissue (stained with hematoxylin and eosin) including NET samples such as lung carcinoid (A), pheochromocytoma (B) and insulinoma (C), as well as NEC samples such as synaptophysin (D), gastric NEC (E) and cutaneous NEC (F).

The classification of NENs into NECs and NETs has only recently been proposed by the World Health Organization (WHO) with the aim of allowing specialists to manage these diseases consistently, regardless of anatomical location, thus reducing inconsistencies and contradictions

Figure 2. Sections of different types of NEN cancerous tissue



among the organ-specific system previously in use (Rindi, et al., 2018). This classification, based on a consensus conference held at the International Agency for Research on Cancer (IARC) in 2017 and subsequently enriched in 2019 (Nagtegaal, et al., 2019), suggests distinguishing NENs based on their degree of differentiation, which is inferred from the appearance of cancer cells observed under the microscope. In particular:

- NETs are well-differentiated neoplasms classified into three levels as G1, G2 and G3, corresponding to low, intermediate, and high grade.
- NECs are poorly differentiated neoplasms and are always high grade i.e., G3.

As reported in Table 1, the grade and cell differentiation depend, in turn, on other factors such as mitotic count and Ki-67 cell labeling index. Furthermore, according to the type of cells that characterize them, NECs can in turn be divided into small- and large-cell type NECs. As described in section 4, this information was used to retrieve and characterize the NENs within the medical information sources used by the proposed integration model.

Despite the ongoing standardization process, the current nomenclature on NENs, while including established and accepted definitions, still presents variants related to the different anatomical sites. This heterogeneity in terminology and classification creates confusion and hinders the integration of information from different data sources. This justifies the creation of a tool, such as the one proposed in this work, aimed at supporting researchers and specialists who collect, organize, and systematically analyze the existing biomedical data on these diseases (which, being rare, are almost ignored the industrial sector). The aim is to provide them with a global and unified view of this data in a single, well-harmonized knowledge base.

3.2. Biomedical Information Sources

Integrated information sources include existing biomedical ontologies and databases, storing heterogeneous data, often overlapping and poorly connected (or not connected at all), about diseases (cancers and rare diseases), genes, gene products, biological processes, and molecular functions as well as known gene-disease and disease-disease associations. Relevant information for study and research on NENs has been extracted from these sources (which are currently only accessible independently) and connected into a single multi-layered knowledge model accessible through an easy-to-use client-server application. The list of integrated sources is shown in Table 2 with a reference to the official website, more details are provided below.

The National Cancer Institute Thesaurus (NCIT) is an open-source medical ontology aimed at providing a controlled vocabulary usable by researchers and specialists in the various subdomains of oncology (Kumar & Smith, 2005). It provides stable and unique codes for biomedical concepts, preferred terms, synonyms, research codes and other information. At a high level, the NCIT ontology includes the concept of disease that describes several cancers with their properties including cellular,

Table 1. Classification and grading criteria of NENs

Class	Subclass	Differentiation	Grade	Mitotic rate	Ki-67 index
NET	G1	Well-differentiated	Low	< 2	< 3%
	G2		Intermediate	2–20	3–20%
	G3		High	> 20	> 20%
NEC	Small-cell type	Poorly differentiated	High	> 20	> 20%
	Large-cell type				

Table 2. List of integrated biomedical information sources

Information Source	Format	Website
National Cancer Institute Thesaurus (NCIT)	OWL, OBO	ncithesaurus.nci.nih.gov
Orphanet Rare Disease Ontology (ORDO)	OWL	www.ebi.ac.uk/ols/ontologies/ordo
Disease Ontology (DO)	OWL, OBO	disease-ontology.org
Mondo Disease Ontology (MONDO)	OWL, OBO	mondo.monarchinitiative.org
Gene Ontology (GO)	OWL, OBO, CSV	geneontology.org
MedGen database	CSV	www.ncbi.nlm.nih.gov/medgen
DisGeNet database	RDF, CSV	www.disgenet.org

anatomical, morphological, and clinical characteristics. It also includes molecular concepts such as genes, proteins, pathways, expression of fusion proteins and chromosomal translocations, used both for data encoding and as links that relate diagnostic and therapeutic concepts. It includes over 100,000 definitions and over 400,000 cross-links between concepts and is updated frequently by a team of experts (Merabti, Joubert, Lecroq, Rath, & Darmoni, 2010).

The Orphanet Rare Disease Ontology (ORDO) provides information on rare diseases (i.e., occurring in less than 1 in 2000 people) and aims to help improving their diagnosis and treatment. It captures relationships between the following main classes: clinical entities (diseases, groups of disorders, syndromes, etc.), epidemiology (annual incidence, cases/families etc.), genetic material (genes with proteins product, non-coding RNA, etc.), geography (diffusion in countries), and inheritance (autosomal dominant, autosomal recessive, etc.). It includes a multi-hierarchical thesaurus consisting of more than 7,000 entries and 4,000 synonyms (Vasant, et al., 2014).

The Disease Ontology (DO) provides consistent, reusable, and sustainable descriptions of human disease terms, phenotype characteristics and disease concepts. DO terms are well defined and use references to established terminologies including NCIT and the Unified Medical Language System (UMLS) thesaurus (Bodenreider, 2004). DO includes concepts representing types of diseases, anatomical entities, cells, phenotypes, symptoms associated with anatomical areas, inheritance patterns, and transmission processes. It currently includes more than 10,000 terms (Schriml, et al., 2012).

The Mondo Disease Ontology (MONDO) is a semi-automatically constructed ontology aimed at harmonizing disease definitions across different data sources (including NCIT, ORDO and DO) to address the lack of unified disease terminology. Its ontological scheme provides a hierarchical structure that can be used to classify diseases. It provides the representation of various concepts that identify the disease (acute, degenerative, etc.), its characteristics (rare or common, syndromic, or isolated, etc.), and its relative susceptibility (e.g., inheritance).

The Gene Ontology (GO), part of the larger Open Biomedical Ontologies (OBO) project, aims to develop and maintain a controlled vocabulary (including more than 40,000 terms) for describing genes and gene products in all species. It consists of three domains: cellular components (describes the parts of a cell, or its extracellular environment, where a gene product is active), molecular functions (describes the elementary activities of a gene product at the molecular level, such as ligand or catalysis), and biological processes (describes the biological molecular operations or events to which a gene or gene product contributes). Each GO-term (an ontology class) has a unique identifier, a definition, and several relations to other terms. The additional GO Annotation File (GOF) is a CSV file including more than 8 million statements about the function of genes and gene products, resulting from genome annotation studies (Ashburner, et al., 2000).

The MedGen database organizes medical genetic information, including terms and their relations, through stable unique identifiers. It is a comprehensive resource for accessing essential information on phenotypic health topics related to human medical genetics, gathered from established, high-quality sources. Among the functions performed by MedGen is the mapping of terms from different ontologies and data sources (including NCIT, ORDO and UMLS). As described in the next section, this is the main MedGen feature that was used in this work (Louden, 2020).

The DisGeNet database integrates and standardizes data on disease-associated genes and variants from multiple sources including scientific literature. It covers the entire spectrum of human diseases as well as normal and abnormal traits, including over 30,000 diseases and traits, 20,000 genes, and 1 million gene-disease associations. DisGeNet can be used to study the molecular basis of human diseases and their comorbidities, the properties of the disease genes, the hypotheses on the therapeutic action and adverse effects of drugs, etc. It is also useful for extrapolating gene-disease, variant–disease and disease-disease associations i.e., similarities between diseases based on shared genes and variants (Piñero, et al., 2020).

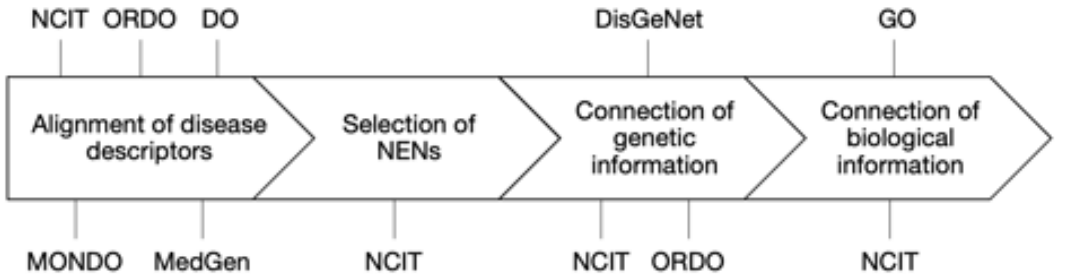
4. INTEGRATION AND HARMONIZATION APPROACH

The first aim of this research was to obtain a uniform classification framework for NENs within the three main ontologies currently in use in most medical applications: NCIT, DO and ORDO (see section 3.2). This was done by reducing the inconsistencies and contradictions between schemes and instances through a linked data approach that also relies on the integration of further sources, namely MedGen and MONDO. Then, on the harmonized model, additional information sources covering NENs-related genetic and molecular information (i.e., DisGeNet and DO) were integrated through a semantic-based multilayer network model.

The diagram in Figure 3 summarizes the proposed harmonization approach with integrated (above) and supporting (below) data sources for each step. First, the alignment of the NCIT, ORDO and DO disease descriptors is carried out using the information included in MONDO and MedGen. Then (step 2) only the information relating to the NENs is selected, exploiting the properties of NCIT. In the third step, the genetic information related to the selected diseases (coming from DisGeNet) is integrated with the information of the same type already gathered from NCIT and ORDO. Finally (fourth step), also the biological information (coming from GO) is integrated with the information of the same type collected from NCIT, thus obtaining the complete integrated model.

Based on this approach, in this section we formally characterize the layered structure of the defined integrated knowledge model (section 4.1), then we describe the composition and integration issues related to the definition of each layer (sections 4.2 onwards).

Figure 3. Flow diagram of the proposed integrated approach



4.1. Multilayer Network Model

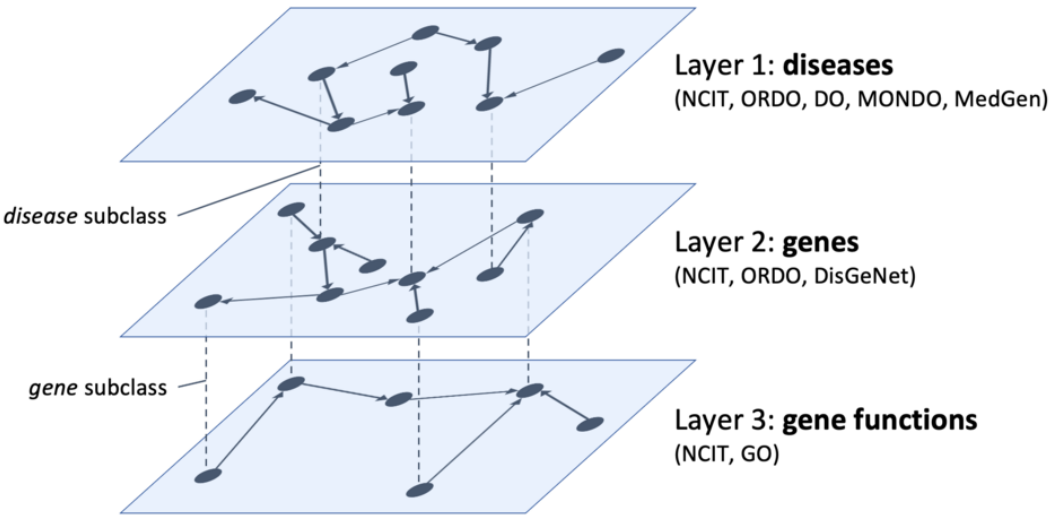
Complex domains, like the one we are modelling, are characterized by heterogeneous entities related in different ways and include multiple subsystems and levels of connectivity. Multilayer networks are an emerging formalism able to represent this complexity through a generalization of the graph structure where the nodes and edges are distributed on different layers, each representing an "aspect" of the domain (Kivelä, et al., 2014). Multilayer networks are particularly effective for modeling biological systems such as gene co-expression networks, protein-protein interaction networks or pathways. Their importance in biomedicine has been thoroughly investigated in (Hammoud & Kramer, 2020) where several variants of this model have also been formalized.

According to (Boccaletti, et al., 2014) we define a multilayer network as a triple $M = \left(V, E, L\right)$ where the sets V and E represent, respectively, the nodes and edges of the network while $L = \left\{L_1, \ldots, L_d\right\}$ is the set of network layers. In turn, each $L_i \in L$ is a subgraph $L_i = \left(V_i, E_i\right)$ composed by the nodes V_i and the edges E_i such that $V = \bigcup_{i=1}^{d} V_i$ and $E = \bigcup_{i=1}^{d} E_i$.

In this version (also called multilevel or multidimensional network) each node can appear in several layers but have direct connections only with the nodes of the same layer. Therefore, there are no explicit interlayer connections, but these are implicitly represented by the projections of the same node in different layers. A further constraint of our model is that each pair of adjacent layers shares at least one node while non-adjacent layers have no shared nodes i.e., $L_i \cap L_j \neq \varnothing$ iff $\left|i - j\right| = 1$. This allows to better encapsulate the entities and relations of a domain aspect into a single layer using shared nodes as bridges between related aspects.

As shown in Figure 4, our model is made of three interconnected layers that represent, respectively, information on diseases, affected genes, and their functions (i.e., biological processes and molecular functions). A linked data approach was used to harmonize information at each level across the selected data sources (shown in the figure) within a single ontology. This process is detailed in the following sections. Then specific concepts are used as bridges between layers (i.e., shared nodes). In particular, the disease class (and its subclasses) is used to move from layer 1 to layer 2 while gene class (and their subclasses) are used to move from layer 2 to layer 3.

Figure 4. Visual representation of the defined three-level network model



Interlayer relations, which link the same concept in different ontologies, are implemented with the equivalent-class OWL statement. The same statement has been used to link the defined ontologies with the original sources (when these are ontologies themselves). Navigation and search within and between layers have been implemented at the application level as described in section 5.

4.2. First Layer: Diseases

The first layer was obtained by extrapolating and harmonizing information on NENs from NCIT, ORDO and DO ontologies (see section 3.2). A first issue of this task concerns the absence, for most of the diseases described by NCIT and ORDO, of a uniform identification code. The lack of such a code hinders the integration between these sources and with external information. To solve the problem, disease names of these ontologies were mapped to UMLS codes using two additional external resources: the MONDO ontology and the MedGen database (see section 3.2). Luckily, the DO ontology already includes the UMLS codes of the described diseases.

A first mapping of the codes was carried out using the MONDO ontology (whose purpose is precisely to unify the nomenclatures of known diseases). All the subclasses of the OWL class cell-proliferation-disorder were retrieved from MONDO. Then, UMLS and NCIT codes associated with the resulting neoplasms were retrieved and used to build such mapping. Unfortunately, this search was not exhaustive: in fact, several NCIT NETs and NECs still lacked UMLS code. The MedGen database was used to obtain the missing disease identification codes. In addition to providing UMLS code mappings, this database also includes weekly updated information on the validity of those codes. Data within the names archive was used to obtain the non-suppressed UMLS codes associated with NCIT codes, data within the ordo-cui-history archive was used to obtain the non-suppressed UMLS codes associated with ORDO codes. Archive schemas are summarized in Table 3.

Once the neoplasm descriptors were aligned, only the NENs were extrapolated from the three ontologies starting from their properties and molecular characteristics. Given that only NCIT contains this detailed information, reflecting the results of more recent studies, NENs are first retrieved on NCIT and then, thanks to the obtained alignment, retrieved also on ORDO and DO.

For the extraction of the NETs from NCIT, the following properties of the disease class were used: disease-has-abnormal-cell (which allows to carry out a first filtering considering only neoplasms that present anomalies in the neuroendocrine cells) and disease-has-finding (which allows to select only neoplasms that present the appearance of tumor cells with well-differentiated lesions). In this way all NETs were extrapolated, regardless of their grade (which ranges from 1 to 3). For the extraction of the NECs from NCIT the following properties of disease were used: disease-has-finding (which allows to carry out a first filtering by considering only neoplasms that have a poorly differentiated

Table 3. MedGen archives and fields used for identifier harmonization

Archive	Field	Description	Example
names	CUI	UMLS disease or medical resource identifier	C4054192
	name	Name that the disease or medical resource has within the data source indicated in the source field	Peritumoral Brain Edema
	source	Data source name (e.g., SNOMED CT, NCIT, OMIM, GTR)	NCIT
	suppress	Flag indicating whether the UMLS code-resource association is still valid or has been suppressed (Y or N)	N
ordo-cui-history	ORDO-id	Name that the disease or medical resource has in ORDO	Orphanet_247353
	CUI	UMLS disease or medical resource identifier	C0343055
	is-current	Flag indicating whether the association is currently in use (0 or 1)	1

appearance under the microscope as well as high mitotic rate) and disease-is-grade (which specifies the degree of cell diffusion and proliferation that, in the case of NECs, must be 3).

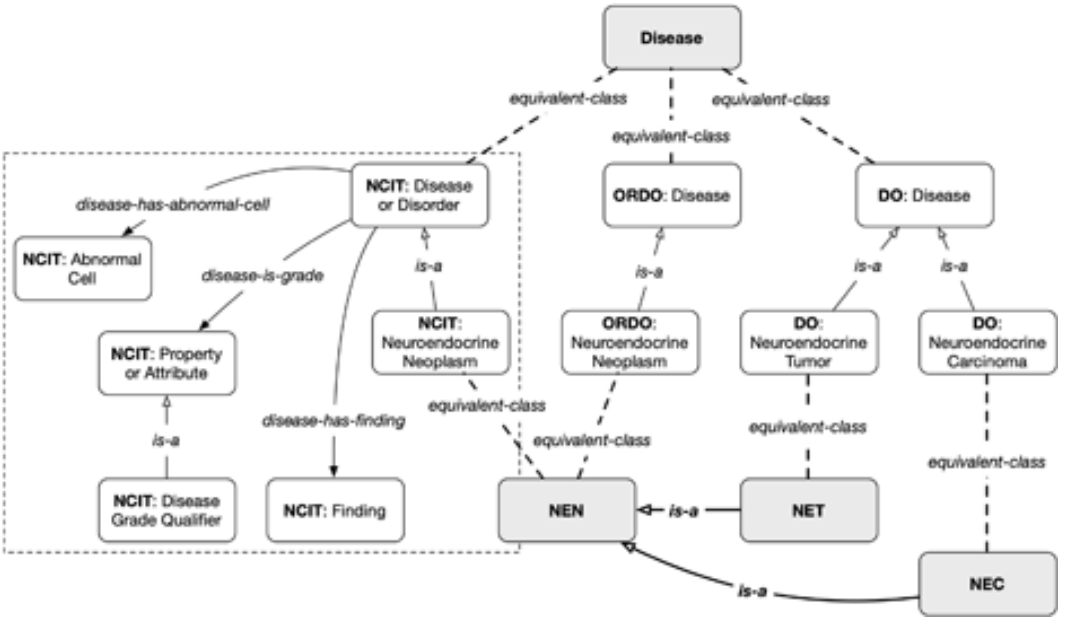
The retrieved diseases were annotated as NETs and NECs, respectively, and the associated UMLS codes were used to retrieve and associate the same diseases in ORDO and DO. During this association, the presence of some neuroendocrine carcinomas not correctly classified in the ORDO ontological scheme was found. Figure 5 represents some high-level classes and relations of the harmonized model. The NCIT classes inside the dotted box were used for the identification of NETs and NECs, the gray classes, and the relations in bold were introduced by the level 1 integration schema to group and reorganize NETs and NECs under the harmonized ontologies.

4.3. Second Layer: Genes

The second layer provides information on the variations in the human genome that lead to the NENs described at the first level, such as permanent mutations or changes in the structure of a gene. Various genetic information is collected in this layer, including: the genes involved in the pathogenesis of the diseases and their characteristics; the cytogenetic anomalies, the potential molecular anomalies, the single-nucleotide polymorphism (SNP) or variants (SNV) i.e., the variations of the genetic material in a single nucleotide. Disease-disease associations (DDA) are also included which represent the result of studies that relate human diseases through their molecular causes based on the network of associations between genes, proteins, environmental factors, etc. This information may also be useful to specialists in assessing the comorbidity index.

Part of this information is already included in some of the biomedical data sources considered (i.e., NCIT and ORDO), further information has been extracted from DisGeNet (see section 3.2). Using the disease-mapped-to-gene property applied on NENs within NCIT, it was possible to extrapolate the genes associated with these diseases. Additional information of each gene was then retrieved including the observed anomalies (gene-has-abnormality property) and the related cytogenetic anomalies (disease-has-cytogenetic-abnormality property). From the latter information (if present), the chromosome to which the anomaly is connected was also obtained (cytogenetic-abnormality-

Figure 5. First layer - some relevant classes and relations of the harmonized model



involves-chromosome property) as well as the related molecular anomalies (disease-has-molecular-abnormality property), if present. The left part of Figure 6 shows the top-level classes and relations involved, as well as the mappings with that introduced by layer 2 integration schema (in gray). A similar process was also applied for the extraction of genetic information from ORDO where all the properties, that lead from a disease to the connected genes, have been considered.

As anticipated, information from NCIT and ORDO has been connected with information from the DisGeNet database which includes and integrates data on genes and diseases from various sources such as biomedical ontologies and scientific articles. The main archive fields used to this aim are summarized in Table 4. To obtain the gene-disease associations, the DisGeNet gene-disease archive was considered by extracting, starting from the UMLS codes of NENs, the genes whose association with the disease has been confirmed in at least 6 studies (n-of-pmids-association field) from 2010 on. In addition, only the associations reporting an evidence index of 1 were considered, thus indicating that all papers support the association.

The genetic variation-disease associations were obtained from the corresponding archive in a similar way, including the variation that causes the onset, the chromosome, and the chromosomal position in which it is located. In this case only variations confirmed by at least 30 field-based studies with an evidence index of 1 were considered. Through the genetic-variation-gene archive it was also possible to obtain the gene associated with this variation starting from the variant identifier (snp-id field). Finally, from the corresponding archive, the disease-disease associations for NENs were also obtained by considering diseases that have at least 3 genes in common (a similar association could also have been obtained considering the number of variations in common).

The right part of Figure 6 represents the top-level classes (in gray) used to encode the information gathered from DisGeNet and connected to the other classes of the layer 2 integration schema (in gray). At the end of this phase, the sets of genes most involved in the onset of NETs and NECs are also obtained. To this end, genes were considered whose association with these diseases has been confirmed since 2010 by at least 30 scientific publications or biomedical data sources. These sets constitute a useful reference for researchers and sector specialists and their finding is an original result of this data integration project.

4.4. Third Layer: Gene Functions

The third layer provides additional information on genes and gene products responsible for the onset of NENs including their molecular functions (the elementary activities of a gene product at the molecular

Figure 6. Second layer - some relevant classes and relations of the harmonized model

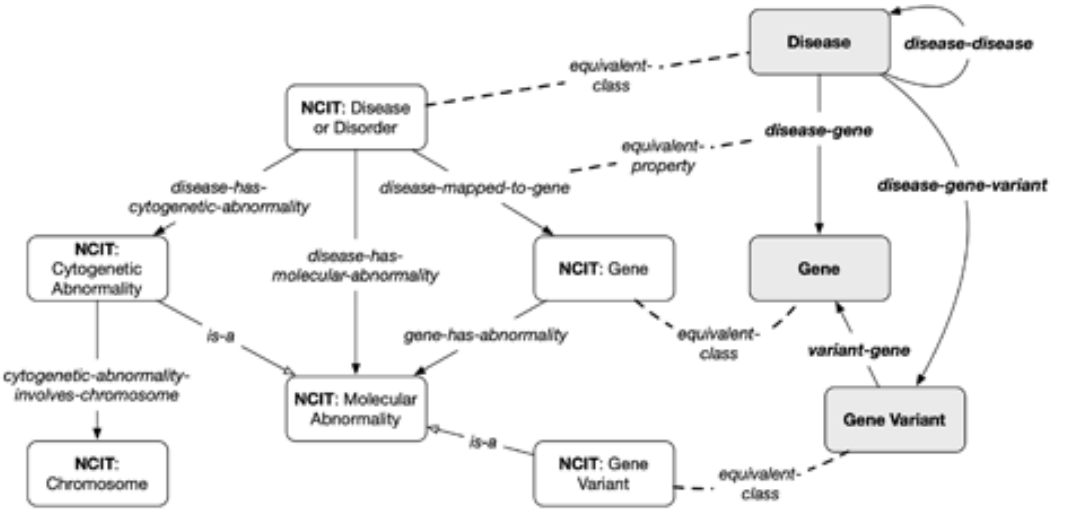


Table 4. DisGeNet archives and main fields used for the extraction of NEN genetic information.

Archive	Field	Description	Example
gene-disease	gene-id	NCBI Entrez gene identifier	4221
	gene-symbol	Gene symbol	MEN1
	disease-id	Disease UMLS code	C0238462
	n-of-pmids-association	Total number of publications reporting the gene-disease association	4
	ei	Evidence index for the gene-disease association	1
genetic-variation-disease	snp-id	dbSNP variant Identifier	rs794728640
	chromosome	Chromosome of the variant	11
	position	Position in chromosome	64807914
	disease-id	Disease UMLS code	C0025267
	n-of-pmids-association	Total number of publications reporting the variant-disease association	6
	ei	Evidence index for the gene-disease association	1
genetic-variation-gene	snp-id	dbSNP variant Identifier	rs794728640
	gene-id	NCBI Entrez gene identifier	4221
disease-disease-association	disease-id-1	UMLS code of the first disease	C0238462
	disease-id-2	UMLS code of the second disease	C0007131
	n-genes	Genes in common among diseases	1
	n-variants	Gene variants in common among diseases	1

level, such as binding or catalysis) and biological processes (operations or sets of events relevant to the functioning of integrated living units: cells, tissues, organs, and organisms). Molecular functions correspond to the activities that can be performed by single gene products (i.e., a protein or RNA) or by molecular complexes composed of several gene products. Biological processes include both specific processes such as glucose transmembrane transport and broad processes such as DNA repair. It should be noted that the relations between gene products (or groups of gene products) and biological processes or molecular functions are one-to-many, reflecting the biological reality that a particular protein can function in different processes, contain domains that perform different molecular functions and participate in multiple interactions with other proteins, organelles or locations in the cell (Magee, 2011).

The information used to build the third layer comes from NCIT and GO (see section 3.2). In a first step, the products encoded by genes are retrieved on NCIT following the gene-encodes-product relation. Then, the GO Annotation File (GOF) is used which contains a large set of statements associating gene products with molecular functions and biological processes. Table 5 shows the main fields used to this end: in particular, the db-object-symbol field identifies a gene product while the GO-id field identifies the activity within the GO ontology. The aspect field qualifies this association as a biological process or a molecular function while the db-object-type field provides a description of the gene product. It should be noted that GOF statements must be semantically interpreted. In fact, the association described so far can be modified based on the optional value of the qualifier field which can be one of the following:

- The value "not" indicates that it has been experimentally demonstrated that a gene product does not perform a particular activity, or it has been shown to have had a loss of function over the

Table 5. The main fields of the GO Annotation File (GOF) used for gene product association

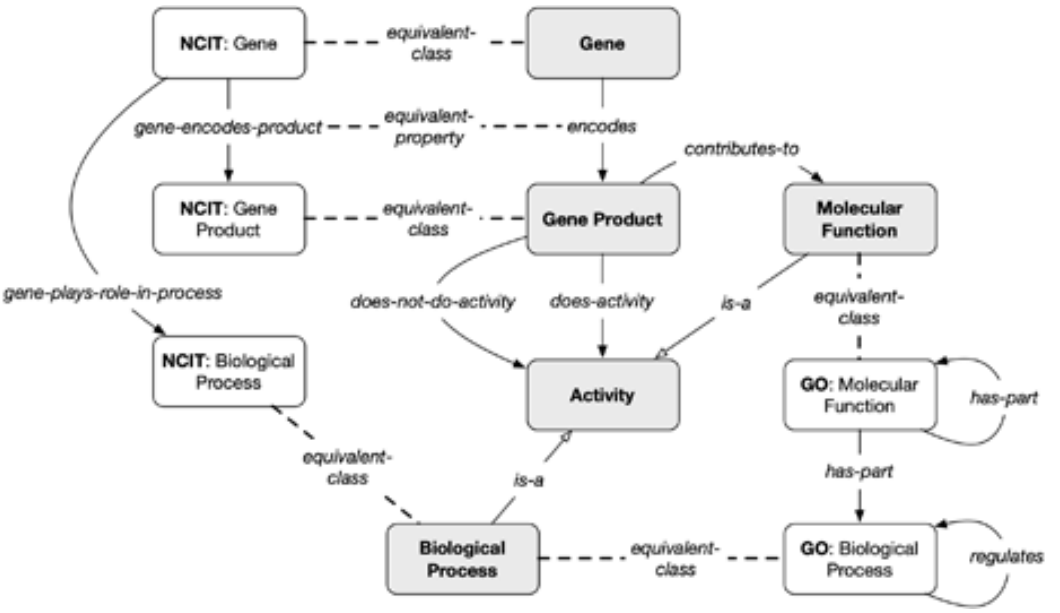
Archive	Field	Description	Example
annotation	db-object-symbol	A unique and valid symbol identifying the gene product	PHO3
	db-object-name	Gene product name	Toll-like receptor 4
	db-object-type	Gene product type description	protein
	GO-id	Unique identifier of the GO ontology representing the activity associated to the gene product	GO:0003993
	aspect	Activity type (P for biological process, F for molecular function)	F
	qualifier	Optional item that modifies the interpretation of an annotation (not, contributes-to)	not

course of evolution. The entry must therefore be interpreted as a "non-association" between the gene and the GO-term.

- The value "contributes-to" is used only for molecular functions when a function of a protein complex is facilitated, but not directly performed by one of its subunits. It is particularly useful for annotating molecular functions in cases where a complex has an activity, but not all the individual subunits are involved.

Once the GO-terms corresponding to a function or process are retrieved, the main GO ontology is queried to obtain information on such activities, including the name, the type, and the description. For molecular functions, the processes, or functions of which it is a part are also extrapolated through the part-of property. Then, the regulated biological processes are obtained through the regulates property. Figure 7 represents some high-level classes and relations of the harmonized model obtained for the

Figure 7. Third layer - some relevant classes and relations of the harmonized model



third layer. It also takes care of harmonizing GO biological processes with those described by NCIT and connected to genes through the gene-plays-role-in-process property.

5. DEVELOPED PROTOTYPE

This section describes the client-server application developed to browse and search for information on the integrated model described in section 4. The system architecture is described in section 5.1 as well as the functions provided and the user interface. Section 5.2 discusses system performance and includes some considerations on validating the system and the underlying knowledge model.

5.1. Architecture and User Interface

The Virtuoso Universal Server was selected as the middleware to store the original biomedical ontologies and databases; it is an open-source solution able to manage different data formats and access protocols simultaneously. The original datasets are copied to the server, which is also in charge of updating them periodically, starting from the original endpoints (which were not directly used for better performance). Moreover, the same server also includes the defined multilayer integration model.

End users can access server information via a lightweight Java desktop application. Based on a query specified via an easy-to-use visual interface, a SPARQL query sequence is generated and forwarded to the server via HTTP. Then, the results obtained by the server are used to compose the answer that is shown graphically to the user. The Java client uses the Jena framework to manage RDF graphs and query them via SPARQL. OWL APIs were also used for the client-side manipulation of OWL ontologies. Figure 8 summarizes the main components of the system architecture.

Figure 9 shows the "diseases" section of the client application. It is the first view presented to the user and allows to obtain the classification of NETs and NECs in each of the three ontologies considered (NCIT, ORDO and DO). The diseases obtained for each ontology are the result of the integration work described in section 4.2 which also considers information from MONDO and MedGen. The interface has been designed to provide the user with both a broad view of the information and a partial view: in fact, he will be able to select only some pathologies that will be stored for subsequent steps.

Figure 10 shows the "genetic information" section of the application. It allows to obtain information relating to the genes involved in the diseases selected in the first phase, including: gene-disease, variation-disease and disease-disease associations, cytogenetics anomalies, molecular

Figure 8. Client-server architecture of the developed prototype

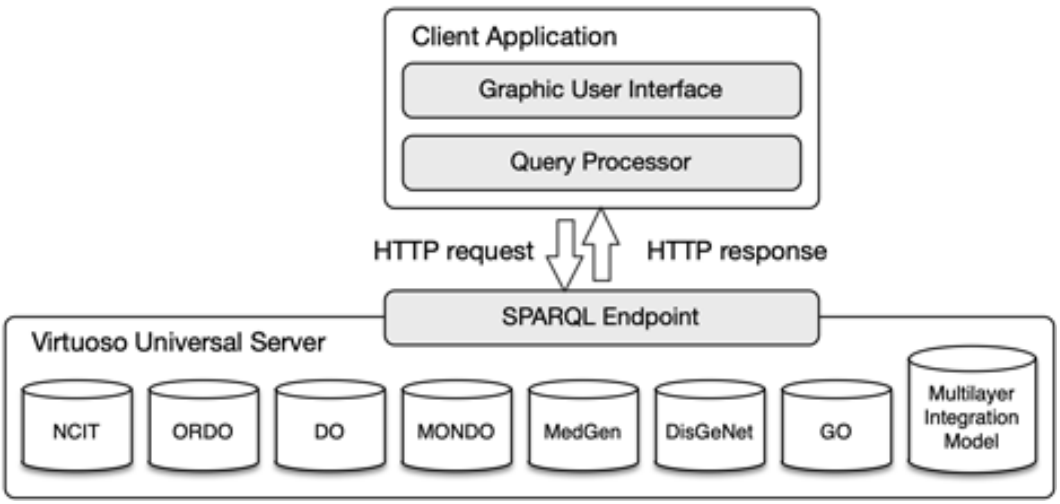
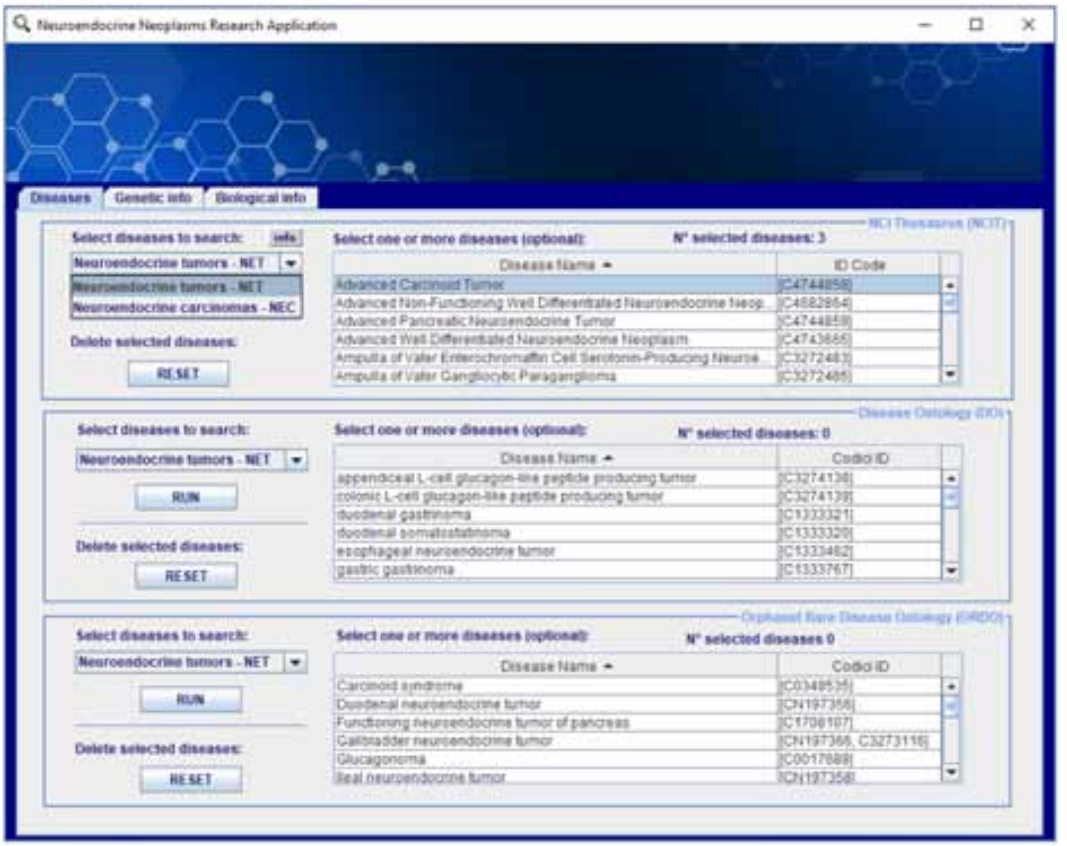


Figure 9. "Diseases" section of the client application



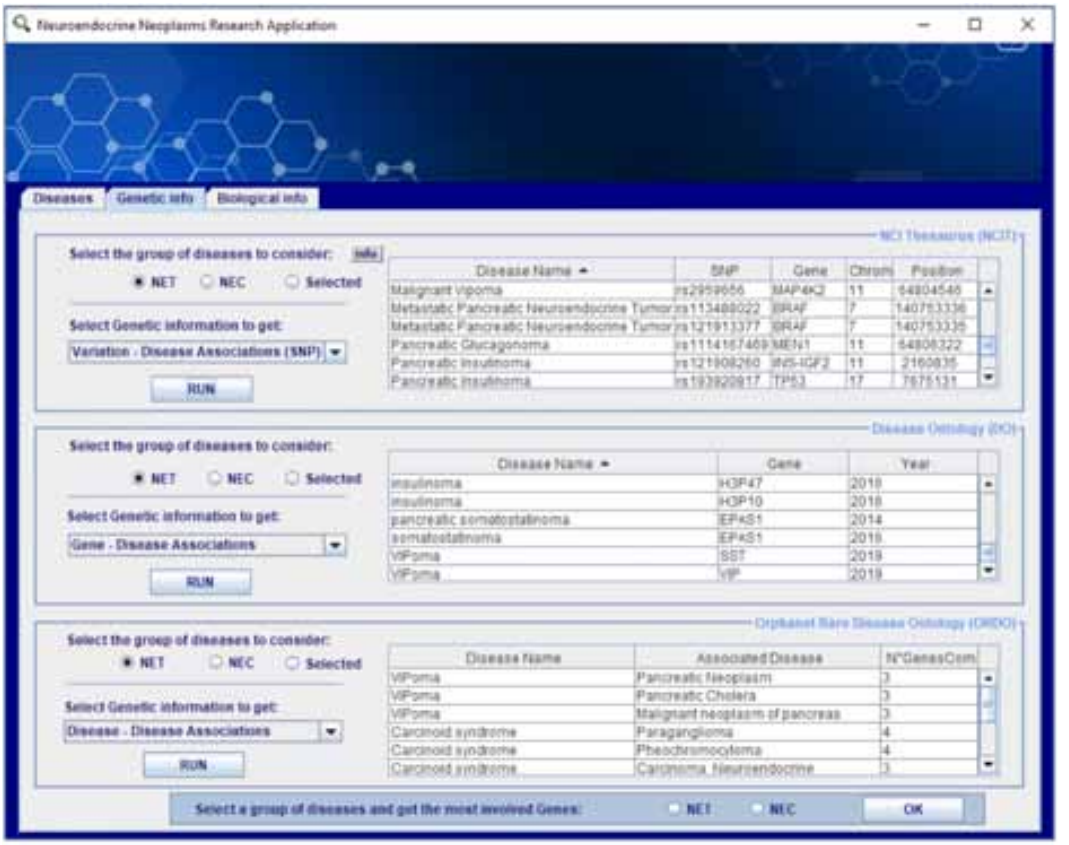
anomalies, etc. The obtained information is the result of the integration work described in section 4.3 between NCIT, ORDO and the DisGeNet database. The interface also allows to obtain the genes most involved in NETs and NECs or the genes whose association with diseases has been confirmed in at least 30 data sources or research studies.

Figure 11 shows the "biological information" section of the application. It allows to obtain information on molecular functions and biological processes associated with genes whose variation causes the diseases selected in the preceding phases. The information obtained is the result of the integration work between NCIT and GO described in section 4.4. The interface allows to obtain a detailed description of the activities of the selected gene or gene product and to move on the related activities following the existing has-part and regulates properties, thus navigating the GO graph.

5.2. Performance and Validation

A server-side test installation was set on an Ubuntu machine with a 2.3GHz quad-core Intel Core i7 processor and 16Gb of RAM. With this hardware configuration, most queries are answered by the server in a split second and just the most complex (mixing information from semantic and non-semantic sources) take longer, rarely more than 2 seconds. These results are in line with the recent benchmarks on RDF stores (Atemezing & Amardeilh, 2018) that rank Virtuoso Universal Server as one of the fastest triple stores both for instant queries (i.e., those used to generate dynamic views on the client) as well as for analytical ones (i.e., those used for validation and mapping purposes).

Figure 10. "Genetic info" section of the client application

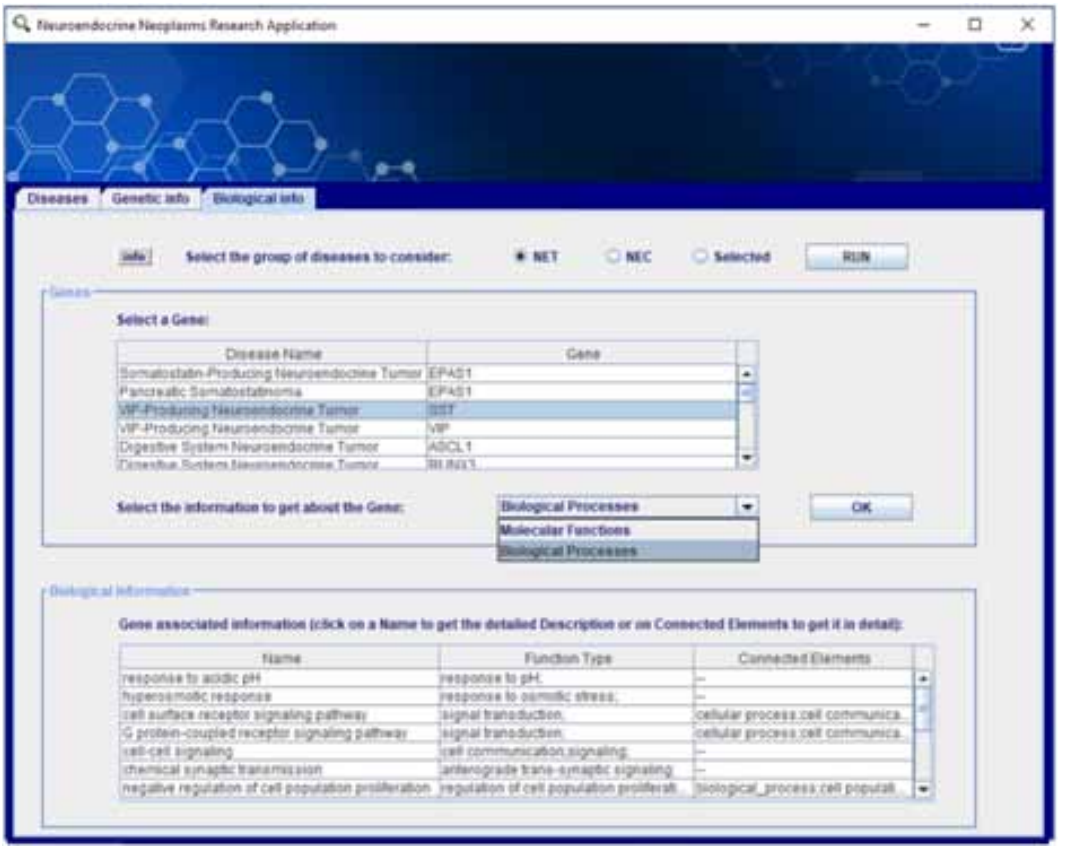


The system validation was performed qualitatively by involving a domain expert with the aim of verifying the consistency and correctness of the ontological knowledge as well as the quality of the alignment between the data sources. An iterative approach was adopted in which the expert was asked to use the system and provide feedback which was in turn used to improve the level of alignment (Dragisic, et al., 2016). In the specific case, two validation iterations were enough to obtain a satisfactory result.

To measure the quality of the integrated knowledge model, the metrics defined in (Tartir & Arpinar, 2007) were also considered. In this regard, it should be noted that, during the harmonization process, no new classes were added with respect to those already included in the source models. The only exception is the Activity class introduced in the third layer to group the related concepts of Molecular Function and Biological Process (already existing in NCIT and GO). This also applies to most relations, except those introduced in the second and third layers for the incorporation of non-ontological data sources (i.e., disease-disease, disease-gene, variant-gene, and disease-gene-variant from DisGeNet as well as does-activity, does-not-do-activity and contributes-to from GOF).

These properties ensure that the main quality metrics of the source schemes, such as Relationship Richness, Attribute Richness, and Inheritance Richness, are only marginally affected by the introduced semantic elements. The same is also true for Instance Metrics since no instances have been added or changed from the original ones. On the other hand, we found that the evaluation of such metrics on the whole integrated model is challenging task given its hybrid nature, including ontological and

Figure 11. "Biological info" section of the client application



non-ontological information, the latter integrated "on the fly" based on user requests. As also reported in section 6, exploring this issue could be a promising direction for future research.

6. CONCLUSION AND FURTHER WORK

In this paper, we have described a research work aimed at designing and implementing a domain-specific linked data application for the analysis, aggregation, and study of existing information on neuroendocrine neoplasms: a type of rare tumors. The application uses a knowledge base obtained by aligning and integrating existing semantic and non-semantic biomedical sources within a single multilayer network model.

Beyond the specific domain, the paper analyzes how to aggregate the results from the most recent studies with omics databases, genomic data repositories, data sources expressed in an ontological language, and traditional database schemes. The work is capable of being adapted to other domains, thus facilitating the rapid integration of heterogeneous data sets, reducing the time spent on data management and prioritizing its analysis.

The directions of extension of the proposed system are manifold. The design of additional information layers of the model is already underway, with the aim to integrate more domain aspects. In particular, an additional layer would be responsible for adding information about disease related phenotypes including morphology, development, biochemical and physiological properties, etc. Indeed, phenotypic data, combined with ever-increasing amounts of genomic data, have enormous

potential to accelerate the identification of clinically viable prognostic or therapeutic implications and to improve our understanding of rare diseases.

Additional layers may include human tissue information associated with disease-causing genes. Such information can help to find common features between different organs and further elucidate the function of genes associated with neuroendocrine neoplasms. An additional layer may also include information on drugs currently approved and in use for neuroendocrine neoplasms. Extrapolating the information on this point could be very complex because the treatment of this type of rare tumors still represents an important clinical problem. They are in fact biologically heterogeneous and contain subpopulations of cells with different angiogenic, invasive and meta-static properties. Therefore, their response to therapeutic agents is also heterogeneous.

In general, thanks to the approach based on linked data, there is no limit to the possible aggregation of omics information. Each information level would broaden the field of applicability of the system, making it increasingly complete and useful for supporting researchers and specialists in the biomedical sector. On the other hand, the multi-layer organization would help to deal with this vastness of information in an organized and governable way.

As anticipated in section 5, another promising research direction is the extension of existing ontology quality metrics to hybrid, workflow-based knowledge bases, as the one proposed. Moreover, to help anticipating the evolutions of the integrated schemas, the possibility of incorporating automatic ontology alignment approaches (Abayomi-Alli, et al., 2021) and methods for learning taxonomic and non-taxonomic relations (Hassan, Ali, Fathalla, Kholief, & Hassan, 2021) will be also explored.

ACKNOWLEDGMENT

This work is partially supported by the RarePlatNet project on diagnostic and therapeutic innovations for neuroendocrine and endocrine tumors and for glioblastoma through an integrated technological platform of clinical, genomic, ICT, pharmacological and pharmaceutical skills, funded by the Campania region, Italy under the grant on POR CAMPANIA FESR 2014/2020, axis 1, objective 1.2.

ADDITIONAL FUNDING INFORMATION

The publisher has waived the Open Access Article Publication fee.

REFERENCES

Abayomi-Alli, A., Arogundade, O., Misra, S., Akala, M., Ikotun, A., & Ojokoh, B. (2021). An Ontology-Based Information Extraction System for Organic Farming. *International Journal on Semantic Web and Information Systems*, *17*(2), 79–99. doi:10.4018/IJSWIS.2021040105

Aryan, P., & Ekaputra, F. (2020). *Sparqlgpviz: SPARQL Graph Pattern Visualization*. OSF Preprints.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A. e., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556 PMID:10802651

Atemezing, G., & Amardeilh, F. (2018). Benchmarking commercial RDF stores with publications office dataset. In *European Semantic Web Conference* (pp. 379-394). Cham: Springer. doi:10.1007/978-3-319-98192-5_54

Belleau, F., Nolin, M., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, *41*(5), 706–716. doi:10.1016/j.jbi.2008.03.004 PMID:18472304

Bensz, W., Borys, D., Fujarewicz, K., Herok, K., Jaksik, R., Krasucki, M., & Ochab, M. e. (2016). Integrated system supporting research on environment related cancers. In *Recent Developments in Intelligent Information and Database Systems* (pp. 339–409). Springer. doi:10.1007/978-3-319-31277-4_35

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., & Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, *544*(1), 1–122. doi:10.1016/j.physrep.2014.07.001 PMID:32834429

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), D267–D270. doi:10.1093/nar/gkh061 PMID:14681409

Capuano, N., Gaeta, A., Guarino, G., Miranda, S., & Tomasiello, S. (2016). Enhancing augmented reality with cognitive and knowledge perspectives: A case study in museum exhibitions. *Behaviour & Information Technology*, *35*(11), 968–979. doi:10.1080/0144929X.2016.1208774

Capuano, N., Longhi, A., Salerno, S., & Toti, D. (2015). Ontology-driven generation of training paths in the legal domain. *International Journal of Emerging Technologies in Learning*, *10*(7), 14–22. doi:10.3991/ijet.v10i7.4609

Childs, L., Mamlouk, S., Brandt, J., Sers, C., & Leser, U. (2016). SoFIA: A data integration framework for annotating high-throughput datasets. *Bioinformatics (Oxford, England)*, *32*(17), 2590–2597. doi:10.1093/bioinformatics/btw302 PMID:27187206

Dumontier, M., Callahan, A., Cruz-Toledo, J., Ansell, P., Emonet, V., Belleau, F., & Droit, A. (2014). Bio2RDF release 3: a larger connected network of linked data for the life sciences. In *12th Semantic Web Conference* (pp. 401–404). Aachen: ACM.

Fathalla, S. (2018). Detecting Human Diseases Relatedness: A Spreading Activation Approach Over Ontologies. *International Journal on Semantic Web and Information Systems*, *14*(3), 120–133. doi:10.4018/IJSWIS.2018070106

Grigoris Effraimidis, U. K.-R. (2021). Multiple endocrine neoplasia type 1 (MEN-1) and neuroendocrine neoplasms (NENs). Seminars in Cancer Biology.

Guérin, E., Marquet, G., Burgun, A., Loréal, O., Berti-Equille, L., Leser, U., & Moussouni, F. (2005). Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW. In *2nd International Workshop on Data Integration in the Life Sciences* (pp. 158-174). Springer. doi:10.1007/11530084_14

Hammoud, Z., & Kramer, F. (2020). Multilayer networks: aspects, implementations, and application in biomedicine. Big Data Analytics, 5(5).

Hasan, M., Kousiouris, G., Anagnostopoulos, D., Stamati, T., Loucopoulos, P., & Nikolaidou, M. (2021). CISMET: A Semantic Ontology Framework for Regulatory-Requirements-Compliant Information Systems Development and Its Application in the GDPR Case. *International Journal on Semantic Web and Information Systems*, *17*(1), 1–24. doi:10.4018/IJSWIS.2021010101

Hassan, M., Ali, M., Fathalla, S., Kholief, M., & Hassan, Y. (2021). Learning Non-Taxonomic Relations of Ontologies: A Systematic Review. *International Journal on Semantic Web and Information Systems*, *17*(1), 97–122. doi:10.4018/IJSWIS.2021010105

Kamdar, M. (2018). Mining the Web of Life Sciences Linked Open Data for Mechanism-Based Pharmacovigilance. In WWW'18: Companion Proceedings of the The Web Conference (pp. 861-865). ACM. doi:10.1145/3184558.3186576

Kamdar, M., Fernández, J., Polleres, A., Tudorache, T., & Musen, M. (2019). Enabling Web-scale data integration in biomedicine through Linked Open Data. *Digital Medicine*, *2*(90). PMID:31531395

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J., Moreno, Y., & Porter, M. (2014). Multilayer networks. *Journal of Complex Networks*, *2*(3), 203–271. doi:10.1093/comnet/cnu016

Kumar, A., & Smith, B. (2005). Oncology Ontology in the NCI Thesaurus. In *Artificial Intelligence in Medicine* (pp. 213–220). Springer. doi:10.1007/11527770_30

Livingston, K., Bada, M., Baumgartner, W. Jr, & Hunter, L. (2015). KaBOB: Ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*, *16*(126), 126. doi:10.1186/s12859-015-0559-3 PMID:25903923

Louden, D. (2020). Medgen: Ncbi's portal to information on medical conditions with a genetic component. *Medical Reference Services Quarterly*, *39*(2), 183–191. doi:10.1080/02763869.2020.1726152 PMID:32329672

Magee, L. (2011). Describing knowledge domains: a case study of biological ontologies. In *Towards a Semantic Web* (pp. 289–301). Chandos Publishing. doi:10.1016/B978-1-84334-601-2.50010-6

Masseroli, M., Canakoglu, A., & Ceri, S. (2016). Integration and querying of genomic and proteomic semantic annotations for biomedical knowledge extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(2), 209–219. doi:10.1109/TCBB.2015.2453944 PMID:27045824

Merabti, T., Joubert, M., Lecroq, T., Rath, A., & Darmoni, S. (2010). Mapping biomedical terminologies using natural language processing tools and UMLS: Mapping the Orphanet thesaurus to the MeSH. *IRBM*, *31*(4), 221–225. doi:10.1016/j.irbm.2010.04.003

Messaoudi, C., Fissoune, R., & Hassan, B. (2016). A Survey of Semantic Integration Approaches in Bioinformatics. *International Journal of Computer and Information Engineering*, *10*(12), 2058–2063.

Nagtegaal, I., Odze, R., Klimstra, D., Paradis, V., Rugge, R., Schirmacher, P., Washington, K. M., Carneiro, F., & Cree, I. (2019). The 2019 WHO classification of tumours of the digestive system. *Histopathology*, *76*(2), 182–188. doi:10.1111/his.13975 PMID:31433515

Piñero, J., Ramírez-Anguita, J., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, *48*(D1), D845–D855. PMID:31680165

Rindi, G., & Inzani, F. (2020). Neuroendocrine neoplasm update: Toward universal nomenclature. *Endocrine-Related Cancer*, *27*(6), 211–218. doi:10.1530/ERC-20-0036 PMID:32276263

Rindi, G., Klimstra, D., Abedi-Ardekani, B., Asa, S., Bosman, F., Brambilla, E., Busam, K. J., de Krijger, R. R., Dietel, M., El-Naggar, A. K., Fernandez-Cuesta, L., Klöppel, G., McCluggage, W. G., Moch, H., Ohgaki, H., Rakha, E. A., Reed, N. S., Rous, B. A., Sasano, H., & Cree, I. A. et al. (2018). A common classification framework for neuroendocrine neoplasms: An International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal. *Modern Pathology*, *31*(12), 1770–1786. doi:10.1038/s41379-018-0110-y PMID:30140036

Schriml, L., Arze, C., Nadendla, S., Chang, Y., Mazaitis, M., Felix, V., Feng, G., & Kibbe, W. (2012). Disease Ontology: A backbone for disease semantic integration. *Nucleic Acids Research*, *40*(D1), D940–D946. doi:10.1093/nar/gkr972 PMID:22080554

Tartir, S., & Arpinar, I. (2007). Ontology evaluation and ranking using OntoQA. In *Proceedings of the International Conference on Semantic Computing (ICSC 2007)* (pp. 185-192). Irvine, CA: IEEE. doi:10.1109/ICSC.2007.19

Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., . . . Rath, A. (2014). ORDO: An Ontology Connecting Rare Disease, Epidemiology and Genetic Data. 22nd Annual International Conference on Intelligent Systems for Molecular Biology, Boston, MA.

ENDNOTES

1 lod-cloud.net
2 bioportal.bioontology.org
3 www.obofoundry.org
4 www.bioinformatics.deib.polimi.it/GPKB-advance

Nicola Capuano is an assistant professor of computer science at the School of Engineering of the University of Basilicata, Italy. His research interests include Computational Intelligence, Artificial Intelligence in Education, Knowledge-Based Systems and Cognitive Robotics. He is the author of more than 100 publications in scientific journals, conference proceedings and books on these topics. He is Associate Editor for the Springer's "Journal of Ambient Intelligence and Humanized Computing" and for "Frontiers in Artificial Intelligence". He is scientific referee and member of editorial boards for international journals and conferences. He is member of the SmartLearn research group at the Open University of Catalonia. He worked as Project Manager and Scientific Consultant for research organizations and private companies. He was principal investigator of RTD projects co-funded by the European Commission. He played scientific and management roles in several other projects. He is a certified Project Management Professional (PMP). Master Degree with Honors in Computer Engineering at the "Federico II" university of Naples (Italy), PhD in Computer Engineering and Electronics at the same university.

Pasquale Foggia has been researcher at the "Federico II" university of Naples from 1999 to 2004. From 2004 to 2008 associate professor at the same university. From 2008 to 2019 he was associate professor at the University of Salerno, in the scientific sector "Systems for Information Processing" (ING-INF/05), while since February 2019 he is a full professor at the same university. His scientific interests have focused on the areas of Pattern Recognition and Computer Vision. More specifically, his research activity has been on methodologic aspects such as classification and learning algorithms based on graphs, soft computing techniques, and applications in fields such as image and video analysis, intelligent video surveillance, biomedical image analysis and diagnosis support systems, robot vision.

Luca Greco received his Master Degree cum laude in Electronic Engineering in 2008 and his PhD in Information Engineering in 2013, both at the University of Salerno. Since April 2015 he is Researcher at the Department of Information and Electrical Engineering and Applied Mathematics (DIEM) and belongs to the research group in Artificial Vision (MIVIA Lab), characterized by national and international collaborations. In July 2018 he obtained the National Scientific Qualification (ASN) for the functions of second-tier university professor in sector 09 / H1 (Information Processing Systems). His scientific interests concern the field of Information Retrieval, Sentiment Analysis and semantic technologies. He's also involved in research concerning Machine Learning techniques applied to Fog / Edge computing systems. He is author of over 50 scientific papers, including journal articles and international conference proceedings.

Pierluigi Ritrovato is Full Professor of Processing Systems at the Department of Information and Electrical Engineering and Applied Mathematics of the University of Salerno. Since 2015 he has been a member of the research group on Intelligent Machines for the Video, Images and Audio analysis (MIVIA), where he initially dealt with aspects related to the application of semantic technologies for intelligent video analysis. He is currently working on the definition of IoMT (Internet of Medical Things) architectures for the collection of data from wereable sensors and the use of Artificial Intelligence and semantic web technologies for the analysis of physiological and clinical data for the prevention and monitoring of chronic diseases and tumors. He is CEO of the spin-off AI4Health srl, founded at the end of 2017. He has carried out research activities in the areas of Software Engineering with particular reference to development processes and service oriented architectures (SOA), large-scale distributed systems (grids and cloud computing), Adaptive Learning and Knowledge Management systems using semantic technologies. He has been technical and scientific coordinator of several applied research and technological development projects at national and European level in the fields of technology Enhanced Learning, knowledge management, software engineering and Grid Computing.