“D3.5”

Final set of curated publicly available multimodal and multimedia data

Version: 1.0

Last Update: 23/06/20

Distribution Level: *PU*

Distribution level

PU = Public

RE = Restricted to a group of the specified Consortium

PP = Restricted to other program participants (including Commission Services)

CO= Confidential, only for members of the ExaMode Consortium (including the Commission Services)

Project supported by European Union

Horizon 2020 grant agreement 825292

The ExaMode Project Consortium groups the following Organizations:

Partner Name	**Short name**	**Country**
HAUTE ECOLE SPECIALISEE DE SUISSE OCCIDENTALE	HES-SO	Switzerland
UNIVERSITA DEGLI STUDI DI PADOVA	UNIPD	Italy
SIRMA	SIRMA AI	Bulgaria
STICHTING KATHOLIEKE UNIVERSITEIT	Radboudumc	Netherlands
MICROSCOPEIT SP ZOO	MICROSCOPEIT	Poland
AZIENDA OSPEDALIERA PER L'EMERGENZA CANNIZZARO	AOEC	Italy
SURFSARA BV	SURFSARA BV	Netherlands

Document Identity

Creation Date: Last Update:	29/04/2020 23/06/2020

Revision History

Version	**Edition**	**Author(s)**	**Date**
0	0.1	Manfredo Atzori	30/04/2020
Comments:	Deliverable outline		
0	0.2	Sebastian Otálora	10/06/2020
Comments:	Deliverable draft, including graphs & description for the data curation pipeline		
0	0.3	Manfredo Atzori	11/06/2020
Comments:	Proofread draft, final editing, added some sections		
0	0.4	Sebastian Otálora	12/06/2020
Comments:	Addressing comments, completing table of external dataset sources.		
0	5	Francesco Ciompi	18/06/2020
	Comments on version 0.4.		
1	0	Manfredo Atzori, Sebastian Otalora	23/06/2020
	Answered to the comments of the reviewer, updated tables and performed the final control		

Executive summary

While the first set of curated publicly available multimodal and multimedia data (described in deliverable 3.1) was supposed to include at least 500 images and related text extracted from scientific literature and the web, the final set of curated publicly available data (described in this deliverable) was supposed to include at least 5000 images and related text.
As described in this report, also in this case we accomplished the deliverable. The final set of publicly available multimodal and multimedia data includes over 9900 images from the medical literature and the associated text, in the form of image captions, article title, abstract and text. The images are available to the consortium via Cartesius (SURFSARA BV). In addition, 250 WSIs from publicly available clinical data sources and 11 additional publicly available datasets (including over 12'000 WSIs and over 2'000 annotated image patches) were identified as resources to train and test knowledge extraction algorithms on the ExaMode use cases, leading to a total of over 12'000 WSIs and over 12'000 images from publicly available sources. Several experiments were performed to use strong and weak annotations provided with such datasets, as described in several publications submitted to peer-reviewed international conferences and journals.

Table of Contents

List of Figures

List of tables

List of abbreviations

WSI	Whole Slide Image
DNN	Deep Neural Networks
GB	Giga bytes
WP	Work package
TMA	Tissue Micro Array
WSI	Whole Slide Image
H&E	Hematoxylin and Eosin
CSV	Comma Separated Values
GDCV	Genomic Data Commons
TCGA	The GDC Cancer Genome Atlas
FFPE	Formalin-Fixed Paraffin-Embedded

1 ExaMode datasets overview

ExaMode's overall aim is to allow easy and fast, weakly supervised knowledge discovery of exa-scale heterogeneous data, also in highly specific domains (for instance in the medical sciences).
Exa-scale volumes of diverse data from distributed sources are continuously produced. Healthcare data stand out because of size, heterogeneity, the knowledge included in the data and its potential commercial value.
Several limits prevent extracting knowledge and value from medical data, such as the following ones:

- Training machine learning algorithms requires data annotations from specialized medical doctors, which are difficult to obtain.
- Data heterogeneity strongly limits the training of Deep Neural Networks (DNN), making them difficult to generalize well to images acquired with different setups and in different hospitals.

ExaMode solves the mentioned problems by developing a weakly supervised knowledge discovery system to extract multimodal information from highly heterogeneous and unstructured data. In the original plan (Figure 1), ExaMode data included proprietary clinical data and scientific literature data. In the concrete advancement of the project, ExaMode data include also publicly available clinical data (a set of whole slide images selected from TCGA, section 3.1) and additional publicly available data sources (section 3.3), that allow to increase the heterogeneity and robustness of the trained deep learning models.
The idea of including such datasets was originally motivated by the need of data to start running experiments as soon as possible (i.e. before solving ethics and data management requirements). However, in the end we decided to continue adding similar data sources because they can be extremely useful for the project, since they can be included into weakly supervised learning procedures. Even if the overall objective of ExaMode is to bypass as much as possible the interaction with experts by using weakly supervised learning systems, our experiments in [5] show in fact that models trained on small sets of annotations made on different datasets can be fine-tuned using weak labels in order to teach algorithms more quickly and with better results. Finally, the availability of some of these datasets (particularly the one described in section 3.1) will also allow to test the models on highly heterogeneous datasets that resemble multiple clinical conditions.

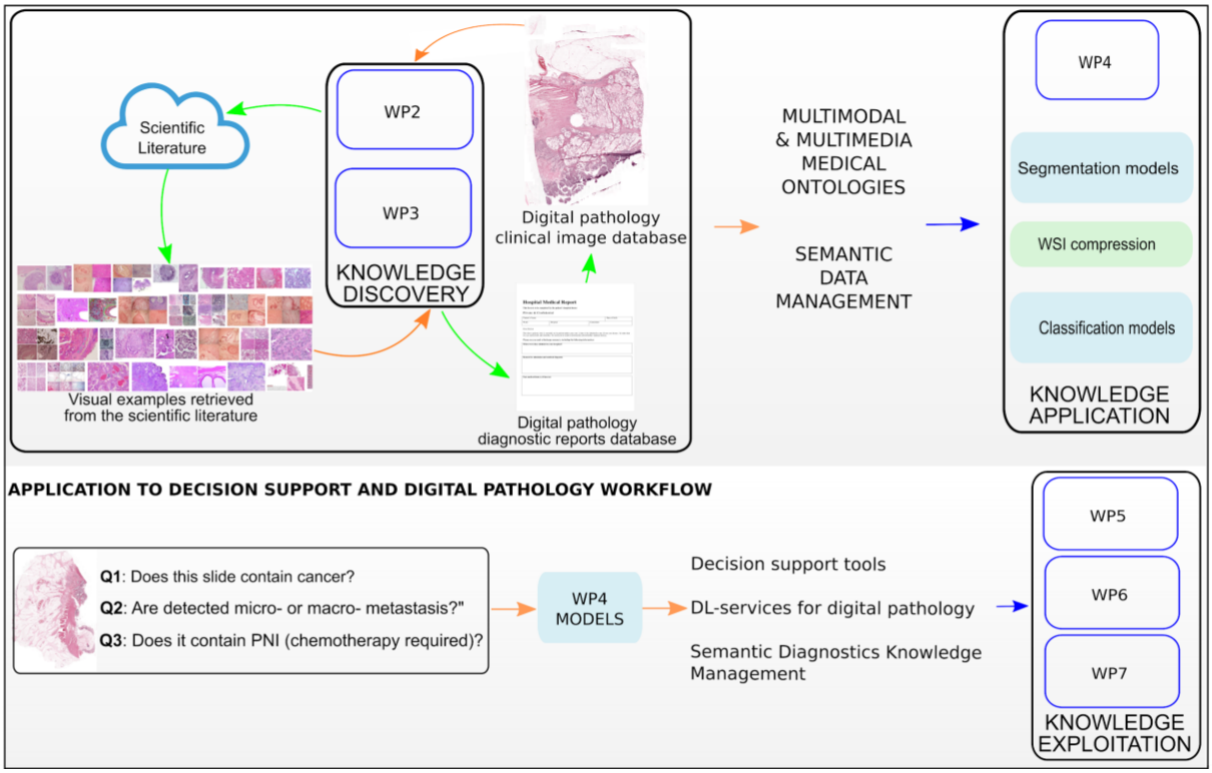

Figure 1 ExaMode pipeline, highlighting the data types that are expected to be used for training, i.e. Scientific literature data and digital pathology clinical data.

Table 1 ExaMode multimodal and multimedia data *(updated on 23/06/2020)*.

	TASK	First set of proprietary data				Final set of cured publicly available multimodal and multimedia data					
						Publicly available clinical data			Data from scientific literature		
		WSIs	TMA Images	Text	Source	Whole Slide	Text	Source	Images	Text	Source
COLON	Adenocarcinoma. Detection of cancer in polyps (in screening population).	2000		Diagnostic report, structured (table)	AOEC	50	Structured (table)	TCGA	2699	Image caption and article text	PMC Central
		2500		Synoptic report, structured (table)	Radboudumc						
		40	80	Structured (table)	Bern University						
UTERINE CERVIX	Squamous cell carcinoma	2000		Diagnostic reports, structured (table)	AOEC	45	Structured (table)	TCGA	962	Image caption and article text	PMC Central
		50		Synoptic report	Radboudumc						
LUNG	Classification/detection of growth patterns related to cancer aggressiveness, prognosis	2000		Diagnostic report, structured (table)	AOEC	100	Structured (table)	TCGA	4151	Image caption and article text	PMC Central
CELIAC DISEASE	Celiac disease detection in duodenal biopsies	2000		Diagnostic report, structured (table)	AOEC				165	Image caption and article text	PMC Central
		50		Synoptic report	Radboudumc						
PROSTATE	Gleason grading					50	Structured (table)	TCGA	1925	Image caption and article text	PMC Central
Additional data sources from publicly available datasets (Table 2)						12441		Various	2156		Various
TOTAL		**10140**	**80**			**12686**			**12085**		**Various**

2 First set of proprietary ExaMode data

The first set of proprietary data is thoroughly described in D7.5 “First set of data curated and available” and in D.5.3 “Example datasets”. The aim of this section is to briefly summarize them.

2.1 First set of data curated and available

The first set of histopathology data was collected and curated during the first year of the project, it includes diagnostic reports and Whole Slide Images (WSI).
These data are provided by the hospital members of the consortium: AOEC and RADBOUDUMC. The selected tissue types include images of several histological subtypes of Non-Small-Cell Lung Cancer, benign and malignant biopsies of the Colon and of the Uterine Cervix. For Coeliac Disease, biopsies of duodenal tissue were collected, and every case contained two slides: the H&E and the CD3 immunohistochemical staining.

2.2 Example datasets

While the data in the deliverable 7.5 focused on collecting a first set of curated data, the deliverable 5.3 is an example dataset that includes 200 WSIs annotated by pathologists and over 1100 data annotations. The data annotations of the example datasets cover all the cases included in the priority list of tissue & cancer types provided in D5.1 and they follow the indications provided in the list of the essential knowledge requirements for each tissue presented in D5.2.
The set of annotations defined in deliverable 5.2 are now being used by the partners in the consortium to perform data analysis pipelines that include heterogenous sources of data and annotations.
Over 1000 data annotations were done using the ASAP tool provided by RADBOUDUMC. The remaining annotations were performed using the VIRTUM tool provided by MICROSCOPEIT, allowing to test the system. Currently, in the context of Deliverable 7.7 "First set of annotated digital pathology data", data annotations are being increased by AOEC and RADBOUDUMC. All the annotations are now performed using VIRTUM, improving the ExaMode exploitation.
The data annotations are fundamental to develop and test the weakly supervised knowledge extraction tools targeted in WP2 “Semantic knowledge discovery and visualization”, WP3 “Image content-based knowledge discovery” and WP6 “Multimodal knowledge management”. The data annotations allow to train the computer aided diagnosis algorithms for digital pathology that will be developed in WP4 “Computational Pathology” and WP5 “Decision support and image enrichment”.

3 Final set of cured publicly available multimodal and multimedia data

3.1 TCGA publicly available clinical data sources

The TCGA publicly available data are described in detail in the submitted D3.1 "First set of publicly available multimodal and multimedia data".
The ExaMode publicly available clinical pathology dataset is designed to train and test the ExaMode tools in order to guarantee robustness even when data are highly heterogeneous. The dataset currently includes three diseases (colon cancer, lung cancer and uterine cervix cancer). For each disease, the dataset includes up to 100 samples. The data are selected from the public portal "National Cancer Datasets Genomic Data Commons" (GDC). The data are highly heterogeneous in terms of acquisition center while they are well balanced in the diagnosis.
An important feature of the TCGA data is the availability of clinical metadata, such as the surgical pathology reports and different cancer gradings. This clinical multimodal data source poses an interesting resource for testing the computational models in scenarios where both sources of information are used, e.g., for finding correlations between semantic diagnostic words and relevant regions in the images. An example of this is shown in Figure 2. Members of the consortium have started working with the multimodal data and have reported encouraging results [6], allowing to consider it as a valuable resource for the advancement of the project.

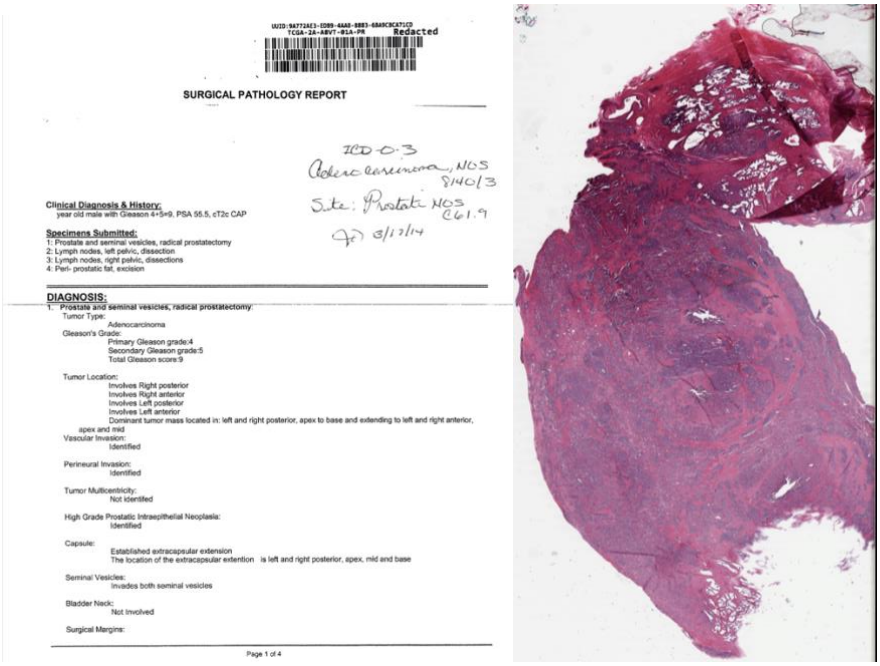

Figure 2 Example of a prostate cancer report with its corresponding whole slide image from the TCGA publicly available data repository.

3.2 Data from scientific literature

As shown in Figure 1, images and text from the scientific literature represent one of the data sources that are planned to be used to train models in the ExaMode project.
Privacy constraints and the high costs of medical imaging often prevented the access of medical imaging research to large collections of images. Global science support programs now encourage data sharing, thus increasing the availability of publicly available archives. However, data and annotations performed by specialists are still difficult to obtain.
The medical literature has been a knowledge resource for many years. While until now most works focused on text analyses, ExaMode aims at providing ways to extract multimodal knowledge also from the scientific literature. The first set of data from scientific literature was included in D3.1 and it is now updated in deliverable 3.5. The members of the consortium also started analyzing multimodal data from the scientific literature, reporting encouraging results [4,6,7].

3.2.1 Data curation pipeline

An overview of the steps involved in the data curation pipeline is shown in Figure 3.

1. In the first step, automatic downloading via FTP[1] of the full open access PMC database[2] (more than 10 million articles with a size > 11 terabytes) is done for all the articles, including: original text (PDF), abstract, images in the article, and an XML file containing the full text, image captions and other metadata. Downloading all the data can take several weeks, for this reason we use a previously mirrored database and update it only with the new/changed articles. The data gathered is updated with the latest version of PMC, that corresponds to the year 2020.
2. In the second step, we filter those images that are histopathology images, which are identified as light microscopy or DMLI, as identified in PMC. Since there are many misclassified (other microscopy modalities) and images from gross specimens (macroscopic tumors) we further filter the images to obtain histopathology images with high confidence. Using a convolutional neural network designed for this problem [8] that uses the image and its caption we obtain a robust model that allow us to filter ~80,000 histopathology images.
3. As third step, we obtain only the histopathology images that concerns the four ExaMode use cases and, in addition, prostate. Within the classified histopathology images, there are images from all organs (brain, lungs, prostate, cervix, skin, among others). For this we use the MeSH medical subject headings service[3] to select only those from colon/rectum, cervix, prostate, celiac disease, and lung. A similar strategy was developed by our partners in [7] for finding images and data from articles describing rare cancers.
4. Finally, the fourth step is the exploitation of the filtered set of images, where they can be used for augmenting the existing clinical databases or for training and testing deep learning models with challenging heterogenous data. While the idea of including images from scientific literature to enhance the diagnostic process by pathologists was originally proposed by HES-SO [4], now is also used by researchers around the world,

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/
[2] https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_package/
[3] https://meshb.nlm.nih.gov/search/

for example, in a recent article [9], the authors show how PMC images can be used to train multimodal machine learning models and interact with pathologists.

3.2.2 Extracted data

The images and text from the scientific literature is a subset of images from the PubMed Central (PMC) dataset of biomedical open access literature and it represents an extension of the dataset presented in D3.1.
The images are available to the consortium via Cartesius (SURFSARA BV) at the following paths:

1. /projects/0/examode/CeliacDisease/PMC/
2. /projects/0/examode/Colon/PMC/
3. /projects/0/examode/Lung/PMC/
4. /projects/0/examode/Prostate/PMC/
5. /projects/0/examode/UterineCervix/PMC/

As described in Table 1, currently the scientific literature ExaMode dataset includes 2699 images related to colon and colon cancer, 962 images related to uterine cervix and uterine cervix cancer, 4151 images related to lung and lung cancer, 165 images related to celiac disease and 1925 images related to prostate and prostate cancer, leading to a total of 9'902 images. For each case, a tab separated values (.tsv) file contains the reference to the image name, the light microscopy confidence score, the caption, PMC identifier and the article abstract.
The scientific literature ExaMode dataset is strongly different from the datasets described in the section 2. It is not composed of clinical images (such as whole slide images or tissue microarrays). Instead, it consists of figures extracted from articles contained in a variety of medical journals, together with the associated text (e.g. image caption, title and abstract of the article, and the reference to the corresponding scientific paper, enabling knowledge extraction as well as text-based search and retrieval). This dataset is expected to grow further during the ExaMode project lifespan, since scientific articles are continuously published.
This part of the ExaMode dataset has a huge potential, but it is also the most difficult to analyze and to include in the knowledge discovery pipeline. In fact, the images of the scientific literature data are characterized by extremely high variability (in scale, resolution and color) while the associated text is also characterized by high variability (particularly due to how different authors and journals refer to sub-images in compound images separation).

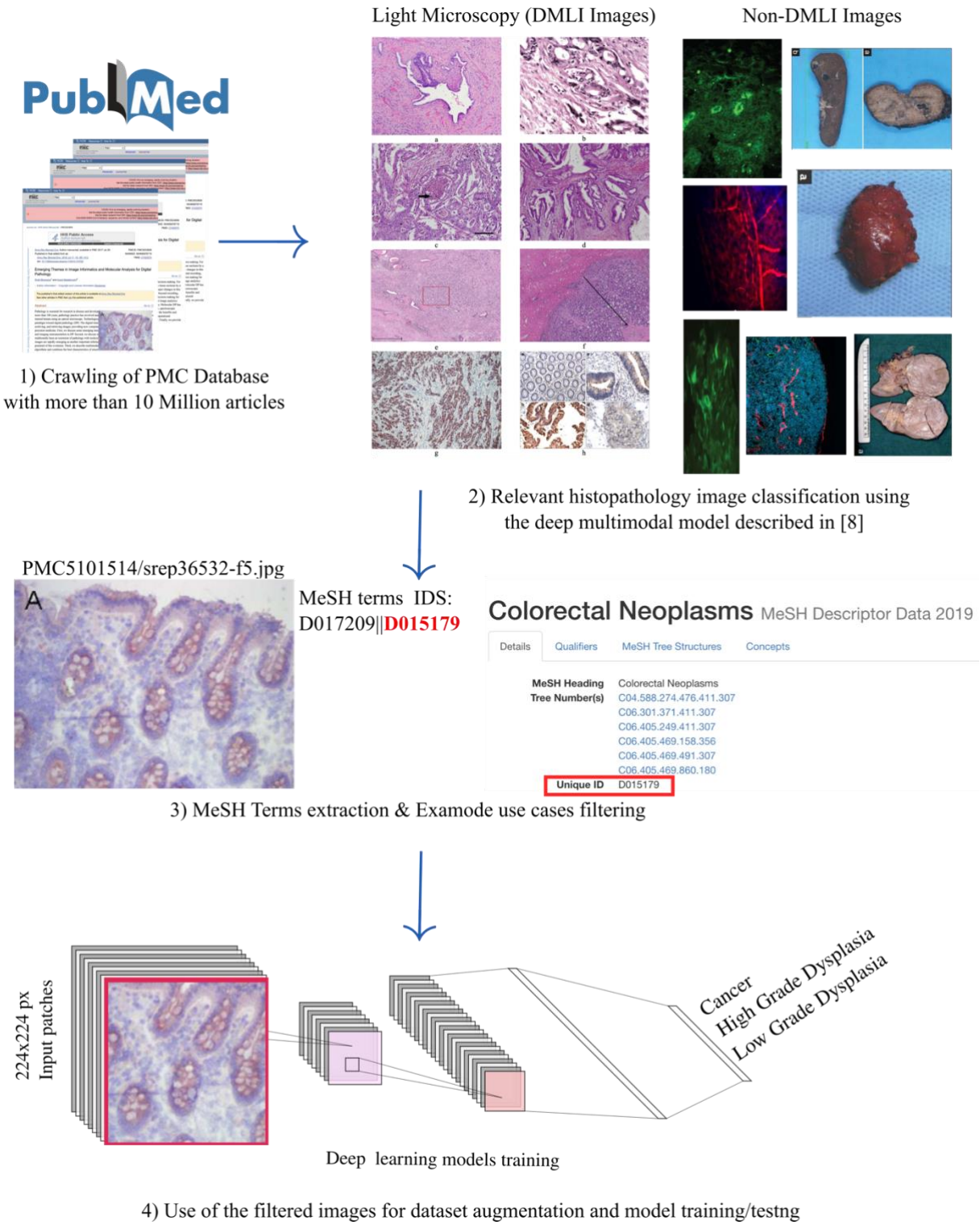

Figure 3 Data curation pipeline for images from the scientific literature

3.3 Additional publicly available data sources

This section includes additional publicly available data sources, selected from datasets that have been used for challenges and competitions, or simply that were released together with scientific articles in the last years.
Such publicly available datasets can be extremely useful for the project, since they can be included into weakly supervised learning procedures. One of the motivations motivating the ExaMode project is that getting image annotations from highly specialized personnel can be extremely difficult and expensive. The overall objective of ExaMode is thus to bypass as much as possible the interaction with experts by using weakly supervised learning systems. Our experiments show that pre-made strong annotations on different datasets can be fine-tuned using weak labels in order to teach algorithms more quickly and with better results [5], thus targeting the original ExaMode objective.
Table 2 summarizes the list of datasets from additional publicly available sources. Notably, two of the datasets (colon and prostate) contain a large number (>5'300) of annotated images and whole slide images (>10'000), that could be used jointly with the annotated data from the first set of proprietary data to have more data for training and testing the developed algorithms in realistic scenarios characterized by high heterogeneity.

Table 2 Other publicly available data sources *(updated on 23/06/2020)*.

Use case	Source	Task	WSIs	Images	Image patches	Associated text	Annotations (Y/N)	URL
Colon	Tissue bank of the National Center for Tumor diseases (NCT, Heidelberg)	Colon tissue image classification into the classes: Adipose, background, debris, lymphocytes, mucus, smooth muscle, normal.	86			Diagnostic report, structured (table)	Y	https://zenodo.org/record/1214456#.XuD-zmr7Q8M
Colon	Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, Marx A, Zollner F: Multi-class texture analysis in colorectal cancer histology (2016), Scientific Reports (in press)	Colon tissue image classification into the classes: Tumor, stroma.			5'000	Diagnostic report, structured (table)	Y	https://zenodo.org/record/53169#.XuD1_Wr7Q8N
Colon	Bokhorst et al., Learning from sparsely annotated data for semantic segmentation in histopathology images, MIDL 2019.	Classification of colorectal tissue images in the following classes: 1) tumor, 2) desmoplastic stroma, 3) necrosis and debris, 4) lymphocytes, 5) erythrocytes, 6) muscle, 7) healthy stroma, 8) fatty tissue, 9) mucus, 10) nerve, 11) stroma lamina propria, 12) healthy glands, 13) background.	70			Diagnostic report, structured (table)	Y	* This dataset is not publicly available but it is released by RADBOUDUMC which is part of the consortium for the other member.
Colon	K. Sirinukunwattana, D.R.J. Snead, N.M. Rajpoot, "A Stochastic Polygons Model for Glandular Structures in Colon Histology Images," in IEEE Transactions on Medical Imaging, 2015.	Segmentation of glandular objects.		237		N/A	Y	https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/download/
Colon	MILD-Net: Colorectal Adenocarcinoma Gland (CRAG) Dataset	Segmentation of glandular objects.		165		N/A	Y	https://drive.google.com/file/d/1p3dZXpgeA1IcGO6vXhStbVLMku-fZTmQ/view
Lung	School of Medicine, Stanford University, USA (Centers information not listed)	Classification/detection of growth patterns related to cancer aggressiveness, prognosis		868		N/A	N	https://tma.im/cgi-bin/viewArrayBlockList.pl
Lung	TCGA LUAD	Classification of Lung Adenocarcinoma tissue	585			Diagnostic report, unstructured pathology report	N	https://portal.gdc.cancer.gov/projects/TCGA-LUAD
Lung	ACDC@LUNGHP		200			N/A	Y	https://acdc-lunghp.grand-challenge.org/Download/
Prostate	Radboud University Medical Center and Karolinska Institute	Gleason grading	11'000			Diagnostic report, structured (table)	Y	https://www.kaggle.com/c/prostate-cancer-grade-assessment/data
Prostate	TMAZ	Gleason grading		886		N/A	Y	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OCYCMP

Prostate	TCGA PRAD	Gleason grading	500			Diagnostic report, unstructured pathology report	N	https://portal.gdc.cancer.gov/projects/TCGA-PRAD
TOTAL			**12'441**	**2'156**	**5'000**			

4 Conclusion

Deliverable 3.5 "Final set of cured publicly available multimodal and multimedia" was expected to include at least 5'000 images and related text extracted from scientific literature and the web. As described in this report, the deliverable was well accomplished. The final set of curated publicly available multimodal and multimedia includes over 9'900 images from the scientific literature and the associated text (in the form of image captions, article title, abstract and text) and 245 WSIs from TCGA. Finally, a list of additional publicly available sources was presented in table 2, including over 12'000 WSIs and 2'000 image regions with annotations, leading to a total of over 12'000 WSIs and over 12'000 images from publicly available sources (Table 1).

The scientific literature images and the TCGA images are available to the consortium via Cartesius (SURFSARA BV). Several experiments were performed to use the publicly available datasets to train models. The experiments and results are described in several publications submitted to peer reviewed international conferences and journals, suggesting that the publicly available multimodal and multimedia data represent a useful resource for ExaMode.

The collected data comes with a high degree of heterogeneity, which will allow to test the algorithms developed in WP3 and WP4 in challenging scenarios in order to also build more useful and reliable prototypes for WP5, WP6 and WP7.

Project supported by European Union

Horizon 2020 grant agreement 825292

5 References

[1] F. Fraggetta, S. Garozzo, G. F. Zannoni, L. Pantanowitz, E. D. Rossi, et al., "Routine digital pathology workflow: The Catania experience," *J. Pathol. Inform.*, vol. 8, no. 1, p. 51, 2017.

[2] A. M. Shebl, K. R. Zalata, M. M. Amin, A. K. El-Hawary, "An inexpensive method of small paraffin tissue microarrays using mechanical pencil tips". *Diagn. Pathol.*, 2011.

[3] I. Zlobec, G. Suter, A. Perren, A. Lugli, "A Next-generation Tissue Microarray (ngTMA) Protocol for Biomarker Studies". *J. Vis. Exp.*, 2014.

[4] R. Schaer, S. Otálora, O. Jimenez-del-Toro, M. Atzori, H. Müller, "Deep learning-based retrieval system for gigapixel histopathology cases and open access literature". *J. Pathol. Inform.*, vol. Accepted, 2019.

[5] S. Otálora, N. Marini, M. Atzori, H. Müller, "Transfer Learning for Gleason Pattern Classification by Combining Weak and Strong Supervision". *IEEE Journal of Biomedical and Health Informatics*. Revision submitted, 2020.

[6] S. Otálora, M. Atzori, A. Khan, O. Jimenez-del-Toro, V. Andrearczyk, H. Müller. "A systematic comparison of deep learning strategies for weakly supervised Gleason grading". Medical Imaging 2020: Digital Pathology (Vol. 11320, p. 113200L). International Society for Optics and Photonics.

[7] A. Dhrangadhariya, O. Jimenez-del-Toro, V. Andrearczyk, M. Atzori, H. Müller. "Exploiting biomedical literature to mine out a large multimodal dataset of rare cancer studies". In Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications (Vol. 11318, p. 113180A). International Society for Optics and Photonics.

[8] V. Andrearczyk, H. Müller. "Deep multimodal classification of image types in biomedical journal figures". In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 3-14). Springer, Cham.

[9]. A. Schaumberg, W. Juarez, S. Choudhury, L. Pastrián, B. Pritt, M. Pozuelo, ..., S. Yip. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. Nature Modern Pathology (2020).

Project supported by European Union

Horizon 2020 grant agreement 825292