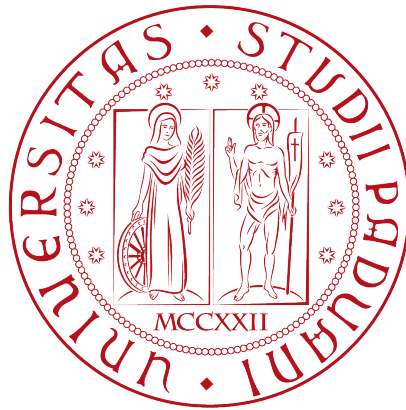


Bridging Information Access and Visual Analytics Methods for Supporting the Decision Process in the Digital Pathology Domain



Fabio Giachelle

Supervisor: Prof. Gianmaria Silvello

Department of Information Engineering
University of Padua

This dissertation is submitted for the degree of

Doctor of Philosophy

School on Information Engineering

Curriculum: Information and Communication Science and Technologies

Cycle: XXXV

September 2022

To my family

Acknowledgements

First, I would like to thank my supervisor, Prof. Gianmaria Silvello, and express my deepest appreciation for the patience and foresight with which he guided me throughout this short but intense journey. I am grateful for the insightful conversations that gave me the opportunity to grow not only from the scientific side but also, and most importantly, from the human one.

Then, I would like to acknowledge all the senior members of the Information Management Systems (IMS) Research Group, Prof. Maristella Agosti, Prof. Nicola Ferro, and Prof. Giorgio Maria Di Nunzio. It was a great pleasure to have the opportunity of working with them; it was a profound learning moment.

I am also grateful to all of my past and present colleagues at the IMS group. I would like to thank all of them for the stimulating discussions, the innovative ideas, and for all the difficult and joyful moments we shared in the last three years. In particular, I would like to thank Ornella and Stefano for their precious contributions to the design and development of MedTAG and SKET respectively.

A special thanks to all of my friends, for giving me their time, trust, and enthusiasm with which to move forward.

I will always be grateful to my parents Lucia and Franco, and my sister, Laura. I am deeply indebted with them for always motivating, assisting, and inspiring me. Hence, this thesis is dedicated to my family. Finally, I would like to thank Elena, for the constant support and love that accompanied me all along this wonderful journey.

Of course, these few lines are not enough to thank everyone for the friendship, affection, and help I received. Hence, I hope to share with you soon new amazing moments like those of the last three years.

Abstract

In recent years, deep-learning approaches for digital pathology have proven to be effective in image analysis tasks such as classification. Despite the promising results, the adoption of such approaches in clinical practice is still limited due to two major issues. First, there is a lack of annotated datasets required to train and evaluate deep learning algorithms. Annotating large datasets is expensive and hard to achieve especially for the scarcity of expert pathologists willing to do such a time-consuming task. Secondly, the outcomes of deep-learning approaches are difficult to comprehend and assess, due to the black-box nature of the models involved. Nevertheless, in the digital pathology domain, pathologists should be able to understand why a specific outcome has been determined in order to trust model predictions. Moreover, explainable artificial intelligence is not only desirable but is also a mandatory requirement according to recent regulations such as the European General Data Protection Regulation (GDPR). According to recent studies, pathologists prefer visual explanations for algorithms' outcomes, clearly indicating the scientific claims supporting each prediction. Information visualization and visual analytics methods could be used to allow pathologists to visually comprehend machine predictions, by means of intuitive explanation interfaces. Despite other domains, such as radiology, have already benefited from using these techniques, their employment in the digital pathology domain is still limited.

In this thesis, we tackle the former issues by synergically combining different computational, analytics, and visual approaches to support diagnostics and the decision-making process in digital pathology. Firstly, we propose the Semantic Knowledge Extractor Tool (SKET) for the knowledge extraction process from free-text pathology reports. SKET automatically generates weak annotations that are used to train a deep-learning-based image classification system for digital pathology. Secondly, we propose SKET eXplained (SKET X), an explainability tool that exploits visual analytics techniques to visually explain the outputs of SKET. Then, we introduce MedTAG, a customizable annotation tool for clinical reports, with the purpose of facilitating the creation of consistent and permanent ground truth labels and speeding up the burdensome annotation task. To this aim, MedTAG integrates SKET for automatic annotation facilities.

Moreover, we propose NanoWeb for the exploration of the knowledge generated by the interconnected network of scientific facts extracted from the literature and encoded as machine-readable statements within the Linked Data paradigm. Finally, we integrate our contributions into the ExaSURE System for Unified Resource Exploration (ExaSURE) ecosystem for unified access in a web-based fashion. The ExaSURE ecosystem represents a step forward for integrating algorithmic and visual digital tools in clinical practice to support pathologists' work in their daily routines.

Table of contents

| | |
|--|--------------|
| List of figures | xiii |
| List of tables | xxiii |
| 1 Introduction | 1 |
| 1.1 Objectives and contributions | 4 |
| 1.2 Outline | 5 |
| 1.3 Publications | 6 |
| 2 Background | 9 |
| 2.1 Digital pathology and computational pathology | 10 |
| 2.2 Visual analytics and information visualization | 13 |
| 2.3 Visual analytics and information visualization for computational pathology | 19 |
| 2.4 Explainability | 21 |
| 2.4.1 Explainability for computational pathology | 24 |
| 3 ExaSURE | 27 |
| 3.1 Introduction | 27 |
| 3.1.1 ExaSURE ecosystem | 28 |
| 3.2 ExaSURE interface | 31 |
| 3.2.1 Implementation details | 31 |
| 3.3 Conclusions | 33 |
| 4 Knowledge extraction | 35 |
| 4.1 Introduction | 35 |
| 4.2 Material | 38 |
| 4.3 Methods | 39 |
| 4.3.1 Named Entity Recognition | 39 |
| 4.3.2 Entity Linking | 41 |

| | | |
|----------|---|------------|
| 4.3.3 | Data Labeling | 42 |
| 4.3.4 | Graph Creation | 43 |
| 4.4 | Evaluation | 44 |
| 4.4.1 | Tasks | 44 |
| 4.4.2 | Datasets | 44 |
| 4.4.3 | Baselines | 45 |
| 4.4.4 | Results | 46 |
| 4.5 | SKET and the ExaSURE ecosystem | 49 |
| 4.5.1 | SKETUp a web interface for SKET | 50 |
| 4.5.2 | Automatic Report Annotation | 51 |
| 4.5.3 | Pathological Knowledge Visualization | 53 |
| 4.5.4 | WSI Classification | 57 |
| 4.6 | Conclusions and Future Work | 58 |
| 5 | Explainability | 59 |
| 5.1 | Introduction | 59 |
| 5.2 | SKET X Architecture | 60 |
| 5.3 | SKET X Workflow | 63 |
| 5.4 | SKET X Interface | 68 |
| 5.5 | Conclusions and Future Work | 74 |
| 6 | Semantic annotation | 75 |
| 6.1 | Background | 75 |
| 6.2 | Implementation | 79 |
| 6.2.1 | Architecture | 80 |
| 6.2.2 | Installation and customization | 81 |
| 6.2.3 | User interface and interaction | 83 |
| 6.2.4 | MedTAG control panel for statistics and Inter-Annotator Agreement (IAA) | 86 |
| 6.3 | Results and Discussion | 94 |
| 6.3.1 | Biomedical annotation tools comparison | 94 |
| 6.3.2 | Quantitative comparison of biomedical annotation tools | 98 |
| 6.4 | Conclusions | 101 |
| 6.4.1 | Limitations and Future Work | 102 |
| 7 | Knowledge exploration | 103 |
| 7.1 | Introduction | 103 |

| | | |
|----------|---|------------|
| 7.2 | Background | 107 |
| 7.3 | The NanoWeb Architecture | 112 |
| 7.3.1 | Search system | 114 |
| 7.4 | Nanopublication collection statistics | 116 |
| 7.4.1 | Association analysis | 117 |
| 7.4.2 | Scientific Evidences | 117 |
| 7.5 | NanoWeb Graphical User Interface | 121 |
| 7.6 | Expert users survey | 132 |
| 7.6.1 | User feedback | 136 |
| 7.7 | Discussion on maintaining aspects | 137 |
| 7.8 | Conclusions | 138 |
| 8 | Conclusions and Future Work | 141 |
| 8.1 | Conclusions | 141 |
| 8.2 | Directions for Future Work | 143 |
| | References | 151 |

List of figures

| | | |
|-----|---|----|
| 1.1 | SKET as the core of the knowledge extraction process and the downstream applications. SKET extracts labels, mentions, and concepts from free-text clinical reports. Indeed, from the example reported we can notice one label (<i>Adenomatous polyp - low grade dysplasia</i>), two mentions (<i>low-grade dysplasia (mild to moderate)</i> and <i>biopsy sigmoid</i>), and three concepts (<i>Moderate Colon Dysplasia</i> , <i>Mild Colon Dysplasia</i> , and <i>Biopsy of Colon</i>). Users can visualize and edit the annotation data (knowledge extracted) using MedTAG (see Chapter 6). In addition, they can explain the outputs of SKET using SKET X and explore nanopublications of interest (e.g., nanopublications related to colon cancer) using NanoWeb (see Chapters 5 and 7 respectively). | 3 |
| 2.1 | The pyramidal structure of WSI, resulting from different levels of magnification. | 13 |
| 2.2 | VA combines DA, IV, and human analytical reasoning enabling users to visually analyze and comprehend data. VA exploits the human perceptual/cognitive system and analytical reasoning to reveal meaningful patterns in data and foster unexpected insights. | 15 |
| 2.3 | The VA process proposed by Daniel A. Keim [153]. The process takes in input a dataset from S and returns the insight I as the output. I can be determined directly from the set of possible visualizations V or through the confirmation of some hypotheses H by employing automatic data analysis approaches. | 16 |
| 2.4 | The sense-making loop [151] based on the visualization model proposed by van Wijk [292]. | 17 |
| 2.5 | A taxonomy of explainability techniques based on the work proposed by Arrieta et al. [26]. | 24 |

-
- 3.1 ExaSURE ecosystem encompassing several biomedical tools for the digital pathology domain, namely, *ExaNet*, *MedTAG*, *NanoWeb*, *SKETUp* and *SKET X*. The figure describes the typical workflow in the ExaSURE ecosystem which consists of: (i) data ingestion carried out by *SKETUp* which processes the clinical reports and stores the mentions, concepts, and labels extracted in the common database; (ii) reports annotation carried out by human experts using *MedTAG*; (iii) data enrichment with provenance using *NanoWeb*; (iv) knowledge exploration using *ExaNet* and (v) explainability using *SKET X*. 29
- 3.2 ExaSURE homepage providing access to all the tools and services requiring authentication, that is, *ExaNet*, *MedTAG*, *SKETUp* and *SKET X*. For each service, a dedicated button provides direct access to the related resource without requiring the users to authenticate again. In particular, the buttons depicted allow the users to: (A) enter *MedTAG* to consult and annotate clinical reports; (B) access the *MedTAG* dashboard to monitor the annotations for each report and visualize its graph representation using the integrated features of *ExaNet*; (C) enter *SKETUp* to run SKET on a set of user-provided reports; (D) access the *SKETUp* dashboard for monitoring multiple SKET executions and download the resulting outputs; (E) enter *SKET X* explainability interface to comprehend SKET outputs and the rules used in the underlying machine decision process; (F) access the *SKET X* dashboard reporting all the different executions of SKET and the parameters involved. 32
- 3.3 *MedTAG* dashboard for managing reports and the annotation process. The dashboard provides several facilities: (A) visualize report information; (B) search for reports; (C) visualize reports' annotations; (D) visualize the majority vote ground-truth; (E) download reports' annotations; (F) visualize the report's graph representation and (G) annotate a report quickly. In addition, users can go back to the ExaSURE interface by clicking on the button (H). . 33
- 4.1 A qualitative example of the knowledge extraction process carried out by SKET. From a free-text report, SKET extracts key information, namely, *labels*, *mentions*, and *concepts*. For instance, we can observe that the label *Adenomatous polyp - low grade dysplasia* has been generated since the related concepts *Moderate Colon Dysplasia* and *Mild Colon Dysplasia* have been identified in the given report starting from the mention *low-grade dysplasia (mild to moderate)*. Another example is the concept *Biopsy of Colon* which has been identified from the associated mention *biopsy sigmoid*. 36

| | | |
|------|---|----|
| 4.2 | SKET architecture. SKET main components are: (A) Named Entity Recognition, (B) Entity Linking, (C) Data Labeling, and (D) Graph Creation. . . . | 40 |
| 4.3 | SKET is the core of the ExaSURE ecosystem. From clinical reports, a suite of different applications relying on SKET process, analyze, explore, and explain the knowledge contained within reports – also providing weak supervision to train cancer assisted diagnosis tools. | 50 |
| 4.4 | SKETUp upload interface for a single report (left side) and the serialization in JSON format of the SKET output for the concepts identified in the knowledge extraction process (right side). We can observe that the concepts identified in the diagnosis field are <i>Colon Adenocarcinoma</i> , <i>Mild Colon Dysplasia</i> , and <i>Severe Colon Dysplasia</i> . For what concerns instead the procedure field, the <i>Biopsy of Colon</i> concept has been identified. | 52 |
| 4.5 | SKETUp configuration interface for the SKET parameters. We can observe that the threshold used for the pruning phase of the knowledge extraction process is set to 1.8 (the default value). Moreover, the models selected for the EL phase are the Gestalt Pattern Matching (GPM) and the neural model, as described in section 4.3.2. | 53 |
| 4.6 | SKETUp dashboard reporting the information concerning each execution (i.e., task) of SKET. The interface is continuously updated so that users can monitor SKET executions and check their status. When a task ends up correctly (i.e., <i>SUCCESS</i> status), users can download the outputs of SKET and annotate a batch of reports with MedTAG (as long as the batch is saved in the database). | 54 |
| 4.7 | ExaNet interface for report search and visualization. Users can search for reports using column filters. Moreover, users can use the action buttons for: (A) visualizing the graph representation of a report; (B) downloading a report in JSON format and (C) visualizing the report in JSON format without having to download it. | 55 |
| 4.8 | ExaNet visualization of the report-level knowledge graph produced by SKET for a pathology report about colon hyperplastic polyp. | 55 |
| 4.9 | ExaNet visualization of a clinical report, in JSON format, in an interactive interface where users can expand/collapse the keys in the JSON report at will. | 56 |
| 4.10 | Reports' statistics table of MedTAG integrating ExaNet functionalities. Users can click on the <i>Graph</i> button to visualize the graph representation of the medical report selected. | 56 |

-
- 5.1 SKET X acts as an explanation interface to visually comprehend why SKET has produced a certain output and realize whether it is correct or not based on the models/rules employed in the machine decision process. Experts can provide feedback/suggestions to improve the system and, in turn, the effectiveness of the SKET knowledge extraction process. 61
- 5.2 SKET X architecture and technologies adopted. The figure is divided in three sections: (i) *Presentation layer* concerning the front-end and the web interface developed using `React.js`, `HTML5`, and `CSS3`; (ii) *Business layer* where is reported the back-end logic implemented with `Python`, `Django`, and `Celery`; (iii) *Data layer* concerning the data to save either temporary (i.e., cache data saved using `Redis`) or persistently (using `PostgreSQL`). The interface communicates with the business logic via REST API requests that are satisfied asynchronously in the order determined by the queue manager. Then, the outputs of each request are saved in the database. 62
- 5.3 SKET X upload form enabling users to provide the diagnostic reports to process (A) and other information including: (B) the language of the reports (i.e., Dutch, English, and Italian); (C) the report use case (i.e., cervix, colon, and lung cancer); and (D) a description of the current pipeline execution. Users can take advantage of drag and drop facilities to specify the reports to process either in CSV or JSON format (A). 63
- 5.4 SKET X dashboard providing information about the executed SKET pipelines - i.e., pipeline id, use case, pipeline status, start timestamp, end timestamp, description, pipeline parameters. Users can view the parameters of each pipeline by clicking on the dedicated button (A). Similarly, users can access pipeline data by clicking on the dedicated button (B). When the execution of a pipeline ends, its outputs become available for download (C). 64
- 5.5 (A) SKET X *Overview* tab for the translation phase, (B) the reports in the original language (input), (C) the translated reports (output) (C), and (D) the parameters and settings for the current phase. 65

- 5.6 SKET X *Analytics* tab for the EL phase: (A) reports section, the users can change the current report using the left/right buttons; (B) SKET rules for the NER task; and, (C) list of mentions and concepts produced by the knowledge extraction process. Each concept and related mentions are highlighted with the same color in (A) and (C). By clicking/hovering on a specific concept, it is possible to highlight the relevant rules in the Sankey diagram that determined the concept and the related mentions in the report text. On the left side of the Sankey diagram are reported the rules *triggers*, which are boolean expressions tested on each mentioned text. If one or more mentions satisfy a rule trigger, then the related concepts on the right side of the Sankey diagram are highlighted and listed in (C). 66
- 5.7 SKET X *Analytics* tab for the classification phase: (A) reports section to select the current report via left/right buttons; (B) SKET rules for determining the labels visualized with a Sankey diagram; and, (C) list of labels, mentions, and concepts determined by SKET. Each concept and the related mentions are highlighted with the same color in (A) and (C). The Sankey diagram highlights the relevant rules by clicking/hovering on a specific label. On the left side of the Sankey diagram are reported the rules *triggers*. If one or more concepts satisfy a rule trigger, then the related label is highlighted on the right side of the Sankey diagram and also listed in (C). The mentions and concepts involved in the classification task are the *key* mentions/concepts (C), while the *excluded* ones are reported in (D). 67
- 5.8 SKET X *Output* tab showing the SKET outputs for the classification phase (i.e., labels). These are arranged in tabular form so that users can take advantage of column filters to search and visualize specific report information. 68
- 5.9 SKET X *Params* tab showing the parameters for the EL phase. The figure highlights that the current pipeline uses both the GPM and the neural model together with the default SKET threshold of 1.8. Users can change the value of these parameters and then click on the *Run SKET again* button to re-run SKET on the same set of reports but with the new parameter values. Finally, the results of the new SKET run are saved in a new pipeline and the user is asked to provide a description for it. 69

- 5.10 SKET X *Compare* tab for the EL phase showing the comparison interface for the two pipelines specified for the comparison. The interface is organized in four parts: (A) the reports section displaying information about the current report and two buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier and its description, and parameters (e.g., the models used for EL phase and the threshold); (C) first pipeline outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline outputs for the phase selected. The mentions in common, and the related concepts, are highlighted both in the report text (A) and also in the mention/concept lists for each pipeline (C) and (D). Hence, we can observe that there is a mention *injury-free resection margin* and a concept *Resection* that are not highlighted since they have been identified only by the second pipeline (D). Nevertheless, the concepts *Rectal mucous membrane* and *Adenoma* have been identified only by respectively the first pipeline (C) and the second one (D), but since both are associated with the same common mention – i.e., *adenomatous* – they are highlighted as well. 72
- 5.11 SKET X *Compare* tab for the classification phase showing the comparison interface for the two pipelines specified for the comparison. The interface is organized in four parts: (A) the reports section displaying information about the current report and two buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier, the description, and its parameters; (C) first pipeline outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline outputs for the phase selected. The mentions/concepts considered for determining the report labels are regarded as *key* mentions/concepts and are differentiated by the *excluded* ones. Here, two concepts are identified in the first pipeline, namely, *Colon Hyperplastic Polyp* and *Severe Colon Dysplasia*, while in the second one only *Severe Colon Dysplasia* has been identified. Nevertheless, *Colon Hyperplastic Polyp* and *Sigmoid colon* are negligible concepts (i.e., false positives) both associated with the *polyp sigmoid* mention. In contrast, *Severe Colon Dysplasia* is correct since it has been identified using a SKET rule verified by the *severe dysplasia key* mention. 73

- 6.1 Overview of annotation tools and their functionalities. The annotation tools considered come from a recent extensive review of tools for manual annotation of documents [219]. In addition, we consider also TeamTat [141] and INCEpTION [162] and report our judgements. The annotation tools are assessed with 22 criteria, defined in the latter review study, among three categories: *Data* (D), *Functional* (F) and *Technical* (T). The fulfillment of each criterion is indicated with a color in a three levels scale: white (feature absent or not met), light blue (feature partially satisfied), blue (feature satisfied). 76
- 6.2 MedTAG Architecture. The data layer comprises two relational databases, namely, *MedTAG data* and *Log data* to store all the information concerning the annotation process (e.g., concepts, labels, reports, users and their annotations) and logging data such as notifications of malformed clinical reports. The business layer comprises two business units: *Business logic* and *REST API* which jointly control the whole information flow from the front-end to the database and vice-versa. The presentation layer provides the MedTAG front-end, a web interface allowing users to annotate medical reports and download their ground truths. 82
- 6.3 MedTAG sidebar provides the *Configure* option, indicated by the orange arrow, to set up a new custom configuration. 83
- 6.4 MedTAG new configuration interface allows the user to save the current data before creating a new configuration. To guide the user in providing the new configuration files needed (i.e. reports/documents, labels and concepts), MedTAG provides both example and template files. In particular, users can use the example files to test MedTAG without providing their own data. Instead, users can use the template files as a reference to structure their own configuration files. 84
- 6.5 MedTAG main interface for data configuration. Users can provide their own CSV files for the reports/documents to annotate and the concepts and labels to use for the annotation process. Moreover, MedTAG detects automatically the document fields and allows users to specify which of them to annotate and/or display in the interface, as shown in the orange box (1). 88

-
- 6.6 MedTAG main interface in test mode with default configuration: clinical case set to “Colon cancer”, reports’ language set to English, reports’ institute/hospital set to “default_hospital” (the real name has been anonymized) and the annotation mode set to manual. The annotation type active is the *Labels* one. Three labels have been checked: (i) *Cancer*; (ii) *Adenomatous polyp - low grade dysplasia* and (iii) *Hyperplastic polyp*. 89
- 6.7 MedTAG main interface in test mode with default configuration: clinical case set to “Colon cancer”, reports’ language set to English, reports’ institute/hospital set to “default_hospital” (the real name has been anonymized) and the annotation mode set to manual. The annotation type active is the *Linking* one. Three mentions have been identified and linked to the corresponding concepts: (i) *hyperplastic adenomatous polyp* is linked to *Colon Hyperplastic Polyp*; (ii) *mild dysplasia* is linked to *Mild Colon Dysplasia*; and (iii) *tubular adenoma* is linked to *Colon Tubular Adenoma*. 90
- 6.8 MedTAG tutorial interface. To reach the tutorial section, users can click on the *Tutorial* link in the sidebar, indicated by the orange arrow. 90
- 6.9 MedTAG control panel concerning the reports’ statistics. The reports are organized in an interactive table enabling the admin user to: (i) access report data; (ii) delete one or more reports; (iii) download report data including manual and automatic annotations and (iv) access the information concerning IAA and manage the majority vote procedure. 91
- 6.10 MedTAG control panel concerning the team members’ statistics. The ring charts report the annotation work carried out by each team member, so that the admin can keep track of the advancements regarding the whole annotation process. 92
- 6.11 MedTAG *My Statistics* panel, providing information about the user annotation work in terms of documents annotated for each use-case. 92
- 6.12 MedTAG majority vote interface. The admin can overview the selected report and choose the options of interest for the majority vote procedure, including: (i) the annotation mode; (ii) the annotation type and (iii) the team members (annotators) to consider. 93
- 6.13 MedTAG majority vote output for the *Labels* annotation type. The admin can visualize the annotations resulting from the majority vote procedure, together with the corresponding authors. In addition, the admin can download the annotations or change the current majority vote configuration. 93

| | | |
|------|--|-----|
| 7.1 | (A) RDF (trig) representation of the nanopublication encoding the assertion: <i><activin A receptor type 2A - gene-disease association - Colorectal Cancer></i> ; (B) graphical representation of the four parts of the nanopublications with a human-readable representation of the assertion graph; (C) network of gene-disease associations created by five nanopublications. | 104 |
| 7.2 | NanoWeb system architecture. | 112 |
| 7.3 | DisGeNET ontology: number of assertions (yellow) for each DisGeNET association type. | 118 |
| 7.4 | NanoWeb search interface with user-provided query: <i>colorectal cancer</i> . . . | 120 |
| 7.5 | Information layer for the nanopublication. | 122 |
| 7.6 | Data record for the nanopublication with title: <i>mutL homolog 1 - Colorectal Carcinoma</i> | 123 |
| 7.7 | Graph layer for the nanopublication clicked by the user. | 126 |
| 7.8 | Graph exploration: the information window for <i>mutS homolog 6 - Carcinogenesis</i> is displayed as a result for the user click on the edge. | 127 |
| 7.9 | Graph exploration: search for <i>mutL homolog 1 (MLH1)</i> connected entities. . | 127 |
| 7.10 | Advanced search: search for nanopublications regarding the <i>mutL homolog 1</i> gene. | 129 |

List of tables

| | | |
|-----|--|----|
| 4.1 | Data size. For each medical center, we report the number of diagnostic reports associated with each use-case. The “–” symbol represents the lack of reports for a given use-case. | 38 |
| 4.2 | Number of annotated diagnostic reports for each use-case. Label counts are independent of each other except for “Non-informative” in Colon, “Normal squamous” and “Normal glands” in Cervix, and “No cancer” in Lung, which only occur when none of the others does. | 45 |
| 4.3 | Entity linking results on colon, cervix, and lung cancer pathology reports. The considered measures are subset accuracy, micro F1, and weighted F1. Bold values represent the highest scores achieved for each measure. | 47 |
| 4.4 | Text classification results on colon, cervix, and lung cancer pathology reports. The considered measures are subset accuracy, micro F1, and weighted F1. The [†] symbol represents the statistical difference of SKET from unsupervised FastText- and BERT-based approaches – verified using a paired t-test with a p-value < 0.01. Bold values represent the highest scores achieved for each measure. | 49 |
| 4.5 | Number of labels, concepts, mentions, and links automatically annotated by SKET within MedTAG. Statistics are reported for each use-case and globally. | 53 |
| 4.6 | Convolutional Neural Network (CNN) Colon cancer performance when trained with SKET weak labels (CNN-SKET) and with manual ones (CNN-GT). Results refer to Whole Slide Image (WSI) classification on Cannizzaro Hospital (AOEC) and Radboud University Medical Center (RUMC) data. For each considered measure, we report the average obtained through 10-fold cross-validation. Bold values represent the highest scores achieved for each measure. | 57 |
| 6.1 | Number of diagnostic reports annotated per language and use-case. | 94 |

| | | |
|-----|---|-----|
| 6.2 | Number of labels, concepts, mentions and links (mention - concept) automatically annotated per use-case. | 95 |
| 6.3 | Document-level annotation performance analysis in terms of number of actions (e.g. mouse clicks and keys pressed) and elapsed time required to complete the whole annotation process. | 101 |
| 6.4 | Mention-level annotation performance analysis in terms of number of actions (e.g. mouse clicks and keys pressed) and elapsed time required to complete the whole annotation process. | 101 |
| 7.1 | Number of nanopublications per platform. | 117 |
| 7.2 | Assertion numbers for association types: "protein-coding gene expression in tissue" and "protein expression in tissue". | 119 |
| 7.3 | Number of evidences per database. | 119 |

Chapter 1

Introduction

In the last decades, histopathology has experienced an unprecedented revolution; it evolved from a traditional analogical setting to Digital Pathology (DPATH). Histopathology is the science of analyzing tissue specimens using a microscope in order to assess the presence of a disease and eventually evaluate its grade [121]. In the 17th century, the introduction of a new kind of microscope with enhanced magnification/resolution power by Anton van Leeuwenhoek opened new prospects for measuring microscopic objects [203]. Another important step towards histopathology was the introduction of the first microtome adequate for tissue sectioning in the 19th century [280]. In the same period, Johannes Müller published the first book on histopathology [280], entitled *On the Nature and Structure Characteristics of Cancer*. Müller is considered the father of histopathology for his pioneering use of the microscope for pathology. Traditionally, histopathology has been in practice a manual process carried out by pathologists. Specifically, the specimens are fixed on glass slides and are dyed with a staining procedure in order to highlight the precise characteristics of tissue specimens and enhance tissue contrast for the examination with an optical microscope. To automate this burdensome and time-consuming process, automated tissue processors were introduced starting in 1945. Despite these instruments improved tissue processing, a big step forward the modern image analysis was the introduction of digitized histology slides - i.e., WSIs - in 1999 by Wetzel and Gilbertson [311]. This major advancement combined with other technical equipment such as slide scanners, image storage, and workstations marked the advent of DPATH. Hence, a disruptive change in the traditional pathologist's workflow occurred from direct observation on a microscope to digitized glass slides that are visualized on wide screens with different levels of magnification, according to diagnostic requirements. In the last decades, the unprecedented availability of computational resources coupled with the advancements in digitalization led to the new field of Computational Pathology (CPATH). CPATH is a branch of pathology that exploits algorithmic approaches

to process WSIs for several diagnostic purposes including image analysis, classification, and information extraction. Since examining WSIs is a burdensome, time-consuming, and error-prone process CPATH methods are designed to assist pathologists, by providing automatic facilities for supporting diagnostics and the medical decision process. In light of the impressive advancements of Artificial Intelligence (AI) in recent years, state-of-the-art CPATH approaches for image analysis are powered by Machine Learning (ML) and Deep Learning (DL) methods such as CNNs [187].

The context for the present thesis is the ExaMode¹ H2020 European project which aims to develop tools to support the decision-making process in the DPATH domain, leveraging exascale multimodal medical data. In particular, the ExaMode project aims to develop an automatic classification system for WSIs with respect to three use cases: colon, lung, and uterine cervix cancer. To this aim, the idea is to employ a weakly supervised approach that exploits *labels* - i.e., weak annotations - to train a CNN for automatic image classification [199]. To efficiently generate the weak annotations, the proposed solution is to leverage the rich information contained in free-text reports, associated with the images, provided by Laboratory Information Systems (LISs). To extract meaningful information from the textual reports, the Semantic Knowledge Extractor Tool (SKET) has been introduced [197].

SKET is an unsupervised tool that exploits a hybrid approach that couples a rule-based system with pre-trained ML models for knowledge extraction from pathology reports [197]. Figure 1.1 shows qualitatively the knowledge extraction process, having SKET as the core, and the downstream applications. Specifically, SKET extracts from the reports several key information, namely, *labels*, *mentions*, and *concepts*. For instance, in Figure 1.1, we can observe that the label *Adenomatous polyp - low grade dysplasia* has been generated since the related concepts *Moderate Colon Dysplasia* and *Mild Colon Dysplasia* have been identified in the given report starting from the mention *low-grade dysplasia (mild to moderate)*. Moreover, the concept *Biopsy of Colon* has been identified from the associated mention *biopsy sigmoid*. In addition, users can explain the outputs of SKET using SKET X and annotate the reports using MedTAG, a customizable annotation tool for ground-truth creation (see Chapters 5 and 6 respectively). It is worth noting, that SKET has been integrated both in MedTAG and SKET X to provide respectively automatic annotation facilities and the capability of running SKET again with different user-provided parameters.

Despite DL-based approaches for CPATH proven to be effective for image classification tasks [46], they are data-hungry and require large annotated datasets for their training phase. Moreover, the manual annotation of histopathology images is a tedious process that can take up to one hour per single image. In this scenario, Computer-Assisted Diagnostic (CAD)

¹<https://www.examode.eu>

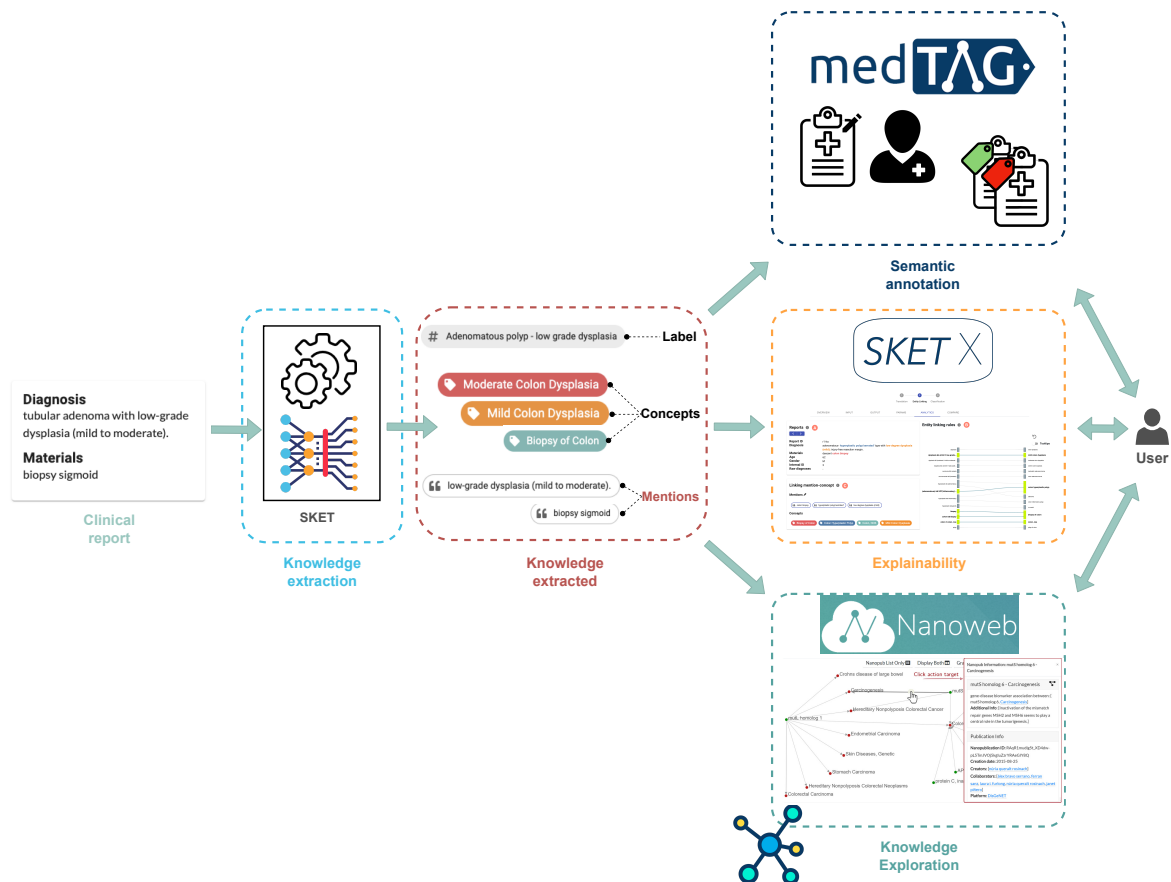


Fig. 1.1 SKET as the core of the knowledge extraction process and the downstream applications. SKET extracts labels, mentions, and concepts from free-text clinical reports. Indeed, from the example reported we can notice one label (*Adenomatous polyp - low grade dysplasia*), two mentions (*low-grade dysplasia (mild to moderate)* and *biopsy sigmoid*), and three concepts (*Moderate Colon Dysplasia*, *Mild Colon Dysplasia*, and *Biopsy of Colon*). Users can visualize and edit the annotation data (knowledge extracted) using MedTAG (see Chapter 6). In addition, they can explain the outputs of SKET using SKET X and explore nanopublications of interest (e.g., nanopublications related to colon cancer) using NanoWeb (see Chapters 5 and 7 respectively).

tools are essential to support pathologists and experts in the annotation process. Indeed, pathologists using CAD tools can save time by taking advantage of automatic facilities such as the detection of the Region of Interests (ROIs) in the tissue specimens to annotate. Hence, the synergic combination of human and machine efforts can improve both the effectiveness and the efficiency of the overall diagnostic process. However, DL methods albeit effective are difficult to explain due to their black-box nature. Therefore, there is an urgent need for eXplainable Artificial Intelligence (XAI) to make AI applications for CPATH comprehensible to pathologists and to promote trust in their predictions [137]. According to recent studies and

interviews, pathologists prefer visual explanations supported by key findings to comprehend algorithm predictions [72, 99]. Moreover, they consider such explanations not a mere “*source of truth*”, but an indication to adjust algorithms’ parameters and steer further analyses. To this aim, Information Visualization (IV) and Visual Analytics (VA) techniques can be employed to provide visual explanations about model outcomes through interactive interfaces. Thus, pathologists can understand why a specific prediction has been determined and the rationale supporting the decision. Despite other domains, such as radiology, have already benefited from using VA techniques [86], their employment in the DPATH domain is still limited.

1.1 Objectives and contributions

The research work presented in this thesis aims to investigate the application of Information Access (IA) and VA methods for supporting the decision-making process in the DPATH domain. In this regard, we developed several tools to assist pathologists and experts to search, access, explore, and annotate clinical reports. Since the automatic annotation of free-text reports is achieved using an unsupervised knowledge extraction tool - i.e., Semantic Knowledge Extractor Tool (SKET) - we employ VA techniques to visually explain why each weak annotation is obtained, so that users can manipulate some model’s parameters, through intuitive interfaces, and refine the knowledge extraction process. The end goal is to improve the effectiveness of the knowledge extraction process and, as a consequence, of the image classification algorithms for the DPATH domain, that are trained on the weak annotations produced.

Therefore, we contributed to multiple lines of research, including:

- **Semantic annotation** is the Natural Language Processing (NLP) task of annotating entity mentions in textual documents with well-defined ontological concepts. The annotated data produced is necessary to train/evaluate ML-based algorithms and the tools performing semantic annotation are known as *semantic annotators*. In this regard, we introduce MedTAG, a customizable annotation tool for creating ground-truths and annotated data (e.g., labels, mentions, and concepts) [107]. Specifically, MedTAG integrates SKET for automating the annotation of clinical reports related to colon, lung, and uterine cervix cancers.
- **Explainability** is the ability to provide human-comprehensible explanations for the predictions of a model. To this extent, we propose SKET X, a tool for explaining the outcomes of SKET by employing VA techniques [197]. SKET X supports pathologists in the analysis of SKET outputs, by means of VA interfaces that allow the experts to

comprehend the models, the parameters, and the rules that determine a specific output. Specifically, SKET X is not limited to static explanations, instead, it allows the users to change model parameters and re-run SKET accordingly so that they can compare the different outputs generated. The purpose is to make pathologists aware of how SKET produces a certain output so that they can provide feedback to continuously improve its effectiveness, the quality of the weak annotations produced, and in turn, also the predictions of the image classification system trained over the weak annotations of SKET.

- **Knowledge exploration** enables the discovery of new findings from the exploration of interconnected data, such as in graph-based structures, in a serendipity-oriented perspective. In this regard, we propose NanoWeb [106], a tool enabling users to search, access, and explore nanopublications using natural language queries. According to the Linked Open Data (LOD) principles, the nanopublication model is machine-readable and supports scientific evidence assertions based on data provenance. Thus, the exploration of nanopublication graphs can enrich domain knowledge and point out interesting findings such as gene-disease connections. In the DPATH domain, this knowledge could be exploited to easily discover, for instance, which genes cause a specific disease, pointing out also other diseases potentially implicated as well as the tissues that may be affected. More importantly, this kind of information is always coupled with the related supporting claims coming from scientific papers, thanks to data provenance.

Finally, we integrate our contributions into the ExaSURE ecosystem, which encompasses the different tools providing a common communication layer and a unified access point for the users interested in using these tools, that are provided as web-based services (see Chapter 3).

1.2 Outline

The present thesis is organized as follows. In Chapter 2, we provide the necessary background on DPATH, CPATH, VA, and the explainability approaches for CPATH, highlighting that the use of VA techniques for explainability purposes in this domain is still limited. In Chapter 3, we describe the ExaSURE ecosystem and how each of the tools integrated contributes to the overall workflow. In Chapter 4, we introduce SKET and we describe how the knowledge extraction process is carried out in order to produce the weak annotations for training image classification algorithms for CPATH. We introduce also SKETUp, as a web-based tool

enabling users to interact with SKET through an intuitive interface. In addition, we describe the ExaNet tool which allows the users to filter clinical reports and visualize their graph-based representation. In Chapter 5, we discuss the explainability of the knowledge extraction process and we introduce SKET X as a VA tool for explaining the outputs generated by SKET. In Chapter 6, we discuss semantic annotation as a crucial practice to produce the annotated datasets - i.e., ground-truths - required for algorithm training and evaluation. In this regard, we propose MedTAG as a customizable annotation tool for ground-truth creation backed up with automatic annotation facilities. In Chapter 7, we discuss knowledge exploration as the process of discovering unexpected findings in linked data. We present nanopublications as a graph of interconnected scientific evidence - i.e., assertions - sustained by data provenance. Hence, we propose NanoWeb as a tool for exploring such knowledge, with the aim of pointing out interesting relations such as gene-disease connections. Finally, in Chapter 8, we draw some general conclusions and discuss future research directions.

It is worth noting, that all the key chapters presented in this thesis are self-contained so that readers can look into a specific chapter and get all the necessary notions for a fruitful understanding. Nevertheless, the background chapter is provided to gather and organize all the relevant concepts in a unique point of reference.

1.3 Publications

The research work presented in this thesis has resulted in the following key publications, ordered by publication date:

- F. Giachelle, D. Dosso, and G. Silvello (2021). Search, access, and explore life science nanopublications on the Web, *PeerJ Computer Science*, February 2021, DOI: <https://doi.org/10.7717/peerj-cs.335>.

This paper describes NanoWeb, a web-based tool enabling the visual exploration of the knowledge represented by the interconnected network of scientific facts encoded in the form of nanopublications. In particular, the nanopublication model is based on the Resource Description Framework (RDF) format that is machine-readable but not easily readable by non-expert users. Thus, the main contribution of NanoWeb is to allow users to search, inspect, and visually explore both life science and biomedical nanopublications through a user-friendly interface. Specifically, NanoWeb allows users to search for nanopublications of interest using natural language queries. Moreover, the visual exploration of the graph representation associated with nanopublications can lead to unexpected meaningful findings, such as gene-disease connections, in

a serendipity-oriented perspective. Finally, NanoWeb allows users to easily cite nanopublications.

- F. Giachelle, O. Irrera and G. Silvello (2021). MedTAG: A Portable and Customizable Annotation Tool for Biomedical Documents, *BMC Medical Informatics and Decision Making*, 21, 352 (2021). DOI: <https://doi.org/10.1186/s12911-021-01706-4>

This paper presents MedTAG, a customizable annotation tool for creating ground-truths and richly annotated data in the biomedical domain. MedTAG integrates SKET - i.e., an unsupervised tool for extracting knowledge from pathology reports - to allow the automatic annotation of clinical pathology reports concerning colon, lung, and uterine cervix cancers. In addition to the automatic annotation mode, MedTAG allows the experts also to manually annotate data from scratch or to revise the automatic predictions of SKET. The annotated data produced with MedTAG are essential to train and evaluate weakly supervised approaches for image classification in the DPATH domain.

- S. Marchesin, F. Giachelle, N. Marini, M. Atzori, S. Boytcheva, G. Buttafuoco, F. Ciompi, G. M. Di Nunzio, F. Fraggetta, O. Irrera, H. Müller, T. Primov, S. Vatrano, G. Silvello (2022). Empowering Digital Pathology Applications through Explainable Knowledge Extraction Tools, *Journal of Pathology Informatics*, 100139, (2022). DOI: <https://doi.org/10.1016/j.jpi.2022.100139>

This paper describes SKET, its downstream applications, and the explainability interface provided by SKET X. The main contributions of this work are: (i) SKET a knowledge extraction tool specifically suited for pathology reports; (ii) SKET X a web-based tool designed to explain SKET predictions in terms of rules and parameters involved at each stage of the knowledge extraction process (e.g., entity linking, classification) by leveraging VA techniques. Pathologists can use SKET X to visually comprehend why a certain prediction has been determined, evaluate the impact of the different parameters involved, and eventually provide useful feedback to improve the overall effectiveness of the knowledge extraction process.

Chapter 2

Background

Histopathology is the science of analyzing tissue specimens and cells in order to identify signs suggesting the presence of a particular disease and evaluate its progression [121]. Specifically, the specimens are collected with biopsies or other surgical interventions on patients. Then, the specimens are fixed on glass slides to be analyzed using a microscope. Moreover, the specimens under examination are dyed using a staining procedure (e.g., Hematoxylin and Eosin (H&E) stain [9]) to visualize and distinguish their different components. Traditionally, the examination was mostly a manual process requiring the pathologists to observe the specimens in the glass slides directly on optical microscopes. In the last decades, the introduction of digitalization processes in histopathology has led to the advent of DPATH enabling pathologists to visualize the digitized glass slides on large screens and with different levels of magnification. Thus, improving image analysis and making the whole diagnostic process more effective. In particular, the purpose of image analysis is to “*obtain meaningful information from images in an objective and reproducible manner*” [5].

The large amounts of digitized glass slides available, namely WSIs, have paved the way for the employment of algorithms for image analysis and for the rise of CPATH. Indeed, AI approaches and, specifically, DL algorithms for image analysis have obtained promising results and demonstrated to be effective, especially for image classification tasks [188, 92, 290, 138, 74, 274, 46, 199]. Despite the promising results, the adoption of AI algorithms and other CAD tools in clinical settings is still limited. Moreover, most of the CAD tools proposed in DPATH rely on DL algorithms that are in general effective but difficult to explain and trust due to their black-box nature [281]. Hence, pathologists are still facing information overload while carrying out burdensome, time-consuming, and error-prone processes. Thereby, there is an urgent need for explainable CAD tools solution for reducing the human cognitive effort. In particular, these tools should not only ease image analysis activities but also integrate the knowledge necessary to comprehend the predictions and

for their assessment. The work we present in this thesis tries to tackle these issues by synergistically combining different computational, analytics, and visual approaches to support diagnostics and the medical decision process with respect to the DPATH domain. Hence, in this chapter, we provide the necessary background on DPATH and CPATH, as well as on visual analytics and its use in DPATH for information access, exploration, and explainability. In this way, the reader can have a comprehensive view of how visual analytics, explainability, and knowledge exploration can support and speed up CPATH tasks and in turn improve the overall pathologist workflow. The rest of this chapter is organized as follows. In Section 2.1 we provide a description of DPATH and CPATH. In Section 2.2 we provide a comprehensive overview of VA, IV, and their applications. In Section 2.3 we describe current state-of-the-art VA applications for the DPATH domain and we envision future scenarios where VA methods could be used for CPATH. In Section 2.4 we discuss explainability in the context of CPATH and its crucial role in promoting human comprehension and trust with respect to algorithms' results and predictions.

2.1 Digital pathology and computational pathology

The Digital Pathology Association (DPA)¹, a non-profit organization involving interdisciplinary experts such as pathologists and scientists, has provided definitions for both DPATH and the emerging field of CPATH. DPATH has been defined by the experts of the DPA as “*a blanket term that encompasses tools and systems to digitize pathology slides and associated meta-data, their storage, review, analysis, and enabling infrastructure*” [1]. Moreover, DPA defines CPATH as “*a big-data approach to pathology, where multiple sources of patient information including pathology image data and meta-data are combined to extract patterns and analyze features*” [1]. Another definition proposed for CPATH is “*a branch of pathology that involves computational analysis of a broad array of methods to analyze patient specimens for the study of disease*” [1]. CPATH involves the application of computational-intensive techniques such as AI approaches for diagnostics and supporting the decision-making process in the DPATH domain. In this regard, ML and DL algorithms are employed to power image analysis such as image classification and segmentation. In the histopathology domain, glass slides have been extensively used traditionally for diagnostic purposes [3]. Historically, histopathology has been mostly a manual process requiring pathologists to examine tissue specimens and glass slides using brightfield microscopy. In the last decades, the advent of DPATH and the spread of digitization practices for glass slides have paved the way for new computer-aided scenarios, where automatic systems can support pathologists in both diag-

¹<https://digitalpathologyassociation.org>

nostic and clinical activities. A WSI is the digital representation of a whole histopathological glass slide, obtained by employing slide scanners. Typically, the inspection of each WSI is possible at different levels of magnifications (e.g., $20\times$ and $40\times$), using appropriate visualization software. The different levels of magnification induce a "pyramidal" structure, as shown in Figure 2.1. Thus, enabling the pathologists to visualize WSIs on a commercial monitor using the level of detail required for diagnostic purposes. Hence, when pathologists are not interested in conducting their analyses with a high level of detail, down-sampled images with lower magnification are provided so that the computer Random Access Memory (RAM) is not overloaded with the cumbersome amount of data of whole higher magnification images. Instead, when they are interested in a small specific portion of the specimen, that is analyzed using a high level of magnification is available.

The analysis of WSIs is a burdensome and time-consuming activity since a single WSI can require a pathologist even more than an hour of work [172]. In this context, CAD tools integrating DL algorithms, capable of learning from data, can reduce the workload of pathologists and support the medical decision-making process. Indeed, CPATH aims at developing CAD tools designed to automatically analyze DPATH images. DL state-of-the-art approaches usually involve CNNs, which demonstrated to be effective in solving different CPATH tasks. CNNs usually require to be trained on large annotated datasets due to the high morphological and technical variability in images, equipment, and protocols. For instance, the same glass slide digitized with two different scanners may produce images rendered with slightly different colors. Nevertheless, preserving the captured colors during the whole DPATH workflow is considered a challenge [3]. In recent years, predictive models for medical image analysis have evolved from models relying on manually engineered feature extraction to supervised methods based on DL. In contrast to the first models which were characterized by low performance inadequate for clinical applications, the second ones have demonstrated to be effective especially in solving image classification tasks [98]. For the time being, two remarkable supervised approaches have been proposed: (i) classification of the small tiles/patches within a WSI [150] and (ii) weakly supervised WSI classification based on the Multiple Instance Learning (MIL) assumption [169, 46, 199]. The MIL assumption is based on the idea that a learner receives a set of labeled *bags*, where each bag contains several instances. In the binary classification case, this translates into a labeled negative bag only when all the bag instances are negative. Instead, when a bag is labeled positive it means that at least one instance is positive as well [83]. The MIL assumption in the DPATH domain is instantiated as follows: a single whole slide-level diagnosis produces a weak label that applies to all the related tiles within the same WSI. Thereby, if a slide is negative then all the

tiles are negative as well, conversely, when a slide is positive then at least one tile indicates the presence of cancer.

Despite DL approaches for the supervised classification of small WSI tiles have achieved promising results [31], they require extensive and time-consuming pixel-wise annotations, thus limiting their application only to small curated datasets. Instead, weakly supervised methods based on the MIL assumption can promote the application of computational approaches on large-scale datasets with a higher level of generalization. Indeed, weakly supervised approaches can leverage the rich textual information provided in clinical reports, Electronic Health Records (EHRs), and LISs to train classification models. The vast amount of valuable data present in such systems can power DL algorithms without requiring further expensive and tedious annotation activities. In addition, when the data archived in LISs are provided in a structured format, it is easier to retrieve specific information. Nevertheless, in the case of unstructured free-text clinical reports, knowledge extraction algorithms need to be employed to obtain the weak annotations [199]. In this regard, an unsupervised knowledge extraction tool called SKET is specifically suited for the DPATH domain [197]. To be precise, SKET combines rules based on the experience of pathologists with pre-trained ML models to automatically extract key labels (i.e., weak annotations) and concepts from diagnostic reports (see Section 4). This approach can not only avoid experts to annotate free-text clinical reports but also dramatically reduce the time needed to get the annotations, especially in the case of large datasets, shifting from hundreds of human hours to just a few hours [199]. Despite the promising results obtained by DL-based approaches, these models are often difficult to explain due to their black-box nature. In particular, humans may not understand the reason why a specific prediction or output has been determined. Nevertheless, in-depth comprehension of these models is not only desirable but also urgent, especially in the DPATH domain, in order to trust models and their predictions [135]. An approach to support the comprehension of the former models involves the application of VA and IV techniques to allow the users to visually comprehend the underlying machine decision process by leveraging interactive interfaces enabling punctual inspection of specific aspects of interest [197].

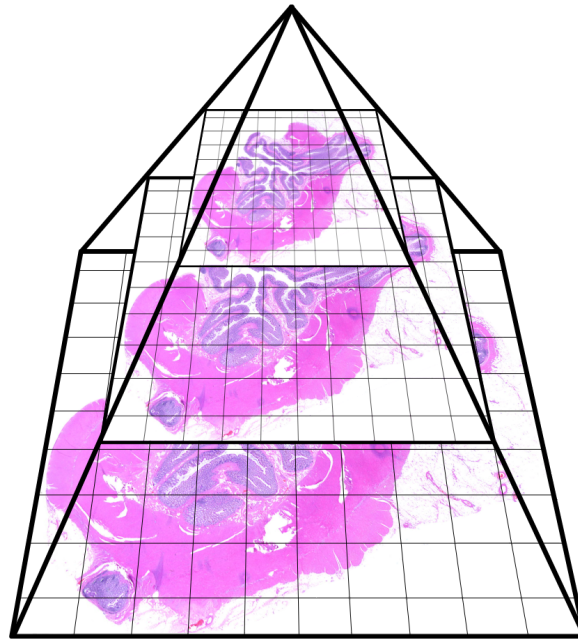


Fig. 2.1 The pyramidal structure of WSI, resulting from different levels of magnification.

2.2 Visual analytics and information visualization

The first appearance of the term “*visual analytics*” was probably in the special issue “Guest Editors’ Introduction-Visual Analytics” of IEEE Computer Graphics and Applications in 2004 [301]. Nevertheless, before this term was coined several related research activities were conducted with the following terms: *Exploratory Data Analysis* [285], *Scientific Visualization* [244], *Data Driven Discovery & Knowledge Discovery in Databases (KDD)* [101, 220], *Information Visualization* [272], and *Visual Data Mining* [269]. Exploratory Data Analysis (EDA) is a term coined by John W. Tukey for describing the act of *looking at data to see what it seems to say* [213, 285]. EDA is a branch of data analysis that tries to identify interesting patterns and insights in data from direct observations. In contrast to Confirmatory Data Analysis (CDA), which assumes a model for data and tries to verify it using statistical hypothesis tests, EDA does not make any assumptions on data and does not try to evaluate any pre-defined hypotheses; it focuses instead on identifying interesting and not evident patterns on data through the observation leveraging exploratory visual tools, such as box plots, and human intuition. Moreover, in contrast to CDA which employs visualization techniques just to present results, EDA takes advantage of interactive visualizations to manipulate how the data are presented, thus producing different views that could foster understanding and new insights. This implies a direct integration of the users in the data analysis process and their fundamental roles in knowledge discovery. This paradigm shift was the first important step

toward the research field of VA. The purpose of visualization is to “*convey salient information about underlying data and processes*” [122]. As reported in [122], the 1987 National Science Foundation’s Visualization in Scientific Computing Workshop report provides the following definition for visualization:

Visualization is a method of computing. It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights.

Robert Kosara defined three criteria and requirements for any visualization: (i) it is based on (non-visual) data; (ii) it produces an image and (iii) the result is readable and recognizable [166]. The meaning of the former criteria is that any visualization process involves data, coming in general from different datasets, that are visualized as readable and comprehensible images. These criteria apply also to scientific visualization and IV as two branches of visualization. The first one is related to the visualization of the data produced in scientific processes such as three-dimensional and spatial data as occurs in computational chemistry [244]. Instead, a definition proposed for IV is: “*the use of computer-supported, interactive, visual representations of abstract data to amplify cognition*” [53]. Hence, IV is not strictly bounded to the visualization of scientific data, as occurs instead for scientific visualization; it is rather related to the act of visualizing abstract data (e.g., business data), regardless of their kind, to convey meaningful information and insights [266]. Visual data mining combines automatic analysis techniques such as data mining and statistical approaches with IV. Ankerst defined visual data mining as “*a step in the KDD process that utilizes visualization as a communication channel between the computer and the user to produce novel and interpretable patterns*” [23]. Simoff proposed another definition for visual data mining as “*the process of interaction and analytical reasoning with one or more visual representations of abstract data that leads to the visual discovery of robust patterns in these data that form the information and knowledge utilized in informed decision making*” [269]. Hence, visual data mining involves three key elements: (i) data mining and computational approaches to unveil not evident patterns; (ii) interactive IV approaches enabling (iii) users to produce different visualization of data with the aim of promoting knowledge discovery and support the decision making process. The combination of these aspects represented another important step toward VA.

VA is the science of analytical reasoning facilitated by visual interactive interfaces [279]. After the tragic events of September 11, 2001, the US Department of Homeland Security (DHS) established the National Visualization and Analytics Center (NVAC) in 2004 with the purpose of countering future terrorist attacks in the US and worldwide [279]. To this

aim, NVAC defined a long-term agenda for supporting research and development of VA as a tool for analyzing potential terrorist threats. The term *Visual Analytics*, originally coined by James J. Thomas in the research and development agenda of VA “Illuminating the path” [279], refers now to an interdisciplinary field that combines different research areas including Data Analysis (DA), Data Mining (DM), Human Computer Interaction (HCI) and IV [155, 154, 153]. The grand challenge of VA is to turn information overload [307] and data deluge [127] into an opportunity to analyze huge amounts of data, such as in a continuous stream, leveraging the human visual/perceptual system and expertise to make effective decisions [155]. In this regard, Keim et al. pointed out that relying only on interactive visualization methods is clearly not enough to deal with massive and ever-increasing amounts of data [152]. Indeed, the synergic combination of IV, DM, and HCI can support data analysis and pattern recognition, thus improving the decision-making process. After the first definition proposed by James J. Thomas, several other definitions for VA have been proposed [77]. All the definitions proposed may be qualitatively synthesized by the diagram in Figure 2.2.

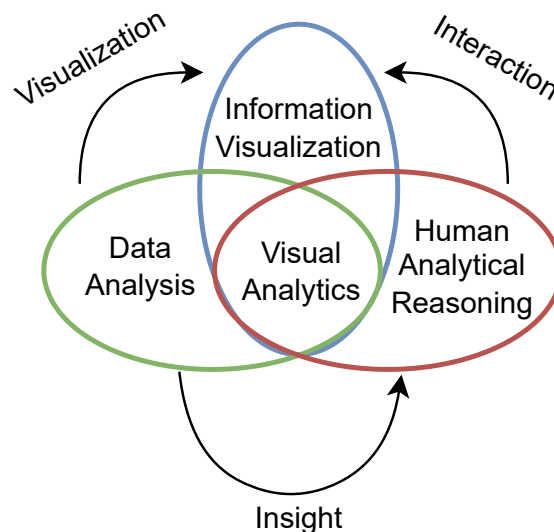


Fig. 2.2 VA combines DA, IV, and human analytical reasoning enabling users to visually analyze and comprehend data. VA exploits the human perceptual/cognitive system and analytical reasoning to reveal meaningful patterns in data and foster unexpected insights.

The user has a prominent role in VA since human judgment is a crucial part of the VA process. Indeed, VA takes advantage of the human-in-the-loop paradigm to integrate human action/feedback to interactively manipulate data, refine parameters and processes, and explore results [234, 115, 283, 97]. The famous Shneiderman’s mantra “*Overview first, filter and*

zoom, details on demand” is a guideline for the design of IV interfaces and systems [266]. The meaning of this mantra is that user interfaces should be designed to show to the user only the relevant information for the current activity, providing auxiliary information when needed. This is desirable to avoid overwhelming users with unnecessary or redundant information. The same principle applies also for VA, for which Daniel A. Keim adapted Shneiderman’s mantra as follows: “*Analyze first, show the important, zoom, filter and analyze further, details on demand*” [151, 153]. This mantra is formally defined by Daniel A. Keim in the visual analytics process [153] illustrated in Figure 2.3. The visual analytics process takes in input a set of datasets $S = S_1, \dots, S_m$ and returns the insight I as the output.

We can achieve the insight I through two diagram branches: (i) U_{CV} which indicates an insight generated by taking advantage of the set of possible visualization V for the data; (ii) U_{CH} which occurs when some hypotheses H are confirmed using automatic data analysis approaches. Visualization can either interest data V_S or hypotheses V_H . Similarly, hypotheses can be generated from data H_S or from visualization H_V . Moreover, user interactions can induce a self-loop either on visualization U_V or in hypotheses U_H . In particular, U_V indicates an interaction that changes the current visualization, such as mouse selection or zoom. Instead, U_H describes an interaction that generates new hypotheses from existing ones. The last self-loop concerns D_W which represents a data transformation such as data pre-processing, cleaning, and selection. Finally, the feedback loop indicates that this is a continuous process where user interactions refine, for instance, the parameters/hypotheses of a model to visually comprehend the effect of these variations in the final output. VA methods

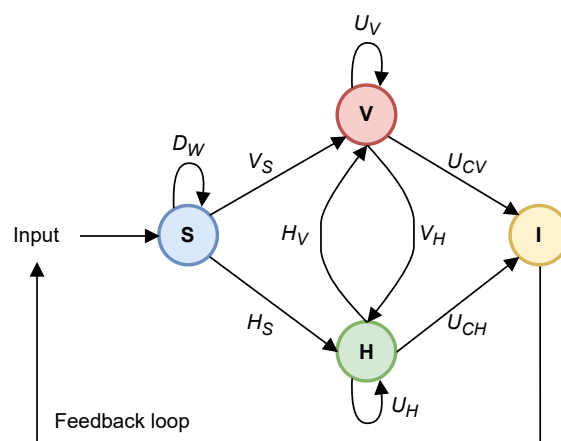


Fig. 2.3 The VA process proposed by Daniel A. Keim [153]. The process takes in input a dataset from S and returns the insight I as the output. I can be determined directly from the set of possible visualizations V or through the confirmation of some hypotheses H by employing automatic data analysis approaches.

allow the users to enter into a sense-making loop where data are interactively manipulated to grasp insights and foster the whole knowledge discovery process [253]. Figure 2.4 shows the sense-making loop based on the visualization model proposed by van Wijk [292]. The sense-making loop describes how data are interactively visualized, leveraging the human perceptual/cognitive system, to discover new valuable knowledge and gain insights. Thereby, new hypotheses may be formulated and new analyses may be conducted also changing the specifications and the kind of visualization to refine the overall knowledge discovery process. VA techniques allow the experts to exploit interactive interfaces to conduct dynamic analysis

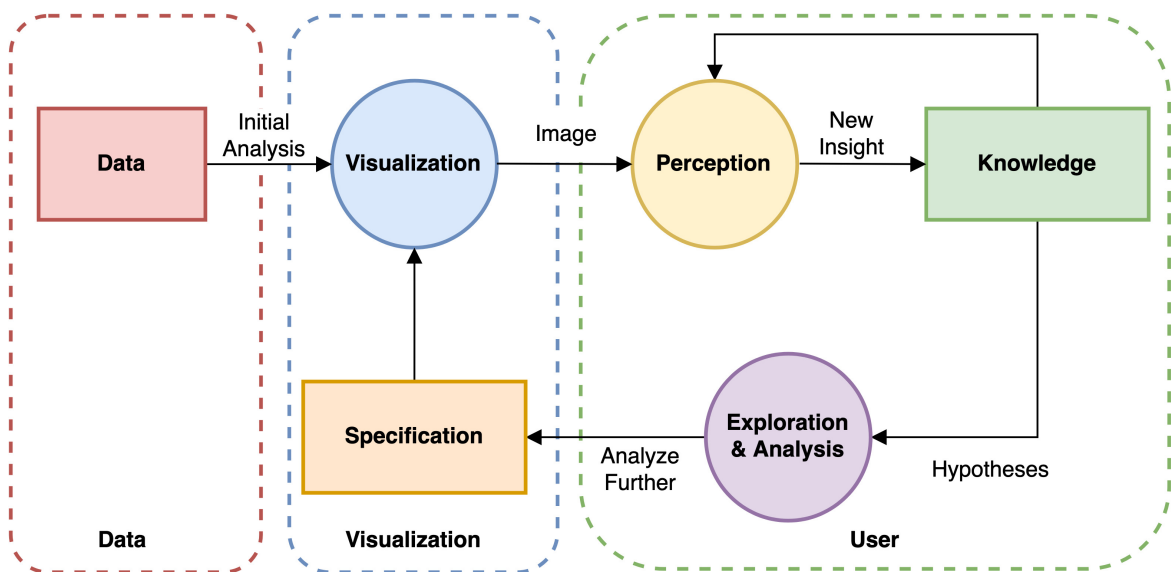


Fig. 2.4 The sense-making loop [151] based on the visualization model proposed by van Wijk [292].

of data and identify not-evident patterns so that they can make decisions accordingly. VA techniques are particularly useful when dealing with on-going processes whose intermediate outputs need to be analyzed as they proceed asynchronously in the background. VA is not limited to the visualization of static data, as occurs for Data Visualization (DV), it is focused instead on providing an insightful visual representation of dynamic data, such as intermediate algorithm results, to adjust the algorithm's parameters in order to continuously refine its outputs. This process is also known as Progressive Visual Analytics (PVA) [261, 22].

VA methods have been applied in several application domains including (i) monitoring and ensuring public safety and security; (ii) studying of environment and climate change; (iii) analyzing social media information; (iv) engineering; (v) business and financial analysis; (vi) software analytics and (vii) biology and health [151, 153]. For instance, Andrienko et al. [12, 11, 13] investigate the use of VA to analyze huge amounts of spatial data in order to understand movement behaviors and identify mobility patterns. In particular, the authors

present a VA framework enabling users to identify the most visited places by leveraging a combination of interactive visual interfaces and computational approaches, such as clustering methods. In the context of Information Retrieval (IR), Angelini et al. propose to employ VA methods to support the exploration of experimental data, failure analysis, and evaluation of IR models, which is a crucial task in this domain [19, 20, 18]. Instead, for what concerns the financial domain, several VA applications have been proposed to analyze financial data [315], including (i) the FinDEX system proposed by Keim et al. [156] which allows the users to visually compare the performance of a fund with respect to the market and according to user-specified time intervals; (ii) the Smartmoney system proposed by Wattenberg [299] was designed to monitor the stock market using a layout based on the Shneiderman's treemap diagram [265], that allows point-wise inspection as well as an overview of the bigger picture; (iii) the FinVis tool proposed by Rudolph et al. [252] allows non-expert users to interpret the return, risk, and other aspects of financial data in order to make decisions for their personal financial plan; (iv) the EVA tool proposed by Leite et al. for the detection of fraudulent events [184]. Social media represents another domain where VA techniques have been applied successfully for different tasks including sentiment analysis, monitoring information spreading, and understanding user behaviors [305]. For what concerns security, VA methods demonstrated to be effective for supporting analyses and monitoring of digital threats such as cyber-attacks [112, 21, 16, 14]. Furthermore, biology and medicine offer several scenarios of applicability for VA techniques [45, 302]. For instance, in the context of computational biology, Ding et al. employed VA techniques for supporting biomedical researchers in the exploration of complex genomics data [84]. Specifically, they used unsupervised clustering along with interactive interfaces to visually compare different combinations of gene expression and their impact on clinical outcomes and survival predictions. Another example, in the computational biology domain, is the OmicsView system proposed by Casey et al. [56]. OmicsView allows the users to visually identify key biomarkers for a specific disease and enable disease-versus-normal gene expression comparisons. For what concerns the healthcare domain, VA techniques have been applied for different purposes including (i) analyzing healthcare processes and their conformance to existing guidelines [28]; (ii) improving healthcare education [288]; (iii) supporting physicians in the exploration of disease patterns and analysis of the interactions with patients' characteristics, by leveraging the information content of EHRs [250]; (iv) exploring complex network medicine information such as gene-gene, gene-disease, and drug-gene interactions, in order to support explorative analysis and hypotheses validation [15]; (v) supporting the medical decision-making process [124]; (vi) supporting the interactive visualization and exploration of COVID-19 EHRs [140] and (vii) the progressive visualization of the epidemiological models [17].

In recent years, VA methods have attracted increasing interest for their application to the DPATH and CPATH [270]. For instance, these methods are used to support diagnostic activities and reporting.

2.3 Visual analytics and information visualization for computational pathology

In the last decades, the workload of pathologists has increased and it is expected to continue steadily on this trend in the future [277]. Indeed, a “*fully digital*” pathology workflow [275, 103] typically entails a higher digitalization throughput, thus producing large amounts of images to be analyzed. Hence, information overload is a challenge that needs to be addressed. To mitigate this, CAD tools could support and optimize the pathologist’s workflow [102]. In recent years, the importance of employing VA methods within DPATH has been recognized by pathologists as these methods could support them in diagnostic tasks [72]. However, limited literature has been presented so far. Thus, highlighting that the application of these techniques in the DPATH domain is still pioneering. In this regard, Corvò et al. [73] suggest four key areas where VA methods can be employed in DPATH practice:

1. Support for WSIs quality assurance. The authors suggest assessing the quality of the slides digitized using a dashboard integrating VA features. For instance, the users should be able to visualize the slides according to multiple filters such as acquisition date, morphologic structure, and staining type. Pathologists and technical experts should be also able to interactively judge the quality of the WSIs, pointing out the presence of artifacts.
2. Support advances in Image Management System (IMS). The authors suggest that VA features should be integrated also into IMSs to enable search, exploration, and manipulation of large-scale histopathology images.
3. Support in diagnostics. The recent advancements of AI and DL for image analysis and classification [38] suggest that these computational approaches will be integrated into future CAD tools for supporting pathologists in diagnostic activities. In this regard, VA methods should be exploited to allow users to visualize the features used in image analysis and related tasks. In addition, the authors indicate that the next-generation CAD tools should be explainable using VA techniques that allow answering why a specific prediction has been made. This is a mandatory requirement in the DPATH

domain since pathologists need to understand and trust machine decisions in order to adopt such algorithms in clinical practice [137].

4. Support in Reporting. Reviewing tissue slides and reporting are the two major activities of pathologists [72]. For this reason, the future CAD tools should support pathologists in reporting by showing relevant findings and their provenance [240]. This can be achieved by employing VA methods to combine findings, communicate them, and foster insights.

In recent years, Corvò et al. have proposed *PathoVA* [71], which is a VA tool supporting pathologists for diagnostic and reporting activities. Specifically, *PathoVA* is a CAD tool that allows the users to (i) visualize the glass slides digitized; (ii) visualize the diagnostic trace, that is, a visual log of the interactive activity of pathologists with the tissue slide viewer; (iii) generate a final report enriched with the information concerning slide image portions and the related findings. *PathoVA* is particularly suited for tissue examination purposes including tubule detection, nuclei detection, and mitotic cell counting, which are important factors to assess cancer progression [204]. In addition, *PathoVA* allows the user to quantify tumor areas, boundaries, and compute the Nottingham Histologic Score (NHS), which is employed to assess the cancer grade and its progression [96]. Finally, *PathoVA* combines image analysis results with interactive report writing, so that textual annotations can refer to the specific portions of interest. In contrast, previous works such as the *GRAPHIE* tool proposed by Ding et al. only focused on the exploration and annotation of large collections of histology image datasets, without providing specific reporting features [85]. In light of this limitation, the authors of *PathoVA* pointed out the importance of integrating image analysis and reporting in a unique interface in order to improve pathologists' productivity. In this regard, Cervin et al. [58] proposed the first tool integrating both aspects with the aim of speeding up report annotation using a structured report format that opened for automatic extraction of report findings. Another advancement has been proposed by Corvò et al. with the introduction of the *IComPath VA* tool [70]. *IComPath* allows pathologists and experts to conduct hypotheses exploration as the investigation of diagnostic biomarkers and the assessment of candidate new features for the image analysis feature space in *CPATH*. In addition, *IComPath* provides support for the visual creation of groups of patients (cohorts) characterized by similar properties, in order to investigate common patterns, compare quantitative information/measurements, and verify hypotheses. Recently, new tools have been introduced for the efficient visualization of high-dimensional multi-channel microscopy images of specimens [221, 128]. These images consist of billions of pixels (10^9 or more) and are multiplexed in multiple channels (e.g., 60 or more) to potentially represent millions of individual cells, thus requiring up to hundreds of gigabytes per image. The introduction of

visualization tools supporting the efficient visualization of this kind of complex and heavy images paved the way for the development of IV and VA tools designed to support image analysis and diagnostic activities for DPATH in a web browser setting. For instance, the *Facetto* tool proposed by Krueger et al. combines VA with semi-automated analysis of cell types and states [171]. In particular, experts can (i) visualize the histological images where cells are classified using clustering algorithms; (ii) adjust and steer the clustering process by leveraging interactive visualization and multiple coordinated views. In contrast, *histoCAT* is an IV toolbox for MATLAB² providing analysis of spatial features (e.g., cell size and shape) for single-cell data extracted from histologic images [258]. Specifically, *histoCAT* enables the exploration of individual cell phenotypes, cell-cell interactions, and morphological structures; it leverages multidimensional reduction techniques, such as t-SNE [291], to represent data with different visualizations (e.g., scatter plots, box plots, and histograms). Another tool, *Scope2Screen* proposes the exploration and annotation of high-plexed WSIs – i.e., size: 100+ GB; resolution: $\geq 30k \times 30k$ – by providing interactive lensing for focused analysis [142]. *Scope2Screen* allows pathologists to store and organize snapshots of annotated ROIs so that they can search, access, and restore quickly specific image locations.

It is worth mentioning that, all the previous tools have been positively received by pathologists and experts, especially for supporting them in diagnostic and routine tasks. However, the former tools do not integrate explainability aspects to support pathologists in the comprehension of algorithm predictions, thus limiting insights and clinicians' trust. To mitigate this, VA methods could be used to visually explain machine decisions and predictions. Indeed, other medical fields such as radiology have already benefited from the employment of visual explainability methods in CAD tools. For instance, Dmitriev et al. propose a VA approach for supporting the decision-making process of a CAD tool designed for the classification of pancreatic cystic lesions [86]. Despite the urgent need for explainable AI solutions, especially for biomedical and diagnostic purposes as for CPATH, to the best of the author's knowledge, there are currently limited applications of VA for explainability in CPATH.

2.4 Explainability

Explainability is the ability to provide “*explanations as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans*” [118]. Explainability is strictly connected with the concepts of *interpretability*, *comprehensibility*, and *understandability* [26]. Interpretability is defined as

²<https://www.mathworks.com/products/matlab.html>

the “*ability to explain or to provide the meaning in understandable terms to a human*” [89]. Comprehensibility is the “*ability of a learning algorithm to represent its learned knowledge in a human understandable fashion*” [76, 111]. Instead, understandability “*denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally*” [210]. In the last years, the rise of AI based on ML and DL has pointed out the essential need of XAI to face the shortcomings of interpretability and explainability concerning the underlying decision mechanisms of these models [149]. The lack of interpretability is due to the fact that these models take advantage of statistical and probabilistic approaches to solve complex problems, typically dealing with a large number of dimensions in the feature space. Thereby, the structure and the decision process of these models are opaque and difficult to interpret by humans. Indeed, Yoshua Bengio (a pioneer in the research of DL) stated “*As soon you have a complicated enough machine, it becomes almost impossible to completely explain what it does*” [44]. Nevertheless, explainability is not only desirable but also a requirement as delineated by the *High-Level Expert Group on Artificial Intelligence*, established by the European Commission, in order to set the guidelines for trustworthy AI. Specifically, experts identified explicability as a guiding principle for ethical AI, by stating: “*algorithmic processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions explainable to those affected both directly and indirectly*” [256]. This is also in compliance with the European GDPR³ which requires transparency for algorithms to promote fairness and protect against discrimination and biases. To promote interpretability and XAI, the Defense Advanced Research Projects Agency (DARPA) started the XAI program in May 2017 [120]. Even though there are multiple definitions for XAI, a common goal is delineated, that is, providing human-comprehensible explanations so that users can understand and trust the decision-making process of a model. In other words, XAI is “*a research field that aims to make AI systems results more understandable to humans*” [2]. In accordance with this purpose, Gunning [119] proposes the following definition for XAI:

XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

For the time being, in light of the crucial need for interpretable models, different XAI approaches have been introduced, as shown in Figure 2.5. The first distinction, regarding XAI methods for ML, is between *ante-hoc* and *post-hoc* explainability. The first one refers

³<https://gdpr-info.eu/art-12-gdpr>

to models intrinsically explainable, for instance, by design such as rule-based systems and decision trees. These models are regarded also as *transparent models*. Specifically, a model is considered *transparent* if it is understandable by itself [26]. For instance, a rule-based system is interpretable (transparent) by design, since for each input X satisfying a condition C there is a rule $X \xrightarrow{C} Y$ determining the output Y . In contrast, post-hoc explainability methods are designed to tackle black-box models and their opaque nature (e.g., based on DL architectures). This kind of techniques can be divided into *model agnostic* and *model specific*. Model agnostic explainability techniques do not have any clue about the inner model structure they are applied to. They try to perturb the input for the model in order to observe how and in which measure the output change. This approach is used for instance by Local Interpretable Model-Agnostic Explanation (LIME), a tool proposed by Ribeiro et al. [245] capable of building a local linear model that mimics the predictions of a black-box model to explain it. To this aim, LIME exploits two post-hoc explainability techniques, namely, *explanation by simplification* and *local explanation*. The first one refers to the act of creating a simplified model (e.g., a decision tree) around an opaque one, with the aim of generating the same predictions for a given input. Thus, the simplified model can be studied to comprehend the decision process of the complex one and gain insights. Instead, local explanation methods target the local behavior of a model, by understanding the output predicted by the model for specific input values. Then, the knowledge obtained on a focused part of the model can be exploited to figure out how the whole model works. Another post-hoc explainability technique is *feature relevance explanation*, which consists of measuring the relevance of each feature in determining a specific model prediction [26]. This approach is employed by the SHapley Additive exPlanations (SHAP) post-hoc technique to calculate a feature importance score as the sum of individual feature contributions. The last post-hoc technique used also for model-agnostic explainability is *visual explanation*. Visual explanation allows the users to visualize the behavior of a model. To this aim, visual explanation methods are usually coupled with dimensionality reduction techniques enabling model interpretability by means of intuitive visualization in a low-dimensional space. In this regard, Cortez et al. present a set of visualization techniques based on sensitivity analysis to support the explanation of black-box models such as neural networks [68, 69]. In contrast to model-agnostic techniques, model-specific methods can leverage the knowledge about a complex model's inner structure/architecture to adopt post-hoc explainability strategies that better explain its predictions. For instance, we can observe in Figure 2.5 that for CNNs, Recurrent Neural Networks (RNNs), Multi-Layer Neural Network (MLNN) we can take advantage of the *architecture modification* technique to comprehend the internal decision mechanism. Specifically, this technique entails a set of alterations to the model architecture, such as

adding/removing some layers or changing the loss function, to grasp this perturbation's effect on model predictions. Neural networks and specifically CNNs are employed in different domains for a broad set of purposes, including image analysis for CPATH.

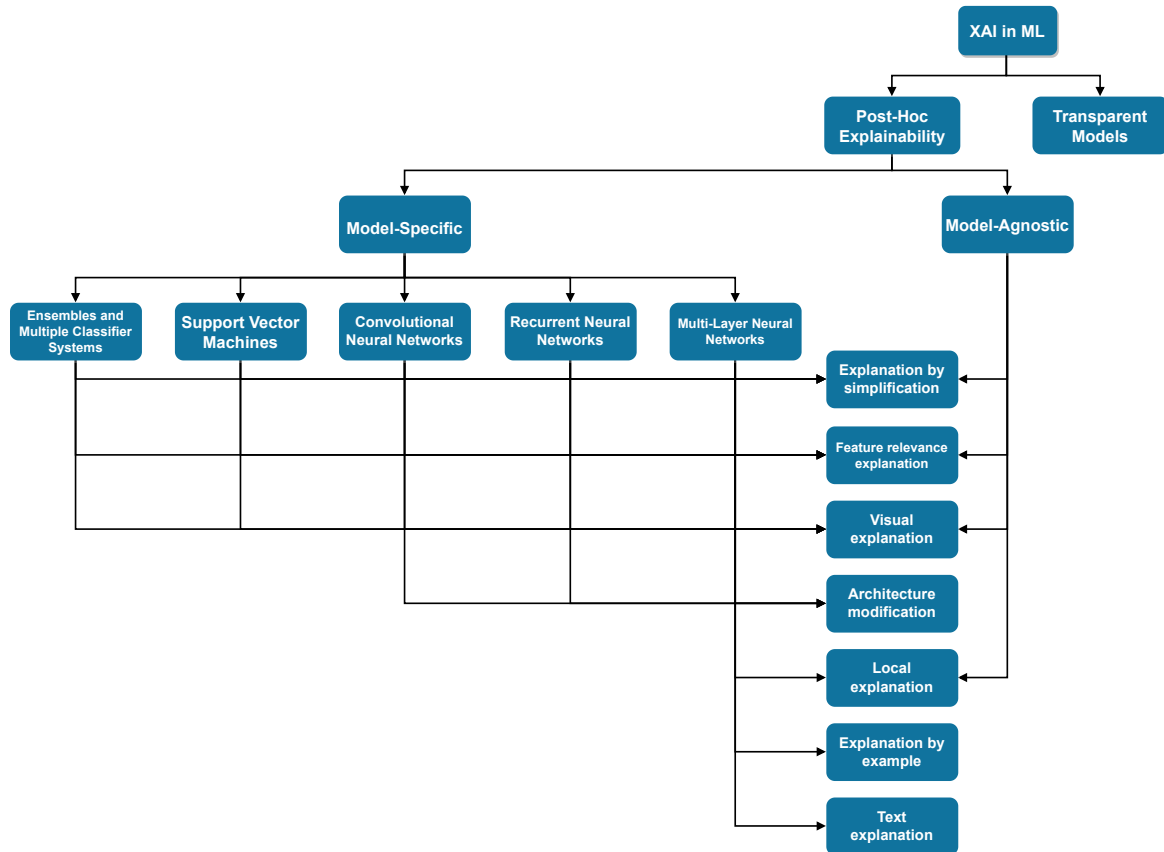


Fig. 2.5 A taxonomy of explainability techniques based on the work proposed by Arrieta et al. [26].

2.4.1 Explainability for computational pathology

Explainability is essential in the medical domain to allow physicians and experts to understand and trust model predictions [134]. Furthermore, it is also a mandatory requirement according to the ethical AI principles invoking transparency for machine decision processes, especially for applications in medicine [256, 176]. However, in the context of CPATH, most of the approaches for image analysis are based on DL techniques (e.g., CNNs) that are effective but difficult for humans to comprehend due to their black-box nature [137]. Hence, there is an urgent need for XAI in the DPATH domain [136]. Moreover, Lake et al. point out that AI systems need a paradigm shift from solving pattern recognition problems to building causal models [179]. In this regard, Holzinger et al. [135] emphasize this concept with respect to

the medical domain where in order to build trustworthy, transparent, and interpretable models is necessary to understand the *causality* [227] in the learned machine decision process. Thus, *causability* is proposed as a measure of explanations' effectiveness in supporting human casual understanding [135].

For what concerns image analysis based on CNNs, post-hoc explainability techniques can be employed to explicate why a specific prediction has been made. Figure 2.5 shows that *explanation by simplification*, *feature relevance explanation*, *visual explanation*, and *architecture modification* are post-hoc techniques that could be exploited to explain black-box models based on CNNs, as well as on RNNs and MLNNs. Moreover, Evans et al. [99] indicate four key categories for XAI targeting image analysis, including:

- **Saliency maps** try to explain predictions individually using visualizations overlays on input images that convey the pixel-wise *saliency* of different RoIs [211]. For instance, saliency could measure the cancer risk associated with a specific tissue area having red, green, and blu colors indicating respectively high, moderate, and low risks [214].
- **Concept attribution** approaches try to explain individual predictions and the underlying decision mechanism of a model using a set of high-level concepts expected to be part of the representation learned by the model or to be relevant for a specific prediction [157, 114].
- **Prototypes** mimic instances of a particular class, feature, or model outcome by means of synthetic visualizations. For instance, Li et al. introduced a *prototype layer* in the proposed neural network architecture, so that the neural network can come naturally with explanations for each prediction, after learning *prototypes* that resemble the encoded inputs [186].
- **Counterfactuals** examples try to explain model predictions by means of different what-if scenarios. For instance, Seah et al. [262] propose for a given image to create a new image that resembles the original one but induces the neural network to classify it with a different class. This idea has been used by Pocevičiūtė et al. [233] for image classification in CPATH, where patches containing tumor cells are transformed into healthy ones and tested with the neural network to compare prediction scores. As a result, it is possible to understand which parts of the images contribute to the correct prediction and which do not.

Evans et al. [99] conducted an interview with six expert pathologists. From the interview, it emerges that pathologists have a clear preference for visual explanations: “*Pathologists are always looking for visual things that match thinking. Anything outside this modality is*

foreign". Moreover, pathologists consider *interactivity* a valuable feature since they desire to interact with the AI interface to manipulate the results.

Müller et al. [214] use visual explanation and feature relevance explanation techniques to explain the predictions of a weak-supervised prognostic model for cancer risk stratification. In particular, the authors use risk saliency maps to highlight different cancer risks and associated areas in the tissue examined. In addition, the authors exploit clustering techniques in order to group patches according to similar scores for each feature. Thus, the authors show that it is possible not only to confirm expected key features but also to learn novel ones.

Tosun et al. [284] propose HistoMaprTM, a proprietary explainability tool for CPATH. HistoMaprTM is designed to support pathologists during ground truth creation and annotation tasks providing them suggestions on potential RoIs to examine. For what concerns explainability, HistoMaprTM integrates a "why?" button in the interface so that users can visualize additional information - i.e., key findings - about the labels assigned to the current image and the associated confidence scores.

Despite the usefulness of the previous XAI tools and approaches, none of them take advantage of VA techniques to explain why a specific prediction has been achieved and which part of the model has contributed to it and in which measure. To mitigate this, we present SKET X, which is a VA tool for explaining the outputs of SKET, a knowledge extractor tool capable of generating weakly supervised labels to train a CNN for image classification in CPATH. To the best of our knowledge, SKET X is the first tool to employ not only IV but also VA techniques for explaining decision-support models for CPATH. Finally, the capabilities of SKET X are integrated with the functionalities provided by the ExaSURE ecosystem.

Chapter 3

ExaSURE

3.1 Introduction

In recent years, huge volumes of healthcare data have been produced. However, the vast majority of them are estimated to be in unstructured formats [216]. For instance, narrative clinical reports and EHRs are usually provided as text-based documents that are human-readable but not machine-readable. Thereby, expensive and time-consuming tasks including data cleaning, pre-processing, and mining are required to extract meaningful information. Thus, AI algorithms may rely on the former error-prone data pre-processing phases, which limit their full potential. In addition, most of the medical data produced to date are buried in detached databases within proprietary software and systems lacking interoperability [183]. The heterogeneity of data coming from multiple sources and the incompatibility of systems limits proper data exchange, analysis, and interpretation. This poses hindrances to effective and efficient secondary use of medical data for several purposes including research, medical communication, and international cooperation [183].

To overcome these limitations, common data exchange standards are an essential prerequisite. In this regard, international organizations such as Health Level Seven International (HL7), Integrating the Healthcare Enterprise (IHE), and openEHR have proposed well-established data standards for medical data, including Fast Healthcare Interoperability Resources (FHIR) and openEHR [223, 33, 182, 148].

Nevertheless, standard data formats are not enough, we need compatible systems providing seamless communication, integration, and data exchange; in short: interoperability. A definition proposed in the literature for interoperability is “*the ability of two or more systems or components to exchange information and to use the information that has been exchanged*” [104]. According to several studies, interoperability is essential to guarantee data access and re-use as well as reduce the slow down in medical processes [183, 237, 180, 230].

Despite the urgent need for healthcare data sharing and interoperability standards, there is still a high heterogeneous landscape highlighting a digital divide and a delay in embracing interoperability solutions [129–131].

To promote interoperability with respect to the tools, systems, and services we have developed for the broad digital pathology domain, we devised ExaSURE. ExaSURE is a holistic ecosystem enabling seamless communication and information access by providing a unified entry-point for the tools developed to support the decision-making process in the digital pathology domain.

The ExaSURE ecosystem is available at <http://w3id.org/exasure>¹.

3.1.1 ExaSURE ecosystem

The ExaSURE ecosystem encompasses all the tools we developed for the broad digital pathology domain, namely, *ExaNet*, *MedTAG*, *NanoWeb*, *SKETUp* and *SKET X*. The typical workflow is depicted in Figure 3.1 which illustrates how digital pathology clinical reports, expressed in natural language, are processed to extract knowledge and generate weak annotations - i.e., labels describing the overall clinical reports such as *Cancer* for indicating the presence of cancer. At the core of the knowledge extraction process, there is SKET which performs the extraction of mentions, concepts, and labels from the textual reports. In particular, the process is articulated in the following steps: (i) data ingestion carried out by SKETUp which processes the clinical reports, invokes SKET, and stores the mentions, concepts, and labels extracted in the common database; (ii) reports annotation and ground-truth creation carried out by human experts using MedTAG; (iii) data stored as RDF triples are enriched with provenance information by leveraging the nanopublication model and NanoWeb; (iv) knowledge exploration of the graph representation of clinical reports using ExaNet and (v) explainability of the outputs generated by SKET using SKET X. The final aim is to generate weak annotations necessary to train image classification algorithms for the digital pathology domain. The grand challenge is to develop a system for the automatic classification of images concerning digital pathology, such as WSIs, pointing out the cancer presence in each image and the eventual dysplasia grade. In this regard, each tool in the ExaSURE ecosystem contributes to supporting the medical decision-making process.

A short description for each tool/service integrated into the ExaSURE ecosystem follows:

- **SKETUp** (available at: <http://w3id.org/sketup>) is a web-oriented tool that allows experts in computational pathology to obtain the machine-readable representation of clinical reports related to the ExaMode use cases (i.e., colon, lung, and uterine cervix

¹Access provided with credentials: demo/demo

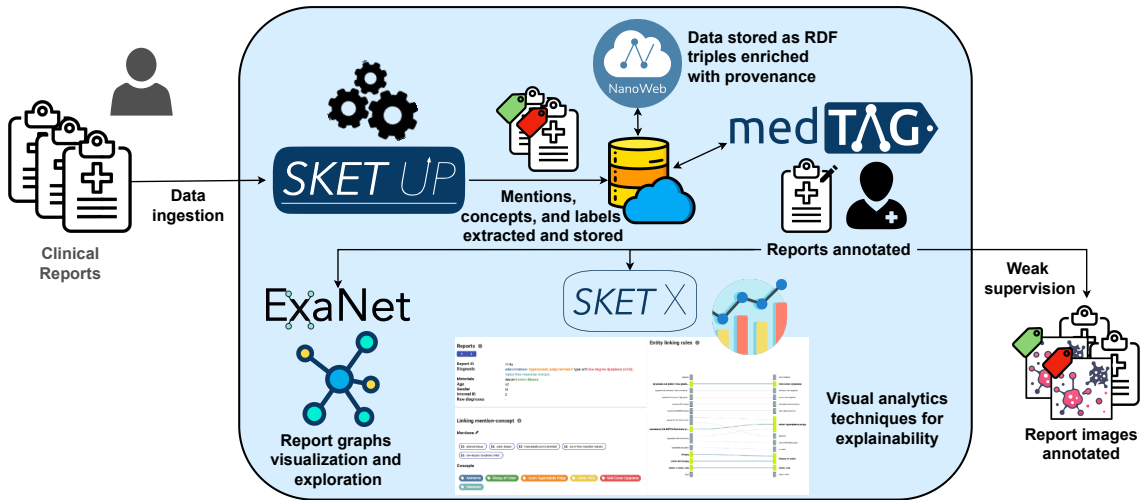


Fig. 3.1 ExaSURE ecosystem encompassing several biomedical tools for the digital pathology domain, namely, *ExaNet*, *MedTAG*, *NanoWeb*, *SKETUp* and *SKET X*. The figure describes the typical workflow in the ExaSURE ecosystem which consists of: (i) data ingestion carried out by SKETUp which processes the clinical reports and stores the mentions, concepts, and labels extracted in the common database; (ii) reports annotation carried out by human experts using MedTAG; (iii) data enrichment with provenance using NanoWeb; (iv) knowledge exploration using ExaNet and (v) explainability using SKET X.

cancer). The clinical reports to upload and process are provided in natural language. SKETUp consists of an online instance of SKET (available at: <http://w3id.org/sketup>); it allows physicians and experts to interact with SKET, execute it multiple times, and download the outputs produced. SKETUp saves the outputs of SKET, and the clinical reports provided, in a database which is in common to all the tools within the ExaSURE ecosystem. Hence, after the outputs are saved, users can annotate the reports using MedTAG. SKETUp allows the users to download the outputs of SKET - i.e., mentions, concepts, and labels extracted together with the machine-readable representation of the clinical reports provided - in JSON format.

- **MedTAG** (available at: <http://w3id.org/medtag>) is a web-based collaborative biomedical annotation tool that has been extensively used in the context of the ExaMode European project to annotate clinical reports. It has been designed to ease the manual annotation process carried out by physicians and experts. MedTAG is a portable and customizable annotation tool designed to allow fast, flexible, and intuitive annotation of biomedical documents. The end goal is to generate the ground-truths/gold standards necessary to evaluate NER+L algorithms. MedTAG relies on the publicly available ExaMode ontology, which is designed to model both cancer diagnoses and the anatom-

ical location that might be affected by the disease. Since the annotation process is an expensive and time-consuming task, that could involve a huge number of documents to annotate, MedTAG integrates SKET for automatic annotation facilities in order to speed-up the overall annotation process. Even if MedTAG has been specifically designed for the broad biomedical domain, it could be used potentially also for the annotation of general purpose documents as long as a proper configuration is provided.

- **ExaNet** (available at: <http://w3id.org/exanet>) allows the users to access and explore the graph representation of the ExaMode clinical reports. It allows physicians and experts to consult, search and visually explore the medical reports and their graph representation using an interactive web interface.
- **SKET X** (available at: <http://w3id.org/sketx>) is a web-oriented tool that allows pathologists and experts to visually comprehend the outputs generated by SKET and the underlying machine decision process [197]. SKET X allows the users to analyze SKET outputs by means of explanation interfaces leveraging VA techniques. The explanation interfaces of SKET X provide a visual interactive analysis of the outputs and the models, parameters, and rules that determine a certain output. Hence, it allows the pathologists to comprehend why a certain output has been achieved and judge whether the decision is correct or not, thus promoting further SKET improvements. SKET X integrates a queue manager to handle multiple concurrent executions of SKET - i.e., pipelines - so that users can execute SKET with different parameters and compare the results obtained.
- **NanoWeb** (available at: <http://w3id.org/nanoweb>) is a tool to search, access and explore life science and biomedical nanopublications [106, 105]. A nanopublication consists of a Resource Description Framework (RDF) graph based on an assertion, that represents a scientific statement extracted, manually or automatically, from a scientific publication. Nanopublications represent scientific facts as machine-readable information tokens, according to the LOD principles. However, nanopublications are not human-readable and there is no way to find relevant nanopublications for a given topic using the natural language. To mitigate this, NanoWeb provides a public web-based search system for life science nanopublications. Using NanoWeb, users can not only search for the relevant nanopublications for a given topic, but also explore their relation network to find new meaningful connections. For instance, this is particularly useful in the case of gene-disease relations, where we are interested to identify also indirect multi-hop relations (e.g., gene A causes disease B; disease C causes disease B; gene D causes disease C; A and D cause the same disease B).

Moreover, NanoWeb allows the users to search for relevant nanopublications not only using queries expressed in natural language, but also using a guided interface that enables users to search for specific entity types (i.e., genes, disease, proteins, tissues), publication authors and provenance. Finally, NanoWeb allows the users to cite a nanopublication, providing also a link to the dedicated landing page.

3.2 ExaSURE interface

Figure 3.2 shows the ExaSURE homepage providing a list of entry points for the tools and services encompassed by ExaSURE. In particular, ExaSURE provides unified access to all the services requiring an authentication mechanism, namely, *ExaNet*, *MedTAG*, *SKETUp* and *SKET X*. Hence, users entering ExaSURE get access to the whole ecosystem, without having to authenticate in each service individually. Using a unique ExaSURE account, users can access all the tools/services and switch among them seamlessly. For instance, users can just click on the buttons (A) and (B) of Figure 3.2 to access respectively the homepage of MedTAG and its dashboard for monitoring the annotations done for each report. Figure 3.3 shows the MedTAG dashboard enabling users to perform several activities including: (A) visualize report information; (B) search for reports; (C) visualize the annotations for each report and the users that have done them; (D) visualize the majority vote ground-truth resulting from all the annotations done by the different users; (E) download the reports' annotation in multiple formats including CSV and JSON; (F) visualize the report's graph representation provided by ExaNet and (G) annotate a report quickly. In addition, users can click on the ExaSURE logo in the top bar (H) of the interface depicted in Figure 3.3 to go back to the ExaSURE homepage and then switch to another service such as SKET X by clicking on the button (E) in Figure 3.2. It is worth mentioning that NanoWeb is not reported in the ExaSURE interface of Figure 3.2 since it is already publicly available and does not require any additional authentication mechanism.

3.2.1 Implementation details

The ExaSURE ecosystem has been implemented as a web-based unified gateway for all the tools/services integrated. The ExaSURE architecture consists of (i) a front-end interface built with React.js², HTML5 and CSS3; (ii) a back-end, implemented using Python and

²<https://reactjs.org/>

Welcome [fabio_giachelle!](#)

Please, choose one of the following services



Fig. 3.2 ExaSURE homepage providing access to all the tools and services requiring authentication, that is, *ExaNet*, *MedTAG*, *SKETUp* and *SKET X*. For each service, a dedicated button provides direct access to the related resource without requiring the users to authenticate again. In particular, the buttons depicted allow the users to: (A) enter *MedTAG* to consult and annotate clinical reports; (B) access the *MedTAG* dashboard to monitor the annotations for each report and visualize its graph representation using the integrated features of *ExaNet*; (C) enter *SKETUp* to run SKET on a set of user-provided reports; (D) access the SKETUp dashboard for monitoring multiple SKET executions and download the resulting outputs; (E) enter *SKET X* explainability interface to comprehend SKET outputs and the rules used in the underlying machine decision process; (F) access the SKET X dashboard reporting all the different executions of SKET and the parameters involved.

Django³, for the gateway functionalities providing a single authentication mechanism for the ExaSURE services; (iii) a common relational database implemented using PostgreSQL.

³<https://www.djangoproject.com/>

The dashboard features a navigation bar with the 'Exa SURE' logo (marked with a red circle 'H') and the 'medTAG' logo. A search bar is present in the top right. The configuration bar includes dropdown menus for 'Colon', 'English', and 'AOEC', along with 'Manual', 'Change', and 'Download' buttons. The main heading is 'REPORTS' OVERVIEW'. Below it, a text block states: 'In this section you can check how many reports have been annotated so far for each use case. You can also delete one or more reports'. A 'Columns' button is located below the text. The table below has columns for 'id_report', 'language', 'usecase', and 'annotations'. The first row of data shows a report with ID 'e3aecee809d8ab4...', language 'English', usecase 'Colon', and 11 annotations (marked with a red circle 'C'). A search bar with a red circle 'B' is at the top right. A row of red circles 'E', 'A', 'D', 'F', and 'G' is overlaid on the table's right side.

| <input type="checkbox"/> | id_report | language | usecase | annotations ↓ |
|--------------------------|--------------------|----------|---------|---------------------|
| <input type="checkbox"/> | e3aecee809d8ab4... | English | Colon | 11 (C) |
| <input type="checkbox"/> | 8d337edee1d0283... | English | Colon | 9 |
| <input type="checkbox"/> | 024904af0078ed8... | English | Colon | 7 |

Fig. 3.3 MedTAG dashboard for managing reports and the annotation process. The dashboard provides several facilities: (A) visualize report information; (B) search for reports; (C) visualize reports' annotations; (D) visualize the majority vote ground-truth; (E) download reports' annotations; (F) visualize the report's graph representation and (G) annotate a report quickly. In addition, users can go back to the ExaSURE interface by clicking on the button (H).

3.3 Conclusions

We propose the ExaSURE ecosystem as a gateway to access all the tools/services, developed for the digital pathology domain, with a unified authentication mechanism. The tools developed - i.e., *ExaNet*, *MedTAG*, *NanoWeb*, *SKETUp*, and *SKET X* - are encompassed by the ExaSURE ecosystem which provides a communication layer that promotes interoperability and efficient switch of tool/service without requiring the users to authenticate again every time the access to a different service is required. ExaSURE has been designed to provide fast, direct, and reliable access to all the tools/services provided to support the decision-making process in the digital pathology domain. ExaSURE enables the full-fledged integration of different tools in a unique workflow designed to process clinical reports provided in natural

language with the aim of automatically extracting meaningful weak annotations describing the overall reports. The mentions, concepts, and labels (i.e., weak annotations) are extracted from the user-provided clinical reports using SKET. Then, the weak annotations are used to train image classification algorithms for the digital pathology domain. The ultimate purpose is to develop a system capable of automatically classifying images from the digital pathology domain, such as WSIs, to determine whether each image presents cancer indications and the eventual grade. This kind of system presents several benefits: (i) reduce the workload of pathologists that could verify the automatic annotations done by the system without starting their analysis from scratch; (ii) speed up the image analysis task since the manual analysis is time-consuming; (iii) reduce missing data and human errors.

The tools/services integrated into the ExaSURE ecosystem contributes to supporting the decision-making process in the digital pathology domain. In this regard, they provide: (i) information access to the clinical reports and the annotation data; (ii) facilities for searching, annotating, and visually exploring the reports; (iii) a user-friendly interface to interact and explain the outputs of SKET and the rules part of the underlying machine decision process.

As future work, we plan to integrate into ExaSURE an active learning system to take advantage of the feedback provided by pathologists and experts to continuously improve the effectiveness of the knowledge extraction process in terms of the quality of the weak annotations produced. In turn, this could improve the effectiveness also of image classification algorithms for the digital pathology domain, since they are trained on the weak annotations.

Chapter 4

Knowledge extraction

4.1 Introduction

Exascale volumes of multimodal data have been produced for decades in the biomedical domain. Biomedical data include patient information, clinical data, biological laboratory data, bio-images, bio-signals, instrumental examinations, and genetic data. Hundred of thousands of reports have been used to communicate diagnoses, encoding vast medical knowledge. Free-text reporting is the standard for communicating the diagnosis, guiding patients' treatment, and other applications, such as cancer registries. Processing high volumes of free-text reports to extract crucial knowledge is usually performed manually. However, this becomes an extremely time-consuming process since reports vary widely between institutions, contain noise, and do not present a standard structure. In this context, NLP methods are central [78, 42, 123, 43, 282, 225, 170, 298] as they empower the efficient automatic processing of thousands of clinical reports and the extraction of key information for several downstream tasks, such as clinical note mining [255, 159] and structuring [109], risk prediction [116], clinical decision support [110], and precision medicine retrieval [248].

In the context of digital pathology, NLP techniques can drive noticeable advances by exploiting the availability of textual pathology reports paired with digital histopathology images (i.e., WSIs) in clinical practice. WSIs are used as a gold standard to diagnose cancer cases and related diseases [63, 4]. Within WSIs, tissue patterns and morphology vary depending on the image magnification level – enabling different tasks such as detection, classification, or segmentation [80]. However, the lack of training datasets containing pixel-wise annotations for entire images [30, 259, 75] limits the effectiveness of supervised ML models [163]. Nevertheless, from the textual pathology reports, it is possible to extract key concepts (e.g., the diagnosis outcome) to annotate the associated WSIs. Although noisy, the extracted concepts can then serve as weak labels to train prediction models for image

classification tasks [46, 52]. However, even though automated solutions involving ML are increasingly being integrated into biomedical domains, NLP applications to digital pathology are less common. Compounding the situation further, the actual use of AI algorithms in digital pathology requires a large amount of data annotations by pathologists. However, they are rarely available in a clinical setting [78, 82].

To overcome such limitation, this work aims at proving the viability of unsupervised NLP techniques to automatically extract critical information from pathology reports and use it for different DPATH applications, such as automatic report annotation, pathological knowledge visualization, and WSI classification. In this regard, we present the SKET, an unsupervised hybrid knowledge extraction system that combines an expert system with pre-trained ML models to extract knowledge from pathology reports. Specifically, SKET is designed to extract key information such as *labels* (i.e., weak annotations), *mentions*, and *concepts* from free-text pathology reports provided by the user. The full list of the labels defined for the knowledge extraction process is reported in Section 4.3.3, whereas the concepts to be extracted are defined in the ExaMode ontology as described in Section 4.2. Figure 4.1 shows qualitatively the knowledge extraction process performed by SKET. For instance, we can observe, that SKET generates the label *Adenomatous polyp - low grade dysplasia* since it identifies the concepts *Moderate Colon Dysplasia* and *Mild Colon Dysplasia* from the mention *low-grade dysplasia (mild to moderate)* in the report text.

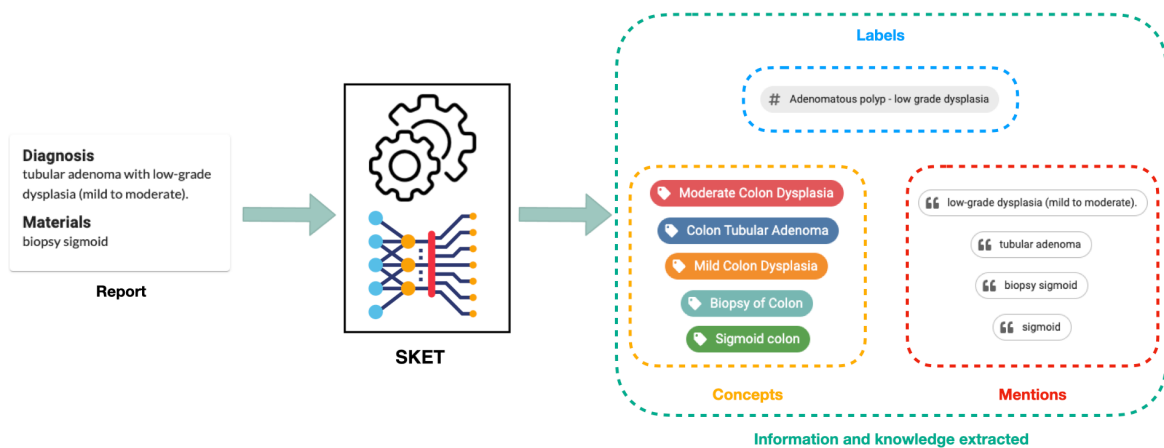


Fig. 4.1 A qualitative example of the knowledge extraction process carried out by SKET. From a free-text report, SKET extracts key information, namely, *labels*, *mentions*, and *concepts*. For instance, we can observe that the label *Adenomatous polyp - low grade dysplasia* has been generated since the related concepts *Moderate Colon Dysplasia* and *Mild Colon Dysplasia* have been identified in the given report starting from the mention *low-grade dysplasia (mild to moderate)*. Another example is the concept *Biopsy of Colon* which has been identified from the associated mention *biopsy sigmoid*.

In recent years, NLP has shifted from using rules to ML approaches [66, 298], which have the advantage of learning regularities from data and generalizing to previously unseen patterns. Moreover, the advent of efficient Neural Language Models (NLMs) [206, 37, 231, 81] paved the way for the pre-training era, where large NLMs trained in a self-supervised fashion on huge datasets are used to develop unsupervised or weakly supervised NLP models for a number of downstream tasks. Nevertheless, similarly to Santus et al. [255], we argue that rule-based techniques capture critical information that should be used together with – and not substituted by – ML to improve performance.

We evaluate SKET effectiveness on entity linking and text classification, considering three different diseases: colon, cervix, and lung cancer. In this regard, we resort on diagnostic reports coming from two medical centers in Italy and The Netherlands. Then, we compare SKET with unsupervised ML approaches to understand the impact that combining rule-based techniques and pre-trained ML models have on the extraction of knowledge from pathology reports. The achieved results highlight the viability of ML methods for information extraction in the pathology domain, but also stress the importance of expert knowledge to reach the high levels of accuracy required to (semi-)automate the clinical practice. Moreover, the applicability of the proposed approach is enhanced by the considered multilingual setting. Besides effectiveness, we must consider that understanding and explaining decisions and outcomes is crucial in clinical practice. However, the black-box nature of many ML models, especially those based on DL, makes it difficult to understand and trace back the underlying decision process. Hence, there is an urgent need for a shift towards XAI [135, 132]. In the biomedical domain, clinicians and domain experts need to understand *why* a specific output has been produced to trust the system and its predictions. To this end, in Chapter 5 we propose SKET X as a tool for enabling pathologists to visually explain the outcomes of SKET. In addition in the present chapter, we also report different digital pathology applications where SKET has been successfully integrated as a core system [107, 199]. In particular, we deepen the use of SKET in such applications and the advantages it entails.

SKET source code is publicly available at <https://github.com/ExaNLP/sket>. Besides, SKET can also be deployed as a Docker container. For information about the Docker version of SKET, please refer to <https://github.com/ExaNLP/sket#docker>.

The rest of the chapter is organized as follows: Section 4.2 describes the considered data resources. Section 4.3 presents SKET. Section 4.4 describes the experimental setup and reports quantitative and qualitative results. Section 4.5 outlines the digital pathology downstream applications empowered by SKET. Finally, Section 4.6 draws some conclusions.

Table 4.1 Data size. For each medical center, we report the number of diagnostic reports associated with each use-case. The “–” symbol represents the lack of reports for a given use-case.

| | Colon | Cervix | Lung |
|------|-------|--------|-------|
| AOEC | 1,704 | 1,777 | 1,902 |
| RUMC | 2,065 | 2,350 | – |

4.2 Material

The data used to develop and evaluate SKET comes from two different medical centers: the Cannizzaro Hospital (AOEC), Catania, Italy and the Radboud University Medical Center (RUMC), Nijmegen, The Netherlands. The AOEC data includes diagnostic reports for colon, cervix, and lung cancer cases, written in Italian and associated with WSIs. All data were collected in the clinical workflow and fully anonymized afterwards. Similarly, the RUMC data consists of diagnostic reports and the associated WSIs for colon and cervix cases, written in Dutch – after the use of speech-to-text tools – and anonymized. For both medical centers, the considered reports span several diagnostic outcomes. Table 4.1 reports the total number of diagnostic reports for each considered use-case and medical center.

Diagnostic reports contain the results of the analyses performed on specific tissues (or cells) to obtain a pathological-clinical diagnosis – i.e. presence or absence of the disease. AOEC and RUMC diagnostic reports follow the College of American Pathologists (CAP) international guidelines¹ for pathology reports [273, 95] and contain the patient’s personal and clinical-specific information, the description of how a specimen appears to the naked eye and at the microscope, and provide the final diagnosis.

As mentioned above, AOEC and RUMC diagnostic reports are written in Italian and Dutch, respectively. However, most of the resources required to develop NLP methods that extract concepts from unstructured text are in English. To overcome this limitation, we first translated diagnostic reports in English and then performed data curation over them. We used the open-source, pre-trained Marian Neural Machine Translation (NMT) models [145], which exhibit a Transformer-based [293] encoder-decoder architecture with six layers in each component. Given the complexity of the task, such an automatic approach introduces systematic translation errors that, if propagated, could hamper the effectiveness of the extraction process. For this reason, we performed a data curation step, in which recurring, manually identified translation errors were corrected through the use of handcrafted rules.

¹<https://www.cap.org/protocols-and-guidelines>

We defined an ontology² for modeling the clinical reports in the digital pathology domain: ExaMode³ ontology. Amongst other aspects not relevant for the current work, the ontology specifically defines the key concepts and properties to model: (i) the diagnosis of colon, cervix, and lung cancer; (ii) the anatomical location where the disease might be located; (iii) the procedure employed to get the tissue and (iv) the tests conducted on the tissue. Despite many medical ontologies focusing specifically on cancer exist, no single ontology comprehensively models all the diseases related to the cases mentioned above, their anatomical location, topography, and pathology laboratory process.

4.3 Methods

SKET adopts a combination of pre-trained Named Entity Recognition (NER) models and unsupervised Entity Linking (EL) methods to extract key concepts (entities) from the diagnostic reports and link them to the reference ontology. The use of pre-trained NER models and unsupervised EL methods makes SKET suitable for weak supervision tasks. In this regard, the pathological concepts extracted from diagnostic reports can serve as weak labels to train prediction models for image classification tasks [46, 52], or as nodes to build report-level knowledge graphs for information retrieval tasks [196].

As reported in Figure 4.2, SKET consists of four components: (A) Named Entity Recognition, (B) Entity Linking, (C) Data Labeling, and (D) Graph Creation. Components (A) and (B) are sequential, whereas components (C) and (D) are parallel. Below, we describe for each component the different methods and techniques we adopted, expanded, or developed.

4.3.1 Named Entity Recognition

NER is the task of identifying and categorizing key information – i.e., entities – within text. An entity can be any word or phrase that consistently refers to the same concept or object of the world. Each identified entity is classified into a pre-defined category, such as disease, protein, gene, cell type, etc.

SKET relies on a combination of pre-trained neural models and rule-based techniques to perform NER. At its core, SKET adopts ScispaCy models [217], which provide full NER pipelines for biomedical data, comprising large medical vocabularies, and Word2Vec [206]

²<https://w3id.org/examode/ontology/>

³ExaMode stands for “Extreme-scale Analytics via Multimodal Ontology Discovery & Enhancement” and is an H2020 project financed by the EU commission. More information can be found at: <http://www.examode.eu/>

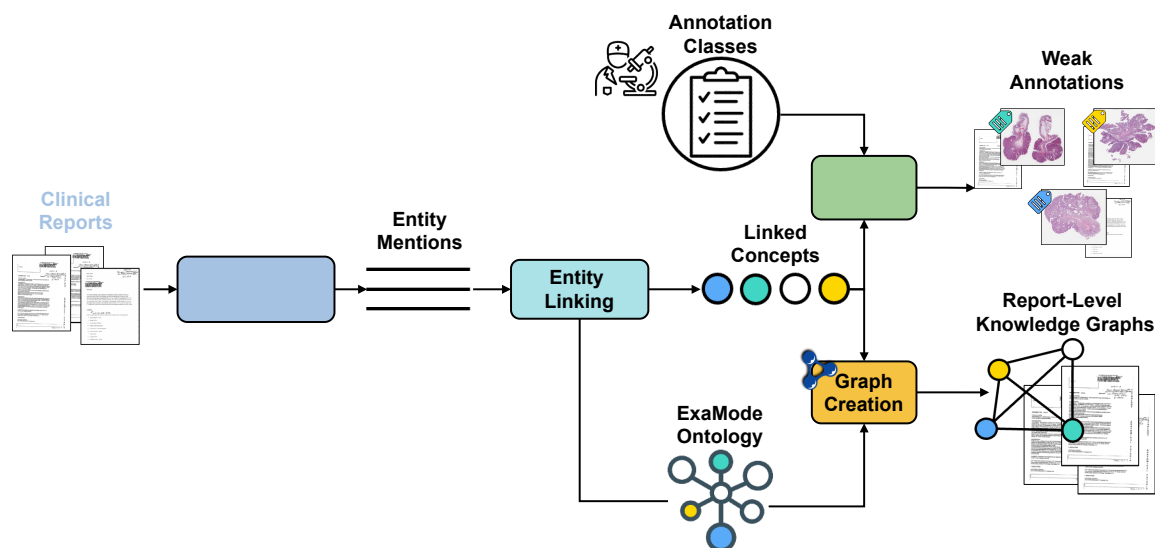


Fig. 4.2 SKET architecture. SKET main components are: (A) Named Entity Recognition, (B) Entity Linking, (C) Data Labeling, and (D) Graph Creation.

word vectors trained on the PubMed Central Open Access Subset [238]. It is worth mentioning that SKET has been designed to be deployed with any of the core models available at: <https://allenai.github.io/scispacy/>.

Then, SKET extends the ScispaCy pipeline with two additional components: Entity Fusion and Negation Detection.

Entity Fusion: SKET extends the NER pipeline with a set of rules used to identify and merge specific entities otherwise regarded as separate by ScispaCy. For instance, “transverse” and “colon” are considered as separate entities, whereas we are interested in “transverse colon” as a unique entity. Hence, we developed regular expressions that identify trigger terms indicative of a set of otherwise potentially separate entities. Once a trigger term is identified, SKET matches the entities extracted by ScispaCy with the candidate terms associated with the trigger. Depending on the trigger term, the match that SKET performs between extracted entities and candidate terms follows different rules based on directional and positional attributes. Directional attributes specify the set of extracted entities to be matched with the candidate trigger terms, and it can assume three values:

- PRE: match with the entities preceding the trigger entity.
- POST: match with the entities succeeding the trigger entity.
- BOTH: match between the entities both preceding and succeeding the trigger entity.

Positional attributes specify the maximum distance allowed between the trigger entity and the other one, and it can assume two values:

- **EXACT**: the matched entity must be right before/after the trigger entity.
- **LOOSE**: the matched entity can be anywhere before/after the trigger entity.

The described set of rules has been developed on a holdout dataset that we do not include in this work. The dataset consists of 50 diagnostic reports for each use-case and medical center, for a total of 250 diagnostic reports. The complete set of rules is available on the SKET GitHub repository⁴.

Negation Detection: To handle negated entities, we extend the NER pipeline with NegEx [61], a negation detection algorithm evaluating whether extracted entities are negated within text. NegEx uses regular expressions to identify the scope of trigger terms that are indicative of negation, such as “no” or “ruled out”. Then, the entities extracted within the scope of a trigger term are marked as negated. In this way, SKET identifies – and removes from the final list of extracted entities – those entities that NegEx regards as negated. For example, if we consider the phrase “free of dysplasia”, NegEx identifies the trigger term “free of” and marks “dysplasia” as negated, which is then removed by SKET.

4.3.2 Entity Linking

EL is the task of assigning unique meanings to entities mentioned within text. In other words, the objective of EL is to determine whether a given entity refers to a specific concept or object within a reference ontology.

SKET employs a combination of ad-hoc and similarity matching techniques to link the extracted entities to unique concepts within the ExaMode ontology. Given an extracted entity, SKET first tries to match it using ad-hoc matching and when it fails SKET employs similarity matching.

Ad-Hoc Matching: SKET uses regular expressions to identify trigger terms indicative of a specific ontological concept. Once a trigger term is identified, SKET matches the entity containing the trigger term with the closest ontology concept. For instance, if an extracted entity contains the term “carcinoma”, then SKET links the entity to the ontology concept “colon adenocarcinoma”. As for Entity Fusion, the ad-hoc matching rules have been developed on the holdout dataset and are available on GitHub.

⁴<https://github.com/ExaNLP/sket/tree/main/sket/nerd/rules/>

Similarity Matching: SKET performs similarity matching using a combination of string and semantic matching techniques. For string matching, SKET relies on the GPM algorithm [243], which computes the similarity of two strings as the number of matching characters divided by the total number of characters in the two strings. Matching characters are those in the longest common subsequence plus, recursively, matching characters in the unmatched region on either side of the longest common subsequence. For semantic matching, SKET exploits the word vectors provided by ScispaCy models [217]. In other words, SKET performs semantic matching as the cosine distance between the vector representations of the extracted entities and the ontology concepts – where vector representations are the mean of the word vectors composing the extracted entities or the ontology concepts.

Both string and semantic matching produce a ranking of ontology concepts ordered by decreasing similarity with a given target entity. To combine the two rankings – and select the concept with the highest rank – SKET performs rank fusion using the CombSUM [264] with min-max normalization. Before selection, a pruning phase is performed on the combined ranking, in which ontology concepts with a similarity score lower than a predetermined threshold are removed. The threshold value has been set empirically to 1.8 using the holdout dataset. The pruning phase aims to increase precision by reducing false positives, which occur when ontology concepts are incorrectly linked to the extracted entities.

4.3.3 Data Labeling

SKET also provides labels as one of its main outputs. Given the set of concepts extracted from each diagnostic report, SKET maps a clinically relevant subset of such concepts to a set of annotation classes defined by AOEC pathologists. For each use-case, we report below the set of annotation classes.

Colon Annotations:

1. Cancer
2. Adenomatous polyp - high grade dysplasia
3. Adenomatous polyp - low grade dysplasia
4. Hyperplastic polyp
5. Non-informative

Cervix Annotations:

1. Cancer - adenocarcinoma in situ
2. Cancer - adenocarcinoma invasive
3. Cancer - squamous cell carcinoma in situ
4. Cancer - squamous cell carcinoma invasive
5. High grade dysplasia
6. Low grade dysplasia
7. HPV infection present
8. Koilocytes
9. Normal squamous
10. Normal glands

Lung Annotations:

1. Cancer - non-small cell cancer, adenocarcinoma
2. Cancer - non-small cell cancer, large cell carcinoma
3. Cancer - non-small cell cancer, squamous cell carcinoma
4. Cancer - small cell cancer
5. No cancer

Thus, the Data Labeling component produces annotations from diagnostic reports that can be used to perform weakly supervised classification tasks.

4.3.4 Graph Creation

SKET also builds report-level knowledge graphs using the extracted concepts as nodes and the semantic relations of the ExaMode ontology as edges. The use of ontology concepts and relations to describe diagnostic reports increases the semantic understanding of the underlying data [6]. Once created, report-level knowledge graphs are encoded in a machine-readable format through RDF.

4.4 Evaluation

4.4.1 Tasks

We evaluate the effectiveness of SKET on two different tasks: entity linking and text classification. The evaluation of SKET on entity linking also serves as a proxy to validate the quality of the RDF graphs it produces. On the other hand, text classification results help understanding the viability of using SKET as an automatic annotator in weak supervision tasks. Between the two tasks, text classification has a prominent role as it provides weak annotations that can be used to reduce the high costs of training cancer assisted diagnosis tools – which prevent unleashing the full potential of digital pathology applications [199].

4.4.2 Datasets

Entity Linking: We evaluate SKET effectiveness to extract concepts from pathology reports on a subset of the proprietary data described in Section 4.2. For each use-case and medical center, 250 reports have been manually annotated by experts using the concepts from the ExaMode ontology. Overall, the total number of annotated reports amounts to 1,250. In terms of annotations, all use-cases have been annotated with a large number of different concepts. For Colon cancer, the number of different concepts that can be found within reports stands at 19, while for Cervix and Lung cancer amounts to 21 and 11, respectively. This large number of different concepts highlights the complexity of the task, both for model predictions and human annotation efforts. In particular, the task can be seen as an extreme multi-label classification problem [59, 251], where the goal is to tag a given report with a subset of the relevant concepts from a large concept list.

Text Classification: To evaluate the effectiveness of SKET to weakly annotate pathology reports, the proprietary data described in Section 4.2 has been manually labeled by experts using the annotation classes defined by AOEC pathologists. For each use-case, AOEC and RUMC reports have been annotated with one or more classes, making the task a multi-label classification problem. Table 4.2 reports the total number of reports annotated for each class in each use-case. Given the multi-label nature of the task, the total number of annotations does not reflect the total number of reports. As a side note, the class imbalance of the datasets reflects a real-case scenario, where certain conditions – e.g., low-grade dysplasia in Colon cases – occur more often than others in the clinical routine.

Table 4.2 Number of annotated diagnostic reports for each use-case. Label counts are independent of each other except for “Non-informative” in Colon, “Normal squamous” and “Normal glands” in Cervix, and “No cancer” in Lung, which only occur when none of the others does.

| Colon | |
|---|-------|
| Cancer | 495 |
| Adenomatous polyp - high grade dysplasia | 510 |
| Adenomatous polyp - low grade dysplasia | 841 |
| Hyperplastic polyp | 508 |
| Non-informative | 1,140 |
| Cervix | |
| Cancer - adenocarcinoma in situ | 125 |
| Cancer - adenocarcinoma invasive | 32 |
| Cancer - squamous cell carcinoma in situ | 638 |
| Cancer - squamous cell carcinoma invasive | 88 |
| High grade dysplasia | 1,544 |
| Low grade dysplasia | 1,053 |
| HPV infection present | 1,221 |
| Koilocytes | 86 |
| Normal squamous | 1,265 |
| Normal glands | 1,266 |
| Lung | |
| Cancer - non-small cell cancer, adenocarcinoma | 961 |
| Cancer - non-small cell cancer, large cell carcinoma | 68 |
| Cancer - non-small cell cancer, squamous cell carcinoma | 528 |
| Cancer - small cell cancer | 144 |
| No cancer | 247 |

4.4.3 Baselines

Entity Linking: We compare SKET with two unsupervised approaches based on BioFastText [37, 314] and BioClinical BERT [81, 8] models. For a fair comparison, both approaches adopt the same NER ScispaCy pipeline used by SKET, but without the extensions introduced with it. Then, the approaches perform EL by computing the cosine distance between the vector representations of the extracted entities and the ontology concepts – obtained with FastText in one case and with BERT in the other. The ontology concept closest to the extracted entity is kept and, when appropriate, mapped to the corresponding annotation

class. Both methods represent a straightforward approach to perform text classification with lack of annotated data.

Text Classification: We compare SKET with the Bio FastText and BioClinical BERT unsupervised approaches described above. Beyond unsupervised approaches, we also use SKET to weakly annotate diagnostic reports and then train FastText and BERT models in a supervised fashion. In this case, we stack a classification layer on top of the pre-trained models and perform end-to-end classification – that is, the models take diagnostic reports as input and directly produce classes as output. Due to the introduction of supervised models, performances on text classification are obtained through 10-fold cross-validation.

4.4.4 Results

Entity Linking: Table 4.3 reports the results obtained by SKET and the considered baselines on entity linking. Overall, we see that SKET achieves high performances for both micro- and weighted-average F1 measures in each use-case. As for accuracy, the performances vary depending on the use-case, and the lowest score is obtained in Colon cancer with a value of 0.6280. In terms of use-cases, the best SKET results are obtained on Lung cancer. Compared to Colon and Cervix cases, Lung cancer presents a lower number of concepts to identify, thus reducing the task complexity. On the other hand, Colon and Cervix use-cases show similar SKET performances, having a comparable number of concepts.

When we compare SKET performances with unsupervised approaches we can see that SKET outperforms them for all measures in each use-case. This result shows the effectiveness of combining ad-hoc rules with ML models, which make SKET both precise and sensitive. Indeed, ad hoc matching makes SKET precise while semantic matching makes it sensitive. To further support this outcome, we observe that the performances of unsupervised baselines – only relying on ML models and semantic matching – have low accuracy values. Given that we consider entity linking as a multi-label task, we resort on subset accuracy – where the set of concepts predicted for a report must exactly match the corresponding set of ground-truth concepts. Thus, accuracy values are more prone to rapidly decreasing with a large number of classes, and less precise models are naturally affected by this behavior.

Text Classification: Table 4.4 reports the results obtained by SKET and the considered baselines on text classification. Overall, we observe that SKET achieves high performance on Colon and Lung cancer use-cases, whereas it shows low accuracy values on Cervix cancer. The motivation behind this drop in performance on Cervix reports can be attributed to the high number of annotation classes (i.e., ten) and the multi-label setting. We recall that we

Table 4.3 Entity linking results on colon, cervix, and lung cancer pathology reports. The considered measures are subset accuracy, micro F1, and weighted F1. **Bold** values represent the highest scores achieved for each measure.

| Colon | | | | |
|--------------|----------|---------------|---------------|---------------|
| Approach | Model | Measures | | |
| | | Accuracy | Micro F1 | Weighted F1 |
| Unsupervised | SKET | 0.6280 | 0.8861 | 0.8694 |
| | FastText | 0.0660 | 0.5000 | 0.6146 |
| | BERT | 0.1840 | 0.3905 | 0.4527 |
| Cervix | | | | |
| Approach | Model | Measures | | |
| | | Accuracy | Micro F1 | Weighted F1 |
| Unsupervised | SKET | 0.7020 | 0.8322 | 0.8368 |
| | FastText | 0.0900 | 0.2802 | 0.3439 |
| | BERT | 0.0720 | 0.2715 | 0.2940 |
| Lung | | | | |
| Approach | Model | Measures | | |
| | | Accuracy | Micro F1 | Weighted F1 |
| Unsupervised | SKET | 0.8624 | 0.9375 | 0.9262 |
| | FastText | 0.2510 | 0.5610 | 0.6506 |
| | BERT | 0.3806 | 0.6804 | 0.8395 |

rely on subset accuracy, which performs exact match between predicted and ground-truth labels – causing performance to drop faster when the number of classes is larger. The higher values for both micro and weighted F1 measures, which do not perform exact match between predicted and ground-truth labels, further support this intuition.

Compared to unsupervised baselines, SKET achieves better performance in both Colon and Cervix use-cases. In particular, the (relative) performance gap between SKET and baselines varies from 20% to 40% across measures. To confirm SKET effectiveness, we conducted a paired t-test and found that there is a statistical difference (p -value < 0.01) between its performance and that of the baselines on all the considered measures. This outcome shows the effectiveness of introducing ad-hoc rules at both NER and EL levels, as well as the soundness of combining different matching techniques together. On the other hand, the unsupervised BERT-based approach outperforms both SKET and FastText in

Lung cancer. In this case, the paired t-test confirmed a statistical difference between BERT performance and that of SKET and FastText. Nevertheless, the performance gap between BERT and SKET never exceeds 5%. This highlights the robustness of SKET across different use-cases and makes it a viable solution in real scenarios, where annotated data are hard and expensive to get (such as in clinical practice). Besides, the Lung cancer use-case presents two major differences with Colon and Cervix ones. First of all, Lung annotation classes all revolve around different, but closely related, cancer types. As a consequence, contextualized NLMs (e.g., BERT [81]) – which are able to properly model the small semantic, contextual variations of such classes – achieve competitive results. Secondly, Lung cancer data only consists of AOEC reports. The lack of RUMC reports makes the dataset more homogeneous and easier than the others, thus reducing classification inconsistencies for baseline models too.

Regarding weakly supervised models, the results reported in Table 4.4 demonstrate the effectiveness of using SKET to weakly annotate diagnostic reports and then train FastText and BERT models in a supervised fashion. In this regard, both weakly supervised FastText- and BERT-based approaches outperform their unsupervised counterparts. The only exception is for BERT on Lung cancer data, where the unsupervised BERT approach achieves top performance. On the other hand, the weakly supervised BERT obtains the best results overall in both Colon and Cervix use-cases. Hence, SKET proves to be effective when used to bootstrap supervised models in absence of manual annotations. Following this procedure, supervised models can first be trained on data automatically annotated by SKET and then fine-tuned on small manually annotated batches, thus reducing annotation times and costs.

Table 4.4 Text classification results on colon, cervix, and lung cancer pathology reports. The considered measures are subset accuracy, micro F1, and weighted F1. The † symbol represents the statistical difference of SKET from unsupervised FastText- and BERT-based approaches – verified using a paired t-test with a p-value < 0.01. **Bold** values represent the highest scores achieved for each measure.

| Colon | | | | |
|-------------------|----------|---------------------------|---------------------------|---------------------------|
| Approach | Model | Measures | | |
| | | Accuracy | Micro F1 | Weighted F1 |
| Unsupervised | SKET | 0.7525 [†] | 0.8386 [†] | 0.8373 [†] |
| | FastText | 0.4146 | 0.5298 | 0.5514 |
| | BERT | 0.5167 | 0.5697 | 0.6587 |
| Weakly Supervised | FastText | 0.7116 | 0.8287 | 0.8276 |
| | BERT | 0.7586 | 0.8432 | 0.8421 |
| Cervix | | | | |
| Approach | Model | Measures | | |
| | | Accuracy | Micro F1 | Weighted F1 |
| Unsupervised | SKET | 0.5281 [†] | 0.7791 [†] | 0.7611 [†] |
| | FastText | 0.2533 | 0.4882 | 0.4445 |
| | BERT | 0.3066 | 0.3962 | 0.4867 |
| Weakly Supervised | FastText | 0.4744 | 0.7542 | 0.7566 |
| | BERT | 0.5397 | 0.7901 | 0.7737 |
| Lung | | | | |
| Approach | Model | Measures | | |
| | | Accuracy | Micro F1 | Weighted F1 |
| Unsupervised | SKET | 0.8137 | 0.8387 | 0.8262 |
| | FastText | 0.5221 | 0.7296 | 0.6853 |
| | BERT | 0.8523[†] | 0.8630[†] | 0.8526[†] |
| Weakly Supervised | FastText | 0.7701 | 0.8313 | 0.8247 |
| | BERT | 0.8127 | 0.8375 | 0.8249 |

4.5 SKET and the ExaSURE ecosystem

SKET has been integrated as a core system into different downstream applications for digital pathology. Figure 4.3 depicts the SKET ecosystem, where SKET UP represents the online access point to interact with SKET, SKET X provides explanations for SKET results,

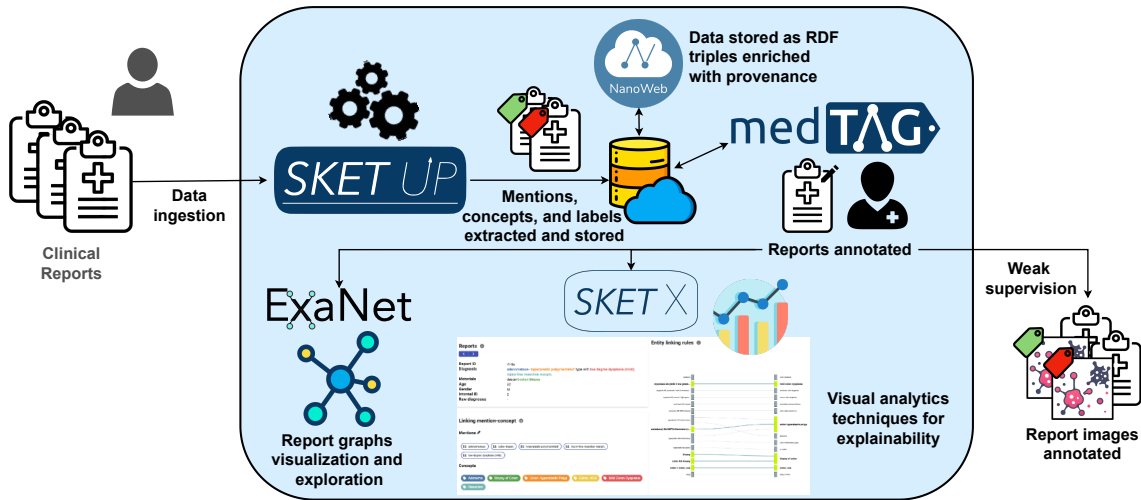


Fig. 4.3 SKET is the core of the ExaSURE ecosystem. From clinical reports, a suite of different applications relying on SKET process, analyze, explore, and explain the knowledge contained within reports – also providing weak supervision to train cancer assisted diagnosis tools.

medTAG [107] integrates SKET automatic annotations to support semi-automatic tagging, and ExaNet [107] allows to visualize and explore SKET report-level knowledge graphs. Moreover, SKET labels can also be used to supervise cancer-assisted diagnosis tools [199].

4.5.1 SKETUp a web interface for SKET

As described in section 4.2, users can execute SKET using either the source code or the Docker container both publicly available on GitHub. In addition, we developed SKETUp, which is a web-based tool providing an interface for interacting with SKET. SKETUp enables pathologists, experts, and not technologically-savvy users to interact with SKET. SKETUp can be used to obtain both SKET outputs and the machine-readable representation of the clinical reports related to the ExaMode use cases (i.e., colon, lung, and uterine cervix cancer). SKETUp consists of an online instance of SKET available at <http://w3id.org/sketup>⁵; it allows physicians and experts to interact with SKET, execute it multiple times, and download the outputs produced. SKETUp saves the outputs of SKET, and the clinical reports provided, in a database which is in common with all the tools within the ExaSURE ecosystem. Hence, after the outputs are saved, users can annotate the reports using MedTAG. SKETUp allows the users to download the outputs of SKET - i.e., mentions, concepts, and labels extracted together with the machine-readable representation of the clinical reports provided - in JSON format. SKETUp allows the users to upload and process clinical reports provided in natural

⁵Access provided with credentials: demo/demo

language. In particular, SKETUp provides two different interfaces for uploading the reports, one for single-report processing and the other for processing a batch of reports. Figure 4.4 shows the interface for single-report processing (left side). It allows the users to specify the report language, use case, institute, and diagnosis text. In addition, the concepts identified by SKET regarding the user-provided report are visible on the right side. Similarly, the batch interface allows the users to specify the same information plus the choice of whether to save the reports or not in the database. It is worth mentioning that users can save in the database only batches of reports, thus single reports are processed on-demand but not saved. Moreover, users can take advantage of drag-and-drop facilities to upload batches of reports in JSON format. As described in section 4.3.2, SKET uses different models and parameters for the NER phase. In this regard, SKETUp allows the users to specify the execution parameters of SKET in a dedicated interface, as shown in Figure 4.5. Each execution of SKET is an asynchronous task executed in the background so that users can execute it multiple times. To this aim, SKETUp integrates a queue manager to schedule multiple executions of SKET as they are requested by the users. Users can monitor their own tasks in the dedicated dashboard, as shown in Figure 4.6. The contents of the dashboard are dynamic and continuously updated with information about the user's tasks, such as the task identifier and status. From the dashboard, users can download the outputs of SKET and annotate a batch of reports saved in the database using MedTAG.

Implementation details

SKETUp has been developed and implemented using the following technologies:

- the front-end interface is built with React.js⁶, HTML5 and CSS3.
- the back-end for web API and services is built with the Python web framework Django⁷. The back-end integrates Celery⁸ as a queue manager for multiple executions of SKET.
- the relational database is implemented using PostgreSQL.

4.5.2 Automatic Report Annotation

SKET has been integrated as an automatic annotator within MedTAG⁹ [107]. MedTAG is a collaborative biomedical annotation tool that provides four annotation types:

⁶<https://reactjs.org/>

⁷<https://www.djangoproject.com/>

⁸<https://docs.celeryq.dev/en/stable>

⁹MedTAG is available at <https://github.com/MedTAG/medtag-core/>

Report information
Please, provide the following parameters:

Report ID: 2cfc22f2-7af0-4a52-90b7-e4bbdb722e05

Language: English

Use Case: Colon

Institute: AOEC

Diagnosis: adenocarcinoma with mild, focally severe dysplasia. Results were obtained with a colon biopsy.

Save

```

1 {
2   "test": {
3     "Anatomical Location": [
4       [
5         "http://purl.obolibrary.org/obo/UBERON_0001155",
6         "Colon, NOS"
7       ]
8     ],
9     "Diagnosis": [
10      [
11        "http://purl.obolibrary.org/obo/MONDO_0002271",
12        "Colon Adenocarcinoma"
13      ],
14      [
15        "http://purl.obolibrary.org/obo/NCIT_C4848",
16        "Mild Colon Dysplasia"
17      ],
18      [
19        "SevereColonDysplasia",
20        "Severe Colon Dysplasia"
21      ]
22    ],
23    "Procedure": [
24      [
25        "http://purl.obolibrary.org/obo/NCIT_C51678",
26        "Biopsy of Colon"
27      ]
28    ],
29    "Test": []
30  ]
31 }

```

Fig. 4.4 SKETUp upload interface for a single report (left side) and the serialization in JSON format of the SKET output for the concepts identified in the knowledge extraction process (right side). We can observe that the concepts identified in the diagnosis field are *Colon Adenocarcinoma*, *Mild Colon Dysplasia*, and *Severe Colon Dysplasia*. For what concerns instead the procedure field, the *Biopsy of Colon* concept has been identified.

- **Labels:** allows the user to assign, by clicking on the check-boxes, one or more labels to a document. The labels indicate some reports' properties (e.g. "Cancer" label indicates the presence of a cancer-related disease).
- **Concepts:** allows the user to specify which concepts are relevant for a document. Users can take advantage of auto-complete functionalities for searching the relevant concepts to assign to each document.
- **Mentions:** shows the list of the mentions identified by the user in the report text.
- **Linking:** allows the user to link the mentions identified with the corresponding concepts. Users can link the same mention to multiple concepts.

For each annotation type, SKET provides automatic annotations for reports associated with Colon, Cervix, and Lung use-cases. At present, MedTAG has been used by experts to produce more than 7,000 annotations. On the other hand, SKET annotations within MedTAG exceed 100,000 units. Table 4.5 reports SKET annotation statistics for each annotation type.

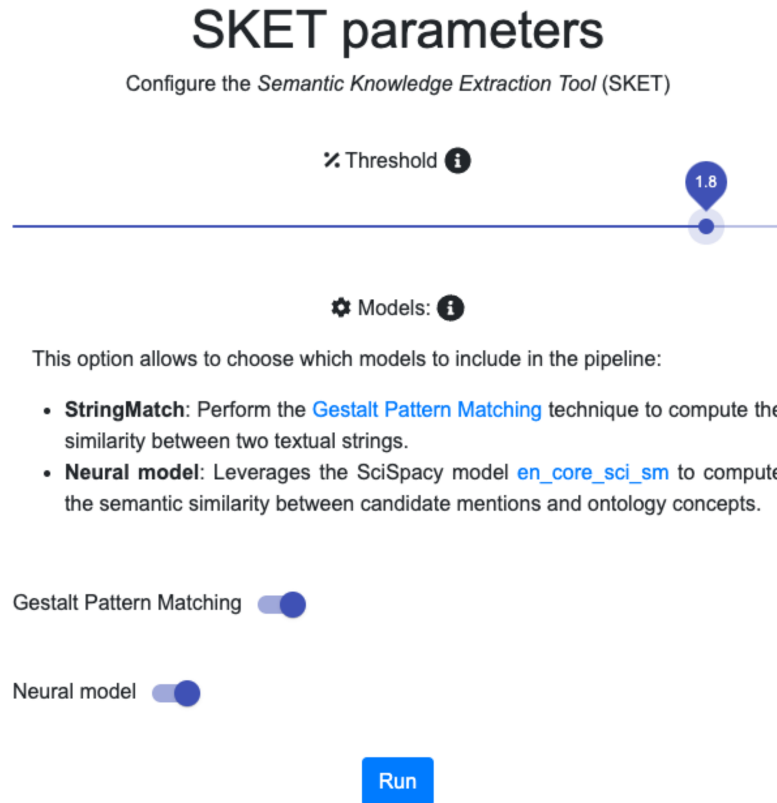



Fig. 4.5 SKETUp configuration interface for the SKET parameters. We can observe that the threshold used for the pruning phase of the knowledge extraction process is set to 1.8 (the default value). Moreover, the models selected for the EL phase are the GPM and the neural model, as described in section 4.3.2.

Table 4.5 Number of labels, concepts, mentions, and links automatically annotated by SKET within MedTAG. Statistics are reported for each use-case and globally.

| Annotation Type | Colon | Cervix | Lung | Total |
|-----------------|---------------|---------------|--------------|----------------|
| Labels | 9,309 | 16,033 | 2,066 | 27,408 |
| Concepts | 11,932 | 12,936 | 2,336 | 27,204 |
| Mentions | 10,926 | 12,070 | 2,336 | 25,332 |
| Linking | 11,932 | 12,936 | 2,336 | 27,204 |
| Total | 44,099 | 53,975 | 9,074 | 107,148 |

4.5.3 Pathological Knowledge Visualization

The report-level knowledge graphs produced by SKET can be explored with ExaNet. ExaNet is available at <http://w3id.org/exanet> or it can be accessed through the “Reports’ stats” functionality of MedTAG, under the “Graph” feature associated with each report that has been

 Your tasks






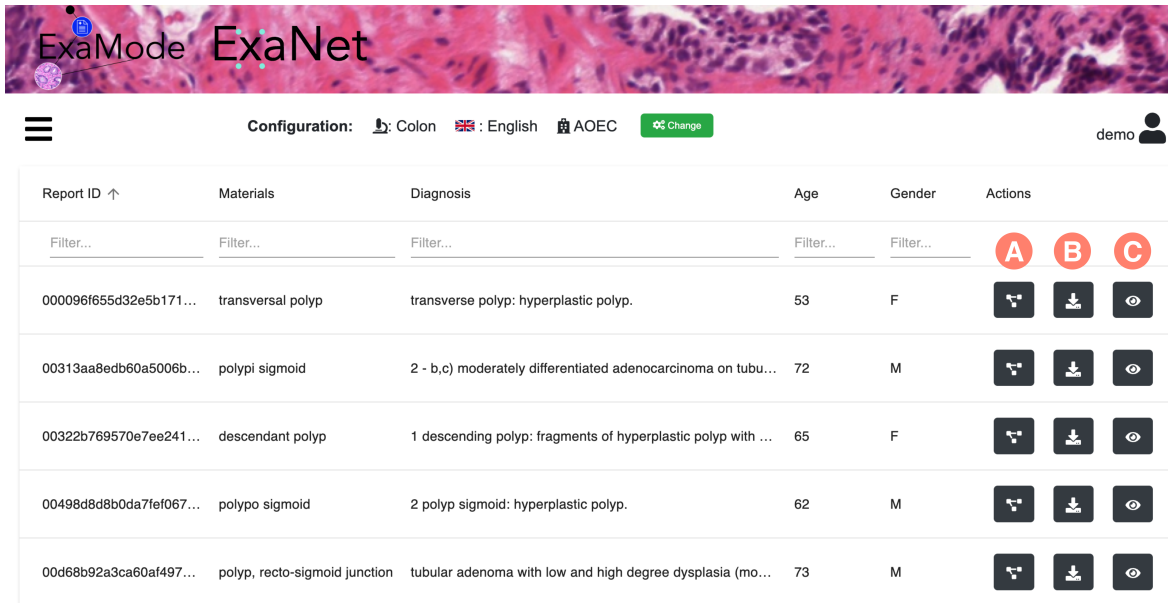
| Task ID | Type | Status | Report/Batch ID | Batch number | Institute | Use case | Start timestamp | End timestamp ↓ | Download | Annotate this batch |
|--------------------------------------|---------------|--|--------------------------------------|--------------|-----------|-----------|---------------------|-------------------------|--------------------------|---|
| Filter... | Filter... | Filter... | Filter... | Filter... | Filter... | Filter... | Filter... | Filter... | | |
| 701d52e6-4b47-4b3d-a26a-64e53da8... | SINGLE_REPORT | SUCCESS | 2cfc22f2-7af0-4a52-90b7-e4b2b722e05 | Not provided | AOEC | Colon | 2022-07-12T13:03:08 | 2022-07-12T13:03:21.276 | Download |  |
| 1188113c-363b-4933-b6b5-d9b4d6015... | SINGLE_REPORT | SUCCESS | 72be2705-bcd4-425b-b660-d367cecc1ede | Not provided | AOEC | Colon | 2022-03-09T18:03:24 | 2022-03-09T18:03:33.301 | Download |  |
| 885783cd-3b60-48be-8748-ed83809bb... | SINGLE_REPORT | SUCCESS | c084814f-5ac3-478e-8011-279ba70ef1a | Not provided | AOEC | Colon | 2022-02-22T16:02:18 | 2022-02-22T16:02:41.412 | Download |  |
| 60f21e95-e856-4938-b90f-fa87486419ab | BATCH | SUCCESS | batch_20220206221635 | Not provided | AOEC | Colon | 2022-02-06T22:16:35 | 2022-02-06T22:16:44.061 | Download |  |
| 664fa6a8-dcc1-466f-ac36-2edc54311e27 | BATCH | SUCCESS | batch_20220206185314 | Not provided | AOEC | Colon | 2022-02-06T18:53:14 | 2022-02-06T18:53:23.355 | Download |  |

Fig. 4.6 SKETUp dashboard reporting the information concerning each execution (i.e., task) of SKET. The interface is continuously updated so that users can monitor SKET executions and check their status. When a task ends up correctly (i.e., *SUCCESS* status), users can download the outputs of SKET and annotate a batch of reports with MedTAG (as long as the batch is saved in the database).

annotated by SKET. ExaNet allows the users to access and explore the graph representation of the ExaMode clinical reports; it allows physicians and experts to consult, search and visually explore the medical reports and their graph representation using an interactive web interface. ExaNet enables users to explore graph connections by leveraging pan and zoom functionalities. On top of this, ExaNet allows users to visualize an interactive JSON serialization of each pathology report, providing also download capabilities.

Conceptually, ExaNet stems from ontology visualization tools. The visualization of ontologies is a fundamental task to assess ontologies and enable users to explore, verify, and understand them and their underlying structures [191, 189, 190, 181]. Nevertheless, compared to ontology visualization, where the focus is primarily on the Terminological Box (TBox) – i.e., definition of classes and properties – ExaNet focuses instead on the Assertional Box (ABox) – that is, individuals and instance data. Furthermore, ExaNet replaces the classes Internationalized Resource Identifier (IRI) with the corresponding literals.

Figure 4.7 depicts the ExaNet main visualization interface for searching and exploring the reports. Users can click on the button in Figure 4.7.A to visualize the graph representation of the report selected, as shown in Figure 4.8. Similarly, users can also download a report in JSON format or just visualize it by clicking respectively on buttons (B) and (C). For instance, in Figure 4.9 we can observe the visualization of a report in JSON format. In particular, the report is visualized in an interactive interface that allows the users to expand/collapse the keys in the JSON report at will. Since MedTAG integrates ExaNet functionalities, users can visualize the graph representation of a report by clicking on the *Graph* button, as shown in Figure 4.10.



The interface shows a table of pathology reports with columns for Report ID, Materials, Diagnosis, Age, Gender, and Actions. The Actions column contains three buttons labeled A, B, and C. The table contains six rows of data.

| Report ID ↑ | Materials | Diagnosis | Age | Gender | Actions |
|-------------------------|-------------------------------|--|-----------|-----------|-------------|
| Filter... | Filter... | Filter... | Filter... | Filter... | A B C |
| 000096f655d32e5b171... | transversal polyp | transverse polyp: hyperplastic polyp. | 53 | F | [A] [B] [C] |
| 00313aa8edb60a5006b... | polypi sigmoid | 2 - b,c) moderately differentiated adenocarcinoma on tubu... | 72 | M | [A] [B] [C] |
| 00322b769570e7ee241... | descendant polyp | 1 descending polyp: fragments of hyperplastic polyp with ... | 65 | F | [A] [B] [C] |
| 00498d8d8b0da7fef067... | polypo sigmoid | 2 polyp sigmoid: hyperplastic polyp. | 62 | M | [A] [B] [C] |
| 00d68b92a3ca60af497... | polyp, recto-sigmoid junction | tubular adenoma with low and high degree dysplasia (mo... | 73 | M | [A] [B] [C] |

Fig. 4.7 ExaNet interface for report search and visualization. Users can search for reports using column filters. Moreover, users can use the action buttons for: (A) visualizing the graph representation of a report; (B) downloading a report in JSON format and (C) visualizing the report in JSON format without having to download it.

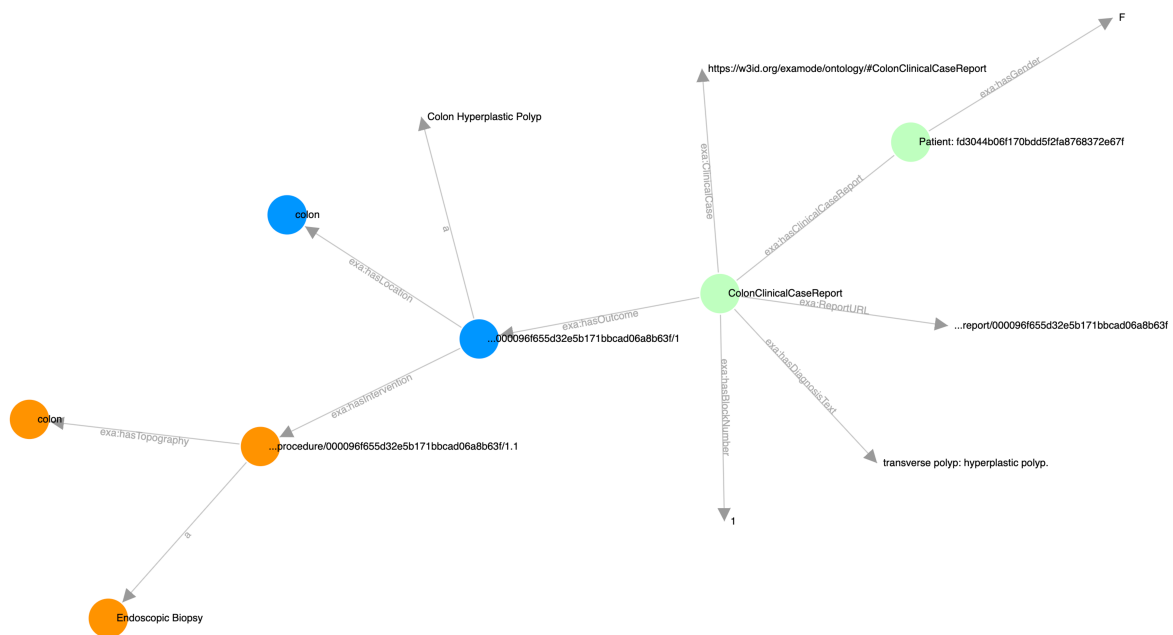


Fig. 4.8 ExaNet visualization of the report-level knowledge graph produced by SKET for a pathology report about colon hyperplastic polyp.

JSON report: [000096f655d32e5b171bbcad06a8b63f](#)

```

▼ "root" : { 8 items
  ▼ "patient" : { 4 items
    "a" : string "http://purl.obolibrary.org/obo/IDOMAL_0000603"
    "hasGender" : string "http://purl.obolibrary.org/obo/NCIT_C46110"
    "PatientURL" : string "https://w3id.org/examode/resource/patient/fd3044b06f170bdd5f2fa8768372e67f"
    "hasGenderLiteral" : string "F"
  }
  "ReportID" : string "000096f655d32e5b171bbcad06a8b63f"
  "ReportURL" : string "https://w3id.org/examode/resource/report/000096f655d32e5b171bbcad06a8b63f"
  ▼ "hasOutcome" : [ 1 item
    ▼ 0 : { 5 items
      "a" : string "http://purl.obolibrary.org/obo/NCIT_C4930"
      ▶ "hasTest" : [ 0 items
      "OutcomeURL" : string "https://w3id.org/examode/resource/000096f655d32e5b171bbcad06a8b63f/1"
      ▼ "hasLocation" : [ 1 item
        0 : string "http://purl.obolibrary.org/obo/UBERON_0001155"
      ]
      ▼ "hasIntervention" : [ 1 item
        ▼ 0 : { 3 items
          "a" : string "http://purl.obolibrary.org/obo/NCIT_C15389"
          ▼ "hasTopography" : [ 1 item
            0 : string "http://purl.obolibrary.org/obo/UBERON_0001155"
          ]
          "InterventionURL" : string "https://w3id.org/examode/resource/procedure/000096f655d32e5b171bbcad06a8b63f/1.1"
        }
      ]
    }
  ]
}

```

Fig. 4.9 ExaNet visualization of a clinical report, in JSON format, in an interactive interface where users can expand/collapse the keys in the JSON report at will.

| id_report_hashed | target_diagnosis | annotations ↓ | | | | | |
|----------------------|---|---------------|---|---|---|---|-------|
| 0e16d79d0acaa309... | tubular adenoma with mild, focally moderate dysplasia ... | 6 | i | ↓ | 👁 | 👤 | Graph |
| 0f69b4ce72f8465b7... | adenocarcinoma (surface fragments). | 6 | i | ↓ | 👁 | 👤 | |
| c323f5d2f717bc221... | villous adenomatous neoplasm with "serrated" and low-... | 6 | i | ↓ | 👁 | 👤 | |

Fig. 4.10 Reports' statistics table of MedTAG integrating ExaNet functionalities. Users can click on the *Graph* button to visualize the graph representation of the medical report selected.

Table 4.6 CNN Colon cancer performance when trained with SKET weak labels (CNN-SKET) and with manual ones (CNN-GT). Results refer to WSI classification on AOEC and RUMC data. For each considered measure, we report the average obtained through 10-fold cross-validation. **Bold** values represent the highest scores achieved for each measure.

| Model | Accuracy | Micro F1 | Weighted F1 |
|----------|---------------|---------------|---------------|
| CNN-SKET | 0.6666 | 0.7741 | 0.7694 |
| CNN-GT | 0.6795 | 0.7866 | 0.7800 |

4.5.4 WSI Classification

The labels produced by SKET are used to reduce supervised-training limitations for Colon cancer assisted diagnosis tools [199] – limitations that prevent the full exploitation of digital pathology applications. In other words, SKET labels serve as weak labels to train a deep image classifier. The proposed model, based on CNNs, makes multi-class predictions at patch-level and then aggregates them through an attention pooling layer [139, 193] to obtain multi-label WSI predictions. The multi-label setting reflects the very nature of the pathology domain, where images (and reports) can highlight multiple findings for the same sample. Therefore, employing models that produce multi-label predictions allows to better approximate real-world pathology scenarios.

The proposed approach has been trained and tested using data composed of Colon WSIs from AOEC and RUMC medical centers. The training set consists of the WSIs associated with the 3769 Colon reports reported in Table 4.1, whereas the test set consists of 227 WSIs from AOEC and 423 from RUMC, for a total of 650 WSIs. Colon cancer was chosen as use-case due to its high social impact and difficulty in diagnosing it. In fact, Colon cancer is the fourth most diagnosed cancer in the world [32]. Besides, the need to identify malignant polyps – which are cell agglomerations protruding from the Colon surface – makes it problematic to diagnose [32]. Thus, to prove the effectiveness of SKET as a weak annotator, we compared the performance of the image classifier trained with SKET labels against its performance when trained using manual labels. Table 4.6 reports the results for subset accuracy, micro-, and weighted-average F1 measures, obtained through 10-fold cross-validation.

The obtained results show the effectiveness of SKET when used as a weak annotator. The performance obtained using weak labels are close to those achieved with manual ones. Precisely, the performance difference between the two CNNs does not exceed 1.3%. Furthermore, we performed the Wilcoxon Rank-Sum test and verified that such performance difference is not statistically significant (p -value < 0.05). Thus, SKET enables the training

of cancer-diagnostics models for DPATH without human intervention, paving the way for the use of ML models in clinical practice.

4.6 Conclusions and Future Work

In this chapter, we presented Semantic Knowledge Extractor Tool (SKET), an unsupervised hybrid knowledge extraction system that combines rule-based techniques with pre-trained Machine Learning (ML) models to extract critical pathological concepts from diagnostic reports. The concepts extracted from diagnostic reports can serve different digital pathology applications, such as automatic annotation, knowledge visualization, discovery, or image classification. A throughout evaluation demonstrated SKET effectiveness in annotating Colon, Cervix, and Lung cancer use-cases – making it a viable solution to reduce pathologists' workload. The results and analyses highlighted the importance of expert knowledge in developing unsupervised systems for specialized medicine. Moreover, the effectiveness of SKET as a weak annotator suggests that it can be used as a first, cheap solution to bootstrap supervised models in the absence of manual annotations. Besides, we highlight the fact that SKET has also been used to empower different digital pathology downstream applications. In particular, SKET labels have been used to reduce training limitations for colon cancer-assisted diagnosis tools. The use of SKET for training deep image classifiers without human intervention paves the way to ML models in the clinical practice [199]. As future work, we plan to extend SKET to other emerging but under-researched use-cases, such as Celiac disease – whose prevalence has significantly increased over the past 20 years [160].

Chapter 5

Explainability

5.1 Introduction

In recent years, the application of AI algorithms in the biomedical domain has experienced unprecedented growth [198, 91, 296] – especially to perform clinical decision support and diagnostic activities [278, 195, 209]. Therefore, there is an urgent need for XAI tools that can help clinicians and domain experts understand algorithm predictions and their underlying rationale. In this regard, explainability techniques highlight decision-relevant aspects of algorithms that contribute to specific predictions, thus trying to answer why a model has made a certain decision [133, 135, 134, 132]. Hence, explainability methods are essential for humans – and in particular for clinicians – to decide whether to trust algorithm predictions and the (underlying) models that generated them. Among its different uses, explainability can be employed to understand the rationale of NER and EL outputs – such as the entity mentions and concepts identified by SKET within clinical reports. However, since most of the data that humans can understand regard objects restricted to the two/three-dimensional space, there is an urgent need not only for explainable models but also for explanation interfaces [135]. To this end, we have developed SKET X,¹ a web-based environment to interact with SKET and get useful insights about the extraction process and the related outputs. Through SKET X, pathologists and domain experts can visually comprehend SKET and the different components activated during the knowledge extraction process – thus getting a point-wise explanation of the outputs obtained for the provided diagnostic reports.

SKET X exploits VA techniques to support domain experts in the visual comprehension of SKET outputs by means of intuitive and interactive interfaces. Such interfaces allow users to inspect and find out non-evident patterns in data and take decisions accordingly [279].

¹<http://w3id.org/sketx> access provided with credentials: demo/demo

In addition, VA techniques enable users to visually comprehend the results of an ongoing task, while it advances asynchronously in the background. Thus, VA techniques are also used to visually adjust the parameters of a model instance running as a background task to continuously refine its outputs [22, 108].

5.2 SKET X Architecture

Figure 5.1 highlights that SKET X acts as an explanation interface for SKET, enabling users to visualize and inspect SKET outputs and the models/rules involved, for an in-depth understanding of the underlying decision process. Hence, experts can realize why SKET produces a certain output so that they can provide suggestions to improve the system accordingly.

SKET X consists of a web application developed using: (i) Django² - i.e., a Python framework for web development - for the back-end and REpresentational State Transfer (REST) Application Program Interfaces (APIs); (ii) React.js³, HTML5, and CSS3 for the front-end; (iii) Celery⁴ - that is, a task queue supporting task scheduling - as a queue manager and scheduler for the incoming requests; (iv) Redis⁵ - that is, a low-latency message queue & broker - as a message broker for Celery and as an in-memory caching system for temporary data; (v) a PostgreSQL database to guarantee the persistence of the SKET output data. Figure 5.2 shows the architecture of SKET X. We can observe that the interface communicates with the business logic through the REST API end-point. Then, the incoming requests requiring the execution of SKET are processed asynchronously in the order determined by the queue manager. Then, the outputs of each request are saved in the database. It is worth mentioning, that asynchronous executions of SKET and a queue manager are necessary to ensure proper execution of SKET, especially in case of large user-provided batches of reports to process. Moreover, asynchronous execution and scheduling enable the users to execute multiple instances of SKET also with different input data.

²<https://www.djangoproject.com>

³<https://reactjs.org>

⁴<https://docs.celeryq.dev>

⁵<https://redis.io>

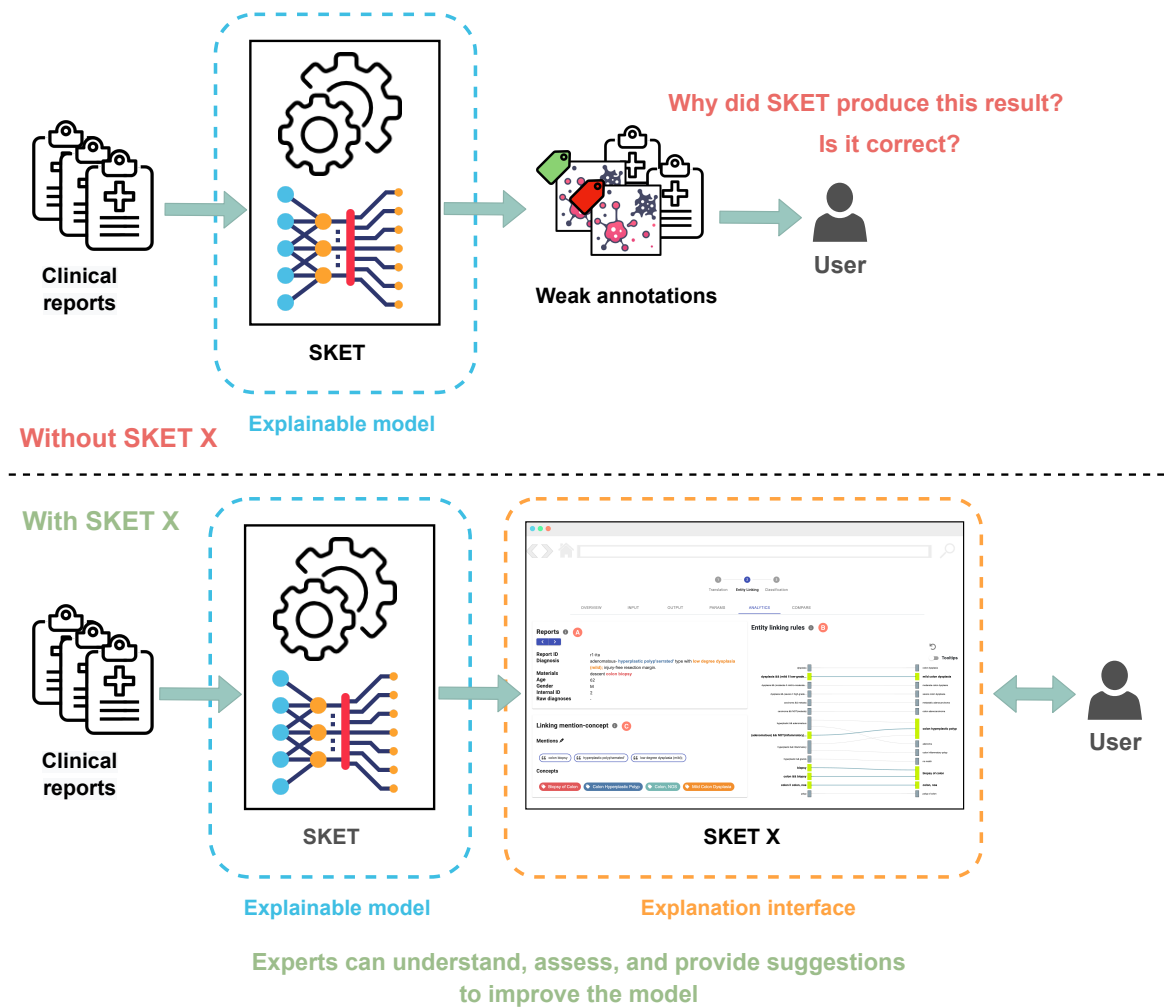


Fig. 5.1 SKET X acts as an explanation interface to visually comprehend why SKET has produced a certain output and realize whether it is correct or not based on the models/rules employed in the machine decision process. Experts can provide feedback/suggestions to improve the system and, in turn, the effectiveness of the SKET knowledge extraction process.

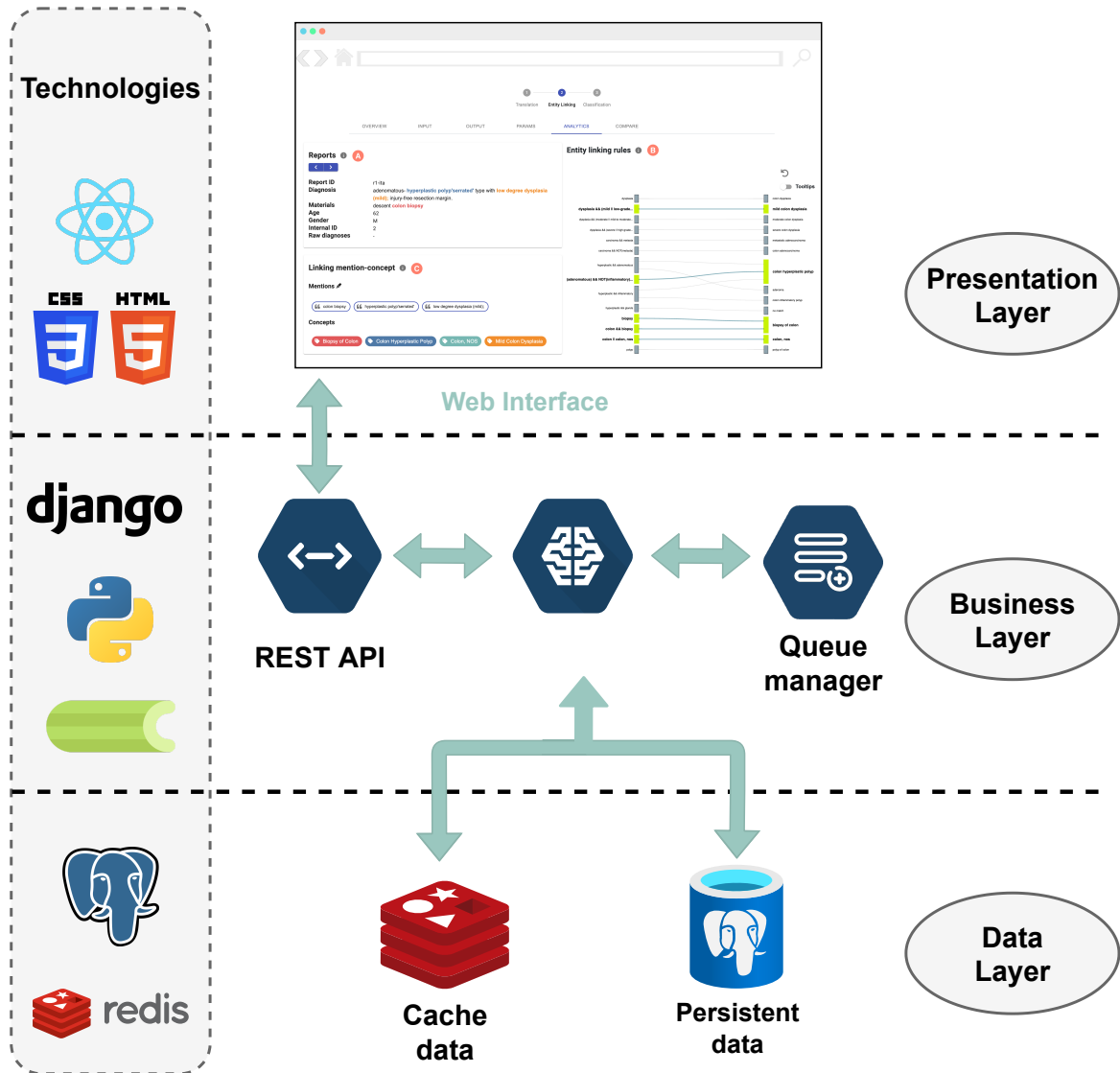


Fig. 5.2 SKET X architecture and technologies adopted. The figure is divided in three sections: (i) *Presentation layer* concerning the front-end and the web interface developed using React . js, HTML5, and CSS3; (ii) *Business layer* where is reported the back-end logic implemented with Python, Django, and Celery; (iii) *Data layer* concerning the data to save either temporary (i.e., cache data saved using Redis) or persistently (using PostgreSQL). The interface communicates with the business logic via REST API requests that are satisfied asynchronously in the order determined by the queue manager. Then, the outputs of each request are saved in the database.

Upload reports

Please, provide the following information:

📁 Reports file ⓘ

A Drag 'n' drop here the report file, or click to select it

🗣️ Language

B Select... | ▾

📁 Use Case

C Select... | ▾

Description ⓘ

D Write a description for your pipeline here...

Upload 📁

Fig. 5.3 SKET X upload form enabling users to provide the diagnostic reports to process (A) and other information including: (B) the language of the reports (i.e., Dutch, English, and Italian); (C) the report use case (i.e., cervix, colon, and lung cancer); and (D) a description of the current pipeline execution. Users can take advantage of drag and drop facilities to specify the reports to process either in CSV or JSON format (A).

5.3 SKET X Workflow

SKET X is an interactive web app that runs SKET on a set of uploaded reports. Users can upload the reports to process using the form depicted in Figure 5.3. SKET X is based on SKET pipelines definable by the user who can customize the parameters and run SKET multiple times to compare the outputs and all the intermediate steps of the process. Each pipeline runs as an asynchronous task, handled by a scheduler with a queue manager. The pipelines are organized for straightforward access in the dashboard interface, shown in Figure 5.4. The dashboard provides information about the SKET pipelines executed by the users and enables access and download of the SKET outputs.

The execution of a SKET X pipeline consists at most of three *phases*, where the currently selected stage is shown on the top of the interface (see Figure 5.5.A).

- *Translation*: the reports are automatically translated from their original language to English. Figure 5.5 reports the information contained in the *Overview* tab of the interface, i.e. the inputs, outputs, and parameters of the translation phase.

Dashboard

Checkout your pipelines' information below i

Your pipelines

| Pipeline ID | Use case | Status | Start Timestamp | End Timestamp | Description | Params | Overview | Download |
|--------------------------|--------------------------|---|--------------------------|--------------------------|---------------------------|---|---|--|
| <small>Filter...</small> | <small>Filter...</small> | <small>Filter...</small> | <small>Filter...</small> | <small>Filter...</small> | <small>Filter...</small> | A | B | C |
| 9e393c1... | Colon | SUCCESS | 2022-07-30T... | 2022-07-30... | Processing of clinical... | | | Download |
| 691bcf1... | Colon | SUCCESS | 2022-07-30T... | 2022-07-30... | Processing of clinical... | | | Download |
| 835ac8a... | Colon | SUCCESS | 2022-07-30T... | 2022-07-30... | Processing of clinical... | | | Download |

Rows per page: 25 ▾ 1-4 of 4

Fig. 5.4 SKET X dashboard providing information about the executed SKET pipelines - i.e., pipeline id, use case, pipeline status, start timestamp, end timestamp, description, pipeline parameters. Users can view the parameters of each pipeline by clicking on the dedicated button (A). Similarly, users can access pipeline data by clicking on the dedicated button (B). When the execution of a pipeline ends, its outputs become available for download (C).

- *Entity Linking*: the entities automatically recognized within the reports are linked to the concepts in the ExaMode ontology. This phase's output consists of the identified mentions and the linked concepts. SKET employs a combination of hand-crafted rules and pre-trained neural models in this phase. The rules relevant to the disease of the given report are shown via a Sankey diagram, where the rules activated for the current report are highlighted. In this context, a rule is activated when one of the identified mentions – e.g., low degree dysplasia (mild) – satisfies one rule trigger – e.g., dysplasia && mild – that implies a link to a specific concept – e.g., mild colon dysplasia – as shown in Figure 5.6.
- *Classification*: SKET exploits mapping rules to decide the appropriate labels for each report. As for the EL phase, the rules relevant for the disease of the considered report are visualized using a Sankey diagram, where the activated rules are highlighted. A rule is activated when one of the identified concepts – e.g., Mild Colon Dysplasia – satisfies one rule trigger – e.g., dysplasia && mild – that implies a specific label – e.g., Adenomatous polyp - low grade dysplasia – as shown in Figure 5.7. The mentions and concepts considered in the classification task are regarded as *key*

A 1 2 3
Translation Entity Linking Classification

OVERVIEW INPUT OUTPUT PARAMS

INPUT B

| Report ID | Diagnosis |
|-----------|-------------------------------|
| r1-ita | Polipo iperplastico-adenoma |
| r2-ita | Adenoma tubulare con displ |
| r3-ita | Neoplasia adenomatosa villo |
| r4-ita | Adenocarcinoma ulcerato (fr |
| r5-ita | Adenoma tubulare con displ |
| r6-ita | Frammenti superficiali di ade |
| r7-ita | Polipo iperplastico-adenoma |

OUTPUT C

| Report ID | Diagnosis |
|-----------|-------------------------------|
| r1-ita | adenomatous- hyperplastic p |
| r2-ita | tubular adenoma with low de |
| r3-ita | villous adenomatous neoplas |
| r4-ita | ulcerated adenocarcinoma (f |
| r5-ita | tubular adenoma with mild, f |
| r6-ita | superficial fragments of ader |
| r7-ita | adenomatous- hyperplastic p |

PARAMS D

Pipeline ID: 835ac8ac-ea96-45a2-8991-f83bc403cf3d
Description: Processing of clinical reports provided in Italian language

| Param | Value | Description |
|-----------------|---------|---|
| input language | Italian | Reports' original language |
| output language | English | Reports' output language |
| usecase | colon | Reports' usecase among Uterine Cervix, Colon and Lung cancers |

Fig. 5.5 (A) SKET X *Overview* tab for the translation phase, (B) the reports in the original language (input), (C) the translated reports (output) (C), and (D) the parameters and settings for the current phase.

mentions/concepts, whereas the ones not satisfying any rule trigger are regarded as *excluded*, as shown in Figure 5.7.C and 5.7.D, respectively. For instance, in Figure 5.7 we can observe that the key concepts identified are Colon Hyperplastic Polyp and Mild Colon Dysplasia, whereas the excluded ones are Biopsy of Colon and Colon, NOS – both related to the same excluded mention colon biopsy.

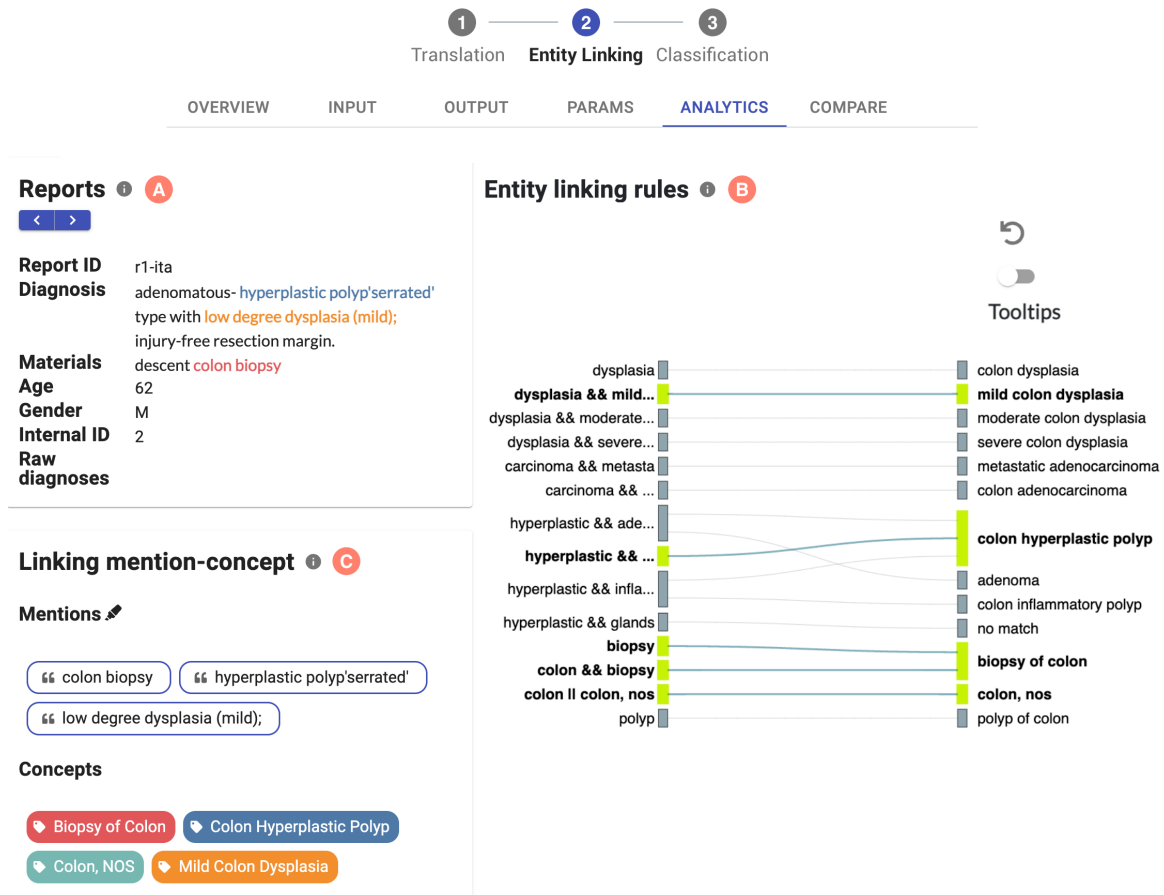


Fig. 5.6 SKET X *Analytics* tab for the EL phase: (A) reports section, the users can change the current report using the left/right buttons; (B) SKET rules for the NER task; and, (C) list of mentions and concepts produced by the knowledge extraction process. Each concept and related mentions are highlighted with the same color in (A) and (C). By clicking/hovering on a specific concept, it is possible to highlight the relevant rules in the Sankey diagram that determined the concept and the related mentions in the report text. On the left side of the Sankey diagram are reported the rules *triggers*, which are boolean expressions tested on each mentioned text. If one or more mentions satisfy a rule trigger, then the related concepts on the right side of the Sankey diagram are highlighted and listed in (C).

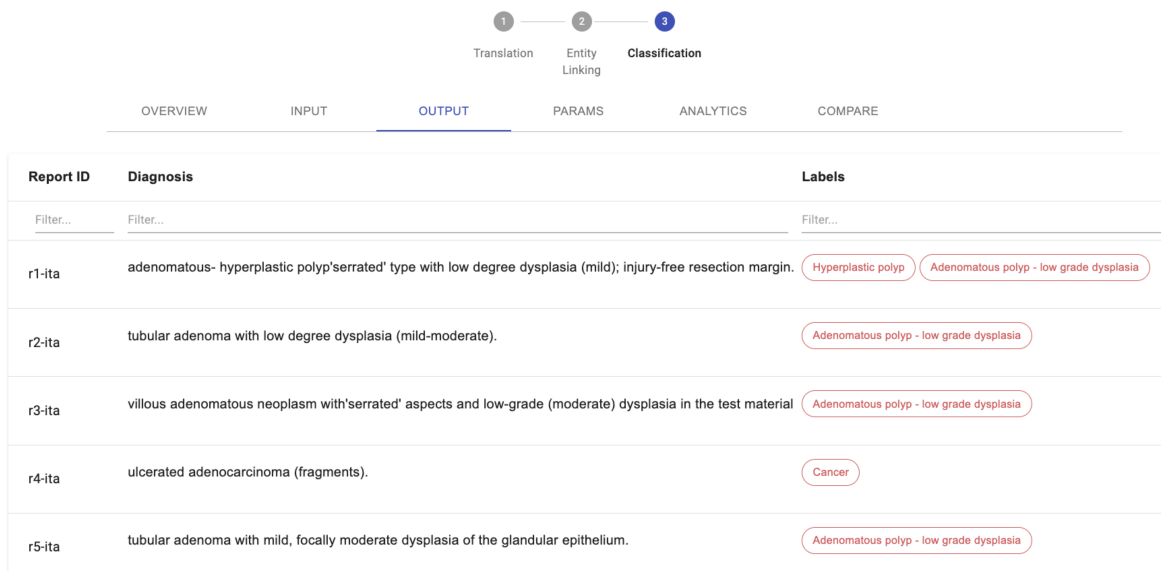


Fig. 5.7 SKET X *Analytics* tab for the classification phase: (A) reports section to select the current report via left/right buttons; (B) SKET rules for determining the labels visualized with a Sankey diagram; and, (C) list of labels, mentions, and concepts determined by SKET. Each concept and the related mentions are highlighted with the same color in (A) and (C). The Sankey diagram highlights the relevant rules by clicking/hovering on a specific label. On the left side of the Sankey diagram are reported the rules *triggers*. If one or more concepts satisfy a rule trigger, then the related label is highlighted on the right side of the Sankey diagram and also listed in (C). The mentions and concepts involved in the classification task are the *key mentions/concepts* (C), while the *excluded* ones are reported in (D).

5.4 SKET X Interface

The interface of SKET X consists of six tabs providing different views of the data, according to the selected phase:

- *Overview tab*: overview of the visual outputs available in the other tabs (i.e., *Input*, *Output*, *Params*, *Analytics*) for the current phase. The contents of the *Analytics* tab are shown in the overview only for the EL and classification phases. Figure 5.5 shows the overview of the inputs, outputs, and parameters for the translation phase.
- *Input tab*: it reports the input data for the current phase. For instance, if the considered phase is translation, this tab shows the reports in the original language, as shown in Figure 5.5.B. Instead, if the considered phase is EL, it shows the translated reports. Similarly, the mentions and the concepts extracted for each report are shown in this tab for the classification phase.
- *Output tab*: it reports the output data for the current phase. For translation, this tab shows the reports translated into English, as shown in Figure 5.5.C. Instead, if the considered phase is EL, it shows the mentions and the concepts extracted for each report. Similarly, the labels generated for each report are shown in this tab for the classification phase, as shown in Figure 5.8.



| Report ID | Diagnosis | Labels |
|-----------|--|--|
| Filter... | Filter... | Filter... |
| r1-ita | adenomatous- hyperplastic polyp'serrated' type with low degree dysplasia (mild); injury-free resection margin. | Hyperplastic polyp Adenomatous polyp - low grade dysplasia |
| r2-ita | tubular adenoma with low degree dysplasia (mild-moderate). | Adenomatous polyp - low grade dysplasia |
| r3-ita | villous adenomatous neoplasm with'serrated' aspects and low-grade (moderate) dysplasia in the test material | Adenomatous polyp - low grade dysplasia |
| r4-ita | ulcerated adenocarcinoma (fragments). | Cancer |
| r5-ita | tubular adenoma with mild, focally moderate dysplasia of the glandular epithelium. | Adenomatous polyp - low grade dysplasia |

Fig. 5.8 SKET X *Output* tab showing the SKET outputs for the classification phase (i.e., labels). These are arranged in tabular form so that users can take advantage of column filters to search and visualize specific report information.

- *Params tab*: it reports the parameters for the current phase, as shown in Figures 5.5.D and 5.9. For instance, for EL, it shows the methods and models used by SKET to perform the linking process between mentions and related concepts. Another important parameter is the *threshold* used by SKET in the pruning phase to reduce false positives and thus increase precision, as described in Section 4.3.2 of Chapter 4. When the phase considered is EL, users can change one or more parameters and then re-run SKET, as shown in Figure 5.9. This is useful to compare two pipelines using different parameters.

Pipeline ID: 835ac8ac-ea96-45a2-8991-f83bc403cf3d
 Description: Processing of clinical reports provided in Italian language

| Param | Value | Description | Edit |
|---------------------------|---------|--|----------|
| reports original language | Italian | Language of the reports provided as input. | disabled |
| Gestalt Pattern Matching | True | Perform the Gestalt Pattern Matching technique to compute the similarity between two textual strings. | |
| Neural model | True | This model leverages the SciSpacy model en_core_sci_sm to compute the semantic similarity between candidate mentions and ontology concepts. | |
| threshold | 1.8 | This threshold regulates the tolerance in the matching function of the Semantic Knowledge Extractor Tool (SKET). The threshold values range from zero to the number of models considered. The models can be at most two: (1) StringMatch: Perform the Gestalt Pattern Matching; (2) Neural model: Leverages the SciSpacy model en_core_sci_sm to compute the semantic similarity between candidate mentions and ontology concepts. | |
| reports language | English | Language of the reports processed by the Semantic Knowledge Extractor Tool (SKET) | disabled |
| usecase | colon | Reports' usecase among Uterine Cervix, Colon and Lung cancers | disabled |

[Run SKET again](#)

Fig. 5.9 SKET X *Params* tab showing the parameters for the EL phase. The figure highlights that the current pipeline uses both the GPM and the neural model together with the default SKET threshold of 1.8. Users can change the value of these parameters and then click on the *Run SKET again* button to re-run SKET on the same set of reports but with the new parameter values. Finally, the results of the new SKET run are saved in a new pipeline and the user is asked to provide a description for it.

- *Analytics tab*: it allows the users to analyze the current report's mentions, concepts, and labels in detail. In particular, if the considered phase is EL, users can inspect the identified mentions and concepts concerning the report textual content, as shown in Figures 5.6.A and 5.6.C. Moreover, by clicking on a mention, the user can inspect the list of associated concepts. At the same time, a user can also do the reverse -

identifying the relevant mentions for a given concept. In addition, if the considered phase is classification, this tab shows the labels determined by SKET and the relations between a label and the concepts from which it derives, as shown in Figure 5.7. To visually explain the rules used by SKET to determine both the concepts and labels, a Sankey diagram is reported on the right side of the interface as depicted in Figure 5.6.B and 5.7.B. On the left side of the Sankey diagram, the rules *triggers* are reported, which are boolean expressions tested on the text of each mention – for the EL phase – and concept – for the classification phase. If one or more mentions/concepts satisfy a rule trigger, then the related concepts/labels on the right side of the Sankey diagram are highlighted.

- *Compare tab*: it allows the users to compare the outputs of two different SKET X pipelines in terms of mentions, concepts, and labels identified for the current report. When the users click on the *compare* tab, they are provided with an initial menu that allows them to specify the two pipelines to compare. After the selection, users can click on the *compare* button to visualize the interface dedicated for the comparison, illustrated in Figure 5.10. The comparison interface is divided into four parts: (A) the reports section displaying information about the current report and two buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier, the description, and its parameters; (C) first pipeline section showing the outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline section with the same structure of (C). In particular, if the considered phase is EL, users can compare the concepts and the mentions identified by each pipeline and deduce which parameters have determined the major differences (e.g., the threshold for the NER task). Moreover, by clicking/hovering on each mention, the users can inspect the list of associated concepts (highlighted in different colors) among the two pipelines. Moreover, the common mentions between the two considered pipelines can be highlighted, thus making them and their related concepts easy to identify. Figure 5.10 shows the outputs of two SKET pipelines that have been executed with different models, where the first pipeline uses only the neural model while the second one uses only GPM. Since the two pipelines considered in Figure 5.10 use different models, they identify different concepts and mentions. Indeed, the common concepts between the two pipelines – i.e., Biopsy of Colon, Colon Hyperplastic Polyp, Colon NOS, and Mild Colon Dysplasia – have been identified using SKET rules, which are used in both pipelines. On the other hand, the disjoint concepts have been identified using the neural model – for Rectal mucous membrane – and GPM – for Adenoma and Resection – respectively. If the considered phase is classification,

users can also compare the labels generated by each pipeline and the *key* concepts associated – that is, the ones from which the labels are determined. For instance, in Figure 5.11, we can observe that the labels generated by SKET are Adenomatous polyp - high grade dysplasia and Hyperplastic polyp for the first pipeline (C) while only Adenomatous polyp - high grade dysplasia for the second one (D). By clicking/hovering on the Hyperplastic polyp label, users can realize that it derives from the Colon Hyperplastic Polyp concept, which, in turn, is associated with the polyp sigmoid mention. Nevertheless, the latter mention does not suggest the presence of a Colon Hyperplastic Polyp. Thus it is a false positive. Similarly, users can do the same with the Adenomatous polyp - high-grade dysplasia label, discovering that it derives from the Severe Colon Dysplasia concept, which is correctly associated, through a SKET rule, with the severe dysplasia mention. Finally, users can also compare the *excluded* mentions and concepts that are not considered for the label generation process, but that can be a good indicator to determine whether the chosen threshold for models produces noisy concepts.

Hence, using SKET X pathologists and domain experts can visually comprehend why a certain concept/label has been extracted. Moreover, by leveraging both inspection and comparison functionalities, users can also understand the impact of different parameters on the obtained outputs, and thus investigate the advantages of combining ad hoc rules with ML models to improve the overall effectiveness of knowledge extraction systems.

1 ——— 2 ——— 3
Translation Entity Linking Classification

OVERVIEW INPUT OUTPUT PARAMS ANALYTICS **COMPARE**

A

Report ID: r1-ita

Diagnosis: adenomatous- hyperplastic polyp'serrated' type with low degree dysplasia (mild); injury-free resection margin.

Materials: descent colon biopsy

Age: 62

Gender: M

B

| Pipeline | Description | Params |
|--------------------------------------|--|-----------------------|
| 691bcf14-7f03-489b-8e1d-cb8a9025162a | Processing of the clinical reports without using GPM (only neural model + rules) | <input type="radio"/> |
| 9e393c1b-ba1b-4595-9403-21876f70fa31 | Processing of the clinical reports without the neural model (only GPM + rules) | <input type="radio"/> |

C

Pipeline: 691bcf14-7f03-489b-8e1d-cb8a9025162a

Pipeline description: Processing of the clinical reports without using GPM (only neural model + rules)

Mentions

adenomatous- colon biopsy hyperplastic polyp'serrated'

low degree dysplasia (mild);

Concepts

Biopsy of Colon Colon Hyperplastic Polyp Colon, NOS

Mild Colon Dysplasia Rectal mucous membrane

D

Pipeline: 9e393c1b-ba1b-4595-9403-21876f70fa31

Pipeline description: Processing of the clinical reports without the neural model (only GPM + rules)

Mentions

adenomatous- colon biopsy hyperplastic polyp'serrated'

injury-free resection margin. low degree dysplasia (mild);

Concepts

Adenoma Biopsy of Colon Colon Hyperplastic Polyp

Colon, NOS Mild Colon Dysplasia Resection

Fig. 5.10 SKET X *Compare* tab for the EL phase showing the comparison interface for the two pipelines specified for the comparison. The interface is organized in four parts: (A) the reports section displaying information about the current report and two buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier and its description, and parameters (e.g., the models used for EL phase and the threshold); (C) first pipeline outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline outputs for the phase selected. The mentions in common, and the related concepts, are highlighted both in the report text (A) and also in the mention/concept lists for each pipeline (C) and (D). Hence, we can observe that there is a mention *injury-free resection margin* and a concept *Resection* that are not highlighted since they have been identified only by the second pipeline (D). Nevertheless, the concepts *Rectal mucous membrane* and *Adenoma* have been identified only by respectively the first pipeline (C) and the second one (D), but since both are associated with the same common mention – i.e., *adenomatous* – they are highlighted as well.

The screenshot displays the SKET X Compare tab for the classification phase. At the top, there are three numbered steps: 1. Translation, 2. Entity Linking, and 3. Classification. Below this, there are tabs for OVERVIEW, INPUT, OUTPUT, PARAMS, ANALYTICS, and COMPARE. The COMPARE tab is active, showing a comparison interface for two pipelines.

Part A: Reports Section

Report ID: r15-ita
 Diagnosis: polyp sigmoid: tubular adenoma with severe dysplasia of the glandular epithelium. no evidence of infiltration of the connective vascular axis. free of resection margins.
 Materials: polyc0-caecum
 Age: 74
 Gender: M
 Internal ID: 1

Part B: Parameters Section

| Pipeline | Description | Params |
|--------------------------------------|--|-----------------------|
| 691bcf14-7f03-489b-8e1d-cb8a9025162a | Processing of the clinical reports without using GPM (only neural model + rules) | <input type="radio"/> |
| 9e393c1b-ba1b-4595-9403-21876f70fa31 | Processing of the clinical reports without the neural model (only GPM + rules) | <input type="radio"/> |

Part C: First Pipeline Outputs

Classification

Pipeline: 691bcf14-7f03-489b-8e1d-cb8a9025162a
 Pipeline description: Processing of the clinical reports without using GPM (only neural model + rules)

Labels

- # Adenomatous polyp - high grade dysplasia
- # Hyperplastic polyp

Key Mentions

- polyp sigmoid:
- severe dysplasia

Key Concepts

- Colon Hyperplastic Polyp
- Severe Colon Dysplasia

Excluded Mentions

- glandular epithelium.
- polyc0-caecum
- tubular adenoma

Excluded Concepts

- Colon Inflammatory Polyp
- Colon Tubular Adenoma
- Rectal mucous membrane

Part D: Second Pipeline Outputs

Classification

Pipeline: 9e393c1b-ba1b-4595-9403-21876f70fa31
 Pipeline description: Processing of the clinical reports without the neural model (only GPM + rules)

Labels

- # Adenomatous polyp - high grade dysplasia

Key Mentions

- severe dysplasia

Key Concepts

- Severe Colon Dysplasia

Excluded Mentions

- glandular epithelium.
- polyc0-caecum
- polyp sigmoid:
- tubular adenoma

Excluded Concepts

- Caecum
- Colon Tubular Adenoma
- Sigmoid colon

Fig. 5.11 SKET X *Compare* tab for the classification phase showing the comparison interface for the two pipelines specified for the comparison. The interface is organized in four parts: (A) the reports section displaying information about the current report and two buttons for switching to the next/previous report; (B) the parameters section displaying pipeline information, such as the identifier, the description, and its parameters; (C) first pipeline outputs for the phase selected (e.g., mentions and concepts) and (D) second pipeline outputs for the phase selected. The mentions/concepts considered for determining the report labels are regarded as *key* mentions/concepts and are differentiated by the *excluded* ones. Here, two concepts are identified in the first pipeline, namely, Colon Hyperplastic Polyp and Severe Colon Dysplasia, while in the second one only Severe Colon Dysplasia has been identified. Nevertheless, Colon Hyperplastic Polyp and Sigmoid colon are negligible concepts (i.e., false positives) both associated with the polyp sigmoid mention. In contrast, Severe Colon Dysplasia is correct since it has been identified using a SKET rule verified by the severe dysplasia key mention.

5.5 Conclusions and Future Work

In this chapter, we presented SKET X, an explainability tool that exploits VA techniques to support pathologists and experts to comprehend SKET outputs as well as the rules, models, and parameters used for determining them. SKET X enables the users to interact with SKET, comprehend its results, and get useful insights concerning the knowledge extraction process. We pointed out that XAI solutions are essential in medicine since physicians need to understand why a specific prediction has been determined to trust model predictions. Thus, the final aim is to increase the awareness of the users interacting with SKET, improve its effectiveness in terms of the quality of the weak annotations produced, and in turn the effectiveness of the image classification system for CPATH that is trained using the weak annotations generated by SKET. We can measure the quality of the weak annotations by comparing them with the ground-truths created with MedTAG, as described in Chapter 4. As future work, we plan to conduct a user study to collect feedback from pathologists and experts in order to improve SKET X accordingly. Specifically, we plan to conduct the user study in an asynchronous fashion, so that they can start it when they prefer. In this regard, we plan of providing the participants with anonymized credentials to access SKET X and a private link to an online form where they can answer a set of predefined questions and provide their feedback. Moreover, we plan to divide the user study into two parts, designed to measure respectively the *learnability* and *usability* of SKET X. The learnability part focuses on assessing the confidence and the awareness of the users with respect to accomplishing a set of predefined tasks such as identifying which concepts are related to a specific mention or the concepts from which a specific label has been generated. Then, the collected answers of each user will be compared with the correct ones, in order to assess the user proficiency with SKET X with respect to explainability purposes. Secondly, we plan to evaluate SKET X in terms of usability and user satisfaction using the System Usability Scale (SUS), which is widely used and is considered an industry standard for assessing usability [41]. Finally, we aim to collect useful opinions and suggestions from pathologists and other experts to identify the key necessities and foster further advancements in the design of transparent and explainable models and algorithms for CPATH.

Chapter 6

Semantic annotation

6.1 Background

In the last decades, exascale volumes of biomedical data have been produced, where the vast majority is available as unstructured text [216]. Health-care professionals traditionally rely on free-text reporting for communicating patient information such as diagnosis and treatments. For instance, narrative clinical reports are usually conceived as free-text reports, which are human-readable but not machine-readable. This brings interoperability issues and limitations to effective secondary reuse of data, essential for medical decision making and support. In order to process the vast amount of unstructured biomedical data from clinical reports and EHRs, Information Extraction (IE) algorithms and NLP techniques have been developed and are currently exploited.

To this aim, significant efforts have been dedicated to applying Named Entity Recognition and Linking (NER+L) methods for entity extraction and semantic annotation [25, 113, 303, 257, 167]. Semantic annotation is the NLP task of identifying the type of an entity and uniquely linking it to a corresponding knowledge base entry [144]; it leverages both text-processing and ML techniques to tackle biomedical information extraction challenges such as terms and abbreviations disambiguation. Furthermore, semantic annotation tasks based on ML methods are often carried out in a supervised context where large-scale training and test annotated corpora are required. Moreover, even in an unsupervised context, NER+L models require annotated datasets for evaluation purposes. However, the lack of manually annotated biomedical datasets poses hindrances to the further development of NER+L systems. In addition, most of the training data available for the biomedical domain covers mainly common entity types (e.g., drugs, genes, and diseases) [212, 185, 88, 168], thus the coverage of some biomedical sub-domains is limited. For these reasons, several attempts to create large annotated biomedical corpora have been conducted [207, 222, 247, 47, 224, 147, 289, 143].

To achieve the high-quality standards required for the biomedical domain, the annotation process demands human-expert supervision. Nevertheless, manual annotation of large datasets is an expensive and time-consuming task requiring plenty of expert annotators with extensive experience in biomedical contents. To support, organize and speed up the annotation process, several annotation tools have been developed [87, 263, 316, 39, 48, 177, 178, 254, 57, 242, 229, 246, 300, 215]. However the biomedical domain is particularly challenging, since biomedical texts contain mentions that are burdensome for semantic annotation, such as the abbreviations of genes and proteins. Moreover, the specificity of some biomedical sub-domains, such as histopathology, requires fine-grained annotation systems designed to be customizable according to physicians' and experts' needs.

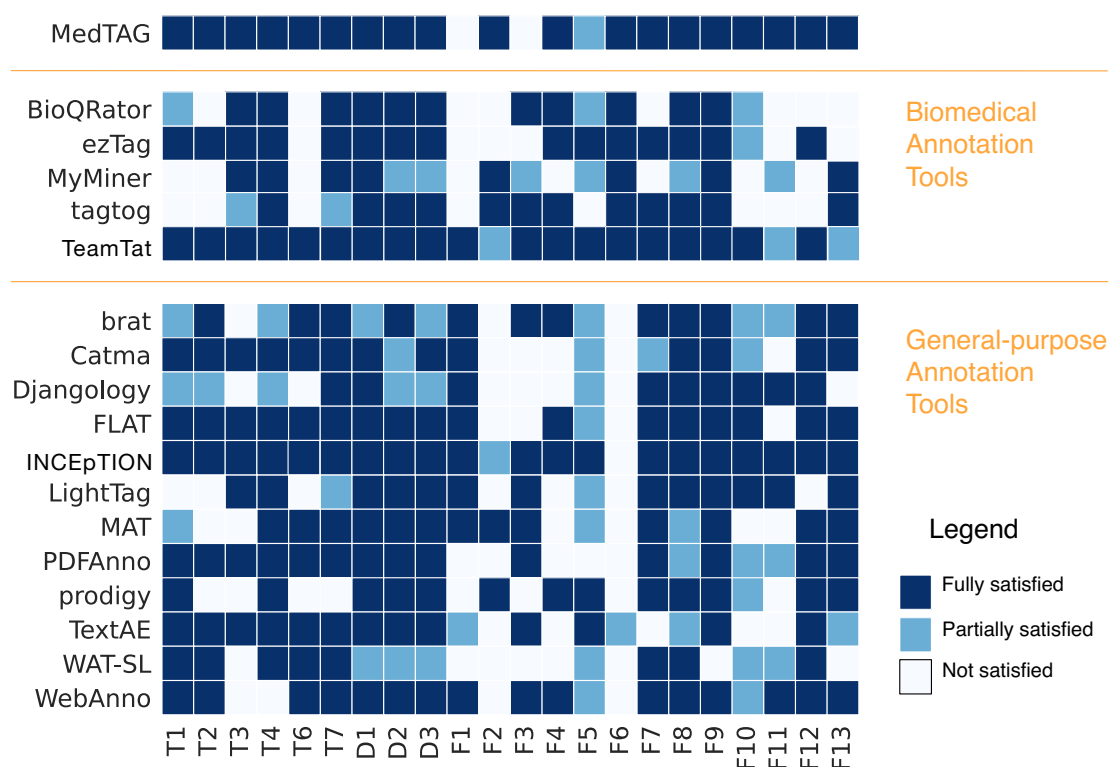


Fig. 6.1 Overview of annotation tools and their functionalities. The annotation tools considered come from a recent extensive review of tools for manual annotation of documents [219]. In addition, we consider also TeamTat [141] and INCEpTION [162] and report our judgments. The annotation tools are assessed with 22 criteria, defined in the latter review study, among three categories: *Data* (D), *Functional* (F) and *Technical* (T). The fulfillment of each criterion is indicated with a color in a three levels scale: white (feature absent or not met), light blue (feature partially satisfied), blue (feature satisfied).

In recent years, several biomedical annotation tools have been released [219, 218]. Motivation for the wide variety of biomedical annotation tools available could be the necessity

of domain-specific functionalities that might be only partially supported or not by other well-established tools. Hence, some tools could be handier than others for a specific task of interest.

A recent extensive review of both general-purpose and biomedical annotation tools provides a detailed comparison of state-of-the-art annotation tools [219]. Some of the common limitations of the available tools are, for instance, the non-availability of the source code or the raised exceptions and failures during the installation process. In addition, even the most popular annotation tools present drawbacks such as a burdensome installation procedure or the lack of documentation. As an example, WebAnno [309] and brat [276] are popular general-purpose annotation tools with a comprehensive set of functionalities, but their installation process is rather complex for the not technology-savvy users. INCEpTION [162, 161] is a more recent general-purpose annotation tool from the authors of WebAnno [309], that mitigates this issue with a web service enabling the users to work online. Moreover, general-purpose annotation tools often do not fulfill the needs of biomedical experts; thus, domain-specific solutions are preferable for this field. Even though brat [276] has been used in several biomedical projects [205, 312, 165, 49, 294], it is designed for general-purpose annotation, thus it provides additional features that are not suited for physicians and experts of the biomedical domain. Since the annotation process is a time-consuming task, biomedical annotation tools should be designed to offer an intuitive streamlined interface that minimizes redundant features, fulfill domain-specific requirements and reduce the annotators workload.

For the in-depth analysis, we focus on the tools specifically designed for biomedical annotations: BioQRator [177], ezTag [178], TeamTat [141], MyMiner [254] and tagtog [57]. Additionally, we also consider two general-purpose annotation tools that are used by the biomedical community as well - i.e., brat [276] and INCEpTION [162].

In Figure 6.1, we can see a heat-map reporting on the functionalities of the current text annotation tools as analyzed by a very recent extensive survey [219]. The provided heat-map is to be used as a visual summary of the features provided by each annotation tool. ¹ In particular, the heat-map considers a list of 15 annotation tools selected according to five major requirements: (i) **Available**: the executable and project source code should be available; (ii) **Web-based**: the tool should be provided as an online web application or as an installable application running in a web browser container; (iii) **Installable**: the installation process should last two hours at most; (iv) **Workable**: it should work for hands-on experiments; (v) **Schematic**: users should be able to configure the annotation schema at will. Hence, several biomedical annotation tools such as Argo [242], Egas [48], Marky [229], ODIN [246],

¹We report details about the main features considered in [219] to ease the comprehension of our analysis.

Pubtator [300] and Textpresso [215] are not considered since they do not satisfy one or more of the previous five requirements.

Moreover, the selected annotation tools are compared according to a set of 22 criteria chosen among the original 26 criteria of the same study [219]. In particular, the criteria are grouped in four categories: (i) *Data*, (ii) *Functional*, (iii) *Publication* and (iv) *Technical*. We excluded the publication criteria (i.e., the four missing criteria) since we are interested in comparing the facilities and functionalities provided by the different tools and not on their coverage in scientific publications.

The *data* criteria are: (D1) format of the schema – whether it is configurable or uses standard formats (e.g. JSON, XML); (D2) input format for documents – whether the input documents are based on standard formats (e.g. JSON, XML) and (D3) output format for annotations – whether the annotations are based on standard formats (e.g. JSON, XML).

The *functionality* criteria are: (F1) support for overlapping mentions/annotations; (F2) support for document-level annotations – users can specify the labels that apply to the whole document (not just for a textual portion); (F3) support for relationship annotations; (F4) support for ontologies and terminologies (i.e. a procedure to import terminology resources is provided); (F5) Support for built-in predictions and active learning from pre-annotated documents; (F6) Integration with PubMed – users can annotate PubMed abstracts just providing a list of PubMed ids; (F7) Suitability for full texts (i.e., tool capable of displaying long text correctly, without compromising readability); (F8) Allowance for saving documents partially (i.e., holding annotations partially to later continue the annotation process); (F9) Ability to highlight parts of the text; (F10) Support for users and teams; (F11) Support for Inter-Annotator Agreement (IAA); (F12) Data privacy (i.e., can be used for private data); (F13) multilingual support (i.e., annotating multilingual documents, that might contain special characters).

The *technical* criteria are (T1) Date of the last version – whether the last version (or commit) has been released within the past five years; (T2) availability of the source code – whether the source code is available in version control platforms; (T3) online availability for use; (T4) easiness of installation – i.e., available online (no installation required) or easy and fast to install (up to half-hour time); (T6) license allowing modification and redistribution; (T7) free of charge. We excluded T5 (quality of documentation) from the technical criteria since we are interested in objective and assessable criteria.

Figure 6.1 shows that several tools lack one of the following functionalities: (i) document-level annotation; (ii) ontology and terminology resources support; (iii) support for multi-label annotation; and (iv) support for collaborative annotations with users and teams. Moreover,

seven over the fifteen selected tools are provided through a license that limits modifications and redistribution.

To mitigate this, we introduce MedTAG, a customizable, collaborative, web-based annotation tool provided as a docker container to enable cross-platform support and quick and easy installation. MedTag provides a step-by-step schema configuration, by which the project/team leader can specify in detail which document parts or document fields can be annotated. We designed MedTAG according to the five primary annotation tools' requirements previously discussed. Besides, we determined the feature coverage provided by MedTAG concerning the former criteria. Figure 6.1 shows that MedTAG satisfies most of the criteria, having a feature coverage of 20 criteria over 22. The rest of the criteria currently not covered by MedTAG, such as the relationship annotations and active learning capabilities, are planned as future work.

6.2 Implementation

MedTAG has been designed to be flexible and customizable, so that users can easily install it and configure the annotation schema at will. Hence, MedTAG is not limited to a specific (sub-)domain (e.g., histopathology), but it can be seamlessly used in different biomedical sub-domains. The key MedTAG functionalities are: (i) a web-based collaborative annotation platform with support for users and roles; (ii) a user-friendly interface with support for click-away mention annotation, mentions highlighting in different colors and automatic saving every time an action is performed; (iii) sorting of documents based on the lexicographic order or the "unannotated-first" policy; (iv) support for mobile devices; (v) download of annotations in several formats (i.e., BioC/JSON, BioC/XML, CSV, JSON); (vi) support for multi-label annotation; (vii) support for document-level annotations; (viii) multilingual support; (ix) support for ontologies/concepts to use for the annotation process; (x) support for IAA; (xi) integration with PubMed; (xii) support for automatic built-in predictions; (xiii) support for schema configuration, so that users can easily import data (i.e., documents, labels and concepts), as CSV files, and choose which document fields to annotate. In order to achieve automatic annotations and built-in predictions, we integrated the SKET² in MedTAG. Note that the support for built-in predictions is currently limited to three cancer use-cases (i.e., cervix, colon, and lung cancer). Nevertheless, we plan to extend the support for automatic built-in predictions also for other use-cases. General-purpose automatic annotation methods are of limited efficacy for the biomedical domain; nevertheless, the integration of SKET paves the road for the integration of other third-party libraries users may want to employ.

²<https://github.com/ExaNLP/sket/>

To exploit the concept linking functionality, MedTAG requires the admin user to specify, during the configuration phase, the CSV file containing all the concepts used for annotating the clinical reports. During the first configuration, the admin user is not defined yet, thus the configuration is handled by the *Test* user in *Test mode*, as described in the *Installation and customization* section. Figure 6.5.2 shows the configuration interface that allows the users to specify the CSV file for the concepts. Moreover, the users can choose whether to use the concepts of the ExaMode ontology³ (necessary for the automatic annotation module using SKET) or a set of concepts from a different ontology. Then, the concepts provided in the CSV file populates the MedTAG database and are integrated in the drop-down menu available to the user to select the concepts. Every concept defined in the provided CSV is uniquely identified with a concept IRI. Thus, users could use concepts defined in different ontologies at the same time. Since the CSV file with the concepts for the annotation process is provided by the admin user, the coherence of the data (e.g., the same concept mapping to more than one IRI from different ontologies) should be checked and enforced by the admin herself. Nevertheless, in the case of the same entity mapping to different ontologies, MedTAG differentiates the concepts in the user interface based on the IRIs and other concept information such as use-cases and semantic areas. Thus, users have the means to disambiguate between potentially similar concepts.

MedTAG source code and the documentation are publicly available at this URL: <https://github.com/MedTAG/medtag-core>.

6.2.1 Architecture

Figure 6.2 illustrates the MedTAG architecture, which consists of three logic layers (i.e., *Data*, *Business* and *Presentation* layer). The data layer concerns information and data management; it consists of two main relational databases realized with PostgreSQL, namely, the *MedTAG data* and the *Log data* databases. The former contains documents, entity concepts/labels, and the relations among them. The latter takes care of logging data such as user-provided information about issues with the documents to be annotated. The business layer controls the whole information flow as the information is displayed in the web interface and stored in the MedTAG database. It consists of two business units, the business logic, and the REST APIs end-point. The first one consists of Python routines and a controller that invokes the proper routine based on the received request. The second one is the back-end entry-point of MedTAG; it handles all the user requests from the web interface, invoking the business logic controller and returning its result to the front-end. The presentation layer provides the

³<http://examode.dei.unipd.it/ontology/>

MedTAG front-end; it consists of a web interface to navigate the documents, annotate them and download the annotations in different formats (i.e., BioC/JSON, BioC/XML, CSV, and JSON).

Figure 6.2 shows the technologies adopted for each logic layer: (i) the front-end interface built with React.js⁴, HTML5 and CSS3; (ii) the back-end for web API and services built with the Python web framework Django⁵; (iii) the *MedTAG data* relational database implemented using PostgreSQL.

Due to the multitude of architecture components, manually installing and configuring each one would be cumbersome and error-prone. To mitigate this, we provide a fast and reliable installation by distributing MedTAG as a docker container.

6.2.2 Installation and customization

Since MedTAG is provided as a Docker container, both *docker*⁶ and *docker-compose*⁷ are required. The detailed installation procedure is described at <https://github.com/MedTAG/medtag-core/tree/main#installation>. We can summarize the MedTAG installation in three steps:

1. Check the Docker daemon - i.e., `dockerd` - is up and running.
2. Download the `MedTAG_Dockerized`⁸ folder from the `medtag-core`⁹ repository, or clone it.
3. Open the `MedTAG_Dockerized` project folder and, on a new terminal session, type `docker-compose up`.

Once the installation process has been completed, MedTAG is available on your browser at `http://0.0.0.0:8000`. At this stage, users can access MedTAG only in *Test mode* – i.e., by using the pre-loaded documents. The pre-loaded documents for the test mode are taken from the histopathology domain because we chose this domain as a use case for introducing and testing MedTAG functionalities.

Users can log into MedTAG and test it with the preloaded medical reports using *Test* as username and password.

⁴<https://reactjs.org/>

⁵<https://www.djangoproject.com/>

⁶<https://docs.docker.com/engine/reference/commandline/docker/>

⁷<https://docs.docker.com/compose/>

⁸https://github.com/MedTAG/medtag-core/tree/main/MedTAG_Dockerized

⁹<https://github.com/MedTAG/medtag-core>

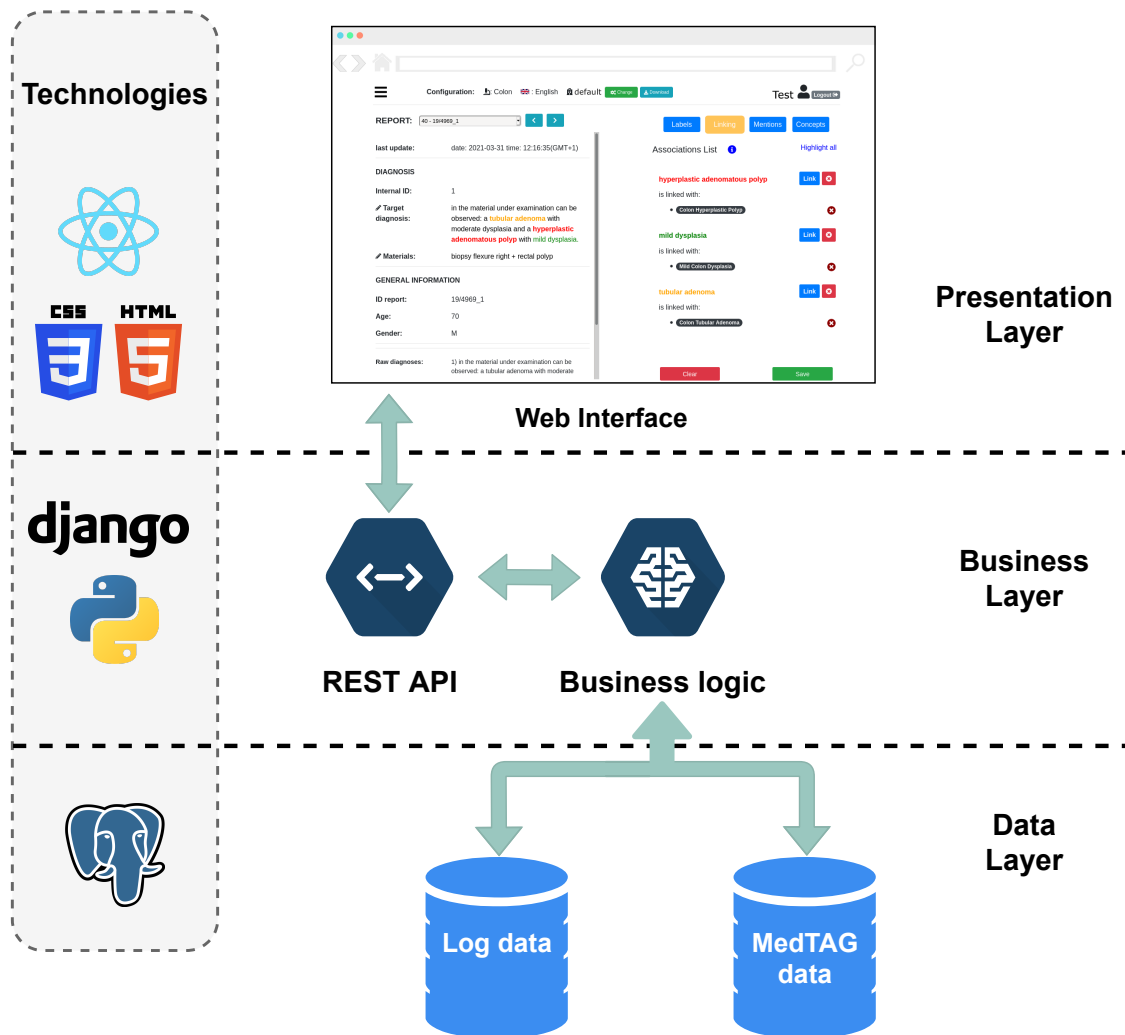


Fig. 6.2 MedTAG Architecture. The data layer comprises two relational databases, namely, *MedTAG data* and *Log data* to store all the information concerning the annotation process (e.g., concepts, labels, reports, users and their annotations) and logging data such as notifications of malformed clinical reports. The business layer comprises two business units: *Business logic* and *REST API* which jointly control the whole information flow from the front-end to the database and vice-versa. The presentation layer provides the MedTAG front-end, a web interface allowing users to annotate medical reports and download their ground truths.

To customize MedTAG, the users need to follow three steps: (i) open the menu and click on *Configure*, as shown in Figure 6.3; (ii) follow the instructions of the guided procedure – i.e., users are asked to provide both the admin user credentials and three CSV files: *concepts_file*, *labels_file* and *reports_file*, as shown in Figure 6.4. The users are provided with CSV templates and with examples containing real data to speed-up the data preparation procedure; (iii) choose which document fields to display and annotate as shown in



Fig. 6.3 MedTAG sidebar provides the *Configure* option, indicated by the orange arrow, to set up a new custom configuration.

Figure 6.5; the *Check* button activates the file compliance procedures that will produce some state messages in different colors to inform the user about whether the CSV files provided are well formatted or not. Figure 6.5 shows the configuration interface that allows the users to specify whether to use the ExaMode concepts (indicated with number two) and labels (indicated with number three) or to upload a new set of concepts from different ontologies. The latter are necessary in case users want to take advantage of automatic annotation features. In addition, users can choose whether to annotate custom documents or PubMed abstracts and titles. In the first case, users are required to provide all the reports to annotate as a CSV file, that is, `reports_file`. Then, users can choose the report fields to annotate at will. In the second case, users have to specify a list of PubMed identifiers as a CSV file. Then, users can annotate both abstract and title of each PubMed article specified.

The detailed customization procedure is available at <https://github.com/MedTAG/medtag-core#customize-medtag>.

6.2.3 User interface and interaction

The MedTAG web interface has been developed based on the positive feedback received from physicians and experts in the digital pathology domain where an instance of MedTAG - i.e., ExaTAG - has been released. Figure 6.6 shows the main MedTag web-interface for the annotation of medical documents or reports. On the top of the web page, there is the

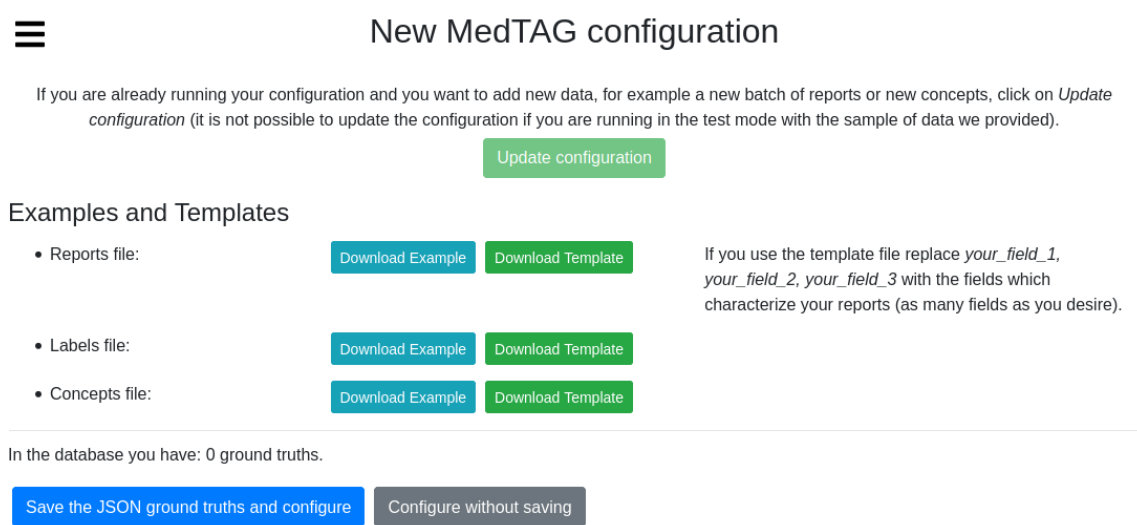


Fig. 6.4 MedTAG new configuration interface allows the user to save the current data before creating a new configuration. To guide the user in providing the new configuration files needed (i.e. reports/documents, labels and concepts), MedTAG provides both example and template files. In particular, users can use the example files to test MedTAG without providing their own data. Instead, users can use the template files as a reference to structure their own configuration files.

header section with the current MedTAG configuration: (i) the clinical case (e.g., Colon cancer); (ii) the report language (e.g., English); (iii) the hospital/institute which provided the report's dataset (e.g., "default _hospital" identifies the institute which provided the datasets of reports pre-loaded in MedTAG in test mode) and (iv) the annotation mode (i.e., manual or automatic) used for the annotation process. In addition, the menu button (left-side) and the user section (right-side) are included in the header as well. It is worth noting that when the automatic annotation mode is active the users visualize the automatic annotations generated by the built-in annotation module. Any user edit concerning the automatic annotations is also replicated in the user profile, available for further edits in manual annotation mode. The user section shows the current username along with the *Logout* button. Below the header, the interface body is divided into two sections: the diagnostic report and the annotation section. The first one (left-side) shows the information regarding the textual document, that in the case of a medical report may contain the diagnosis and the patient's information. Users can navigate between documents using either the keyboard arrows or the *next* and *previous* buttons. The annotation section (right side) shows the information concerning annotation labels, ontological concepts and the mentions identified in the selected document.

MedTag allows the users to use four different annotation types that can be activated alternatively by clicking on the corresponding buttons: (i) *Labels* is a form of document-

level annotation where the reports are classified into predefined categories, (ii) *Mentions* where the user selects words in the text of the reports, (iii) *Linking* where the identified mentions are linked to ontological concepts, and (iv) *Concepts*, another form of document-level annotation, where the reports are annotated with ontological concepts not strictly tied to specific mentions.

In Figure 6.6 the *Labels* action is activated. We can notice three selected labels: “Cancer”, “Adenomatous polyp - low grade dysplasia” and “Hyperplastic polyp”. The labels describe properties or attributes that apply to the whole document, such as the presence or the absence of cancer in the diagnosis of a clinical report. The set of labels used for the document-level annotation process, are provided by the user during the configuration phase, as previously discussed.

In Figure 6.7 the *Linking* action is activated. We can see three selected multi-word mentions in the text: “tubular adenoma”, “hyperplastic adenomatous polyp” and “mild dysplasia”. These mentions are linked to concepts taken from an histopathological ontology: (i) *hyperplastic adenomatous polyp* is linked to *Colon Hyperplastic Polyp*; (ii) *mild dysplasia* is linked to *Mild Colon Dysplasia*; and (iii) *tubular adenoma* is linked to *Colon Tubular Adenoma*.

The ontological concepts linked to the mentions can be selected via a drop-down menu (that, in turn, can be divided into semantic areas) or manually typed in a text field; in the latter case, the user is aided by auto-completion facilities.

To add a new mention, a user can click on any text token. After clicking on a text token, it gets highlighted with a new color, and the neighbor tokens turn highlighted as well, meaning that they could be selected as part of the current mention. All the mentions are highlighted with a different color in the document text and in the list of mentions for fast detection. The users can add, edit and delete the associations at will. Moreover, every time an action is performed, all the concerning information is asynchronously saved in the database; there is also manual saving via the *Save* button. Users can delete (after confirmation) all the annotations related to the current action button selected by clicking on the *Clear* button.

MedTAG enables the team members to collaborate during the annotation process. In particular, users can see anytime the annotations done by other team members for each clinical report by clicking on the button (3) of Figure 6.6. This feature is handy in case of annotation uncertainty (e.g., which concepts to associate to an identified mention). To attain high-quality annotations, users can take advantage of the expertise and work other team members have previously done. In addition, users can visualize the automatic annotations made by the robot user - i.e., the automatic annotation module SKET - by clicking on button (2) of Figure 6.6. Moreover, the users can consult and edit the automatic annotations so that new edits are

automatically copied in the user profiles for further modifications in manual annotation mode. Hence, users can take advantage of automated annotation facilities to reduce the annotation workload. Moreover, the admin can oversee the overall annotation process from the *Team members' statistics* section of the control panel. This section provides the admin user an overview of the annotation work carried out by each team member, providing information such as the number and the percentage of annotated reports for each use-case and annotation type. Hence, the admin can make decisions to coordinate the work of team members and keep track of the advancements in the annotation process.

Finally, users can download their annotations in different formats (i.e., BioC/JSON, BioC/XML, CSV and JSON), by clicking on the *Download* button.

Overall, a detailed graphical tutorial is always available to the users to learn how to use MedTAG; the *Tutorial* link is provided in the sidebar, as shown in Figure 6.8.

6.2.4 MedTAG control panel for statistics and Inter-Annotator Agreement (IAA)

MedTAG provides a unified interface that allows the admin user to access the annotation statistics (e.g., the number of users that annotated each report) and access the information concerning IAA for each report. It is worth noting that only the admin user can consult the statistics concerning the overall annotation process. Instead, other members can only access their statistics in the dedicated menu section *My statistics*. Figure 6.9 shows the control panel information organized in a dynamic table, where the admin can search, access, and filter the reports according to a selection of columns filters. Moreover, the admin can choose anytime which columns to show by clicking on the *Columns* button. The last column provides the following action buttons:

- **Delete:** enables the admin user to remove the corresponding reports.
- **Download:** allows the admin user to download either the original annotations or the ones resulting from the majority vote procedure. Moreover, the admin user can also download the automatic annotations generated by the built-in prediction system. Several download options are provided, including the output file format, the annotation mode (i.e., manual or automatic) and type (i.e., *Labels*, *Concepts*, *Mentions* and *Linking*).
- **Inspect statistics:** allows the admin user to consult the report information as well as the statistics concerning the annotations of the selected report. The annotation statistics regards all the annotation types provided in MedTAG (i.e., *Labels*, *Concepts*, *Mentions*

and *Linking*) and include the number of users that identified each label, mention or concept in the report. In addition to user annotations, the interface shows the automatic annotations highlighted in blue produced by the built-in prediction system.

- **IAA and majority vote:** allows the admin user to access the information concerning IAA for each report. Figure 6.12 shows the pop-up modal by which the admin can specify the options for the majority vote procedure. The admin can choose from a drop-down menu which team members (annotators) to consider, as well as the annotation mode and type. The procedure returns only the annotations that achieved more than fifty percent of agreement among the annotators considered. Then, the admin can download the annotations resulting from the majority vote procedure, as shown in Figure 6.13.

It is worth mentioning, that the online instance of MedTAG¹⁰ integrates also in the control panel the support for the ExaNet visualization interface for the graph representation of clinical reports, as described in Section 4.5.3 of Chapter 4.

Figure 6.10 shows the *Team members' statistics* section of the control panel, which provides the information about the advancements in the annotation work for each team member. Access to this section is restricted to the admin user. The admin can overview the annotation work carried out for each use-case and annotation type using ring charts providing information about the number of annotated reports and the corresponding percentage out of the total. Moreover, Figure 6.11 shows that team members can keep track of their work by consulting the section *My statistics*, where other ring charts visually summarize the personal annotation statistics.

¹⁰<http://w3id.org/medtag> access provided with credentials: demo/demo

Configure MedTAG with your data

PubMed IDs (Mandatory one between PubMed and reports files)

Insert here the .CSV file with the PubMed IDs of the articles you want to annotate. *Abstract* and *Title* sections will be annotable.

Example

Nessun file selezionato

Reports (Mandatory one between PubMed and reports files)

Insert here the .CSV file with the reports to annotate.

Example

clinical_reports.csv

Reports' fields (It is mandatory to set at least one field to display. It is optional to set at least one field to annotate)

Below you can find all the fields which characterize your reports. For each key, you have to decide if you want to display, hide, or display and annotate the value corresponding to that key. Remember that to perform mentions annotation, you must select at least one field checking the button *Display and Annotate*

| | | | |
|-------------------------|----------------------------|--|---|
| age | <input type="radio"/> Hide | <input checked="" type="radio"/> Display | <input type="radio"/> Display and Annotate |
| target_diagnosis | <input type="radio"/> Hide | <input type="radio"/> Display | <input checked="" type="radio"/> Display and Annotate |
| date_of_hospitalization | <input type="radio"/> Hide | <input checked="" type="radio"/> Display | <input type="radio"/> Display and Annotate |
| clinical_notes | <input type="radio"/> Hide | <input type="radio"/> Display | <input checked="" type="radio"/> Display and Annotate |
| clinical_studies | <input type="radio"/> Hide | <input type="radio"/> Display | <input checked="" type="radio"/> Display and Annotate |
| treatments | <input type="radio"/> Hide | <input type="radio"/> Display | <input checked="" type="radio"/> Display and Annotate |

1

Files for CONCEPTS IDENTIFICATION AND LINKING (Optional)

2

If you want to perform concepts identification and linking you must provide the concepts.

You inserted reports for: **Colon**; EXAMODE ontology is available for these secases. **You can use the EXAMODE ontology concepts OR upload your own ones.**

Insert here the .CSV file with the concepts.

Example

Nessun file selezionato

Files for LABELS ANNOTATION (Optional)

3

If you want to perform labels annotation you must provide the labels related to the use cases you are interested in.

You inserted reports for: **Colon**; EXAMODE labels are available for these usecases. **You can insert the EXAMODE labels or upload your own ones.**

Insert here the .CSV file with the annotation labels.

Example

Nessun file selezionato

Fig. 6.5 MedTAG main interface for data configuration. Users can provide their own CSV files for the reports/documents to annotate and the concepts and labels to use for the annotation process. Moreover, MedTAG detects automatically the document fields and allows users to specify which of them to annotate and/or display in the interface, as shown in the orange box (1).

The screenshot displays the MedTAG main interface in test mode. At the top, the configuration is set to: Colon, English, default_hospital, Manual, and Change. A user profile 'Test' is visible with a Logout button. Below the configuration, the 'REPORT:' section shows '1 - report_1' with navigation arrows. The 'Reports' order is set to 'Lexicographical'. The 'last update:' is 'date: 2021-08-28 time: 18:01:19(GMT+1)' and the 'ID:' is 'report_1'. The 'materials:' section contains the text 'octopus sigma'. The 'target_diagnosis:' section contains the text 'in the materials under examination can be observed a tubular adenoma with moderate dysplasia and a hyperplastic adenomatous polyp with mild dysplasia.' On the right, the 'Labels' annotation mode is active, and three labels are checked: Cancer, Adenomatous polyp - low grade dysplasia, and Hyperplastic polyp. The 'Non-informative' label is unchecked. At the bottom, there are 'Clear', 'Save', and three numbered buttons (1, 2, 3).

Fig. 6.6 MedTAG main interface in test mode with default configuration: clinical case set to “Colon cancer”, reports’ language set to English, reports’ institute/hospital set to “default_hospital” (the real name has been anonymized) and the annotation mode set to manual. The annotation type active is the *Labels* one. Three labels have been checked: (i) *Cancer*; (ii) *Adenomatous polyp - low grade dysplasia* and (iii) *Hyperplastic polyp*.

Fig. 6.7 MedTAG main interface in test mode with default configuration: clinical case set to “Colon cancer”, reports’ language set to English, reports’ institute/hospital set to “default_hospital” (the real name has been anonymized) and the annotation mode set to manual. The annotation type active is the *Linking* one. Three mentions have been identified and linked to the corresponding concepts: (i) *hyperplastic adenomatous polyp* is linked to *Colon Hyperplastic Polyp*; (ii) *mild dysplasia* is linked to *Mild Colon Dysplasia*; and (iii) *tubular adenoma* is linked to *Colon Tubular Adenoma*.

Fig. 6.8 MedTAG tutorial interface. To reach the tutorial section, users can click on the *Tutorial* link in the sidebar, indicated by the orange arrow.

REPORTS' OVERVIEW

In this section you can check how many reports have been annotated so far for each use case. You can also delete one or more reports if you want.

Columns

| | id_report | language | usecase | institute | annotations ↓ | |
|--------------------------|--|--|--|--|---|--|
| | <input type="text" value="Filter..."/> | <input type="text" value="Filter..."/> | <input type="text" value="Filter..."/> | <input type="text" value="Filter..."/> | <input type="text" value="Filter..."/> | Search... <input type="text" value="Q"/> |
| <input type="checkbox"/> | e3aecee809d8ab4... | English | Colon | default_hospital | 11 i | 🗑️ ⬇️ 👁️ 👤 |
| <input type="checkbox"/> | 8d337edee1d0283... | English | Colon | default_hospital | 9 i | 🗑️ ⬇️ 👁️ 👤 |
| <input type="checkbox"/> | 024904af0078ed8... | English | Colon | default_hospital | 7 i | 🗑️ ⬇️ 👁️ 👤 |
| <input type="checkbox"/> | 38165a8cb2471a1... | English | Colon | default_hospital | 6 i | 🗑️ ⬇️ 👁️ 👤 |
| <input type="checkbox"/> | 11ff311f98e76f215... | English | Colon | default_hospital | 6 i | 🗑️ ⬇️ 👁️ 👤 |

Rows per page: 5 ▾ 1-5 of 10 < 1 2 >

Fig. 6.9 MedTAG control panel concerning the reports' statistics. The reports are organized in an interactive table enabling the admin user to: (i) access report data; (ii) delete one or more reports; (iii) download report data including manual and automatic annotations and (iv) access the information concerning IAA and manage the majority vote procedure.

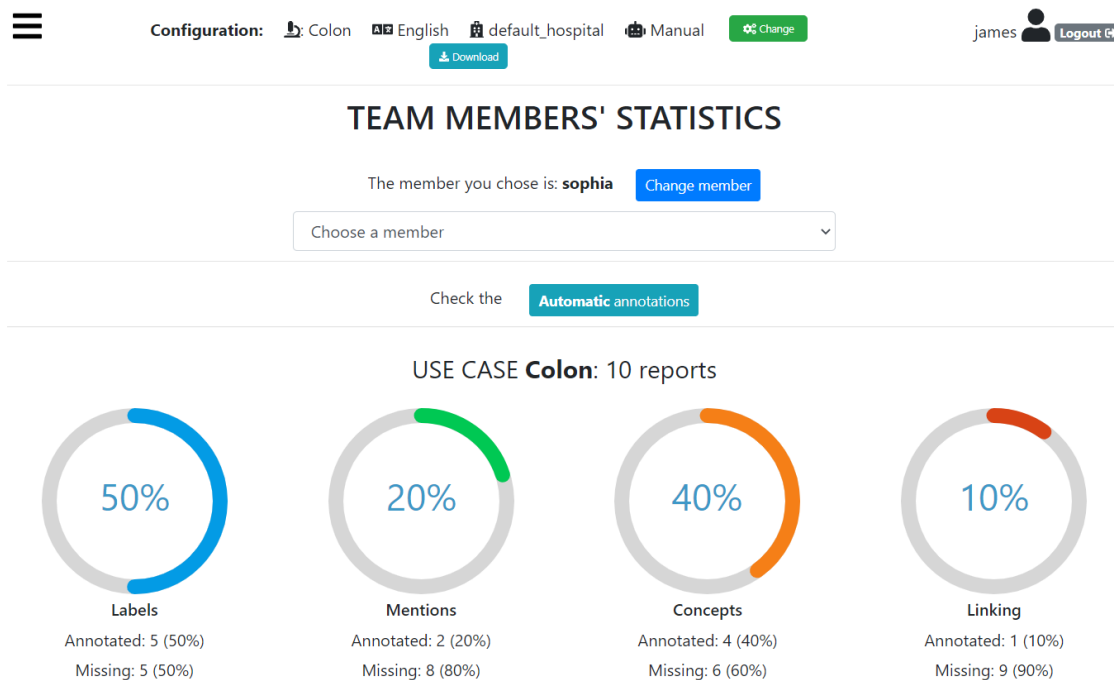


Fig. 6.10 MedTAG control panel concerning the team members' statistics. The ring charts report the annotation work carried out by each team member, so that the admin can keep track of the advancements regarding the whole annotation process.

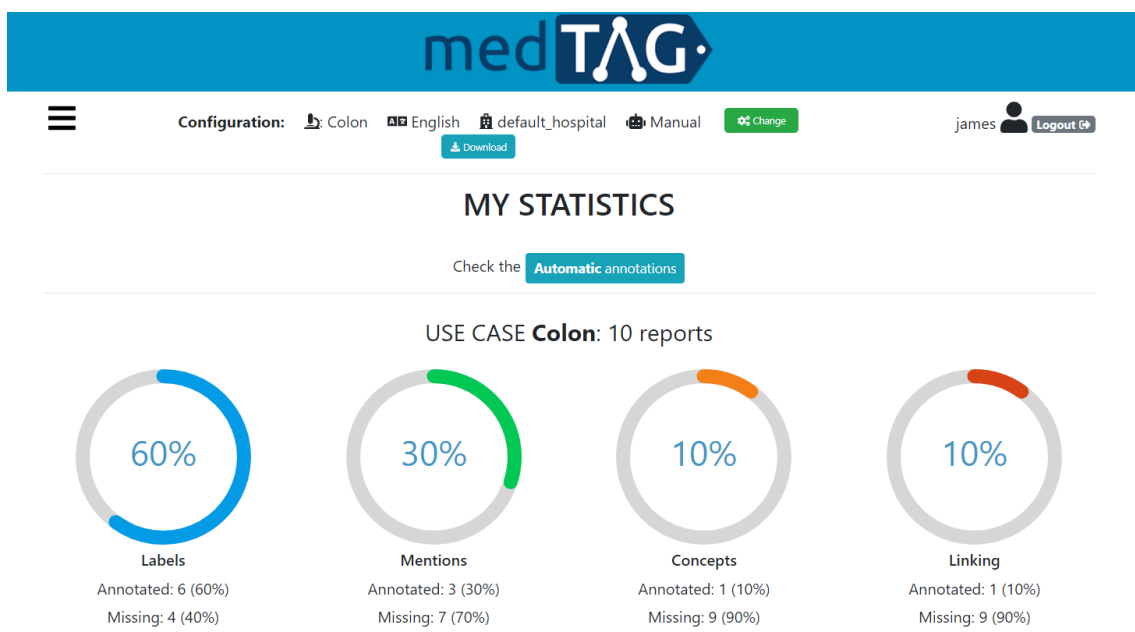


Fig. 6.11 MedTAG *My Statistics* panel, providing information about the user annotation work in terms of documents annotated for each use-case.

Majority vote ground-truth REPORT e3aecee809d8ab4779d869658ad2d5d8 ×

| | | | |
|--------------------------|---|--|---|
| target_diagnosis: | tubular adenoma with low-grade dysplasia (mild) glandular epithelium. | Select an Action | <input type="text" value="Labels"/> |
| materials: | octopus sigma | Select an Annotation Mode | <input type="text" value="Manual"/> |
| age: | 62 | Select at least 2 users you want to consider to create the ground-truth based on majority vote | <input type="text" value="Select a user..."/> |
| report_id: | 33360800475263_19/9779_1 | The users the majority vote ground-truth will be based on are: | <ul style="list-style-type: none"> • james ✕ • sophia ✕ |
| raw_diagnoses: | tubular adenoma with low-grade dysplasia (mild) glandular epithelium. | <input type="button" value="Confirm"/> | |

Fig. 6.12 MedTAG majority vote interface. The admin can overview the selected report and choose the options of interest for the majority vote procedure, including: (i) the annotation mode; (ii) the annotation type and (iii) the team members (annotators) to consider.

Majority vote ground-truth REPORT 024904af0078ed89be428dd5868803ce ×

| | | |
|--------------------------|---|---|
| materials: | right colon polyp | <i>Below you can find the labels belonging to the majority vote ground truth. The labels annotated by the algorithm are highlighted in blue</i> |
| target_diagnosis: | right colon polyps: tubular adenoma with mild dysplasia glandular epithelium. | 2 manual annotations |
| age: | 63 | <ul style="list-style-type: none"> • Cancer This label occurred in: 2 / 2 manual annotations - users: james, sophia • Adenomatous polyp - low grade dysplasia This label occurred in: 2 / 2 manual annotations - users: james, sophia |
| codint: | 19/10077 | <input type="button" value="Change configuration"/> <input type="button" value="Download"/> |
| internalid: | 2 | |
| report_id_hashed: | 024904af0078ed89be428dd5868803ce | |

Fig. 6.13 MedTAG majority vote output for the *Labels* annotation type. The admin can visualize the annotations resulting from the majority vote procedure, together with the corresponding authors. In addition, the admin can download the annotations or change the current majority vote configuration.

| Use-case Language | Cervix cancer | Colon cancer | Lung cancer | Total |
|------------------------------------|---------------|--------------|-------------|--------------|
| Dutch | - | 889 | - | 889 |
| English | 2,361 | - | - | 2,361 |
| Italian | 1,828 | 239 | 2,005 | 4,072 |
| Total | 4,189 | 1,128 | 2,005 | 7,322 |

Table 6.1 Number of diagnostic reports annotated per language and use-case.

6.3 Results and Discussion

MedTAG has been used to annotate diagnostic reports to produce both training and test annotated data. In particular, a specific instance of MedTAG for the ExaMode H2020 EU project¹¹ has been used to generate more than seven thousand annotated reports and more than eight thousand annotations overall. This instance of MedTAG¹² is tailored for the histopathology domain. By connecting to the online instance of MedTAG, users can try its functionalities with real (anonymized) clinical reports from the digital pathology domain without downloading and installing it. The latter has been customized to meet the needs of the physicians and experts concerning the cancer use-cases of the ExaMode project (i.e., cervix, colon, and lung cancer). Pathologists and experts have used MedTAG to annotate the diagnostic reports from two healthcare institutions, namely, AOEC and RUMC. For the time being, ten annotators between physicians and experts have annotated thousands of medical reports in three languages (Dutch, English, and Italian). Table 6.1 reports some statistics about the manual annotation process conducted so far. Instead, Table 6.2 shows the number of automatic annotations done by SKET (i.e. the automatic annotation module) for each annotation type and use-case.

6.3.1 Biomedical annotation tools comparison

The biomedical annotation tools selected for the comparison, according to the five requirements presented above, are BioQRator [177], ezTag [178], MyMiner [254], tagtog [57] and TeamTat [141]. Moreover, we also consider brat [276] and INCEpTION [162] because they are used by the biomedical community in some settings. Figure 6.1 shows that several of

¹¹<https://www.examode.eu/>

¹²<http://w3id.org/exatag/> access granted with credentials: demo/demo

| Annotation type \ Use-case | Cervix cancer | Colon cancer | Lung cancer | Total |
|----------------------------|---------------|---------------|--------------|----------------|
| | Labels | 16,033 | 9,309 | 2,066 |
| Concepts | 12,936 | 11,932 | 2,336 | 27,204 |
| Mentions | 12,070 | 10,926 | 2,336 | 25,332 |
| Linking | 12,936 | 11,932 | 2,336 | 27,204 |
| Total | 53,975 | 44,099 | 9,074 | 107,148 |

Table 6.2 Number of labels, concepts, mentions and links (mention - concept) automatically annotated per use-case.

the considered tools lack T6 (license allowance to modify and redistribute the tool) and F1 (support for overlapping mentions). Almost half of the tools (three out of seven) lacks T2 (availability of the source code), F2 (support for document-level annotation), F11 (support for IAA), F12 (data privacy) and F13 (multilingual support). In contrast, MedTAG satisfies: (T6) MedTAG is provided through the MIT license, permitting the use, modification and distribution of the tool free of charge; (T2) the source code of MedTAG is publicly available¹³; (F12) MedTAG enables the utilization of data on a local system without any sharing with external servers, thus ensuring data privacy; (F2) MedTAG supports two types of document-level annotations, namely, label and concept annotations. The label annotation feature allows the user to tag a document according to a customizable set of labels.

The concept annotation feature allows the users to mark a document as pertinent for one or more ontological concepts. Users can leverage the auto-complete feature to search for the relevant concepts to assign. Note that, as analyzed in [219], only a tiny minority of annotation tools on the market fully support document-level annotation. For instance, MyMiner supports document annotation, but due to limits in the customization process, the annotators must re-define the labels every time new documents are added to the system. Moreover, most of the other annotation tools allow the users to provide document-level annotations only using some workaround such as zero-width annotations and annotations of pre-defined placeholders placed at the beginning or at the end of the document to annotate. However, this practice is additional overhead that further complicates and slows down the annotation process; (F10) MedTAG supports users and roles.

¹³<https://github.com/MedTAG/medtag-core/>

MedTAG is distributed as a Docker container, thus it can seamlessly be deployed in a local environment or a remote cloud solution. Therefore, the administrator can choose whether to grant MedTAG access to annotators only within a local network or “worldwide”; (F13) MedTAG provides multilingual support. It allows the users to annotate the same document (same document identifier) in different languages.

When dealing with thousands of biomedical documents to annotate, time is crucial. Hence, web-based annotation tools provided with the modality of Software as a service (SaaS) are not necessarily the best solution in this context due to possible network delays. For instance, network delays might be experienced when uploading high volumes of data. A local installation can avoid network delays and operate better in the case of large corpora to be annotated. However, several annotation tools present difficulties about the installation process, such as lack of documentation or dependency issues, as stated in [219]. For instance, tagtog can be installed locally only in its commercial version, whereas ezTAG and TeamTat can be installed free of charge. Still, the procedure could be quite complex for the not technology-savvy; ezTag and TeamTat require the user to install and configure some frameworks and software packages manually (e.g., Ruby, Rails, and MySQL) as prerequisites. In contrast, MedTAG provides an easy installation procedure; it requires the user only to execute the `docker -compose up` command (provided that Docker is installed). The MedTAG installation procedure is available and thoroughly described online¹⁴.

Note that TeamTat provides high-level inter-annotator agreement statistics since the project manager can calculate the agreement among annotators. In contrast, MedTAG provides fine-grained statistics by allowing the users to access the information concerning IAA for each report and to download the annotations resulting from the majority vote procedure. For this reason, we consider the criterion (F11) partially satisfied by TeamTat (see Figure 6.1). TeamTat supports the annotation of documents compliant with the Unicode Standard, meaning that documents with special characters are visualized and annotated correctly. However, TeamTat does not provide additional facilities to manage, organize and search documents according to their languages (unless using a specific workaround such as creating language-specific document collections). In contrast, MedTAG allows the users to organize and filter documents according to their languages out-of-the-box; no additional configuration or effort is required. For this reason, we consider (F13) partially satisfied by TeamTat and entirely by MedTAG.

Several biomedical tools let the users upload biomedical documents by using tool-specific procedures and formats. For instance, BioQRator and ezTag only accept medical documents in BioC format. Despite BioC being a well-established file format in the biomedical domain,

¹⁴<https://github.com/MedTAG/medtag-core/tree/main#installation>

adopting it as the only valid format poses hindrances to annotating biomedical documents in other formats. For instance, narrative clinical reports are usually available in an unstructured format, such as plain text. Thus, to use them in BioQRator and ezTag, they need to be converted in BioC format in advance. In contrast, MedTAG allows the users to provide the medical documents as customizable CSV files, letting the user decide and set up which fields to display and annotate. This feature turns out to be helpful, especially when dealing with high volumes of long biomedical documents, where changing data format is not always a feasible or reasonable operation for annotators.

For what concerns the general-purpose annotation tools - i.e., brat and INCEpTION - they are substantially different from MedTAG. For instance, brat [276] is a well-established web-based annotation tool specifically suited for entity and relationship annotations. It has been extensively used for the annotation of biomedical projects [205, 312, 165, 49, 294]. Brat is not available for online use; it requires to be installed locally in a UNIX-like environment. Hence, the procedure could be complex for not technology-savvy users, as stated in [219]. In contrast, MedTAG is provided as a portable and easy-to-run Docker container. Users can configure brat via plain-text schema configuration. Moreover, users can import raw documents and export the annotations in plain-text format. Conversely, MedTAG provides support for several file formats such as BioC/JSON and BioC/XML, which are standard formats for the annotations in the biomedical domain. In addition, MedTAG also provides several other features that brat currently does not support, as (T3) online availability; (F2) support for document-level annotation; and, (F6) integration with PubMed.

INCEpTION is another general-purpose tool used also by the biomedical community [286, 51, 308, 260]. It is an open-source web-based annotation tool both available online and for local installation. For the local use, it requires Java, as described in the online documentation¹⁵. Figure 6.1 shows that INCEpTION covers most of the considered criteria (21 over 22). For instance, it provides active learning facilities to improve suggestions over time in a human-in-the-loop environment and a comprehensive set of features to adapt to different annotation scenarios. However, the INCEpTION interface provides several functionalities not specifically designed for the biomedical domain, which can be perceived as redundant by the biomedical community. Moreover, to achieve annotation flexibility, INCEpTION introduces additional levels of abstraction that increase the complexity of the annotation task, thus resulting potentially not within reach of not technologically-savvy users. For instance, document-level annotation is, at the time of writing, an experimental feature that needs to be explicitly enabled by manually editing a settings file. Moreover, to enable document-level annotations, the user must define a "Document metadata" annotation layer

¹⁵<https://inception-project.github.io/documentation/>

in the project settings. For such a reason, we judge the criterion (F2) as partially satisfied by INCEpTION (see Figure 6.1). In contrast, MedTAG provides document-level annotation facilities off-the-shelf since no additional configuration is required. In addition, MedTAG provides native PubMed integration facilities - i.e., users can annotate PubMed titles and abstracts – whereas INCEpTION employs a third-party tool (i.e., PubAnnotation [158]) to retrieve the documents to annotate from PubMed Central, as stated in [79].

6.3.2 Quantitative comparison of biomedical annotation tools

To quantitatively assess MedTAG performance, we conducted several experiments designed to evaluate MedTAG concerning two annotation tasks: document-level annotation and mention identification. The first one concerns annotations that refer to the whole document, such as labels describing the overall document content (e.g., the “cancer” label may indicate whether a clinical report suggests a cancer condition). Instead, mention identification regards entity mentions identified in the textual content of a document. The annotation tools are compared regarding the number of actions and elapsed time required to complete the overall annotation process. To the best of our knowledge, this is the first available quantitative evaluation of biomedical annotation tools. The analysis we conducted considers a set of web-based biomedical annotation tools - i.e., ezTag, MedTAG, MyMiner, tagtog and TeamTat - evaluated on a sample of one hundred documents, randomly chosen from a real dataset concerning the digital pathology domain (i.e., clinical reports related to colon cancer). For the comparison, we consider only web-based publicly available tools since many biomedical annotation tools are not available for local installation or are not easy to install for not technologically savvy end-users. It is worth noting that our analysis does not focus on usability and HCI aspects (e.g., User Experience (UX) and interface look and feel) that may vary subjectively. Nevertheless, the latter are essential points that should be treated with specific user studies. In contrast, we focused on the annotation work regarding the number of actions (e.g., mouse clicks and keys pressed) and elapsed time to achieve the same annotations in different tools. To perform a fair comparison, we used automatic agents (web robots) designed to annotate using the same annotation speed - i.e., exact time to simulate a mouse click or a key pressed for each annotation tool. The automatic agents have been implemented using the Python Web automation library Selenium¹⁶. The source code of the automated agents used for the experiments is publicly available¹⁷. Since the automatic agents are generally way faster than any human annotator, we introduced a short delay (about two hundred milliseconds) between two consecutive actions, which is also necessary to avoid

¹⁶<https://www.selenium.dev/>

¹⁷<https://github.com/MedTAG/medtag-core/tree/main/benchmark>

overloading the server with too many requests.

Table 6.3 and Table 6.4 show the experimental results in terms of the number of actions and elapsed time for annotating one hundred documents. The elapsed time for each tool was recorded forty times; the resulting mean value and standard deviation are reported in the tables. Table 6.3 shows the performance analysis concerning the document-level annotation task. For the latter task, we considered three tools - i.e., MedTAG, MyMiner, and tagtog - since ezTag does not support document-level annotation, whereas TeamTat provides different document-level annotation facilities. In particular, TeamTat allows us to annotate entities in different documents and then to create relationships between them; this is different from the functional criterion (F2), indicating whether the users can specify labels at the document-level. For this reason, we consider the latter criterion only partially satisfied. The experiments concerning document-level annotation consist of assigning one label for each document to annotate. The labels, mentions, and documents used for testing are publicly available¹⁸ for reproducibility purposes. Table 6.3 shows that MyMiner requires fewer actions than other tools to achieve the same annotations, whereas MedTAG turns out to be the fastest tool in terms of elapsed time. Nevertheless, the time difference between MyMiner and MedTAG is about ten seconds, which is negligible considering different server response times. According to Table 6.3, tagtog requires more actions and time than other tools to complete the annotation process. However, these results are motivated considering that tagtog is one of the most flexible annotation tools and allows to specify whether a document label is *true*, *false* or *unknown*. To this aim, tagtog allows users to choose the correct value from a drop-down menu for a document label. Thus, the users have to click on the drop-down menu two times: the first one to open the pop-up menu and the second for the value selection. In contrast, MyMiner and MedTAG require just one click on a checkbox, based on the assumption that a label may apply for a document or not (the *unknown* state is not allowed). Moreover, MyMiner requires fewer actions than MedTAG to complete the annotation process since it automatically moves on to the following document to annotate after the user selection. However, MyMiner does not allow to specify more labels for a document. In contrast, MedTAG goes beyond this limitation and allows to specify of several labels at the same time for each document. Thus, users can decide on their own when to move on to the next document to annotate.

Table 6.4 shows the performance analysis concerning the mention identification task. The experiments concerning mention identification consist of identifying entity mentions within the documents' textual content. To this aim, we used a set of pre-identified mentions for each of the documents considered. According to Table 6.4, the tools with the lowest number

¹⁸<https://github.com/MedTAG/medtag-core/tree/main/benchmark/datasets>

of actions required are ezTag and TeamTat, whereas MyMiner and MedTAG are the fastest tools in terms of elapsed time. TeamTat and ezTag achieved comparable performance since they are similar in terms of functionalities provided. The experimental results show that MyMiner is the fastest tool in terms of elapsed time. MyMiner provides a neat interface that requires low network resources and bandwidth to work, thus reducing loading time and making the annotation process faster. However, it lacks several functionalities such as (i) support for users and teams, (ii) availability for local installation, and (iii) data privacy (upload of the documents to annotate is required) that could be relevant for the needs of the biomedical community. In contrast, MedTAG is designed to be portable (i.e., local installation is available) and flexible; it provides annotation facilities, such as schema configuration, that allow users to customize the annotation experience. Moreover, MedTAG is faster than other tools, even if it requires more actions. A possible explanation could be the different mention annotation functionality. Indeed, most of the annotation tools allow identifying entity mentions within the text using drag-and-drop facilities. In contrast, MedTAG enables users to annotate mentions with a single click on each text token. The latter facility turns out to be convenient in short mentions, whereas drag-and-drop is more suitable in the case of long ones.

To summarize, we quantitatively compared a set of web-based biomedical annotation tools on two tasks: document-level annotation (one label per document) and mention identification. We conducted several experiments to assess each annotation tool regarding the number of actions and elapsed time required to complete the overall annotation process. From the experimental results emerge that, depending on the task, some tools perform better than others. Despite the higher number of actions required to complete the annotation process, MedTAG turns out to be faster than other tools, especially for the document-level annotation task.

Finally, it is worth noting that the present study focuses on evaluating a set of biomedical annotation tools only on physical aspects such as the number of actions and elapsed time required to annotate all the documents considered. Hence, we do not consider several critical human-centric factors (e.g., UX and HCI) that should be investigated in dedicated usability studies.

| Tool | Number of actions | Elapsed time in seconds (mean) | Standard deviation in seconds |
|-------------|--------------------------|---------------------------------------|--------------------------------------|
| MedTAG | 200 | 46.840 | 0.803 |
| MyMiner | 100 | 56.677 | 0.416 |
| tagtog | 400 | 205.740 | 5.471 |

Table 6.3 Document-level annotation performance analysis in terms of number of actions (e.g. mouse clicks and keys pressed) and elapsed time required to complete the whole annotation process.

| Tool | Number of actions | Elapsed time in seconds (mean) | Standard deviation in seconds |
|-------------|--------------------------|---------------------------------------|--------------------------------------|
| MedTAG | 519 | 159.337 | 0.479 |
| ezTag | 307 | 260.340 | 0.576 |
| MyMiner | 414 | 114.390 | 1.507 |
| tagtog | 404 | 304.692 | 10.067 |
| TeamTat | 307 | 271.577 | 1.542 |

Table 6.4 Mention-level annotation performance analysis in terms of number of actions (e.g. mouse clicks and keys pressed) and elapsed time required to complete the whole annotation process.

6.4 Conclusions

In this chapter, we presented MedTAG, a customizable, portable, collaborative, web-based biomedical annotation tool. We described an instance of MedTAG adopted in the histopathology domain, where MedTAG has been used by physicians to annotate more than seven thousand clinical reports in three languages (Dutch, English and Italian), from two health-care institutions. MedTAG is provided as a docker container to make it distributable, platform-independent, and easy to install/deploy. We designed MedTAG according to the five requirements (i.e. available, distributable, installable, workable, and schematic) defined in a recent extensive review of manual annotation tool [219]. Moreover, MedTAG satisfies 20 over 22 criteria defined in the same study.

The key points of strength of MedTAG are: (i) fast and easy installation because only one command is necessary to install it in less than 10 minutes on a current notebook; (ii) cross-platform support since MedTAG can be installed in every platform supporting docker; (iii) a

collaborative web-based platform supporting users and roles; (iv) broad data formats support including BioC/JSON, BioC/XML, CSV, and JSON; (v) support for schema configuration where the users provide the documents to annotate by using custom CSV files and can decide which fields to display and annotate.

6.4.1 Limitations and Future Work

MedTAG, as the name suggests, is a customizable annotation tool for the biomedical domain; Thus, it is not intended for general-purpose annotations since the users could not exploit domain-specific features such as automatic annotation. It is worth noting that the automatic annotation is currently provided for three cancer use-cases (i.e., cervix, colon, and lung cancer). Nevertheless, we plan to extend the automatic annotation support for other use-cases according to the needs of the biomedical community. The integration of SKET as an automated annotation tool shows the flexibility of MedTAG and how annotation automation may work with MedTAG. Another limitation concerns the file format of the input documents since MedTAG currently supports only plain-text documents. We believe that PDF annotation would be particularly useful, especially when dealing with scientific paper annotation. Hence, we plan to include this feature in the future version of MedTAG. For the time being, MedTAG does not support both overlapping mentions (also known as multi-label annotations) and relationship annotations that are left as future work. Indeed, even if MedTAG allows assigning multiple concept labels to the same mention, it is currently impossible to annotate any sub-mention. Finally, it is worth noting that even if MedTAG is designed for the biomedical domain, it could also be used for general-purpose annotation as long as a suitable schema configuration is provided. As future work, we plan to enrich MedTAG by adding (i) the support for overlapping mentions; (ii) the support for relationship annotations; (iii) the support for active learning capabilities; (iv) the support for PDF annotation; (v) the automatic annotation support for other use-cases relevant for the biomedical community. Thereby, we aim to improve MedTAG according to the biomedical community's needs and foster further developments in this field.

Chapter 7

Knowledge exploration

7.1 Introduction

The scientific world is swiftly becoming data-centric, embracing the principles of the so-called *fourth paradigm of science* [126]. Data are at the center of scientific discovery as well as of scholarship and scholarly communication [40]. The growing role of data is also witnessed by the ever-increasing importance of data science and related research fields concerning the search [60], provenance [62], citation [267], re-use [306], and exploration [241] of data.

There is no “one size fits all” solution when it comes to data search, access, and re-use given the heterogeneity of data representations and models, interoperability issues, and domain-dependent requirements. In the context of scientific data, the *nanopublication model* has been proposed to target some of these issues [117]. Nanopublications exploit the LOD principles [36] to represent scientific facts (*assertions* hereafter) as self-consistent, independent and machine-readable information tokens. A repository of nanopublications is to be thought of as an open and interconnected knowledge graph seamlessly integrated with the supporting scientific literature. Nanopublications can be used to support scientific claims, to explore scientific knowledge by exploiting machine intelligence and as entry points to scientific databases. Hence, this model has been embraced by several scientific fields, especially in the Life Science domain, leading to the creation of more than ten million openly available nanopublications [174].

From the technical viewpoint, a nanopublication is a Resource Description Framework (RDF) graph built around an assertion represented as a triple (subject-predicate-object) and usually extracted, manually or automatically, from a scientific publication. The nanopublication enriches the assertion with provenance and publication information. The RDF

representation format enables interoperability and thus the re-use of data, whereas provenance and publication information eases authorship recognition, credit distribution, and citation.

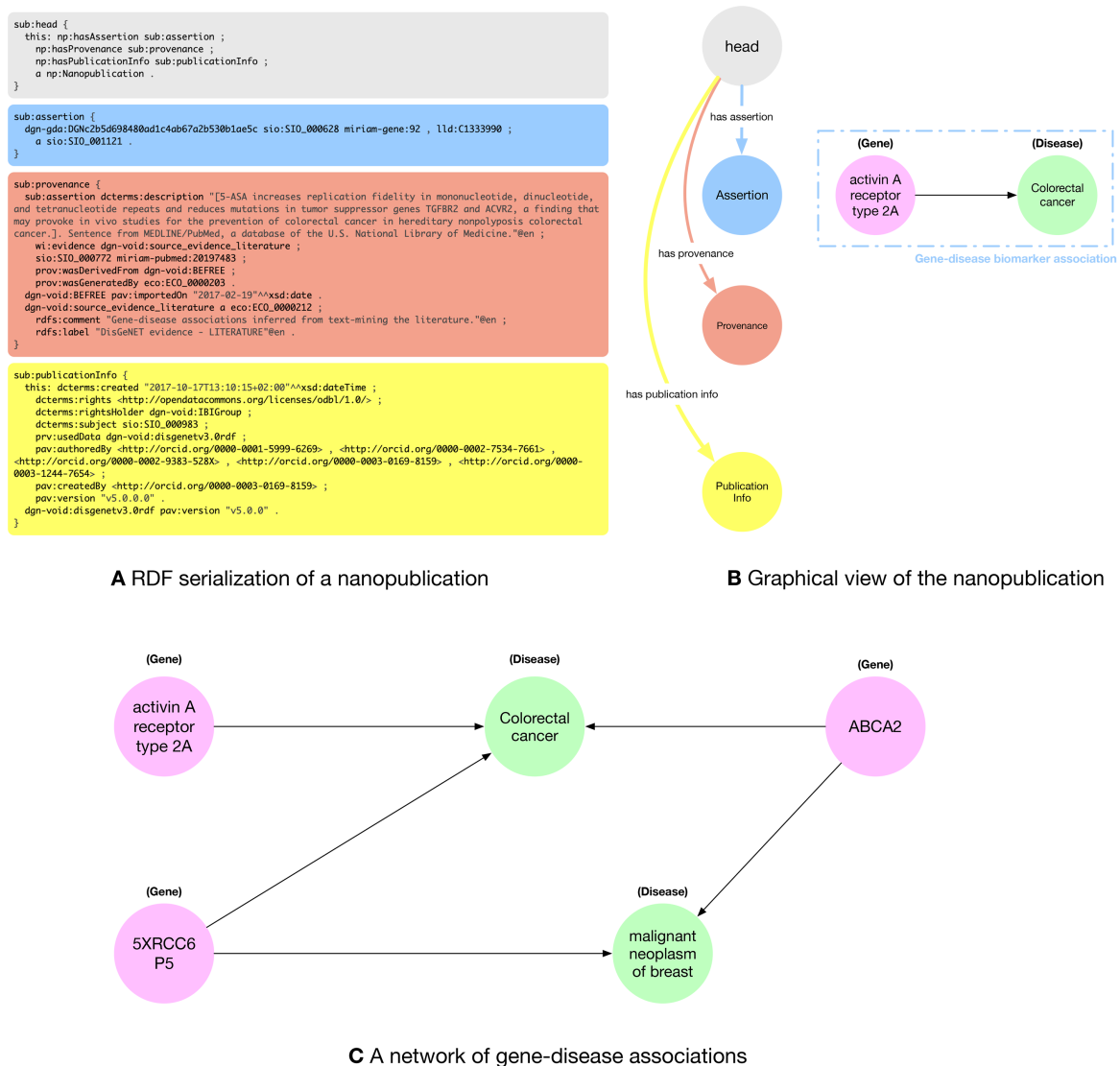


Fig. 7.1 (A) RDF (trig) representation of the nanopublication encoding the assertion: \langle activin A receptor type 2A - gene-disease association - Colorectal Cancer \rangle ; (B) graphical representation of the four parts of the nanopublications with a human-readable representation of the assertion graph; (C) network of gene-disease associations created by five nanopublications.

As an example taken from the biomedical domain, a nanopublication assertion about a gene-disease association is \langle activin A receptor type 2A – gene-disease biomarker association – colorectal cancer \rangle , where *activin A receptor type 2A* is the subject, *gene-disease biomarker association* is the predicate and *colorectal cancer* is the object of the triple. This assertion is extracted from a paper [50], which puts in relation the *activin A receptor type 2A* gene

to the *colorectal cancer* and describes a drug – i.e., *Mesalazine* – that reduces mutations in transforming growth factor of the gene.

In Figure 7.1.a, we can see a snippet of the RDF nanopublication serialization described above. Nanopublications are defined using the compact TriG¹ syntax, that enables to define *prefixes* to avoid to re-write the same IRIs multiple times. In Figure 7.1.a we used some prefixes within the nanopublication assertion, namely: *dgn-gda*, *sio*, *miriam-gene* and *lld*, that are specific of the life science domain. *dgn-gda* identifies a DisGeNET² gene-disease association; *sio* identifies a resource from *Semanticscience Integrated Ontology (SIO)*³, such as the type of a gene-disease association; *miriam-gene* identifies a gene in the National Center for Biotechnology Information (NCBI)⁴ database; *lld* identifies a resource from the Linked Life Data⁵ platform for the Biomedical domain.

The nanopublication is composed of four parts: (i) the *head* that acts as a connector between the other three sub-graphs; (ii) the assertion graph (blue) expressing the relationship between the two concepts of the assertion (the gene-disease association), the relationship of the concepts with external ontologies (the fact that *activin A receptor type 2A* is a gene and *colorectal cancer* is a disease), and possibly a link towards the scientific database storing related data; (iii) the provenance graph (orange) containing metadata about the assertion such as the methods used to generate the assertion and its creators; and, (iv) the publication info graph (yellow) containing the metadata about the evidence paper from which the assertion was extracted and about the nanopublication itself. In Figure 7.1.b, we can see a graphical representation of the four parts of the nanopublications with a human-readable representation of the gene-disease association encoded by the assertion graph.

A key aspect motivating the use of nanopublications is the possibility to exploit LOD features, allowing for exploring relation networks created by connecting related facts encoded in RDF. Indeed, nanopublications create a network of scientific assertions that can be explored to discover connections between facts. In the literature, there is important evidence of using nanopublications as a credible approach for expanding scientific insight, especially in the biomedical domain [64]. As a motivating example, Figure 7.1.c shows a small network of gene-disease associations. We can see that the genes *activin A receptor type 2A* and *5XRCC6P5* are both related to *colorectal cancer*. If we search for other connections, we find another nanopublication relating the *5XRCC6P5* gene to the *malignant neoplasm of breast* disease. Further expanding the relation network, we see that there exist two other

¹<https://www.w3.org/TR/trig/>

²<http://rdf.disgenet.org/>

³<https://github.com/MaastrichtU-IDS/semanticscience>

⁴<https://www.ncbi.nlm.nih.gov/>

⁵<http://linkedlifedata.com/>

nanopublications connecting the *ABCA2* gene with both *colorectal cancer* and *Malignant neoplasm of breast*. Figure 7.1.c presents a small network that shows the relationships between facts extracted from five different papers published in different venues at different times that do not cite each other. This is just a hint about how exploring the nanopublication relation network could lead to finding related concepts and assertions that might not be explicitly connected in the scientific literature and databases.

Nonetheless, despite these premises, nanopublications are not widely used by scientists outside specific circles [226]; they are hard to find and rarely cited. Nanopublications rarely have a human-readable accessible version and cannot be searched via keywords or natural language queries. Although nanopublications are based on LOD principles, there are still no tools that allow the user to explore their connections intuitively and discover if and how one assertion is related to others, as we have done in the example above. Leveraging on the famous *data is the new oil* metaphor [93], we can say that with nanopublications we have a vast oil reservoir but no active refinery, distribution net, and machines to put it into use.

In this work, we target these issues and present the *NanoWeb* application⁶, an open-source and publicly available web service enabling intuitive search, exploration, and re-use of nanopublications. The current version of NanoWeb is tailored for the life science domain, and it is designed to help experts of this domain in their research work. NanoWeb is an extensible tool to be applied to other scientific domains, even though certain customization to do so will be required. NanoWeb is a single entry point to the world of nanopublications enabling the seamless integration of data search, exploration, and re-use services; its central features are:

1. a crawler gathering publicly available nanopublications from the web;
2. two intuitive search functionalities, based respectively on the keyword search and boolean search paradigms;
3. a user-oriented visual interface to consult the nanopublications enriched with information gathered from external authoritative ontologies;
4. a service enabling the graph-based visualization of assertions and the exploration of their relation network;
5. data search functionalities providing entry points to external curated databases storing the scientific facts encoded by the nanopublications as well as to the scientific papers where the assertions were extracted.

⁶<https://w3id.org/nanoweb/>

The rest of the chapter is organized as follows: Section 7.2 presents the background of the nanopublication model and the state of the art of systems based on it. Section 7.3 describes the overall architecture of the NanoWeb application. Section 7.4 reports the statistics about the nanopublications available in NanoWeb. Section 7.5 shows how NanoWeb works and details the functioning of the user interface. Section 7.6 reports the results of the expert users survey conducted on NanoWeb. Section 7.7 discusses the challenges to be faced with maintaining NanoWeb in the medium-long period and how it can scale up to be used in domain others than life science. Finally, Section 7.8 draws some final remarks and outlines future work.

7.2 Background

Basics of Nanopublications. Nanopublications rely on Semantic Web technology. In particular, they are modeled via RDF [117], a widely used standard endorsed by the W3C consortium⁷, adopted for data publishing, accessing and sharing. RDF allows for the manipulation, enrichment, discovery and interoperability of data and it is at the core of the implementation of the LOD paradigm [236].

RDF is based on the concept of *statement*, that presents a <subject, predicate, object> triple-based structure. Within a triple, subject, predicate and object are *resources*. In particular, an RDF dataset can be represented as a graph where, given a triple, the subject and the object are the nodes representing *resources*, while the predicate, the direct edge connecting the two, expresses their *relationship*.

RDF resources can either be *IRIs* (Internationalized Resource Identifiers), *literals* or *blank nodes*. An IRI⁸ is a more general form of URI which can also contain Unicode characters. A literal is a value which can be associated to a specific type of value, such as string, integer, date, time etc. The default value is string. Blank nodes are resources which are labeled with a URI-like string which has validity only inside the database.

In RDF every resource and relationship is labeled. subject and object nodes can be labeled with IRIs, object nodes can also be labeled with literals. Relationships can only be labeled with IRIs. Blank nodes can be subject or object of a triple. A set of RDF triples can also be thought as a directed graph, where subjects and objects are nodes and predicates are the directed edges. Hence, it is also called RDF graph.

In recent years it has been proposed the idea to extend the basic semantic of RDF by using *quads* instead of triples, where an identifier (an IRI) is added. In this way, groups of

⁷<https://www.w3.org/TR/rdf11-primer>

⁸<https://tools.ietf.org/html/rfc3987>

triples may be characterized as belonging to the same subgraph, i.e. to the same *named graph* [55, 54], if they share the same extra URI.

Every nanopublication is made of four basic named graphs as shown in Figure 7.1.a:

1. *Head*: the graph composed of four triples connecting assertion, provenance and publication info graphs together and specifying that the graph at hand is a nanopublication.
2. *Assertion*: the assertion is to be thought of as the minimal unit of thought, a fact or a statement. It can be composed of one or more RDF triples and for this reason, we often call it *assertion graph*.
3. *Provenance*: the named graph made of metadata providing *context* about the assertion. The information contained in the provenance describes how the information expressed in the assertion was created (from some experiment, extrapolate from a paper or article, etc.) and the methods that were used to generate the assertion. It includes information such as authors, institutions, time-stamps, grants, links to evidence papers and other resources.
4. *Publication information*: the graph containing the information about the nanopublication itself, such as its authors, the topic of the assertion, and rights information.

Nanopublication resources and datasets. The website <http://nanopub.org/> is the most comprehensive access point to the world of nanopublications. It collects papers and tools about nanopublications. The central resource to access millions of publicly available nanopublications is the “nanomonitor”⁹. It provides a list of sixteen worldwide distributed servers where nanopublications can be openly accessed and downloaded in several formats. The nanopublications are ordered by identifier, but no full-text or structured search service is available. The nanopublications are accessible in an RDF serialization format. Thus they are machine-readable but not human-readable (see Figure 7.1.a).

[173] describes a Web-based service (i.e., *nanobrowser*) enabling access to human-readable enriched scientific statements extracted from nanopublications. The aim of *nanobrowser* is to enable easy publishing and curation of nanopublications, but unfortunately, at the time of writing, it does not work, even though the source code is publicly available.¹⁰ The nanobrowser had the goal to ease the extraction of facts from scientific papers and to enable the community to curate and revise the statements; its overall objective is different from those

⁹<http://app.tkuhn.eculture.labs.vu.nl/nanopub-monitor/>

¹⁰<https://github.com/tkuhn/nanobrowser>

of NanoWeb even though they share the requirement of making nanopublications human-readable and facilitate access to them. In the same direction, the *whyis* project ¹¹ proposes a knowledge graph infrastructure to support domain-aware management and curation of knowledge from different sources; it leverages on the nanopublication model to represent the facts and handle their provenance in the knowledge base. *whyis* also offers some facilities to allow the users to visually explore the knowledge graph beyond a given entity by using the so-called knowledge explorer [202, 201]; the knowledge explorer shares some similarities with the NanoWeb exploration tool. In particular, they both allow the exploration of the connections between entities in the knowledge graph. Nevertheless, *whyis* does not visualize the scientific assertions encoded by nanopublications. More specifically, the *whyis* project is oriented to the creation and user-based curation of the nanopublications rather than to the search and exploration possibilities connected to them. Hence, NanoWeb is a complementary service rather than a competitor to *whyis*.

[208] advocated for the systematic use of nanopublications to encode scientific facts reported in published papers. They see nanopublications as the key tool to enable reasoning and fact discovery exploiting machine intelligence. Furthermore, they extracted thousands of nanopublications about valuable and hard to discover gene variations and made them publicly available. We enable the search and access to these nanopublications in NanoWeb.

[65] described how they created nanopublications encoding scientific facts associated with more than 38K proteins stored in the neXtProt database. ¹² The main motivation for this work is to exploit nanopublications potential to support end-user research on human proteins enabling machine-reasoning, easy search and access to the protein-related facts. [64] showed how nanopublications as fine-grained annotations answer to complex knowledge discovery queries otherwise challenging to deal with. Also, in this case, queries are performed using the SPARQL structured language confining the use of nanopublications to technical database experts. We crawled and enable keyword-based search over all the publicly available neXtProt nanopublications.

[239] described the process that led to the publication of millions of nanopublications about the pathophysiology of diseases extracted from the scientific literature and backed by curated records in the DisGeNET database. ¹³ The DisGeNET nanopublications are publicly available and accessible via a SPARQL endpoint. NanoWeb collected, indexed all the available DisGeNET nanopublications and made them searchable and human-readable. Each nanopublication is enriched with a URL linking to the related curated record in DisGeNET.

¹¹<http://tetherless-world.github.io/whyis/>

¹²<https://www.nextprot.org/>

¹³<http://rdf.disgenet.org/>

Wikipathways is an online collaborative pathway resource that is made available as RDF and nanopublications [295]. The nanopublications are backed by the Wikipathways curated database and are accessible via a SPARQL endpoint (not available at the time of writing). The resource to convert the RDF triples of Wikipathways to nanopublication is publicly available.¹⁴ We crawled all the Wikipathways nanopublications, that are now searchable and accessible via NanoWeb.

[125] extracted more than 200M assertions about gene-disease associations from the biomedical literature. 7M assertions are explicitly stated in the scientific papers and the rest is implicitly inferred. There is a publicly available dump¹⁵ of the nanopublications shared as additional data for the paper. The website <https://rdf.biosemantics.org/> is intended to share all the nanopublications and to make access to the ontology required to dereference the concepts encoding the assertions. Unfortunately, at the time of writing, the nanopublications as well as the SPARQL endpoints to access them are unavailable.

[10] defined an ontology – VAXMO – for encoding vaccines-related information extracted from scientific literature and used nanopublications to propose a method to store misconceptions about vaccines. Unfortunately, the VAXMO ontology is not accessible as well as the associated nanopublications. Also, [313] recently used the nanopublication model to represent scientific facts manually extracted from the literature about cancer behavioral risk factors. They presented a prototype – AERO – to search and visualize the nanopublications; search is based on SPARQL queries and the visualization is allowed only for the results returned by the SPARQL endpoint. At the time of writing, AERO is not publicly available.

To the best of our knowledge, there is no available tool to visualize nanopublications and explore their connections. The tool which is closer to NanoWeb in terms of semantic search and graph visualization is BioKB [35]. BioKB provides access to the semantic content of biomedical articles through a SPARQL endpoint and a web interface; its goal is to allow the users to search for biomedical entities and visualize their graph of relations. However, BioKB does not account for nanopublications and does not support a multi-level exploration of the graph, enabling an in-depth exploration of the entities relation network.

Overall, the current services for searching nanopublications are all based on sparse SPARQL endpoints. To this end, NanoWeb contributes on two levels. First, it provides a unique online access point to all the publicly available nanopublications from the Life Science domain; and, second, NanoWeb provides advanced services as keyword search, visualization and human-readable access to millions of nanopublications, making them accessible to users without technical expertise in SPARQL and related technologies.

¹⁴<https://github.com/wikipathways/nanopublications>

¹⁵<https://datadryad.org/stash/dataset/doi:10.5061/dryad.gn219>

Search over RDF. RDF graphs can be interrogated through the powerful but complex SPARQL query language [228]. SPARQL is not intuitive for end-users since it presents a complex syntax, far from a natural expression of their information need [304]. It also requires knowledge of the underlying schema of the database, and of the IRIs used in it. This knowledge is often not possessed by the average end-user.

A search paradigm adopted to address the issues related to the use of SPARQL is *keyword search*. Keyword-based methods have gained importance over time both in research and in industry as a paradigm to facilitate the access to structured data [29, 164, 310].

The main difference between SPARQL and keyword search is that, while SPARQL returns the one and only correct answer (or an empty set if there was no answer), keyword search returns a ranking of answers, ordered based on their *relevance* to the information need expressed by the user via the keyword query.

In the literature, keyword query search systems over structured data are mainly focused on relational databases (RDB) [310] but many are also emerging for graph-like databases such as RDF datasets [297, 29]. These systems may be divided into *three* categories.

The first kind of systems is *schema-based*. Examples are [27, 7, 194]. These systems exploit the schema information of the database, be it relational or RDF, to formulate queries in a structured language (SQL or SPARQL depending on the type of the database) designed from the keyword query of the user.

The second category is *graph-based*. Originally born with relational databases [34, 268], the technique at the base of these systems was based on the transformation of the relational database in a graph. These systems are relatively easily translated in the RDF scenario since these databases are already in a graph form. A core challenge of these systems is to deal with the size of big graphs, which can contain tens of millions of nodes, if not more. In several cases, it has been shown that the size makes the task unsolvable by these systems [67].

Stemming from this last class of systems, the last category is the one of the *virtual-document based* systems [146]. First described in [192], this approach relies on the concept of *virtual document* of a graph. Given one graph, RDF or obtained by relational tuples, its corresponding virtual document is obtained by extracting words from it in an automatic way. This produces a “flat” representation of the graph, where its syntax and topology are lost but its semantic and lexical content is somewhat maintained. The virtual document representation is convenient since systems can leverage on efficient state-of-the-art IR methods for indexing and ranking. These methods operate by first extracting subgraphs from the whole database, then converting them in their virtual document representation and ranking these documents with respect to the keyword query. The user receives at the end the ranking of graphs in the order dictated by the ranking on the corresponding documents.

There is no keyword search system for nanopublications, which are always searched via SPARQL endpoints. The complexity of search systems for RDF and their scalability issues have prevented the use of keyword search for RDF data in general and nanopublications in particular. NanoWeb, exploits a very recent advancement in *virtual-document based* systems [90], which enable fast and effective keyword search over RDF and nanopublications.

7.3 The NanoWeb Architecture

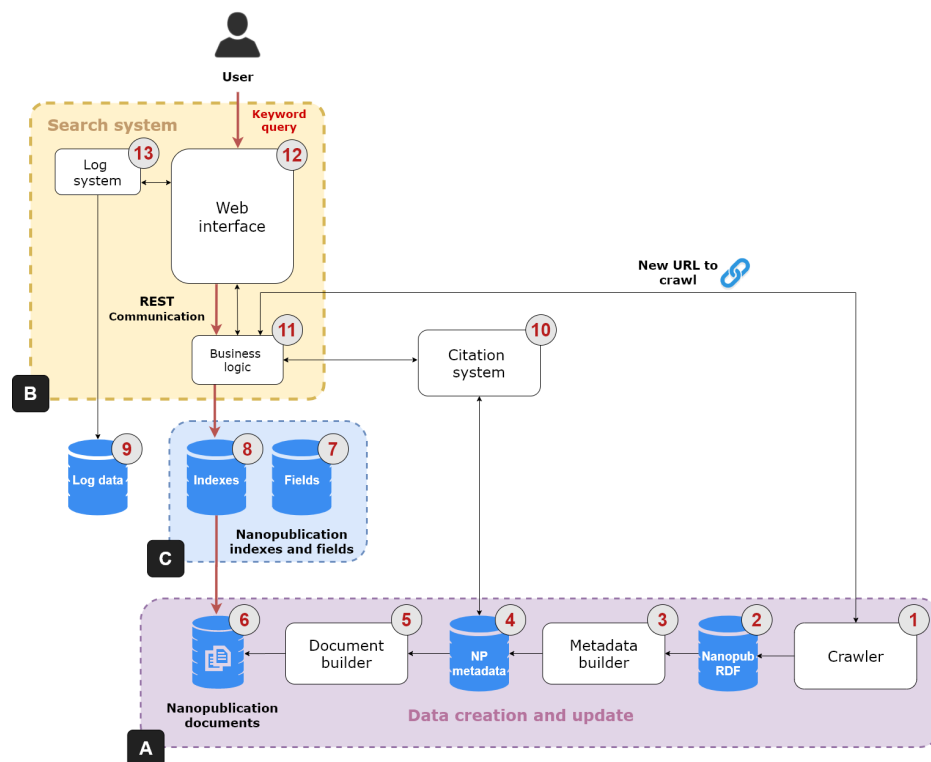


Fig. 7.2 NanoWeb system architecture.

The NanoWeb architecture is composed of four main components: (i) a *crawler* that gathers nanopublications from the Web; (ii) a *search system* that indexes and enables full-text search over the nanopublications; (iii) a nanopublication *citation system*; (iv) a *Web user interface* to search, access and explore the nanopublications and their relation network. Figure 7.2 shows the architecture of the NanoWeb system, which consists of the following areas:

- **Data creation and update** (Figure 7.2, box A):

- **Crawler** (1): it collects nanopublications from different web sources. It considers different types of resources: authoritative ones, such as academic or institutional platforms; and public ones, such as git repositories. Nanopublications are *downloaded* and *stored* in an RDF database (2). The crawler also downloads new nanopublications obtained from URLs that can be provided by the users; this process is handled by the business logic unit (11). The crawler sends a new request for each web source in the list of initial seeds. It parses and scrapes the web pages and produces a list of extracted URLs. Each URL in the list is processed so that direct links to nanopublications are resolved and added to the download queue. Each nanopublication file is downloaded using an independent thread so that requests are handled asynchronously. These files are saved into the RDF database. The links in the URL list that point to other web pages are followed so that these new Web pages are also parsed and scraped in a recursive scraping loop to discover new nanopublications. The crawler is written in Java and it comes with a graphical and a batch mode. The graphical mode allows the user to interact and control crawler activities using a Graphic User Interface (GUI).¹⁶ The batch mode enables a fast and batch-based download using operating systems lacking a GUI.
 - **Metadata builder** (3): the nanopublications are processed to dereference the URLs and to get additional *metadata*; for instance, the nanopublications are enriched with the label of the concepts referring to external ontologies, the names of creators and curators and the title of the evidence papers. These data are saved in a relational database (4).¹⁷
 - **Document builder** (5): The document creation phase occurs after the dereferencing and enrichment phase. The document builder creates “virtual” nanopublication documents, which are saved into a database (6), on which the keyword search system is based.
- **Search system** (Figure 7.2, box B): this system performs keyword search on the nanopublications and it has three components:
 - **Business logic** (11): it is the controller unit of the search system. It performs the orchestration activities such as the coordination of the crawler by feeding it with

¹⁶A demonstration video of the crawler in action, using the graphic mode, is available at <https://bit.ly/2RVIGzl>.

¹⁷All the relational databases are based on PostgreSQL version 10.6 allowing for the table partitioning function; this function enables efficient storage and access to the data.

new nanopublication URLs. It takes the user keyword query as input and returns the relevant nanopublications through the Web interface as output. To perform this task, the business logic unit relies on three databases: the nanopublication documents database (6), the fields (7) and the indexes (8). The indexes database contains the inverted index extracted from the nanopublication documents required to match the query terms with the document terms. The fields database is required to provide fast access to specific nanopublication data such as the authors, curators, and evidence paper metadata.

- **Web interface** (12): it is the front-end allowing the user to search, access, explore and cite nanopublications through an interactive interface. It communicates with the business logic unit using a REST layer that provides public API for accessing nanopublications data in JSON format.
- **Log system** (13): it deals with the logging tasks of the search system and it relies on a specific relational database (9). It communicates with the Web interface to collect relevant user activity information and possible problems.
- **Citation system** (10): it generates the citations text snippet for the nanopublications of interest to the user by relying on the system presented by [100]. Citations are a fundamental tool to give credit to authors and curators of data and publications and help other users to recognize the value of nanopublications. When the business logic unit (11) receives the request to produce a citation for a nanopublication, it sends this request to the citation system, that in turn collects the necessary metadata from the corresponding database (4). Once produced, the citation snippet is returned to the business logic unit and then visualized in the Web interface.

7.3.1 Search system

Let us assume that a user has an information need, and wants to retrieve the nanopublications that satisfy it. Since nanopublications are encoded in RDF, one possibility is to query the graph composed by all the nanopublications via the SPARQL query language, that, as already discusses, presents drawbacks for non-expert users.

We adopt two alternatives to SPARQL, i.e. keyword search and boolean search, both oriented to ease the search process for the users. Boolean search (i.e., advanced search) is adopted for domain-specific searches and it is useful to guide users in query formulation, since they often do not know in advance what they can search. We realized advanced search over the nanopublication metadata database, that allows for searching on specific fields of the indexed data (e.g. genes, diseases, proteins, or authors).

Boolean search enables targeted search functionalities, but it does not allow for general and open full-text search over the nanopublications. To allow users to exploit natural language to search for nanopublications, we realized a keyword search system over RDF data. The system we adopt is based on the *virtual document* strategy, first presented in [192] and used in many other papers about *keyword search* on RDF graphs [90, 94, 200]. The underlying task of these papers is that, given an RDF graph, the user wants to query it, but for some reason, she is unable to use a SPARQL query. Keyword search is an alternative paradigm to using a structured query based on a query made of keywords.

The virtual document strategy is one of the many strategies deployed to face keyword search on RDF graphs. Given an RDF graph, we call its corresponding *virtual document* the textual document obtained from the concatenation of words obtained from the IRIs and Literals contained in the nodes and edges of the graph.

Given a collection of graphs, it is therefore possible to create a corresponding collection of *virtual documents*. Every document is uniquely linked to the graph that generated it since they share the same identifier.

Then, the collection of documents is indexed and, from that moment on, this index can be used to answer keyword queries in the same way it is done in more classic IR scenarios, where the collections are made by “real” documents. In this work, we used a probabilistic model (i.e., BM25 [Robertson et al.]) as a ranking function.

Every time a new query is issued, BM25 uses the virtual document index to create a ranking of documents. The document identifiers are used to retrieve the corresponding graphs, that is, the corresponding nanopublications, from the collection. This list of nanopublications is then returned to the final user in the same order dictated by the ranking.

One may argue that this strategy discards information from the graphs. Since each graph is *flattened* to a document version of itself, information such as its topology and the disposition of words among nodes and edges is lost. This is certainly true, and in fact works such as [90, 94, 200] do not limit themselves to virtual documents, but employ different kinds of heuristics to better leverage on the topology of the graphs.

Moreover, topology-oriented heuristics often rely on the exploration of the graphs, which adds overhead to the whole computation. The more the answers returned by BM25, the bigger this overhead. Therefore, we argue that the use of topology-oriented heuristics does not guarantee a significant improvement in the effectiveness of the rankings obtained by the graphs with respect to the added overhead to the computation.

7.4 Nanopublication collection statistics

In Table 7.1 we report the number of nanopublications per scientific platform currently available in NanoWeb. Currently, we have crawled and indexed nanopublications from the following platforms:

- **DisGeNET:** ¹⁸ “a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases” [232]. DisGeNET is a knowledge management platform integrating and standardizing data about disease-associated genes and variants from multiple sources, including the scientific literature. DisGeNET covers the full spectrum of human diseases as well as normal and abnormal traits. [239] presented the publication of DisGeNET human Gene-Disease Associations (GDAs) as a new Linked Dataset exploiting the nanopublication approach. DisGeNET provides roughly half of the nanopublications, about 5 million, available in NanoWeb.
- **NeXtProt:** ¹⁹ “neXtProt is a protein knowledge platform that aims to support end-user research on human proteins” [65]. [65] converted data from neXtProt into nanopublications to show how they can be used to seamlessly query the data and gain biological insight. In particular, they converted three types of annotations of interest for the biomedical community: variation data, posttranslational modification (PTM), and tissue expression.
- **Protein Atlas:** ²⁰ “A Human Pathology Atlas has been created as part of the Human Protein Atlas program to explore the prognostic role of each protein-coding gene in each cancer type by means of transcriptomics and antibody-based profiling.” [287]. The Human Protein Atlas is an open-access knowledge-base providing the data to allow genome-wide exploration of the impact of individual proteins on clinical outcomes. The Human Protein Atlas (HPA) programme aims to “generate a comprehensive atlas of protein expression patterns in human normal and cancer tissues as well as cell lines.” [235].
- **WikiPathways:** ²¹ “WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways.” [271, 295]. WikiPathways provides rich pathway databases with a focus on genes, proteins and metabolites. The data from WikiPathways have been converted into a dataset of nanopublications as explained in [175].

¹⁸<https://www.disgenet.org/>

¹⁹<https://www.nextprot.org/>

²⁰<https://www.proteinatlas.org/>

²¹<https://www.wikipathways.org/>

| Platform | Number of nanopublication |
|---|---------------------------|
| DisGeNET | 4,717,256 |
| NeXtProt | 4,014,376 |
| Protein Atlas | 1,254,466 |
| Wikipathways | 26,934 |
| <i>Total number of nanopublications</i> | <i>10,013,032</i> |

Table 7.1 Number of nanopublications per platform.

7.4.1 Association analysis

DisGeNET accounts for roughly half the total number of nanopublications in NanoWeb. The assertions encoded by these nanopublications are divided into gene-disease associations of different types. In Figure 7.3, we report the number of assertions in NanoWeb for each association of the DisGeNET ontology. A detailed description of the associations is available in the DisGeNET website.²²

In the same vein, Table 7.2 reports the genes-tissues association types present in NeXtProt nanopublications. In particular, the *protein-coding gene expression in tissue* association describes the relationship between a protein-coding gene in directing the production of proteins expressed in a tissue. Another type of association regarding proteins is the *protein expression in tissue* which describes the expression level (high, low, medium, not detected) of a protein in a tissue. Besides, the *sequence on amino-acid* association describes the relationship between proteins and amino acids. The total number of nanopublication assertions regarding protein associations is over 5 million.

7.4.2 Scientific Evidences

Nanopublication assertions are supported by evidences; an evidence can be a scientific publication, a curated database record or both. The nanopublication evidences in NanoWeb come from several institutional open-access databases such as Bgee²³, Cancer Sanger²⁴, EbiQuickGo²⁵, Gene Expression Omnibus (GEO)²⁶, Protein Atlas²⁰ and UniProt²⁷. We report the evidence databases associated to the nanopublications available in NanoWeb in Table 7.3. The total number of evidences collected from authoritative databases are about

²²<https://www.disgenet.org/dbinfo#section5>

²³<https://bgee.org/>

²⁴<https://cancer.sanger.ac.uk/>

²⁵<https://www.ebi.ac.uk/QuickGO/>

²⁶<https://www.ncbi.nlm.nih.gov/geo/>

²⁷<https://www.uniprot.org/>

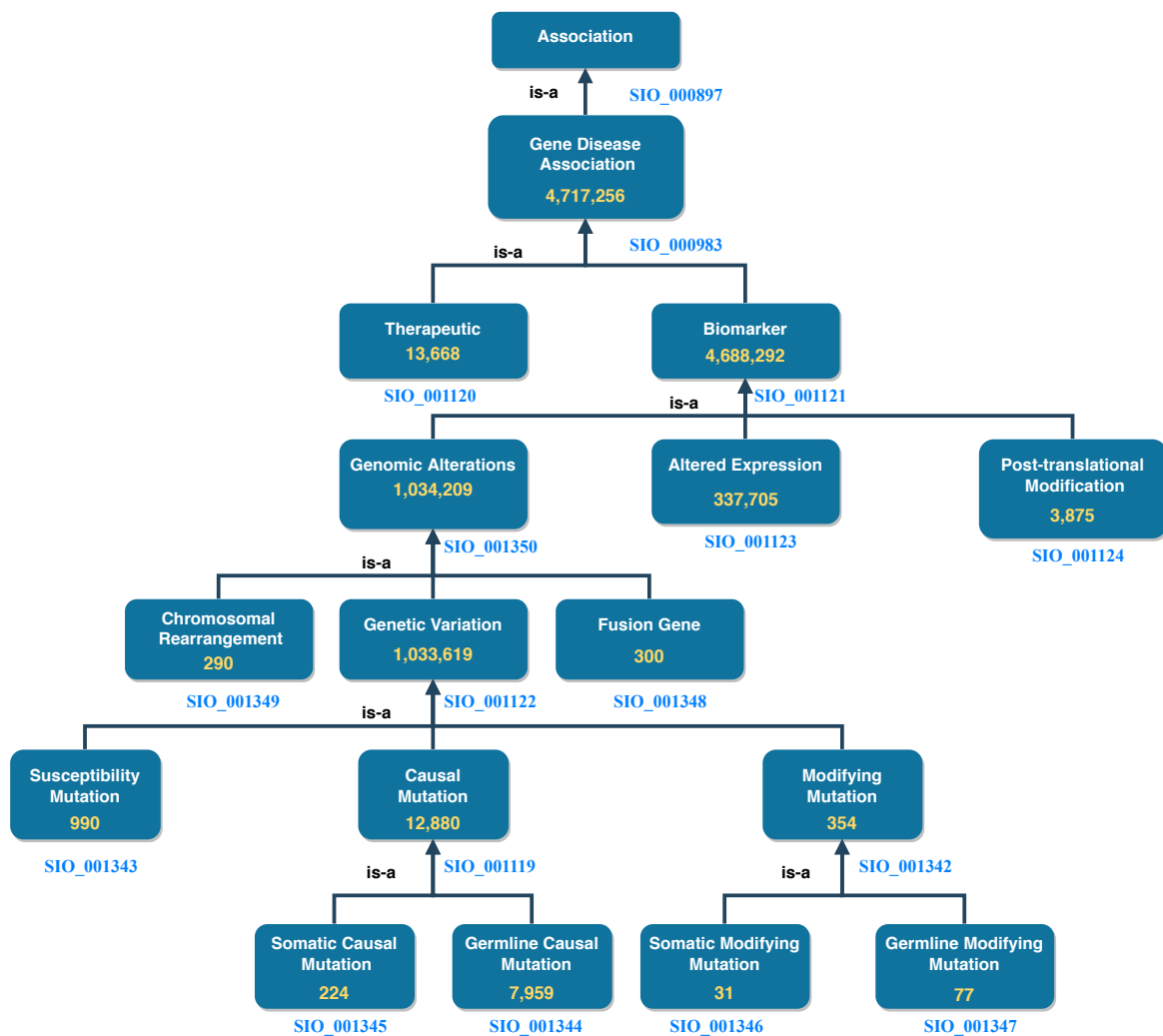


Fig. 7.3 DisGeNET ontology: number of assertions (yellow) for each DisGeNET association type.

11 million, and the evidences coming from publications are more than 6 million. All these publications are available in the PubMed²⁸ database.

²⁸<https://pubmed.ncbi.nlm.nih.gov/>

| Association | Number of assertion |
|--|----------------------------|
| protein-coding gene expression in tissue (generic) | 6 |
| protein-coding gene expression in tissue with quality high | 124,261 |
| protein-coding gene expression in tissue with quality low | 184,615 |
| protein-coding gene expression in tissue with quality medium | 275,241 |
| protein-coding gene expression in tissue with quality negative | 837,144 |
| protein-coding gene expression in tissue with quality not detected | 341,062 |
| protein-coding gene expression in tissue with quality positive | 1,421,203 |
| <i>protein-coding gene expression in tissue (total)</i> | <i>3,183,532</i> |
| protein expression in tissue with level high | 150,366 |
| protein expression in tissue with level low | 241,325 |
| protein expression in tissue with level medium | 361,641 |
| protein expression in tissue with level not detected | 501,133 |
| <i>protein expression in tissue (total)</i> | <i>1,254,466</i> |
| sequence on amino-acid | 739,528 |
| <i>protein associations (total)</i> | <i>5,177,526</i> |

Table 7.2 Assertion numbers for association types: “protein-coding gene expression in tissue” and “protein expression in tissue”.

| Database | Number of evidences |
|----------------------------------|----------------------------|
| Bgee | 5,576,047 |
| Cancer Sanger | 578 |
| EbiQuickGo | 8,876 |
| Gene Expression Omnibus (GEO) | 573,648 |
| Protein Atlas | 4,125,154 |
| UniProt | 628,749 |
| <i>Total number of evidences</i> | <i>10,913,052</i> |

Table 7.3 Number of evidences per database.

The screenshot displays the NanoWeb search interface. At the top, a dark blue header contains a menu icon (3), a search bar with the query 'colorectal cancer' (1) and a 'Search' button, and a pin icon (2). Below the header, there are four filter options: 'Nanopub List Only' (selected), 'Display Both', 'Nanopub Info Only', and 'Show Graph Layer'. The main content area is a table of search results, with the first row highlighted in red. The table lists gene-disease associations for colorectal cancer, including 'mutL homolog 1 - Colorectal Carcinoma' (4a), 'mutL homolog 1 - Colorectal Carcinoma' (4b), 'glutathione S-transferase kappa 1 - Colorectal Cancer', 'mutL homolog 1 - Colorectal Cancer', 'mutS homolog 6 - Colorectal Cancer', 'methylenetetrahydrofolate reductase - Colorectal Cancer', and 'mutS homolog 2 - Colorectal Cancer'. Each row includes a 'Go to: DisGeNET' link (4c) and a graph icon (4d). A 'Load More' button (5) is located at the bottom of the results list.

| Gene-Disease Association | Association Type | Go to: DisGeNET | Graph Icon |
|---|--|-----------------|------------|
| mutL homolog 1 - Colorectal Carcinoma | gene-disease association linked with genetic variation | Go to: DisGeNET | Graph Icon |
| mutL homolog 1 - Colorectal Carcinoma | gene-disease biomarker association | Go to: DisGeNET | Graph Icon |
| glutathione S-transferase kappa 1 - Colorectal Cancer | gene-disease biomarker association | Go to: DisGeNET | Graph Icon |
| mutL homolog 1 - Colorectal Cancer | gene-disease association linked with genetic variation | Go to: DisGeNET | Graph Icon |
| mutS homolog 6 - Colorectal Cancer | gene-disease association linked with genetic variation | Go to: DisGeNET | Graph Icon |
| methylenetetrahydrofolate reductase - Colorectal Cancer | gene-disease association linked with genetic variation | Go to: DisGeNET | Graph Icon |
| mutS homolog 2 - Colorectal Cancer | gene-disease association linked with genetic variation | Go to: DisGeNET | Graph Icon |

5 Load More

Fig. 7.4 NanoWeb search interface with user-provided query: *colorectal cancer*

7.5 NanoWeb Graphical User Interface

The NanoWeb system, available at <http://w3id.org/nanoweb/>, provides an interactive Web interface that the user can use to search, access, explore, and cite nanopublications. A demo video presenting NanoWeb functionalities is available at <https://bit.ly/NWURL2>.

Figure 7.4 shows the NanoWeb search interface. At the top of the page, there is the query input form (1), where the user types the query and searches for nanopublications. There is a button (2) to pin or unpin the query input form on the right side of the query input form. The query input form is unpinned by default; this means that it floats at the top of the page so that it is always visible to the user even when the page is scrolled. The user can press the button to pin the query input form, making it hidden when the page is scrolled. On the left side of the query input form, there is the menu button (3). By clicking on it, the sidebar appears with a list of links to the web app functionalities:

1. **Home:** takes the user to the home page.
2. **Stats:** takes the user to the Web page summarizing the NanoWeb system statistics, such as the number of nanopublications and triples inserted in the database.
3. **About:** takes the user to the page that briefly describes the purpose of the NanoWeb system and summarizes the provided functionalities.
4. **Contacts:** leads to a page with contact information of the authors of this project.

The body of the Web interface consists of three layers displayed alternatively:

- **Nanopublications list** (Figure 7.4.A) A list of nanopublications retrieved for the user query. Each nanopublication is represented with a row in the list, reporting the following information:
 1. The title of the nanopublication (4a).
 2. The assertion of the nanopublication (4b).
 3. A link to the source platform of the data (4c). For instance, in Figure 7.4 the source platform of the data is DisGeNET.
 4. The *graph button* to display the graph associated with the nanopublication (4d). When the user clicks this button, the Graph layer appears to show the nanopublication graph on the right side of the nanopublications list. If the Information layer is displayed, it is replaced with the Graph layer.

The *Load More* button (Figure 7.4.5) loads more relevant nanopublications associated with the query, if any.

As we can see in Figure 7.5, when a user clicks on a specific row, the Information layer is displayed, showing the information regarding the selected nanopublication.

The screenshot displays the DisGeNET interface for the query "colorectal cancer". At the top, there is a search bar with the query and a "Search" button. Below the search bar, there are navigation options: "Nanopub List Only", "Display Both", "Nanopub Info Only", and "Show Graph Layer".

The main content area shows a list of nanopublications. Each row includes the gene name, disease name, and a brief description. A "Go to: DisGeNET" link is provided for each entry. A hand cursor is shown clicking on the first row: "mutL homolog 1 - Colorectal Carcinoma" with the description "gene-disease association linked with genetic variation".

The "Information layer" is expanded for the selected entry, showing the following details:

- 1 mutL homolog 1 - Colorectal Carcinoma**: gene-disease association linked with genetic variation between: [mutL homolog 1, Colorectal Carcinoma]
Additional Info: [Survival of patients with hereditary colorectal cancer: comparison of HNPCC and colorectal cancer in FAP patients with sporadic colorectal cancer.]
- 2 Publication Info**:
Nanopublication ID: RAhxO-EtNeCc-a2TYdr66pMovLwmGWkvY3_hJrdFrnhjE
Creation date: 2017-10-17
Creators: [núria queralt rosinach]
Collaborators: [àlex bravo serrano, ferran sanz, laura i. furlong, núria queralt rosinach, janet piñero]
Platform: DisGeNET
Rights holder: IBIGroup
Nanopub URL: <http://server.nanopubs.lod.labs.vu.nl...>
Get nanopub: [Download](#)
Go to data record: [Data Record](#)
- 3 Provenance**:
Assertion generated by: ECO_0000203
Assertion Generation Description: Gene-disease associations inferred from text-mining the literature. / ECO_0000212
Evidence Source: <http://identifiers.org/pubmed/9935197>
- 4 Cite this nanopub**:
Núria Queralt Rosinach, Àlex Bravo Serrano et al.(5), 2017-10-17, IBIGroup, gene-disease association linked with genetic variation: [Survival of patients with hereditary colorectal cancer: comparison of HNPCC and colorectal cancer in FAP patients with sporadic colorectal cancer.] mutL homolog 1 - Colorectal Carcinoma, DisGeNET, v5.0.0.0, <http://nanocitation.dei.unipd.it...>

At the bottom of the list, there is a "Load More" button.

Fig. 7.5 Information layer for the nanopublication.

- The **information layer** shows information associated with a selected nanopublication, including:

The screenshot shows the DisGeNET web interface. At the top, there is a navigation bar with links for Home, About, Search, Browser, and API, along with the DisGeNET logo. Below this, there are three tabs: Summary of GDAs, Evidences for GDAs, and Summary of VDAs. The main content area displays the title 'Colorectal Carcinoma, C0009402' with a search icon. Below the title, there are two buttons: 'Gene: MLH1' and 'Source: ALL'. A filter box is present with the text 'Filter within current results:'. At the bottom, there is a table with the following data:

| Gene | UniProt | Gene Full Name | Protein Class | N. diseases | DSI |
|------|---------|----------------|----------------------|-------------|-------|
| MLH1 | P40692 | mutL homolog 1 | Nucleic acid binding | 526 | 0.399 |

Fig. 7.6 Data record for the nanopublication with title: *mutL homolog 1 - Colorectal Carcinoma*.

1. **Assertion:** (Figure 7.5.1) This section reports the assertion of the nanopublication of interest and its title. Besides, meaningful entities, such as the disease *Colorectal Carcinoma*, are reported as links to external knowledge bases.
2. **Publication info:** (Figure 7.5.2) This section reports the publication information of the clicked nanopublication. This information includes the creation date, the creators, and the source platform. Moreover, a link to the data record is provided so that the user can be redirected to the data record about the assertion; these links act as entry points to external scientific databases. For instance, Figure 7.6 shows the data record web page for the nanopublication with title: *mutL homolog 1 - Colorectal Carcinoma* in DisGeNET.
3. **Provenance:** (Figure 7.5.3) This section shows the provenance information such as the evidence source and how the nanopublication was generated. It also reports the abstract of the publication, if present.
4. **Cite:** (Figure 7.5.4) This section shows the citation snippet of the nanopublication. The user can copy the citation text by clicking on the *Cite this nanopub* button in the header.

The user can expand/collapse each section by clicking on the title or in the header section.

- **Graph layer:** Figure 7.7 shows the Graph layer displayed on the right side of the nanopublications list after the user click. This layer shows the graph associated with the nanopublication, leveraging on the RDF triple structure. Each graph node corresponds to the subject or the object of an assertion, while the edge represents the predicate. Each assertion is represented with a directed edge.

The figure shows the graph associated with the *mutL homolog 1 - Colorectal Carcinoma* nanopublication. The assertion within this nanopublication has two nodes: *mutL homolog 1* as the subject and *Colorectal Carcinoma* as the object. The subject – a gene – is colored in green, while the object – a disease – is in red. The predicate connecting the two is represented as an oriented grey edge.

There are different ways to interact with the nanopublication graph. For instance, the user can click on a node to expand the relation network and visualize other nodes connected to the nanopublication of interest. The complete list of the user graphic controls available can be consulted by clicking on the *Controls help* button indicated with number three in Figure 7.7. The figure shows a two-levels expansion starting from the subject node *mutL homolog 1* and ending with the expansion of the node associated to the *Colorectal Cancer* disease.

The possible actions that a user can perform on the graph are:

- **Expand/collapse graph network:** When the user left-clicks on an unexpanded node, the graph is expanded. Thus its relation network is shown. Otherwise, if the user clicks on an already expanded node, the graph collapses, and in turn, its relation network is hidden.
- **Show node information:** When the user right-clicks on a node, a dialog modal window appears to show the information concerning that node. For instance, the information window shows the type of entity node clicked, such as *gene* or *disease* in case of nodes coming from nanopublications concerning biological or medical fields.
- **Show edge information:** When the user left-clicks on edge, a dialog modal window appears to show the information regarding the nanopublication. Figure 7.8 shows that when the edge connecting *mutS homolog 6* and *Carcinogenesis* is clicked, the nanopublication information window appears on the right side. The

modal dialog window contains the same information of the Information layer. Still, it has a smaller width and can be dragged anywhere inside the Graph layer, so it is always accessible without covering it.

- **Drag and drop:** The user can drag and move the nanopublication graph by pressing the mouse's left button and moving it around the graph layer. When the desired position has been chosen, the user can release the left button of the mouse to drop the graph.
- **Zoom in/out:** Using the mouse wheel, the user can zoom in or out on the nanopublication graph.
- **Switch between Graph and Information layers:** A button is provided to switch between Graph and Information layers. For instance, when the Graph layer is displayed to go back to the Information layer, the user can click on the *Show Nanopub Info* button (Figure 7.7.1). In the same way, when the Information layer is displayed, the user can switch to the Graph layer by clicking the *Show Graph Layer* button.
- **Rearrange layers:** The Navbar menu manages layers disposition (Figure 7.7.2) and it is provided with the following buttons:
 1. **Nanopub List Only:** It shows a full-screen view of just the nanopublications list layer.
 2. **Display Both:** It opens a two-layers view consisting of the nanopublications list layer and the currently active layer between Graph and Information layers. For instance, Figure 7.7 shows the Graph layer on the right side of the nanopublications list layer.
 3. **Graph Only/Nanopub Info Only:** It shows a full-screen view of the current layer, which can be the Graph layer or the Information layer. For instance, Figure 7.7 shows this button with the text "Graph Only", since the Graph layer is active.

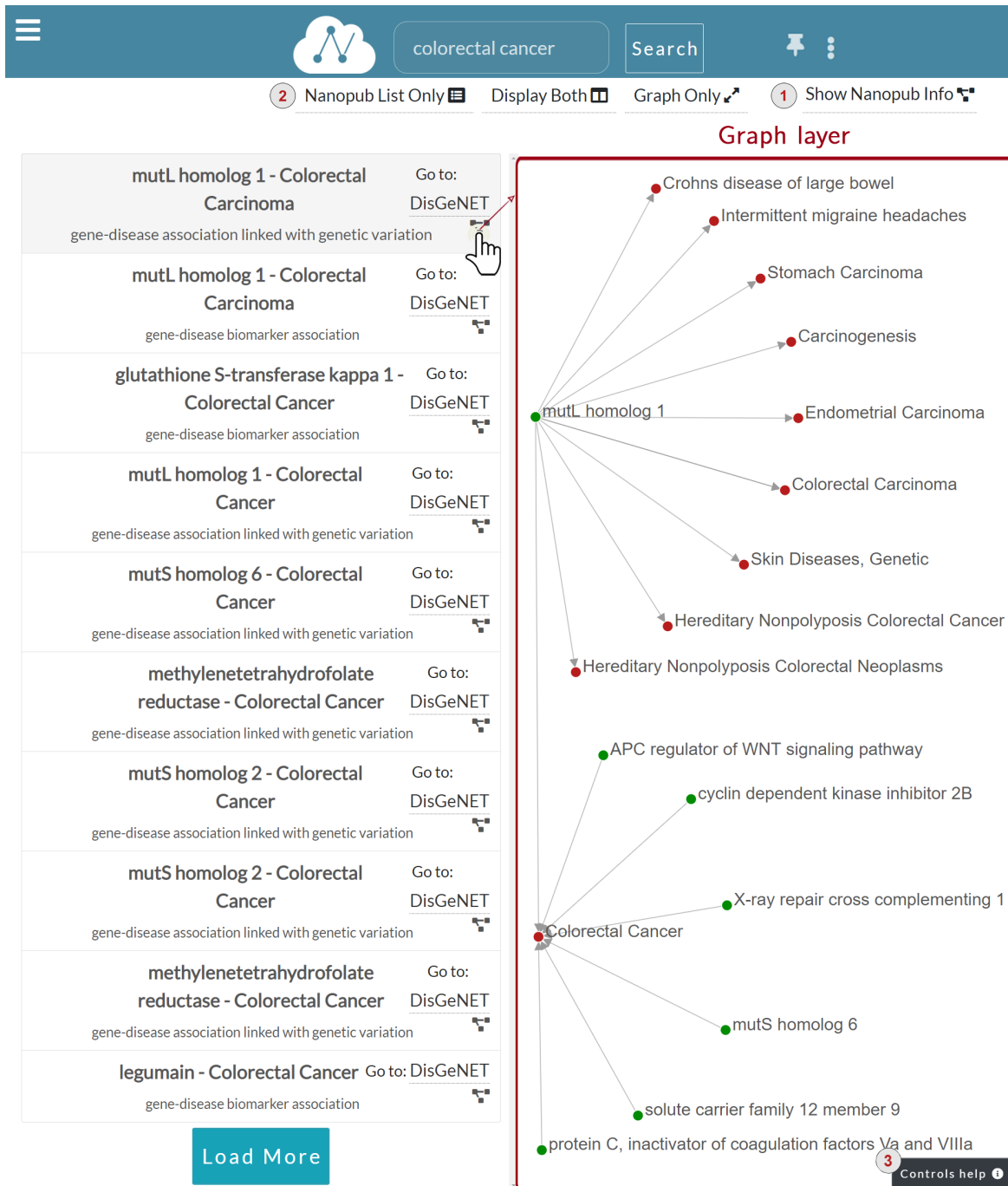


Fig. 7.7 Graph layer for the nanopublication clicked by the user.

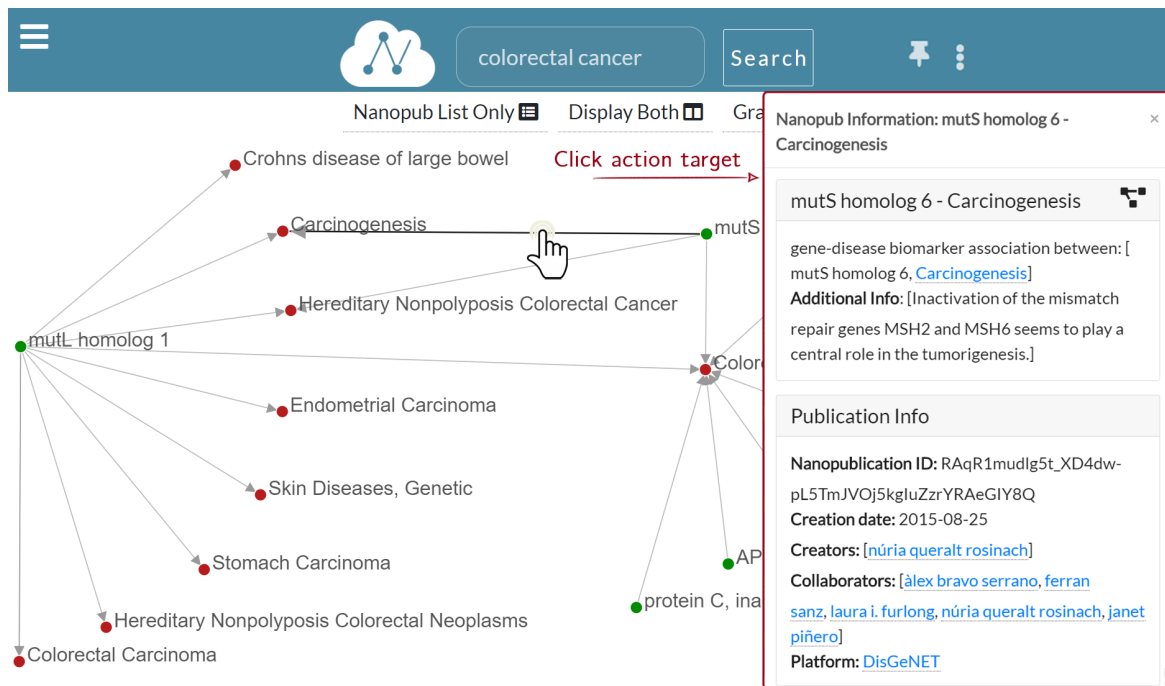


Fig. 7.8 Graph exploration: the information window for *mutS homolog 6 - Carcinogenesis* is displayed as a result for the user click on the edge.

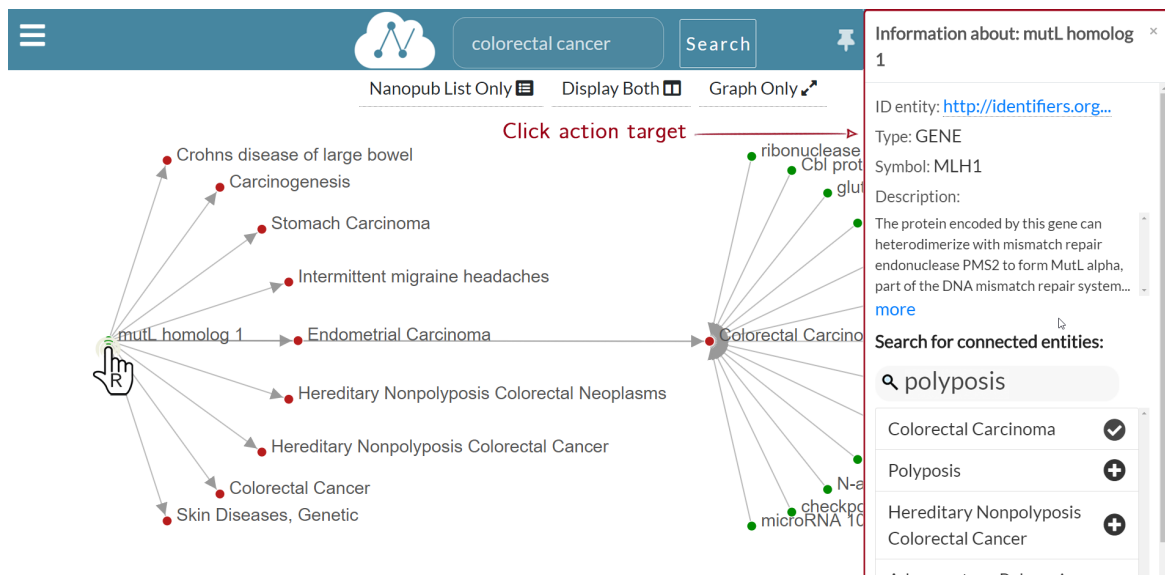


Fig. 7.9 Graph exploration: search for *mutL homolog 1 (MLH1)* connected entities.

Graph exploration

Figure 7.8 shows a multi-level graph exploration for the nanopublication with the title *mutL homolog 1 - Colorectal Carcinoma*, which describes a gene-disease association. This functionality allows the user to explore the relation network of the considered nanopublications. Besides, the graph exploration allows the user to understand how and why different nanopublications are connected. There is no limit to the depth of the exploration, i.e., to the graph's dimension visualized. The user can potentially expand the graph at will until all the nodes connected in the relation network are displayed. In this way, the synthesis power of nanopublications is enhanced by the value of the relation network; it provides a greater information contribution than the sum of the single nanopublications taken separately. Since the graph can have a high density of connections, only a portion of the connected nodes is shown for a new graph expansion request. However, the user could be interested in a specific connection between two nodes, which may not be shown by default. Hence, it is possible to search for specific connections directly on the nanopublication semantic network – we call this functionality “connected entities search”. Figure 7.9 shows the connected entities search in action. In particular, we see the entities connected to the *mutL homolog 1* gene. When the user right-clicks on the node associated with the *mutL homolog 1* gene, the information window is shown on the right side. Inside the information window, there is the “connected entities” input field, where the user can specify the entity name s/he is looking for. For instance, when the user types *polyposis*, a list of matching entities appear, and the user can choose which entities to add to the graph by clicking on the plus button. Using the connected entities search, users can quickly verify whether a direct link between two nodes exists. The “connected entities search” is provided with auto-completion to ease the work of the user.

Implementation specifications

NanoWeb back-end is developed using Django,²⁹ which is a Python-based free and open-source Web framework. The Web app front-end is developed using HTML5, CSS3, Bootstrap framework,³⁰ JavaScript, jQuery,³¹ and the library D3.js.³² In particular, to draw the nanopublication graphs, we used the *D3 Force Layout*,³³ which is specifically designed to implement force-directed graphs. A force-directed graph is a graph where nodes are subjected to forces of two types: attractive and repulsive. These kinds of forces try to simulate physics

²⁹<https://www.djangoproject.com/>

³⁰<https://getbootstrap.com/>

³¹<https://jquery.com/>

³²<https://d3js.org/>

³³https://d3-wiki.readthedocs.io/zh_CN/master/Force-Layout/

scenarios where particles attract or repel each other. Here, the particles are the nodes of the graph, and the edges represent the presence of forces between nodes. When a new instance of a force-directed layout is created, a new D3 simulation starts, and the nodes become subjected to forces. The force-directed layout can be used both for cyclic and acyclic graphs, which can be either directed or not.

To implement the graph exploration, we developed a custom, collapsible force-directed layout where nodes can be expanded or collapsed at will. This layout enables a user-friendly exploration of graphs leveraging on a functional disposition of children nodes around the parents.

In particular, Figure 7.8 shows that children nodes are displayed around parents at evenly spaced angles of an arc. This disposition is designed to facilitate the horizontal expansion of the graph and prevent nodes from overlapping in a multi-level expansion. The custom force-directed layout developed and the NanoWeb code are publicly available³⁴.

The screenshot shows the 'Advanced search' interface. The search bar contains 'mutL homolog 1' under the 'GENE' topic. Below the search bar are four filter buttons: 'Nanopub List Only', 'Display Both', 'Nanopub Info Only', and 'Show Graph Layer'. The results table below shows five entries for 'mutL homolog 1' associated with various diseases, each with a 'Go to: DisGeNET' link.

| | |
|--|---------------------------------|
| mutL homolog 1 - Hereditary Nonpolyposis Colorectal Cancer gene-disease association linked with genetic variation | Go to: DisGeNET |
| mutL homolog 1 - C2020284 gene-disease biomarker association | Go to: DisGeNET |
| mutL homolog 1 - Hereditary Nonpolyposis Colorectal Cancer gene-disease association linked with genetic variation | Go to: DisGeNET |
| mutL homolog 1 - Endometrial Carcinoma gene-disease association linked with altered gene expression | Go to: DisGeNET |
| mutL homolog 1 - Neoplasm Metastasis gene-disease biomarker association | Go to: DisGeNET |

[Load More](#)

Fig. 7.10 Advanced search: search for nanopublications regarding the *mutL homolog 1* gene.

³⁴<https://github.com/giachell/nanoweb>

Advanced search

In addition to keyword search, we introduced the advanced search to guide users in query formulation. The advanced search is based on structured terms that can be general purpose (e.g. nanopublication URLs, author ORCID and scientific evidence identifiers) or domain-specific (e.g. genes, diseases, proteins and tissues). Figure 7.10 shows one of the configurations available in the advanced search interface. The interface is based on filters enabling the users to perform boolean search and restrict the search results. Users can choose the search modality in the *Search by* drop-down menu, marked with number one in Figure 7.10. The interface provides four different search modalities:

1. **Topic:** topic-based search is domain-specific, and it allows the user to find nanopublications for a specific topic. Currently, the available topics are genes, diseases, proteins, and tissues. The user can specify the chosen topic in the *Choose topic* drop-down menu, indicated with number two in Figure 7.10. The user can also specify the name of the entity that s/he is looking for in the *Entity name* input field, marked with number three in Figure 7.10. For instance, in Figure 7.10 the chosen topic is *GENE* and the gene name is *mutL homolog 1*. Since gene and protein names could be quite complex to remember, the *Entity name* input field is provided with an auto-completion functionality. Once the user specifies the details about the topic, the list of related nanopublications is returned, so that the user can visualize and explore them as described for the keyword search interface.
2. **Author:** allows the user to find all the nanopublications related to a nanopublication/evidence author. The provided author could be a nanopublication author or the author of the scientific publications containing the evidence of nanopublication assertions. Users can search for a specific author by providing the author's name or her/his ORCID identifier. The author input field is provided with auto-completion for both author names and ORCID identifiers.
3. **Nanopublication ID:** using this mode, users can search for a specific nanopublication via its identifier/URL. The users can take advantage of the auto-completion feature to search for all the nanopublications.
4. **Evidence:** this mode allows the users to get all the nanopublications extracted from a given scientific publication (i.e., evidence) starting from the publication DOI or PubMed URL (e.g., <http://identifiers.org/pubmed/29970664>).

To define the advanced search interface filters we used structured terms (entities) collected from several public ontologies, databases and terminology resources concerning both life

science and medical domains. For instance, we consider *genes*, *diseases*, *proteins* and *tissues* categories that users can use as filters. The machine-readable versions of the entities are contained in the nanopublications indexed by NanoWeb. To obtain their human-readable version, we leverage on public ontologies and databases. From these resources the associated labels are extracted, stored into the NanoWeb database and then linked to the respective machine-readable entities. To do so, we used some ontologies: *Basic Formal Ontology (BFO)*³⁵, *Chemical Entities of Biological Interest Ontology (CHEBI)*³⁶, *Evidence and Conclusion Ontology (ECO)*³⁷, *Open Biological and Biomedical Ontology (OBO)*³⁸, *Pathway Ontology (PW)*³⁹, *Semanticscience Integrated Ontology (SIO)*⁴⁰, *Sequence Ontology (SO)*⁴¹. Additionally, as terminology resources we employed the *National Center for Biotechnology Information (NCBI)*⁴², *National Cancer Institute Thesaurus (NCIT)*⁴³ and the *Unified Medical Language System (UMLS)*⁴⁴.

The entities extracted from the resources mentioned above are also used for the mapping of nanopublication assertions – originally modeled as machine-readable RDF statements – into a human-readable form. To do so, NanoWeb exploits the entity types to determine the proper visual representation of nanopublication assertions. For instance, in the case of a DisGeNET gene-disease association (dgn-gda), the entity types are *gene* or *disease*. The entities are represented as nodes labeled with the human-readable versions of the corresponding URI used in the RDF serialization of the nanopublication. The nodes are connected together by an oriented edge from *gene* to *disease*. As an example let us consider the assertion of the nanopublication with identifier: *RA3WLHsGFZrDU4kULrSa_pTa0gk8-mwadaj-LZ7kAqpog*:

```
miriam-gene:351 a ncit:C16612 .
lld:C0002395 a ncit:C7057 .
dgn-gda : DGNa4c88520d1a84e659043089fff632d78 sio : SIO_000628 miriam-gene
:351 , lld:C0002395 ;
a sio : SIO_001121 .
```

The assertion describes a *gene-disease association* (dgn-gda) between the NCBI gene *amyloid beta precursor protein* (miriam-gene:351) and the *Alzheimer's disease* (lld:C0002395).

³⁵<https://basic-formal-ontology.org/>

³⁶<https://www.ebi.ac.uk/chebi/>

³⁷<https://www.evidenceontology.org/>

³⁸<http://www.obofoundry.org/>

³⁹<https://rgd.mcw.edu/rgdweb/ontology/search.html>

⁴⁰<https://github.com/MaastrichtU-IDS/semanticscience>

⁴¹<http://www.sequenceontology.org/>

⁴²<https://www.ncbi.nlm.nih.gov/>

⁴³<https://ncit.nci.nih.gov/ncitbrowser/>

⁴⁴<https://www.nlm.nih.gov/research/umls/index.html>

The association type is more specifically a *gene-disease biomarker association* (SIO:001121). NanoWeb enriches the entities with additional information that can be inferred from the RDF graph of the nanopublication. For instance, additional information are the types of the entities – e.g. the fact that first entity (miriam-gene:351) is a gene (ncit:C16612) and that (lld:C0002395) is a disease (ncit:C7057). All these additional information are treated as entity properties that the user can access via the interactive visual representation of the nanopublication. The entity labels *amyloid beta precursor protein* and *Alzheimer's disease* are taken respectively from the NCBI and Linked Life Data platforms. The entity labels are resolved from entity identifiers by relying on public API endpoints such as the *Entrez Programming Utilities (E-utilities)*⁴⁵ provided by NCBI. Nanopublications from the same platform (e.g. DisGeNET, NeXtProt, Protein Atlas, and Wikipathways) use the same authorities to identify entities (e.g. genes, diseases, proteins and tissues). However, when nanopublications from different platforms are visualized, it is sometimes necessary to reconcile different resource identifiers across authorities to link the same entities to others using different identifiers. In the visual representation only one valid identifier is presented for each entity to keep the interface as clean as possible.

7.6 Expert users survey

To better understand the needs of the nanopublication community and improve the critical functionalities of NanoWeb, we conducted an expert users survey to collect feedback from nanopublication and domain experts. We advertised NanoWeb on the nanopublication public mailing lists, on social media targeting the potentially interested communities and private emails to the authors of papers about nanopublications. We asked the nanopublication experts involved in the survey to use NanoWeb, and then to answer a questionnaire. It should be noticed that we did not provide any tutorial to inform the users about NanoWeb functions because we also wanted to investigate how intuitive the system is for first-time users and how steep its learning curve is.

The survey was composed of sixteen questions (Q[1-16]) divided in four sections. The majority of the questions is answered through the Likert five-point scale, ranging from 1 to 5 points, meaning different things depending on the question.

1. **Personal information.** This section is composed of four questions and collects basic information about the participants and their experience with nanopublications:

⁴⁵<https://www.ncbi.nlm.nih.gov/books/NBK25501/>

- **Q1:** *Do you have any experience with nanopublications?*
In this case the answer with 1 point in the Likert scale means: “Not at all” (i.e., I heard someone mentioning nanopublications once), while the 5 points one means: “Quite a lot” (i.e., I created some nanopublications myself)
- **Q2:** *Current Position?*
Single choice between: Academic, Industry, Master Student, PhD Student, Post-Doc.
- **Q3:** *Primary domain of expertise?*
Multiple choices between: Art and architecture, Biology, Chemistry, Communication Science, Computers and the humanities, Computer Science, Economics, Life Science, Linguistics, Mathematics, Medicine, Physics, Psychology, Sociology.

The survey considered fourteen participants in total, counting seven highly-experienced users (5 on the Likert scale) and nine experienced users (4 on the Likert scale). According to the data collected, the majority of the participants (85.7%) are from Academia. Also, according to Q3, the main domains of expertise of the participants are: Computer Science (57.1%), Chemistry (35.7%), Life Science (35.7%), Biology (28.6%), Medicine (14.3%). Computer Science indicates experts in the creation of nanopublications from the technical viewpoint, whereas the others are domain experts who might curate or use nanopublications in their daily work.

2. **The relevance of the addressed problem.** This section explores the existence and quality of other services enabling search, access, exploration, and re-use of nanopublications (all questions are answered according to a 1 (not at all) to 5 (quite a lot) Likert scale):

- **Q4:** *Is searching, accessing, and consulting nanopublications relevant for the stakeholders (e.g., researchers, developers, domain experts)?*
- **Q5:** *To the best of your knowledge, are the currently available tools and services adequate for searching and accessing nanopublications?*
- **Q6:** *To the best of your knowledge, do other tools and services offer interactive visualizations to interact with nanopublications?*
- **Q7:** *To the best of your knowledge, do other available tools and services offer visual exploration possibilities of the nanopublication relation network?*

According to the data collected for questions Q[4-7], the majority of the participants (57%) considers the problem addressed by NanoWeb relevant or very relevant, pointing

out the lack of other tools and services for the interactive visualization and exploration of nanopublications and their relation network. About Q5, 50% of the participants consider the currently available tools and services for searching and accessing nanopublications inadequate (1 or 2 points on the Likert scale) and 42% are not enthusiastic about them (3 points on the Likert scale). 71% of the participants answered that there are no other available tools offering interactive visualizations of nanopublications and 57% say there are no alternative tools to visually explore the nanopublication network. From these answers, we can see that the participants confirm our analysis highlighting the lack of intuitive and visual tools for the access and exploration of the nanopublications despite the confirmed utility of searching and accessing nanopublications for the stakeholders.

3. **NanoWeb - Search Engine and Interface.** The questions of this section are designed to evaluate the search capabilities of NanoWeb and the usability of its interface. This section was answered by twelve participants over fourteen.

- **Q8:** *Is NanoWeb search interface intuitive and easy-to-use?*
- **Q9:** *Is NanoWeb capable of retrieving relevant nanopublications for a given query?*
- **Q10:** *In your opinion, is a search based on keywords an effective way to seek for nanopublications?*
- **Q[11-12]:** *In your opinion, for the not technologically savvy, what is the most effective way to search nanopublications? Q11 and Q12 are the same, but the answers are different since for Q11 the range of answers is from 1: SPARQL end-point to 5: Keyword-based search; whereas, for Q12 the range is from 1: Faceted search to 5: Keyword search.*
- **Q13:** *“Will NanoWeb enhance the productivity of involved stakeholders (researchers, developers, nanopublication experts)?”*

About question Q8, the majority of the participants consider NanoWeb search interface intuitive and easy-to-use (75% answered 4 or above and none answered below 3). There is no accordance instead for Q9 (median = 3, mean = 3.08, STD = 1.04), 42% of the participants answered 3 which means “not sure” and the rest of them is divided into the two other classes “not really” (≤ 2 : 33%) and “quite a lot” (≥ 4 : 25%). One reason that could motivate this kind of distribution might be that participants did not know what they could search in advance, thus many user queries might have not produced the expected results. To address this issue, after the survey we introduced

the advanced search which guides users on NanoWeb search capabilities. Participants are well-distributed for Q10 (median = 3, mean = 3.25, STD = 1.16), there is not a preferred opinion about keyword search; nevertheless, 46% of the participants consider the search based on keywords quite an effective or highly effective (answer 4 or above) way to seek for nanopublications. About Q[11-12], the majority of the participants (58%) consider that keyword-based search is more effective than SPARQL end-point but less effective than faceted search (67%) for the non-technologically savvy. This answer shows how domain experts are more accustomed to use faceted search rather than keyword search for searching structured data as nanopublications are. Keyword search is considered useful, but it should not substitute faceted search as a means to access RDF scientific data. Finally, all the participants believe NanoWeb can moderately (58%) or substantially (42%) enhance the productivity of researchers and nanopublication experts.

4. **NanoWeb - Visual Exploration.** This section of the questionnaire evaluates the experience with the NanoWeb user interface for visual exploration of nanopublications. We designed the questions of this section to investigate whether the visual exploration of nanopublication graphs could lead to the discovery of meaningful relationships and information potentially unknown to the experts. Moreover, we asked the participants to compare NanoWeb with the currently available alternative tools. This section consists of three questions:

- **Q14:** *Do you feel comfortable with the interface for the visual exploration?*
- **Q15:** *Could the visual exploration of the nanopublication graphs lead to the discovery of meaningful relationships and information not known in advance?*
- **Q16:** *Is NanoWeb visual exploration innovative with respect to the currently available alternative tools and techniques?*

With reference to Q14, the majority (64%) of the participants felt very comfortable with the interface for the visual exploration and only 14% gave a score below three points. Moreover, 57% of the participants believe the visual exploration of the nanopublication graphs could lead to the discovery of meaningful relationships and information not known in advance. Finally, half of the participants think that NanoWeb is highly innovative (four or five points) with respect to the state of the art, while only 21% thinks it is only marginally innovative.

7.6.1 User feedback

Finally, we asked the participants to provide some feedback and suggestions to improve NanoWeb. The feedback collected shows that users have appreciated the system:

- *“I very much appreciate the tool, and I think it can be a great push for better accessing and using nanopublications by everyone!”*
- *“I consider the NanoWeb proposal a smart insight for searching nanopublications.”*

We also received useful suggestions to improve the system:

- *“I found the visual exploration innovative, but I think it could be improved by a better UI/UX.”*
- *“Good work! I would suggest that you enable URL-based searching.”*
- *“Consider replacement of keyword search with a concept-based search. This can also be used to enable auto-suggest functionality based on the resources (genes, diseases, etc)”*
- *“I really like the application, but at the end of the day it is dependent on the indexed data. It would be great if there were a possibility to suggest datasets to be included or even better, to be able to add them myself!”*
- *“Downloading of the results as a dataset of nanopublications would be most welcome too. Even better, a Cytoscape plugin that allows me to pull in the full network. I’m looking forward to seeing where you are taking this. Success!”*

We consider the user feedback of great value, so we decided to improve NanoWeb according to the received suggestions. Firstly, we improved both the user interface and experience (UI/UX), providing a responsive mobile device layout. Then, we improved the search system so that a user can perform URL-based searching. Currently, NanoWeb allows the users to find the authors from the ORCID ids; a specific nanopublication from its URL/identifier; and, all the nanopublications related to one particular evidence paper provided its DOI.

The prominent feature we added to NanoWeb, thanks to the user feedback, is the advanced search, as described in Section 7.5. The Advanced search interface is based on structured terms extracted from the life science domain, it enables users to search for nanopublications based on topics (e.g., genes, diseases, proteins, etc.), scientific evidence, and authors. Finally, based on the collected feedback, we planned several further improvements to the system that we discuss as future work in Section 7.8.

7.7 Discussion on maintaining aspects

NanoWeb aims to provide users unified access to nanopublications and to search and explore them through a human-readable interface. Since NanoWeb is tailored for both the life science and medical domains, it is designed to help the experts of these domains in their research work. It also allows users that do not have a prior knowledge about nanopublication to easily interpret and understand the returned content.

Several challenges need to be addressed to maintain a stable, citable system like what NanoWeb aims to be. The major system maintaining challenges are:

1. **Ensure persistent access and re-use of data:** to guarantee persistent and reliable access to data and avoid broken URLs, NanoWeb uses persistent URLs and identifiers to refer to resources. All the indexed nanopublications are directly accessible through a persistent URL provided by the *W3C Permanent Identifier Community Group*⁴⁶. The nanopublication's persistent URL format is: `http://w3id.org/nanoweb/landingpage/<ID>`, where the *ID* in brackets is the nanopublication identifier and satisfies the regular expression: `^[A-Za-z0-9_\-]{43}$`. Nanopublications use persistent identifiers, that allow to access them across different providers. Even if one of the several nanopublication providers is unreachable in a given moment, the others can provide access by using the same identifiers. As for nanopublications, NanoWeb itself is reachable through the persistent URL: `http://w3id.org/nanoweb/`.
2. **Long-term preservation of resources:** every information concerning nanopublications is saved in NanoWeb databases, that are stored in network hard drives using redundancy policies such as *Redundant Array of Independent Disks (RAID)*. The redundancy policies adopted and daily back-up routines are designed to prevent loss of data and ensure long-term preservation.
3. **Ongoing hosting:** NanoWeb is hosted within the cloud architecture of the University of Padova. The institutional cloud architecture and network infrastructure provide a reliable connection service as well as a protection layer from external attacks. A team of system administrators actively control the cloud/network infrastructure and support NanoWeb. NanoWeb is developed in the context of the European project *ExaMode*⁴⁷ which guarantees financial support until 2023. Within the project there are sustainability policies that should guarantee the maintenance of the developed tools well beyond the termination of the project.

⁴⁶<https://w3id.org/>

⁴⁷European Union Horizon 2020 program under Grant Agreement no. 825292

7.8 Conclusions

Scientific and scholarly communications are growing at an incredible speed, and it is hardly possible to keep track of the discoveries and statements presented in the literature, even considering only a specific domain. Moreover, the “redundancy of statements in multiple fora makes it difficult to identify attribution, quality, and provenance” [117]. Hence, the nanopublication model has been proposed to quickly identify, search, and access scientific facts extracted from papers. Nanopublications are represented as graphs centered on a scientific statement (i.e., the assertion) that makes provenance, attribution, and scientific information machine-readable.

Nanopublications are concise noise-free resources characterized by high information density. Leveraging on the semantic-oriented RDF structure, nanopublications efficiently convey information and concepts. Hence, these features make nanopublications particularly suitable for enabling data search, information extraction, and automatic reasoning over scientific facts. Despite the promising features of nanopublications, their use is still restricted to highly-specialized scientific circles.

The central limit to the full exploitation of nanopublications is the lack of services enabling their search, access, exploration, and re-use. Search is limited to the use of structured query languages as SPARQL, and a service to search over all the publicly available nanopublications at once is not available. Nanopublications are machine-readable, but no human-readable counterpart is generated and open to the public. Nanopublications create a vast relation network of scientific facts that could lead to discoveries, but up to now, there are no automatic or manual services enabling graph exploration.

The goal of this work is to provide unified access to Life Science nanopublications in order to allow users to search, access, explore, and re-use them on the Web. To this end, we have designed and developed a web application called *NanoWeb*, that allows the users to (i) search for domain-specific nanopublications using keywords (as they are accustomed to do with Web search engines); (ii) explore their relation network to discover new nanopublications and meaningful connections; (iii) access and understand their content; (iv) connect to the evidence paper and access the related data record in external curated scientific databases; and, (v) easily cite nanopublications when they are re-used in new scientific contexts.

We also presented the benefits of the serendipity-oriented perspective enabled by *NanoWeb* in the Life Science domain. We showed how the exploration of nanopublication graphs could enrich domain knowledge and point out interesting gene-disease connections.

As future work, we plan to extend the system by providing the user with the capability of exploring a new graph generated from an arbitrary set of Life Science nanopublications selected by the user. This functionality represents a significant improvement for the graph

exploration since the initial relation network already considers different nanopublications, instead of starting the graph exploration from a single one. In this way it is possible to highlight, for instance, the set of common diseases due to a selection of genes or, conversely, the set of common genes that cause the disease of interest. Moreover, we plan to crawl and index the Life Science nanopublications that are not currently available on the Web, if not downloading large archive files which are hardly usable.

As future work, we plan to further improve NanoWeb according to the expert users survey's feedback. We will allow the users to add datasets or other domain-specific nanopublication sources to be crawled and indexed by the system. We will add the possibility to select and download custom-made sets of nanopublications. We will propose a customized user experience to save lists of favorite nanopublications, entities, and associations and notify when something new is published.

We will dedicate a fair amount of work to the extension of search functionalities to improve keyword search and to include faceted search which is required by the stakeholders. Indeed, faceted search is commonly adopted solution [24] to search RDF data. A faceted search is particularly useful when it is applied to domain-specific data. For instance, in gene-disease associations, the faceted search can be used to search for specific genes or specific diseases, filtering out all the entities not relevant to the search. Faceted search can be associated with auto-completion functionalities to ease the users' work. Finally, we plan to improve keyword-based searches with ontology and database ID lookups.

Chapter 8

Conclusions and Future Work

In this final chapter, we summarize the key contributions of this work and devise some future research directions.

8.1 Conclusions

In this work, we presented the ExaSURE ecosystem providing users unified access to all the tools/services, we developed for the DPATH domain. The purpose of the presented tools is to support pathologists and experts in the decision-making process for DPATH, enabling pathologists to (i) search, access, annotate, and explore clinical reports and (ii) explain the results of algorithms designed for supporting diagnostic activities. The context of my research work has been the ExaMode¹ H2020 European project, which aims to develop tools to support the decision-making process and knowledge discovery in the DPATH domain, leveraging exascale multimodal medical data. Specifically, the ultimate goal of the project is to develop a system capable of automatically classifying images from the DPATH domain, such as WSIs, to determine whether each image presents cancer indications and the eventual grade. In this regard, a weakly supervised approach is adopted where such DL-based image classification system is trained using weak annotations automatically generated. To this aim, we introduced SKET, an unsupervised knowledge extraction tool for extracting labels (weak annotations), mentions, and concepts from pathology reports provided in natural language. The described image classification system presents several benefits: (i) reduce the workload of pathologists that could verify the automatic annotations done by the system without starting their analysis from scratch; (ii) speed up the image analysis task since the manual analysis is time-consuming; (iii) reduce missing data and human errors. Hence, the

¹<https://www.examode.eu>

work presented in this thesis tackles two major issues: (i) the lack of large annotated datasets for training and evaluating DL-based algorithms for DPATH and (ii) the black-box nature of the models involved, which hinders the human comprehension of models' outcomes and trust on the underlying machine decision process. To this aim, we adopted an approach that contributes to multiple lines of research, including:

- **Semantic annotation:** we introduce MedTAG [107], a customizable annotation tool for ground truth creation from free-text clinical reports. MedTAG aims to facilitate the creation of annotated data, in particular ground truth labels, through intuitive interfaces providing automatic annotation facilities to speed up the burdensome annotation task. To this aim, MedTAG integrates SKET for the knowledge extraction process. Moreover, MedTAG has been used by pathologists and experts to annotate more than seven thousand clinical reports in three languages (Dutch, English, and Italian), from two healthcare institutions (the Cannizzaro Hospital (AOEC), Catania, Italy and the Radboud University Medical Center (RUMC), Nijmegen, The Netherlands). MedTAG is provided as a docker container to make it distributable, platform-independent, and easy to install/deploy. We designed MedTAG according to the five requirements (i.e. available, distributable, installable, workable, and schematic) defined in a recent extensive review of manual annotation tool [219]. In this regard, MedTAG satisfies 20 over 22 criteria defined in the same study. Finally, we conducted a quantitative comparison to evaluate MedTAG in terms of efficiency for the annotation process, against a set of well-established annotation tools. It is worth noting, that the quantitative comparison was conducted using automatic agents to guarantee a fair comparison, that is, only in terms of the number of actions (e.g., mouse clicks) and time required to achieve the same annotation with the different tools. From the analyses conducted, it emerges that MedTAG is faster than other tools involved in the comparison, especially for the document-level annotation task.
- **Explainability:** we presented SKET X [197], an explainability tool that exploits VA techniques to support pathologists and experts in the comprehension of SKET outputs as well as the rules, models, and parameters involved in the knowledge extraction process. SKET X integrates an instance of SKET, in a web-based setting, to enable the users to interact with it through intuitive interfaces that provide visual explanations for SKET outcomes. Users can comprehend its results - e.g., identify visually which concepts are related to a specific mention or label - and get useful insights concerning the knowledge extraction process. Thus, SKET X allows the pathologists not only to understand why a specific outcome has been obtained but also to assess the correctness of the rules involved to foster further improvements for SKET. In addition, We pointed

out that SKET X goes beyond the static visualization of the outcomes of SKET; it allows the users to activate/deactivate a model (e.g., the neural model), change the values for its parameters, and run SKET again with a different configuration. This is useful, for instance, for comparing the outcomes of SKET according to different parameters.

- **Knowledge exploration:** we introduced NanoWeb [106], that is, a unified access point for the world of nanopublications (i.e., concise LOD representations of scientific facts provided as machine-readable graphs based on the RDF structure). NanoWeb allows users to (i) search for nanopublications in natural language (e.g., using keywords); (ii) explore the interconnected network of scientific facts, encoded in the nanopublications' assertions, and discover meaningful connections; (iii) access and understand their content; (iv) reach evidence papers and access the related data record in external curated scientific databases; and, (v) cite nanopublications to re-use them in new scientific contexts. NanoWeb aims to enable the discovery of new meaningful findings from the exploration of scientific interconnected data, in a serendipity-oriented perspective. In the DPATH domain, this translates, for instance, to enriching the clinical reports with scientific evidence coming from the flourishing nanopublications network (e.g., which genes are involved in causing cancer in a specific tissue).

8.2 Directions for Future Work

As future work, we plan to integrate into the ExaSURE ecosystem an active learning system to take advantage of the feedback provided by pathologists and experts to continuously improve the effectiveness of the knowledge extraction process, in terms of the quality of the weak annotations produced. As a direct consequence, since image classification algorithms for DPATH are trained on such weak annotations, their effectiveness could be improved as well. Specifically, we envision a scenario where pathologists interact with SKET X and identify, for instance, a wrong concept predicted for a given report. Then, they inspect the concept to unveil the reason why that specific prediction has occurred. Finally, they provide feedback, through the interface, regarding their discovery and SKET X takes care of propagating it to a supervised model that updates and fixes the predictions accordingly. Moreover, we plan to improve SKET and the other tools integrated with the ExaSURE ecosystem. In particular, we plan to extend SKET to other emerging use-cases, such as Celiac disease, since its prevalence has significantly increased over the last 20 years [160]. For what concerns MedTAG, we plan to improve it by adding the support for (i) overlapping mentions; (ii) relationship annotations; (iii) active learning capabilities; (iv) PDF annotation; (v) automatic annotation also for other

use-cases such as for the Celiac disease. Since usability is a critical issue in clinical practice, we aim to conduct, in addition to the quantitative evaluation yet carried out, also a user study for assessing the usability of MedTAG. Even though we designed its interface so that it is as streamlined as possible, to let the users focus on the annotation task, we believe the precious suggestions of other pathologists and clinicians could improve not only the interface itself but also the features provided for the different annotation tasks. Similarly, we plan to conduct a user study to collect pathologists' and experts' feedback to improve SKET X accordingly. Specifically, we plan to conduct an online-only user study, delivered asynchronously, so that users can start it when they prefer. To this aim, we intend to provide the participants with (i) a video demonstrating SKET X functionalities; (ii) anonymized credentials to access SKET X; and (iii) a private link to an online form where they can answer a set of predefined questions and provide their feedback. Moreover, we plan to divide the user study into two parts, designed to measure respectively the *learnability* and *usability* of SKET X. The first part aims to assess the confidence and the awareness of the users with respect to accomplishing a set of predefined explainability tasks - e.g., identifying the concepts related to a specific mention or label. Then, the collected answers will be compared with the correct ones, in order to assess each user's proficiency with SKET X. Secondly, we plan to evaluate SKET X in terms of usability and user satisfaction using the SUS scale, since it is considered an industry standard for assessing usability [41]. Finally, we aim to collect useful comments and suggestions from pathologists and other experts to identify key needs and foster further advancements in the design of transparent and explainable models and algorithms for CPATH. Instead, for what concerns NanoWeb, we plan to improve it according to the stakeholders' needs and suggestions that emerged from the survey we conducted. Specifically, we plan to (i) integrate faceted search functionalities in the interface (as requested by computational biologists); (ii) improve keyword-based searches with ontology and database ID lookups; (iii) allow the users to load into NanoWeb their own nanopublications datasets or indicate other nanopublication sources to be automatically crawled and indexed by the system; (iv) enable download facilities, such as custom sets of nanopublications of interest; (v) customize the UX by providing users the ability to create personal accounts to save lists of favorite nanopublications, entities, and associations as well as being notified when something new is published (vi) provide users new capabilities of exploring graphs generated from an arbitrary set of nanopublications selected by the user. Indeed, this is not possible in the current version of NanoWeb, since to date graph exploration starts always from a specific nanopublication indicated by the user. This new feature could be used to highlight, for instance, the set of common diseases due to a selection of genes or, conversely, the set of common genes that cause the disease of interest.

Finally, we envision a scenario where the improved integration of the tools/services in the ExaSURE ecosystem could enhance the support for the medical decision-making process by leveraging enriched clinical reports with knowledge coming also from external sources in a holistic-oriented perspective.

List of acronyms

| | |
|--|-------|
| ABox Assertional Box | 54 |
| AI Artificial Intelligence | 2 |
| AOEC Cannizzaro Hospital | xxiii |
| API Application Program Interface | 60 |
| CAD Computer-Assisted Diagnostic | 2 |
| CAP College of American Pathologists | 38 |
| CDA Confirmatory Data Analysis | 13 |
| CNN Convolutional Neural Network | xxiii |
| CPATH Computational Pathology | 1 |
| DA Data Analysis | 15 |
| DARPA Defense Advanced Research Projects Agency | 22 |
| DHS Department of Homeland Security | 14 |
| DL Deep Learning | 2 |
| DM Data Mining | 15 |
| DPA Digital Pathology Association | 10 |
| DPATH Digital Pathology | 1 |
| DV Data Visualization | 17 |
| EDA Exploratory Data Analysis | 13 |
| EHR Electronic Health Records | 12 |
| EL Entity Linking | 39 |
| ExaSURE ExaSURE System for Unified Resource Exploration | viii |
| FHIR Fast Healthcare Interoperability Resources | 27 |
| GDPR General Data Protection Regulation | vii |
| GPM Gestalt Pattern Matching | xv |
| HCI Human Computer Interaction | 15 |
| HL7 Health Level Seven International | 27 |

| | |
|--|-------|
| IAA Inter-Annotator Agreement | 78 |
| IA Information Access | 4 |
| IE Information Extraction | 75 |
| IHE Integrating the Healthcare Enterprise | 27 |
| IMS Image Management System | 19 |
| IR Information Retrieval | 18 |
| IRI Internationalized Resource Identifier | 54 |
| IV Information Visualization | 4 |
| KDD Knowledge Discovery in Databases | 13 |
| LIME Local Interpretable Model-Agnostic Explanation | 23 |
| LIS Laboratory Information System | 2 |
| LOD Linked Open Data | 5 |
| MIL Multiple Instance Learning | 11 |
| ML Machine Learning | 2 |
| MLNN Multi-Layer Neural Network | 23 |
| NER Named Entity Recognition | 39 |
| NER+L Named Entity Recognition and Linking | 75 |
| NHS Nottingham Histologic Score | 20 |
| NLM Neural Language Model | 37 |
| NLP Natural Language Processing | 4 |
| NMT Neural Machine Translation | 38 |
| NVAC National Visualization and Analytics Center | 14 |
| PVA Progressive Visual Analytics | 17 |
| RAM Random Access Memory | 11 |
| RDF Resource Description Framework | 6 |
| REST REpresentational State Transfer | 60 |
| RNN Recurrent Neural Network | 23 |
| RoI Region of Interest | 3 |
| RUMC Radboud University Medical Center | xxiii |
| SaaS Software as a service | 96 |
| SHAP SHapley Additive exPlanations | 23 |
| SKET Semantic Knowledge Extractor Tool | vii |
| SKET X SKET eXplained | vii |
| SUS System Usability Scale | 74 |

| | |
|--|-------|
| TBox Terminological Box | 54 |
| UX User Experience | 98 |
| VA Visual Analytics | 4 |
| WSI Whole Slide Image | xxiii |
| XAI eXplainable Artificial Intelligence | 3 |

References

- [1] Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., et al. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 249(3):286–294.
- [2] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.
- [3] Aeffner, F., Forest, T., Schumacher, V., Zarella, M., and Bradley, A. (2022). Digital pathology and tissue image analysis. In *Haschek and Rousseaux's Handbook of Toxicologic Pathology*, pages 395–421. Elsevier.
- [4] Aeffner, F., Wilson, K., Martin, N. T., Black, J. C., Hendriks, C. L. L., Bolon, B., Rudmann, D. G., Gianani, R., Koegler, S. R., Krueger, J., and Young, G. D. (2017). The gold standard paradox in digital image analysis: Manual versus automated scoring as ground truth. *Archives of pathology & laboratory medicine*, 141(9):1267–1275.
- [5] Aeffner, F., Zarella, M. D., Buchbinder, N., Bui, M. M., Goodman, M. R., Hartman, D. J., Lujan, G. M., Molani, M. A., Parwani, A. V., Lillard, K., et al. (2019). Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *Journal of pathology informatics*, 10(1):9.
- [6] Agosti, M., Di Nunzio, G. M., and Marchesin, S. (2020). A Post-Analysis of Query Reformulation Methods for Clinical Trials Retrieval. In *Proc. of the 28th Italian Symposium on Advanced Database Systems, Villasimius, Sud Sardegna, Italy (virtual due to Covid-19 pandemic), June 21-24, 2020*, volume 2646 of *CEUR Workshop Proceedings*, pages 152–159. CEUR-WS.org.
- [7] Agrawal, S., Chaudhuri, S., and Das, G. (2002). Dbxplorer: A system for keyword-based search over relational databases. In *Proceedings of the 18th International Conference on Data Engineering, ICDE 2002*, pages 5–16. IEEE Computer Society.
- [8] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. *CoRR*, abs/1904.03323.
- [9] Alturkistani, H. A., Tashkandi, F. M., and Mohammedsaleh, Z. M. (2016). Histological stains: a literature review and case study. *Global journal of health science*, 8(3):72.

- [10] Amith, M. and Tao, C. (2018). Representing Vaccine Misinformation using Ontologies. *Journal of Biomedical Semantics*, 9(1):22.
- [11] Andrienko, G. L., Andrienko, N. V., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., Jern, M., Kraak, M., Schumann, H., and Tominski, C. (2010). Space, time and visual analytics. *Int. J. Geogr. Inf. Sci.*, 24(10):1577–1600.
- [12] Andrienko, G. L., Andrienko, N. V., Kopanakis, I., Ligtenberg, A., and Wrobel, S. (2008). Visual analytics methods for movement data. In *Mobility, Data Mining and Privacy*, pages 375–410. Springer.
- [13] Andrienko, N. V. and Andrienko, G. L. (2013). Visual analytics of movement: An overview of methods, tools and procedures. *Inf. Vis.*, 12(1):3–24.
- [14] Angelini, M., Blasilli, G., Bonomi, S., Lenti, S., Palleschi, A., Santucci, G., and De Paoli, E. (2021). Bucephalus: a business centric cybersecurity platform for proactive analysis using visual analytics. In *2021 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 15–25. IEEE.
- [15] Angelini, M., Blasilli, G., Farina, L., Lenti, S., and Santucci, G. (2019a). Nemesis (network medicine analysis): Towards visual exploration of network medicine data. In *VISIGRAPP (3: IVAPP)*, pages 322–329.
- [16] Angelini, M., Bonomi, S., Lenti, S., Santucci, G., and Taggi, S. (2019b). Mad: A visual analytics solution for multi-step cyber attacks detection. *Journal of Computer Languages*, 52:10–24.
- [17] Angelini, M. and Cazzetta, G. (2020). Progressive visualization of epidemiological models for covid-19 visual analysis. In *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications*, pages 163–173. Springer.
- [18] Angelini, M., Fazzini, V., Ferro, N., Santucci, G., and Silvello, G. (2018a). CLAIRE: A combinatorial visual analytics system for information retrieval evaluation. *Information Processing & Management*, in print.
- [19] Angelini, M., Ferro, N., Santucci, G., and Silvello, G. (2014). VIRTUE: A Visual Tool for Information Retrieval Performance Evaluation and Failure Analysis. *J. Vis. Lang. Comput.*, 25(4):394–413.
- [20] Angelini, M., Ferro, N., Santucci, G., and Silvello, G. (2016). A visual analytics approach for what-if analysis of information retrieval systems. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1081–1084.
- [21] Angelini, M., Prigent, N., and Santucci, G. (2015). Percival: proactive and reactive attack and response assessment for cyber incidents using visual analytics. In *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 1–8. IEEE.
- [22] Angelini, M., Santucci, G., Schumann, H., and Schulz, H. (2018b). A review and characterization of progressive visual analytics. *Informatics*, 5(3):31.

- [23] Ankerst, M. (2001). *Visual data mining*. PhD thesis, Ludwig Maximilian University of Munich, Germany.
- [24] Arenas, M., Cuenca Grau, B., Kharlamov, E., Marciuška, S., and Zheleznyakov, D. (2016). Faceted search over RDF-based knowledge graphs. *Journal of Web Semantics*, 37-38:55 – 74.
- [25] Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.
- [26] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.
- [27] Balmin, A., Hristidis, V., Koudas, N., Papakonstantinou, Y., Srivastava, D., and Wang, T. (2003). A system for keyword proximity search on XML databases. In *Proceedings of 29th International Conference on Very Large Data Bases, VLDB*, pages 1069–1072. Morgan Kaufmann.
- [28] Basole, R. C., Park, H., Gupta, M., Braunstein, M. L., Chau, D. H., and Thompson, M. (2015). A visual analytics approach to understanding care process variation and conformance. In *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare*, pages 1–8.
- [29] Bast, H., Buchhold, B., and Haussmann, H. (2016). Semantic search on text and knowledge bases. *Foundations and Trends in Information Retrieval (FnTIR)*, 10(2-3):119–271.
- [30] Bejnordi, B. E., Veta, M., Diest, P. J. V., Ginneken, B. V., Karssemeijer, N., Litjens, G., Laak, J. A. W. M. V. D., Hermsen, M., Manson, Q. F., and Balkenhol, M. (2017a). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA*, 318(22):2199–2210.
- [31] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017b). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.
- [32] Benson, A. B., Venook, A. P., Al-Hawary, M. M., Cederquist, L., Chen, Y. J., Ciombor, K. K., Cohen, S., Cooper, H. S., Deming, D., and Engstrom, P. F. (2018). NCCN guidelines insights: colon cancer, version 2.2018. *Journal of the National Comprehensive Cancer Network*, 16(4):359–369.
- [33] Benson, T. and Grieve, G. (2016). *Principles of health interoperability: SNOMED CT, HL7 and FHIR*. Springer.
- [34] Bhalotia, G., Hulgeri, A., Nakhe, C., Chakrabarti, S., and Sudarshan, S. (2002). Keyword searching and browsing in databases using BANKS. In *Proceedings of the 18th International Conference on Data Engineering*, pages 431–440. IEEE Computer Society.

- [35] Biryukov, M., Groues, V., Satagopam, V., and Schneider, R. (2018). Biokb-text mining and semantic technologies for biomedical content discovery.
- [36] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data – The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.
- [37] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.
- [38] Bokhorst, J., Nagtegaal, I. D., Fraggetta, F., Vatrano, S., Mesker, W., Vieth, M., van der Laak, J., and Ciompi, F. (2021). Automated risk classification of colon biopsies based on semantic segmentation of histopathology images. *CoRR*, abs/2109.07892.
- [39] Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.
- [40] Borgman, C. L. (2015). *Big Data, Little Data, No Data*. MIT Press.
- [41] Brooke, J. et al. (1996). Sus: A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.
- [42] Buckley, J. M., Coopey, S. B., Sharko, J., Polubriaginof, F., Drohan, B., Belli, A. K., Kim, E. M., Garber, J. E., Smith, B. L., Gadd, M. A., Specht, M. C., Roche, C. A., Gudewicz, T. M., and Hughes, K. S. (2012). The Feasibility of Using Natural Language Processing to Extract Clinical Information from Breast Pathology Reports. *J. Pathol Inform*, 3(1):23.
- [43] Burger, G., Abu-Hanna, A., de Keizer, N., and Cornet, R. (2016). Natural Language Processing in Pathology: a Scoping Review. *Journal of Clinical Pathology*, 69(11):949–955.
- [44] Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.
- [45] Caban, J. J. and Gotz, D. (2015). Visual analytics in healthcare—opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262.
- [46] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-Grade Computational Pathology using Weakly Supervised Deep Learning on Whole Slide Images. *Nat Med*, 25:1301–1309.
- [47] Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.-L., and Névéol, A. (2018). A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601.
- [48] Campos, D., Lourenço, J., Matos, S., and Oliveira, J. L. (2014). Egas: a collaborative and interactive document curation platform. *Database*, 2014.

- [49] Campos, D., Matos, S., and Oliveira, J. L. (2013). A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14(1):1–21.
- [50] Campregher, C., Honeder, C., Chung, H., Carethers, J. M., and Gasche, C. (2010). Mesalazine reduces mutations in transforming growth factor β receptor ii and activin type ii receptor by improvement of replication fidelity in mononucleotide repeats. *Clin Cancer Res*, 16(6):1950–1956.
- [51] Canales, L., Menke, S., Marchesseau, S., D’Agostino, A., del Rio-Bermudez, C., Taberna, M., and Tello, J. (2021). Assessing the performance of clinical natural language processing systems: Development of an evaluation methodology. *JMIR Med Inform*, 9(7):e20492.
- [52] Carbonneau, M. A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple Instance Learning: A Survey of Problem Characteristics and Applications. *Pattern Recognit.*, 77:329–353.
- [53] Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization - using vision to think*. Academic Press.
- [54] Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P. (2005). Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM Press.
- [55] Carroll, J. J. and Stickler, P. (2004). RDF triples in XML. In *Proc. of the WWW 2004 Conference (Alternate Track Papers & Posters)*, pages 412–413. ACM Press.
- [56] Casey, F., Negi, S., Zhu, J., Sun, Y. H., Zavodszky, M., Cheng, D., Lin, D., John, S., Penny, M. A., Sexton, D., et al. (2022). Omicsview: Omics data analysis through interactive visual analytics. *Computational and structural biotechnology journal*, 20:1277–1285.
- [57] Cejuela, J. M., McQuilton, P., Ponting, L., Marygold, S. J., Stefancsik, R., Millburn, G. H., Rost, B., Consortium, F., et al. (2014). tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. *Database*, 2014.
- [58] Cervin, I., Molin, J., and Lundström, C. (2016). Improving the creation and reporting of structured findings during digital pathology review. *Journal of pathology informatics*, 7(1):32.
- [59] Chang, W. C., Yu, H. F., Zhong, K., Yang, Y., and Dhillon, I. S. (2020). Taming pretrained transformers for extreme multi-label text classification. In *KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3163–3171. ACM.
- [60] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L. D., Kacprzak, E., and Groth, P. (2020). Dataset search: a survey. *VLDB J.*, 29(1):251–272.
- [61] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *J. Biomed. Informatics*, 34(5):301–310.

- [62] Cheney, J., Chiticariu, L., and Tan, W. (2009). Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474.
- [63] Cheplygina, V., de Bruijne, M., and Pluim, J. P. W. (2019). Not-so-supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis. *Medical Image Anal.*, 54:280–296.
- [64] Chichester, C., Gaudet, P., Karch, O., Groth, P. T., Lane, L., Bairoch, A., Mons, B., and Loizou, A. (2014). Querying neXtProt nanopublications and their value for insights on sequence variants and tissue expression. *J. Web Semant.*, 29:3–11.
- [65] Chichester, C., Karch, O., Gaudet, P., Lane, L., Mons, B., and Bairoch, A. (2015). Converting neXtProt into Linked Data and nanopublications. *Semantic Web*, 6(2):147–153.
- [66] Chiticariu, L., Li, Y., and Reiss, F. R. (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA*, pages 827–832. ACL.
- [67] Coffman, J. and Weaver, A. C. (2014). An Empirical Performance Evaluation of Relational Keyword Search Techniques. *IEEE Trans. Knowl. Data Eng.*, 26(1):30–42.
- [68] Cortez, P. and Embrechts, M. J. (2011). Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348. IEEE.
- [69] Cortez, P. and Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17.
- [70] Corvo, A., Caballero, H. G., Westenberg, M. A., van Driel, M. A., and van Wijk, J. J. (2020). Visual analytics for hypothesis-driven exploration in computational pathology. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):3851–3866.
- [71] Corvo, A., van Driel, M. A., and Westenberg, M. A. (2017). Pathova: A visual analytics tool for pathology diagnosis and reporting. In *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*, pages 77–83. IEEE.
- [72] Corvò, A., Westenberg, M. A., van Driel, M. A., and van Wijk, J. J. (2018). Visual analytics in histopathology diagnostics: a protocol-based approach. In *VCBM@ MICCAI*, pages 23–32.
- [73] Corvò, A., Westenberg, M. A., Wimberger-Friedl, R., Fromme, S., Peeters, M. M., van Driel, M. A., and van Wijk, J. J. (2019). Visual analytics in digital pathology: Challenges and opportunities. *VCBM*, pages 129–143.
- [74] Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525.

- [75] Courtiol, P., Tramel, E. W., Sanselme, M., and Wainrib, G. (2018). Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. *CoRR*, abs/1802.02212.
- [76] Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks*. The University of Wisconsin-Madison.
- [77] Cui, W. (2019). Visual analytics: A comprehensive overview. *IEEE Access*, 7:81555–81573.
- [78] Davenport, T. and Kalakota, R. (2019). The Potential for Artificial Intelligence in Healthcare. *Future Healthc J.*, 6(2):94–98.
- [79] de Castilho, R. E., Ide, N., Kim, J.-D., Klie, J.-C., and Suderman, K. (2019). Towards cross-platform interoperability for machine-assisted text annotation. *Genomics & Informatics*, 17.
- [80] del Toro, O. J., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., Müller, H., and Atzori, M. (2017). Analysis of Histopathology Images: From Traditional Machine Learning to Deep Learning. In *Biomedical Texture Analysis*, pages 281–314. Academic Press.
- [81] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019*, pages 4171–4186. ACL.
- [82] Dhrangadhariya, A., Otálora, S., Atzori, M., and Müller, H. (2021). Classification of noisy free-text prostate cancer pathology reports using natural language processing. In *ICPR, Artificial Intelligence for Digital Pathology Workshop (AIDP)*.
- [83] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.
- [84] Ding, H., Wang, C., Huang, K., and Machiraju, R. (2014). igpse: A visual analytic system for integrative genomic based cancer patient stratification. *BMC bioinformatics*, 15(1):1–13.
- [85] Ding, H., Wang, C., Huang, K., and Machiraju, R. (2015). Graphie: graph based histology image explorer. *BMC bioinformatics*, 16(11):1–11.
- [86] Dmitriev, K., Marino, J., Baker, K., and Kaufman, A. E. (2019). Visual analytics of a computer-aided diagnosis system for pancreatic lesions. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2174–2185.
- [87] Dobbie, S., Strafford, H., Pickrell, W. O., Fonferko-Shadrach, B., Jones, C., Akbari, A., Thompson, S., and Lacey, A. (2021). Markup: A web-based annotation tool powered by active learning. *Frontiers Digit. Health*, 3:598916.

- [88] Doğan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.
- [89] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [90] Dosso, D. and Silvello, G. (2020). Search text to retrieve graphs: A scalable RDF keyword-based search system. *IEEE Access*, 8:14089–14111.
- [91] Dzobo, K., Adotey, S., Thomford, N. E., and Dzobo, W. (2019). Integrating artificial and human intelligence: A partnership for responsible innovation in biomedical engineering and medicine. *Omics : a journal of integrative biology*.
- [92] Echle, A., Rindtorff, N. T., Brinker, T. J., Luedde, T., Pearson, A. T., and Kather, J. N. (2021). Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696.
- [93] Economist (2017). The world’s most valuable resource is no longer oil, but data. *The Economist: New York, USA*. <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> [online October 2020].
- [94] Elbassuoni, S. and Blanco, R. (2011). Keyword Search over RDF Graphs. In *Proc. of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011*, pages 237–242. ACM Press, New York, USA.
- [95] Ellis, D. W. and Srigley, J. (2016). Does standardised structured reporting contribute to quality in diagnostic pathology? the importance of evidence-based datasets. *Virchows Arch.*, 468(1).
- [96] Elston, C. W. and Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410.
- [97] Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., and Andrews, C. (2014). The human is the loop: new directions for visual analytics. *J. Intell. Inf. Syst.*, 43(3):411–435.
- [98] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.
- [99] Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., and Holzinger, A. (2022). The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, 133:281–296.
- [100] Fabris, E., Kuhn, T., and Silvello, G. (2019). A framework for citing nanopublications. In *Proc. of the 23rd International Conference on Theory and Practice of Digital Libraries, TPD L 2019*, pages 70–83.
- [101] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

- [102] Fine, J. L. (2014). 21st century workflow: a proposal. *Journal of pathology informatics*, 5(1):44.
- [103] Fraggetta, F., Garozzo, S., Zannoni, G. F., Pantanowitz, L., and Rossi, E. D. (2017). Routine digital pathology workflow: The catania experience. *Journal of pathology informatics*, 8(1):51.
- [104] Geraci, A. (1991). *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*. IEEE Press.
- [105] Giachelle, F., Dosso, D., and Silvello, G. (2021a). Nanoweb: Search, access and explore life science nanopublications on the web (discussion paper). In *Proceedings of the 29th Italian Symposium on Advanced Database Systems, SEBD 2021, Pizzo Calabro (VV), Italy, September 5-9, 2021*, pages 506–513.
- [106] Giachelle, F., Dosso, D., and Silvello, G. (2021b). Search, access, and explore life science nanopublications on the web. *PeerJ Comput. Sci.*, 7:e335.
- [107] Giachelle, F., Irrera, O., and Silvello, G. (2021c). MedTAG: a portable and customizable annotation tool for biomedical documents. *BMC Medical Informatics Decis. Mak.*, 21(1):352.
- [108] Giachelle, F. and Silvello, G. (2019). A progressive visual analytics tool for incremental experimental evaluation. In *Proceedings of the 10th Italian Information Retrieval Workshop, Padova, Italy, September 16-18, 2019*, pages 2–5.
- [109] Giannaris, P., Al-Taie, Z., Kovalenko, M., Thanintorn, N., Kholod, O., Innokenteva, Y., Coberly, E., Frazier, S., Laziuk, K., Popescu, M., Shyu, C. R., Xu, D., Hammer, R., and Shin, D. (2020). Artificial Intelligence-Driven Structurization of Diagnostic Information in Free-Text Pathology Reports. *Journal of Pathology Informatics*, 11(1):10.
- [110] Glaser, A. P., Jordan, B. J., Cohen, J., Desai, A., Silberman, P., and Meeks, J. J. (2018). Automated Extraction of Grade, Stage, and Quality Information From Transurethral Resection of Bladder Tumor Pathology Reports Using Natural Language Processing. *JCO Clinical Cancer Informatics*, (2):1–8.
- [111] Gleicher, M. (2016). A framework for considering comprehensibility in modeling. *Big data*, 4(2):75–88.
- [112] Goodall, J. R. and Sowul, M. (2009). Viassist: Visual analytics for cyber defense. In *2009 IEEE conference on technologies for homeland security*, pages 143–150. IEEE.
- [113] Gorrell, G., Song, X., and Roberts, A. (2018). Bio-yodie: A named entity linking system for biomedical text. *arXiv preprint arXiv:1811.04860*.
- [114] Graziani, M., Andrearczyk, V., Marchand-Maillet, S., and Müller, H. (2020). Concept attribution: Explaining cnn decisions to physicians. *Computers in biology and medicine*, 123:103865.
- [115] Green, T. M., Ribarsky, W., and Fisher, B. D. (2008). Visual analytics for complex concepts using a human cognition model. In *IEEE VAST*, pages 91–98. IEEE Computer Society.

- [116] Gregg, J. R., Lang, M., Wang, L. L., Resnick, M. J., Jain, S. K., Warner, J. L., and Barocas, D. A. (2017). Automating the Determination of Prostate Cancer Risk Strata From Electronic Medical Records. *JCO Clinical Cancer Informatics*, (1):1–8.
- [117] Groth, P., Gibson, A., and Velterop, J. (2010). The Anatomy of a Nanopublication. *Inf. Serv. Use*, 30(1-2):51–56.
- [118] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.
- [119] Gunning, D. (2017). Explainable artificial intelligence (xai). Technical report, Defense advanced research projects agency (DARPA).
- [120] Gunning, D., Vorm, E., Wang, J. Y., and Turek, M. (2021). Darpa’s explainable ai (xai) program: A retrospective.
- [121] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., and Yener, B. (2009). Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171.
- [122] Hansen, C. D. and Johnson, C. R., editors (2005). *The Visualization Handbook*. Academic Press / Elsevier.
- [123] Hassanpour, S. and Langlotz, C. P. (2016). Information Extraction from Multi-Institutional Radiology Reports. *Artif. Intell. Medicine*, 66:29–39.
- [124] Heart, T., Padman, R., Ben-Assuli, O., Gefen, D., and Klempfner, R. (2022). On intelligence augmentation and visual analytics to enhance clinical decision support systems. In *HICSS*, pages 1–10.
- [125] Hettne, K. M., Thompson, M., van Haagen, H., van der Horst, E., Kaliyaperumal, R., Mina, E., Tatum, Z., Laros, J. F. J., van Mulligen, E. M., Schuemie, M., Aten, E., Li, T. S., Bruskiwich, R., Good, B. M., Su, A. I., Kors, J. A., den Dunnen, J., van Ommen, G.-J. B., Roos, M., ‘t Hoen, P. A., Mons, B., and Schultes, E. A. (2016). The Implicitome: A Resource for Rationalizing Gene-Disease Associations. *PLOS ONE*, 11(2):1–21.
- [126] Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, USA.
- [127] Hey, T. and Trefethen, A. (2003). The data deluge: An e-science perspective. *Grid computing: Making the global infrastructure a reality*, 72:809–824.
- [128] Hoffer, J., Rashid, R., Muhlich, J. L., Chen, Y.-A., Russell, D. P. W., Ruokonen, J., Krueger, R., Pfister, H., Santagata, S., and Sorger, P. K. (2020). Minerva: a light-weight, narrative image browser for multiplexed tissue images. *Journal of open source software*, 5(54).
- [129] Holmgren, A. J., Everson, J., and Adler-Milstein, J. (2022). Association of hospital interoperable data sharing with alternative payment model participation. In *JAMA Health Forum*, volume 3, pages e215199–e215199. American Medical Association.

- [130] Holmgren, A. J. and Ford, E. W. (2018). Assessing the impact of health system organizational structure on hospital electronic data sharing. *Journal of the American Medical Informatics Association*, 25(9):1147–1152.
- [131] Holmgren, A. J., Patel, V., and Adler-Milstein, J. (2017). Progress in interoperability: measuring us hospitals’ engagement in sharing patient data. *Health Affairs*, 36(10):1820–1827.
- [132] Holzinger, A. (2018). From machine learning to explainable ai. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 55–66.
- [133] Holzinger, A. (2021). Explainable AI and multi-modal causability in medicine. *i-com*, 19(3):171–179.
- [134] Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017a). What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923.
- [135] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining Knowl. Discov.*, 9(4).
- [136] Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., and Zatloukal, K. (2017b). Machine learning and knowledge extraction in digital pathology needs an integrative approach. In *Towards Integrative Machine Learning and Knowledge Extraction*, pages 13–50. Springer.
- [137] Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., and Zatloukal, K. (2017c). Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. *CoRR*, abs/1712.06657.
- [138] Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., and Tsuneki, M. (2020). Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific reports*, 10(1):1–11.
- [139] Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based Deep Multiple Instance Learning. In *Proc. of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proc. of Machine Learning Research*, pages 2132–2141. PMLR.
- [140] Ipenza, J. C. C., Romero, N. M. L., Loreto, M., Júnior, N. F., and Comba, J. L. D. (2022). Qds-covid: A visual analytics system for interactive exploration of millions of covid-19 healthcare records in brazil. *Applied Soft Computing*, page 109093.
- [141] Islamaj, R., Kwon, D., Kim, S., and Lu, Z. (2020). TeamTat: a collaborative text annotation tool. *Nucleic Acids Research*, 48(W1):W5–W11.
- [142] Jessup, J., Krueger, R., Warchol, S., Hoffer, J., Muhlich, J., Ritch, C. C., Gaglia, G., Coy, S., Chen, Y.-A., Lin, J.-R., et al. (2021). Scope2screen: Focus+ context techniques for pathology tumor assessment in multivariate image data. *IEEE transactions on visualization and computer graphics*, 28(1):259–269.

- [143] Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- [144] Jovanović, J. and Bagheri, E. (2017). Semantic annotation in biomedicine: The current landscape. *Journal of Biomedical Semantics*, 8(1):1–18.
- [145] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proc. of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 116–121. ACL.
- [146] Kadilierakis, G., Fafalios, P., Papadakos, P., and Tzitzikas, Y. (2020). Keyword Search over RDF Using Document-Centric Information Retrieval Systems. In *The Semantic Web*, pages 121–137, Cham. Springer International Publishing.
- [147] Kafkas, S., Lewin, I., Milward, D., van Mulligen, E. M., Kors, J. A., Hahn, U., and Rebholz-Schuhmann, D. (2012). Calbc: Releasing the final corpora. In *LREC*, pages 2923–2926.
- [148] Kalra, D., Beale, T., and Heard, S. (2005). The openehr foundation. *Studies in health technology and informatics*, 115:153–173.
- [149] Kamath, U. and Liu, J. (2021). *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer.
- [150] Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759.
- [151] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008a). *Visual Analytics: Definition, Process, and Challenges*, pages 154–175. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [152] Keim, D. A., Mansmann, F., Oelke, D., and Ziegler, H. (2008b). Visual analytics: Combining automated discovery with interactive visualizations. In *Discovery Science, 11th International Conference, DS 2008, Budapest, Hungary, October 13-16, 2008. Proceedings*, pages 2–14.
- [153] Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008c). Visual analytics: Scope and challenges. In *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, pages 76–90. Springer.
- [154] Keim, D. A., Mansmann, F., Stoffel, A., and Ziegler, H. (2009). Visual analytics. In *Encyclopedia of Database Systems*, pages 3341–3346. Springer US.
- [155] Keim, D. A., Mansmann, F., and Thomas, J. (2010). Visual analytics: How much visualization and how much analytics? *SIGKDD Explor. Newsl.*, 11(2):5–8.
- [156] Keim, D. A., Nietzschmann, T., Schelwies, N., Schneidewind, J., Schreck, T., and Ziegler, H. (2006). A spectral visualization system for analyzing financial time series data. In *EuroVis*, pages 195–202. Eurographics Association.

- [157] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.
- [158] Kim, J. and Wang, Y. (2012). Pubannotation - a persistent and sharable corpus and annotation repository. In Cohen, K. B., Demner-Fushman, D., Ananiadou, S., Webber, B. L., Tsujii, J., and Pestian, J., editors, *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, BioNLP@HLT-NAACL Montréal, Canada, June 8, 2012*, pages 202–205. Association for Computational Linguistics.
- [159] Kim, Y., Lee, J. H., Choi, S., Lee, J. M., Kim, J. H., Seok, J., and Joo, H. J. (2020). Validation of Deep Learning Natural Language Processing Algorithm for Keyword Extraction from Pathology Reports in Electronic Health Records. *Sci Rep*, (10):1–9.
- [160] King, J. A., Jeong, J., Underwood, F. E., Quan, J., Panaccione, N., Windsor, J. W., Coward, S., deBruyn, J., Ronksley, P. E., and Shaheen, A. A. (2020). Incidence of Celiac Disease is Increasing over Time: a Systematic Review and Meta-Analysis. *Official journal of the American College of Gastroenterology*, 115(4):507–525.
- [161] Klie, J. (2018). Inception: Interactive machine-assisted annotation. In *Proc. of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems (DESIREs 2018)*, volume 2167 of *CEUR Workshop Proceedings*, page 105. CEUR-WS.org.
- [162] Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.
- [163] Komura, D. and Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16:34–42.
- [164] Koplíku, A., Pinel-Sauvagnat, K., and Boughanem, M. (2014). Aggregated search: A new information retrieval paradigm. *ACM Comput. Surv.*, 46(3):41:1–41:31.
- [165] Kors, J. A., Clematide, S., Akhondi, S. A., Van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.
- [166] Kosara, R. (2007). Visualization criticism - the missing link between information visualization and art. In *IV*, pages 631–636. IEEE Computer Society.
- [167] Kraljevic, Z., Bean, D., Mascio, A., Roguski, L., Folarin, A., Roberts, A., Bendayan, R., and Dobson, R. (2019). Medcat—medical concept annotation tool. *arXiv preprint arXiv:1912.10166*.
- [168] Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., et al. (2015). The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

- [169] Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.
- [170] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., and Botsis, T. (2017). Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review. *J. Biomed. Informatics*, 73:14–29.
- [171] Krueger, R., Beyer, J., Jang, W.-D., Kim, N. W., Sokolov, A., Sorger, P. K., and Pfister, H. (2019). Facetto: Combining unsupervised and supervised learning for hierarchical phenotype analysis in multi-channel image data. *IEEE transactions on visualization and computer graphics*, 26(1):227–237.
- [172] Krupinski, E. A., Graham, A. R., and Weinstein, R. S. (2013). Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human pathology*, 44(3):357–364.
- [173] Kuhn, T., Barbano, P. E., Nagy, M. L., and Krauthammer, M. (2013). Broadening the Scope of Nanopublications. In *Proc. of the Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013*, volume 7882 of LNCS, pages 487–501. Springer.
- [174] Kuhn, T., Meroño-Peñuela, A., Malic, A., Poelen, J. H., Hurlbert, A. H., Ortiz, E. C., Furlong, L. I., Queralt-Rosinach, N., Chichester, C., Banda, J. M., Willighagen, E. L., Ehrhart, F., Evelo, C. T. A., Malas, T. B., and Dumontier, M. (2018). Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. In *14th IEEE International Conference on e-Science, e-Science 2018*, pages 83–92. IEEE Computer Society.
- [175] Kuhn, T., Willighagen, E., Evelo, C., Queralt-Rosinach, N., Centeno, E., and Furlong, L. I. (2017). Reliable granular references to changing linked data. In d’Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., and Heflin, J., editors, *The Semantic Web – ISWC 2017*, pages 436–451, Cham. Springer International Publishing.
- [176] Kundu, S. (2021). Ai in medicine must be explainable. *Nature medicine*, 27(8):1328–1328.
- [177] Kwon, D., Kim, S., Shin, S.-Y., and Wilbur, W. J. (2013). Bioqrator: a web-based interactive biomedical literature curating system. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 241–246.
- [178] Kwon, D., Kim, S., Wei, C.-H., Leaman, R., and Lu, Z. (2018). eztag: tagging biomedical concepts via interactive learning. *Nucleic acids research*, 46(W1):W523–W529.
- [179] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40.
- [180] Lancet, T. (2018). Making sense of our digital medicine babel.

- [181] Lanzenberger, M., Sampson, J., and Rester, M. (2009). Visualization in ontology tools. In *2009 International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009, Fukuoka, Japan, March 16-19, 2009*, pages 705–711. IEEE Computer Society.
- [182] Lehne, M., Luijten, S., genannt Imbusch, P. V. F., and Thun, S. (2019a). The use of fhir in digital health—a review of the scientific literature. *GMDS*, (September):52–58.
- [183] Lehne, M., Sass, J., Essenwanger, A., Schepers, J., and Thun, S. (2019b). Why digital medicine depends on interoperability. *NPJ digital medicine*, 2(1):1–5.
- [184] Leite, R. A., Gschwandtner, T., Miksch, S., Kriglstein, S., Pohl, M., Gstrein, E., and Kuntner, J. (2017). Eva: Visual analytics to identify fraudulent events. *IEEE transactions on visualization and computer graphics*, 24(1):330–339.
- [185] Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wieggers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.
- [186] Li, O., Liu, H., Chen, C., and Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- [187] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- [188] Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., and Van Der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):1–11.
- [189] Lohmann, S., Link, V., Marbach, E., and Negru, S. (2014a). Webvowl: Web-based visualization of ontologies. In *Knowledge Engineering and Knowledge Management - EKAW 2014 Satellite Events, VISUAL, EKMI, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers*, volume 8982 of *LNCS*, pages 154–158. Springer.
- [190] Lohmann, S., Negru, S., Haag, F., and Ertl, T. (2014b). VOWL 2: User-oriented visualization of ontologies. In *Proc. of the 19th International Conference on Knowledge Engineering and Knowledge Management EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, volume 8876 of *LNCS*, pages 266–281. Springer.
- [191] Lohmann, S., Negru, S., Haag, F., and Ertl, T. (2016). Visualizing ontologies with VOWL. *Semantic Web*, 7(4):399–419.
- [192] Lopez-Veyna, J. I., Sosa, V. J. S., and López-Arévalo, I. (2012). A Virtual Document Approach for Keyword Search in Databases. In *DATA*, pages 39–48. SciTePress.
- [193] Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2020). Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images. *CoRR*, abs/2004.09666.

- [194] Luo, Y., Wang, W., Lin, X., Zhou, X., Wang, J., and Li, K. (2011). SPARK2: top-k keyword query in relational databases. *IEEE Trans. Knowl. Data Eng.*, 23(12):1763–1780.
- [195] Magrabi, F., Ammenwerth, E., McNair, J. B., de Keizer, N. F., Hyppönen, H., Nykänen, P., Rigby, M. L., Scott, P. J., Vehko, T., Wong, Z. S.-Y., and Georgiou, A. (2019). Artificial intelligence in clinical decision support: Challenges for evaluating ai and practical implications. *Yearbook of Medical Informatics*, 28:128 – 134.
- [196] Marchesin, S. (2018). Case-Based Retrieval Using Document-Level Semantic Networks. In *Proc. of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, page 1451. ACM.
- [197] Marchesin, S., Giachelle, F., Marini, N., Atzori, M., Boytcheva, S., Buttafuoco, G., Ciompi, F., Di Nunzio, G. M., Fraggetta, F., Irrera, O., Müller, H., Primov, T., Vatrano, S., and Silvello, G. (2022). Empowering digital pathology applications through explainable knowledge extraction tools. *Journal of Pathology Informatics*, page 100139.
- [198] Marchesin, S. and Silvello, G. (2022). TBGA: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinform.*, 23(1):111.
- [199] Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., van Rijthoven, M., Aswolinskiy, W., Bokhorst, J. M., Podareanu, D., Petters, E., Boytcheva, S., Buttafuoco, G., Vatrano, S., Fraggetta, F., der Laak, J., Agosti, M., Ciompi, F., Silvello, G., Muller, H., and Atzori, M. (2022). Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *npj Digital Medicine*, 5(1).
- [200] Mass, Y. and Sagiv, Y. (2016). Virtual Documents and Answer Priors in Keyword Search over Data Graphs. In *Proc. of the Workshops of the EDBT/ICDT 2016 Joint Conference*, volume 1558 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- [201] McCusker, J., Rashid, S. M., Agu, N., Bennett, K. P., and McGuinness, D. L. (2018). The Whyis Knowledge Graph Framework in Action. In *Proc. of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018)*, volume 2180 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- [202] McCusker, J. P., Dumontier, M., Yan, R., He, S., Dordick, J. S., and McGuinness, D. L. (2017). Finding melanoma drugs through a probabilistic knowledge graph. *PeerJ Computer Science*, 3:e106.
- [203] Meijer, G., Beliën, J., Van Diest, P., and Baak, J. (1997). Origins of... image analysis in clinical pathology. *Journal of clinical pathology*, 50(5):365.
- [204] Meuten, D., Moore, F., and George, J. (2016). Mitotic count and the field of view area: time to standardize.
- [205] Mihăilă, C., Ohta, T., Pyysalo, S., and Ananiadou, S. (2013). Biocause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics*, 14(1):1–18.

- [206] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of the 27th Annual Conference on Neural Information Processing Systems 2013, NIPS, Lake Tahoe, Nevada, United States, December 5-8, 2013*, pages 3111–3119.
- [207] Mohan, S. and Li, D. (2019). Medmentions: a large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.
- [208] Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B., den Dunnen, J. T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R. and Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., and Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4):281–283.
- [209] Montani, S. (2008). Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. *Appl. Intell.*, 28(3):275–285.
- [210] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15.
- [211] Morch, N. J., Kjems, U., Hansen, L. K., Svarer, C., Law, I., Lautrup, B., Strother, S., and Rehm, K. (1995). Visualization of neural networks using saliency maps. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 2085–2090. IEEE.
- [212] Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., et al. (2008). Overview of biocreative ii gene normalization. *Genome biology*, 9(2):1–19.
- [213] Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44.
- [214] Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., and Zatloukal, K. (2022). Explainability and causability for artificial intelligence-supported medical image analysis in the context of the european in vitro diagnostic regulation. *New Biotechnology*, 70:67–72.
- [215] Müller, H.-M., Van Auken, K. M., Li, Y., and Sternberg, P. W. (2018). Textpresso central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC bioinformatics*, 19(1):1–16.
- [216] Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13):1351–1352.
- [217] Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proc. of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 319–327. ACL.
- [218] Neves, M. and Leser, U. (2014). A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics*, 15(2):327–340.

- [219] Neves, M. and Ševa, J. (2021). An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163.
- [220] Nguyen, T. M., Tjoa, A. M., and Trujillo, J. (2005). Data warehousing and knowledge discovery: A chronological view of research challenges. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 530–535. Springer.
- [221] Norgan, A. P., Shah, K. K., Juskewitch, J. E., and Maleszewski, J. J. (2018). Open-source whole slide image preparation and viewing pipeline. *Archives of Pathology & Laboratory Medicine*, 142(12):1454–1455.
- [222] Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I. J., Nenkova, A., and Wallace, B. C. (2018). A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.
- [223] Oemig, F. and Snelick, R. (2016). *Healthcare Interoperability Standards Compliance Handbook - Conformance and Testing of Healthcare Data Exchange Standards*. Springer.
- [224] Ohta, T., Pyysalo, S., Tsujii, J., and Ananiadou, S. (2012). Open-domain anatomical entity mention detection. In *Proceedings of the workshop on detecting structure in scholarly discourse*, pages 27–36.
- [225] Oliwa, T., Maron, S. B., Chase, L. M., Lomnicki, S., Catenacci, D. V. T., Furner, B., and Volchenboun, S. L. (2019). Obtaining Knowledge in Pathology Reports Through a Natural Language Processing Approach With Classification, Named-Entity Recognition, and Relation-Extraction Heuristics. *JCO Clinical Cancer Informatics*, (3):1–8.
- [226] Page, R. (2018). Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature. *Biodiversity Data Journal*, 6:e27539.
- [227] Pearl, J. (2009). *Causality*. Cambridge university press.
- [228] Pérez, J., Arenas, M., and Gutierrez, C. (2009). Semantics and Complexity of SPARQL. *ACM Trans. Database Syst.*, 34(3):1–45.
- [229] Pérez-Pérez, M., Glez-Peña, D., Fdez-Riverola, F., and Lourenço, A. (2015). Marky: a tool supporting annotation consistency in multi-user and iterative document annotation projects. *Computer methods and programs in biomedicine*, 118(2):242–251.
- [230] Perlin, J. B. (2016). Health information technology interoperability and use for better care and evidence. *Jama*, 316(16):1667–1668.
- [231] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018*, pages 2227–2237. ACL.

- [232] Piñero, J., Ramírez-Angueta, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L. I. (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855.
- [233] Pocevičiūtė, M., Eilertsen, G., and Lundström, C. (2020). Survey of xai in digital pathology. In *Artificial intelligence and machine learning for digital pathology*, pages 56–88. Springer.
- [234] Pohl, M., Smuc, M., and Mayr, E. (2012). The user puzzle - explaining the interaction with visual analytics systems. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2908–2916.
- [235] Pontén, F., Jirstrom, K., and Uhlen, M. (2008). The human protein atlas—a tool for pathology. *The Journal of Pathology*, 216(4):387–393.
- [236] Pound, J., Mika, P., and Zaragoza, H. (2010). Ad-hoc object retrieval in the web of data. In *Proc. of the 19th International Conference on World Wide Web, WWW 2010*, pages 771–780. ACM Press, New York, USA.
- [237] Pronovost, P. J. (2018). *Procuring interoperability: achieving high-quality, connected, and person-centered care*. nam. edu.
- [238] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. (2013). Distributional Semantics Resources for Biomedical Text Processing. *Proc. of LBM*, pages 39–44.
- [239] Queralt-Rosinach, N., Kuhn, T., Chichester, C., Dumontier, M., Sanz, F., and Furlong, L. I. (2016). Publishing DisGeNET as Nanopublications. *Semantic Web*, 7(5):519–528.
- [240] Ragan, E. D., Endert, A., Sanyal, J., and Chen, J. (2015). Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics*, 22(1):31–40.
- [241] Rahman, P., Jiang, L., and Nandi, A. (2020). Evaluating Interactive Data Systems. *VLDB J.*, 29(1):119–146.
- [242] Rak, R., Rowley, A., Black, W., and Ananiadou, S. (2012). Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database*, 2012.
- [243] Ratcliff, J. W. and Metzener, D. E. (1988). Pattern Matching: the Gestalt Approach. *Dr Dobbs Journal*, 13(7):46.
- [244] Rhyne, T.-M., Tory, M., Munzner, T., Ward, M., Johnson, C., and Laidlaw, D. H. (2003). Information and scientific visualization: Separate but equal or happy together at last. In *Visualization Conference, IEEE*, pages 115–115. IEEE Computer Society.
- [245] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- [246] Rinaldi, F., Clematide, S., Schneider, G., Romacker, M., and Vachon, T. (2010). Odin: an advanced interface for the curation of biomedical literature. *Nature Precedings*, pages 1–1.

- [247] Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., and Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966.
- [248] Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., and Lazar, A. J. (2018). Overview of the TREC 2018 precision medicine track. In *Proc. of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*, volume 500-331 of *NIST Special Publication*. NIST.
- [Robertson et al.] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. Okapi at TREC-3. pages 109–126.
- [250] Rostamzadeh, N., Abdullah, S. S., and Sedig, K. (2021). Visual analytics for electronic health records: a review. In *Informatics*, volume 8, page 12. MDPI.
- [251] Ruas, P., Andrade, V. D. T., and Couto, F. M. (2021). Lasige-biotm at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on spanish biomedical documents. In *Proc. of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 324–334. CEUR-WS.org.
- [252] Rudolph, S., Savikhin, A., and Ebert, D. S. (2009). Finvis: Applied visual analytics for personal financial planning. In *2009 IEEE symposium on visual analytics science and technology*, pages 195–202. IEEE.
- [253] Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., and Keim, D. A. (2014). Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613.
- [254] Salgado, D., Krallinger, M., Depaule, M., Drula, E., Tendulkar, A. V., Leitner, F., Valencia, A., and Marcelle, C. (2012). Myminer: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, 28(17):2285–2287.
- [255] Santus, E., Schuster, T., Tahmasebi, A. M., Li, C., Yala, A., Lanahan, C. R., Prinsen, P., Thompson, S. F., Coons, S., Mynderse, L., Barzilay, R., and Hughes, K. (2020). Exploiting Rules to Enhance Machine Learning in Extracting Information From Multi-Institutional Prostate Pathology Reports. *JCO Clinical Cancer Informatics*, (4):865–874.
- [256] Sartor, G. (2020). The impact of the general data protection regulation (gdpr) on artificial intelligence. Technical report, Panel for the Future of Science and Technology (STOA).
- [257] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.
- [258] Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanutelli, V. R., Schulz, D., Giesen, C., Catena, R., Varga, Z., and Bodenmiller, B. (2017). histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, 14(9):873–876.

- [259] Schaumberg, A. J., Juarez, W., Choudhury, S. J., Pastroián, L. G., Pritt, B. S., Pozuelo, M. P., Sánchez, R. S., Ho, K., Zahra, N., and Sener, B. D. (2018). Large-Scale Annotation of Histopathology Images from Social Media. *BioRxiv*, page 396663.
- [260] Schulz, C., Meyer, C. M., Kiesewetter, J., Sailer, M., Bauer, E., Fischer, M. R., Fischer, F., and Gurevych, I. (2019). Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *ACL*.
- [261] Schulz, H.-J., Angelini, M., Santucci, G., and Schumann, H. (2015). An enhanced visualization process model for incremental visualization. *IEEE transactions on visualization and computer graphics*, 22(7):1830–1842.
- [262] Seah, J., Tang, J., Kitchen, A., and Seah, J. (2018). Generative visual rationales. *arXiv preprint arXiv:1804.04539*.
- [263] Searle, T., Kraljevic, Z., Bendayan, R., Bean, D., and Dobson, R. (2019). Medcat-trainer: A biomedical free text annotation interface with active learning and research use case specific customisation. *arXiv preprint arXiv:1907.07322*.
- [264] Shaw, J. A. and Fox, E. A. (1994). Combination of Multiple Searches. In *Proc. of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225, pages 105–108. NIST.
- [265] Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG)*, 11(1):92–99.
- [266] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *VL*, pages 336–343. IEEE Computer Society.
- [267] Silvello, G. (2018). Theory and Practice of Data Citation. *Journal of the American Society for Information Science and Technology (JASIST)*, 69(1):6–20.
- [268] Simitsis, A., Koutrika, G., and Ioannidis, Y. E. (2008). Précis: from unstructured keywords as queries to structured databases as answers. *VLDB J.*, 17(1):117–149.
- [269] Simoff, S. J., Böhlen, M. H., and Mazeika, A. (2008). Visual data mining: An introduction and overview. In *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, pages 1–12. Springer.
- [270] Sirintrapun, S. J. (2015). Building a pipeline for pathology informatics. *Critical Values*, 8:48–51.
- [271] Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S. L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L. M. T., Evelo, C. T., Pico, A. R., and Willighagen, E. L. (2017). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667.
- [272] Spence, R. (2014). *Information Visualization - An Introduction*. Springer.

- [273] Strigley, J. R., McGowan, T., Maclean, A., Raby, M., Ross, J., Kramer, S., and Sawka, C. (2009). Standardized synoptic cancer pathology reporting: a population-based approach. *J. Surg. Oncol.*, 99(8):517–524.
- [274] Srinidhi, C. L., Ciga, O., and Martel, A. L. (2021). Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813.
- [275] Stathonikos, N., Veta, M., Huisman, A., and van Diest, P. J. (2013). Going fully digital: Perspective of a dutch academic pathology lab. *Journal of pathology informatics*, 4(1):15.
- [276] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.
- [277] Stewart, B. W. and Wild, C. P. (2014). *World cancer report 2014*.
- [278] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3.
- [279] Thomas, J. J. and Cook, K. A. (2005). Illuminating the path: The research and development agenda for visual analytics.
- [280] Titford, M. (2006). A short history of histopathology technique. *Journal of Histotechnology*, 29(2):99–110.
- [281] Tizhoosh, H. R. and Pantanowitz, L. (2018). Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9(1):38.
- [282] Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., and Bowles, K. H. (2019). Mining Fall-Related Information in Clinical Notes: Comparison of Rule-Based and Novel Word Embedding-Based Machine Learning Approaches. *J. Biomed. Informatics*, 90.
- [283] Tory, M. and Möller, T. (2004). Human factors in visualization research. *IEEE Trans. Vis. Comput. Graph.*, 10(1):72–84.
- [284] Tosun, A. B., Pullara, F., Becich, M. J., Taylor, D., Chennubhotla, S. C., and Fine, J. L. (2020). Histomapr™: An explainable ai (xai) platform for computational pathology solutions. In *Artificial Intelligence and Machine Learning for Digital Pathology*, pages 204–227. Springer.
- [285] Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley series in behavioral science : quantitative methods. Addison-Wesley.
- [286] Tutubalina, E., Alimova, I., Miftahutdinov, Z., Sakhovskiy, A., Malykh, V., and Nikolenko, S. I. (2021). The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinform.*, 37(2):243–249.

- [287] Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J. M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.-H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., and Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, 357(6352).
- [288] Vaitsis, C., Nilsson, G., and Zary, N. (2014). Big data in medical informatics: improving education through visual analytics. In *e-Health—For Continuity of Care*, pages 1163–1167. IOS Press.
- [289] Van Auken, K., Schaeffer, M. L., McQuilton, P., Laulederkind, S. J., Li, D., Wang, S.-J., Hayman, G. T., Tweedie, S., Arighi, C. N., Done, J., et al. (2014). Bc4go: a full-text corpus for the biocreative iv go task. *Database*, 2014.
- [290] Van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784.
- [291] van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605.
- [292] Van Wijk, J. J. (2005). The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86. IEEE.
- [293] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Proc. of the 31st Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- [294] Verspoor, K., Jimeno Yepes, A., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., and Plazzer, J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*, 2013.
- [295] Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E. L., Evelo, C. T., and Pico, A. R. (2016). Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. *PLOS Computational Biology*, 12(6):1–11.
- [296] Wang, F. and Preininger, A. M. (2019). Ai in health: State of the art, challenges, and future directions. *Yearbook of Medical Informatics*, 28:16 – 26.
- [297] Wang, H. and Aggarwal, C. C. (2010). A survey of algorithms for keyword search on graph data. In *Managing and Mining Graph Data*, pages 249–273. Springer.
- [298] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., and Liu, H. (2018). Clinical Information Extraction Applications: A Literature Review. *J. Biomed. Informatics*, 77:34–49.
- [299] Wattenberg, M. (1999). Visualizing the stock market. In *CHI Extended Abstracts*, pages 188–189. ACM.
- [300] Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593.

- [301] Wong, P. C. and Thomas, J. (2004). Guest editors' introduction—visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 24(5).
- [302] Wu, D. T., Chen, A. T., Manning, J. D., Levy-Fix, G., Backonja, U., Borland, D., Caban, J. J., Dowding, D. W., Hochheiser, H., Kagan, V., et al. (2019). Evaluating visual analytics for health informatics applications: a systematic review from the american medical informatics association visual analytics working group task force on evaluation. *Journal of the American Medical Informatics Association*, 26(4):314–323.
- [303] Wu, H., Toti, G., Morley, K. I., Ibrahim, Z. M., Folarin, A., Jackson, R., Kartoglu, I., Agrawal, A., Stringer, C., Gale, D., et al. (2018). Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*, 25(5):530–537.
- [304] Wu, W. (2013). Proactive Natural Language Search Engine: Tapping into Structured Data on the Web. In *Proc. of the Joint 2013 EDBT/ICDT Conferences*, pages 143–148. ACM Press, New York, USA.
- [305] Wu, Y., Cao, N., Gotz, D., Tan, Y., and Keim, D. A. (2016). A survey on visual analytics of social media data. *IEEE Trans. Multim.*, 18(11):2135–2148.
- [306] Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., and Traweek, S. (2012). Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices. pages 19–22.
- [307] Yang, C. C., Chen, H., and Hong, K. (2003). Visualization of large category map for internet browsing. *Decision support systems*, 35(1):89–102.
- [308] Yi, Y., Shen, Z., Bompelli, A., Yu, F., Wang, Y., and Zhang, R. (2020). Natural language processing methods to extract lifestyle exposures for alzheimer's disease from clinical notes. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2.
- [309] Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.
- [310] Yu, J. X., Qin, L., and Chang, L. (2010). Keyword Search in Relational Databases: A Survey. *IEEE Data Eng. Bull.*, 33(1):67–78.
- [311] Zarella, M. D., Bowman, D., Aeffner, F., Farahani, N., Xthona, A., Absar, S. F., Parwani, A., Bui, M., and Hartman, D. J. (2019). A practical guide to whole slide imaging: a white paper from the digital pathology association. *Archives of pathology & laboratory medicine*, 143(2):222–234.
- [312] Zerva, C., Batista-Navarro, R., Day, P., and Ananiadou, S. (2017). Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*, 33(23):3784–3792.

- [313] Zhang, H., He, X., Harrison, T., and Bian, J. (2019a). Aero: An Evidence-based Semantic Web Knowledge Base of Cancer Behavioral Risk Factors. In *Proc. of the 4th International Workshop on Semantics-Powered Data Mining and Analytics co-located with the 18th International Semantic Web Conference (ISWC 2019)*, volume 2427 of *CEUR Workshop Proceedings*, pages 7–11. CEUR-WS.org.
- [314] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019b). ERNIE: Enhanced Language Representation with Informative Entities. In *Proc. of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019*, pages 1441–1451. ACL.
- [315] Ziegler, H., Nietzsche, T., and Keim, D. A. (2008). Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments. In *12th International Conference on Information Visualisation, IV 2008, 8-11 July 2008, London, UK*, pages 287–295.
- [316] Zvára, K., Tomecková, M., Peleška, J., Svátek, V., and Zvárová, J. (2017). Tool-supported interactive correction and semantic annotation of narrative clinical reports. *Methods of information in medicine*, 56(03):217–229.