“D7.5”

First set of data curated and available

Version: 1.0

Last Update: 29/12/19

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
ONTOTEXT AD	ONTOTEXT AD	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:	02/12/2019
Last Update:	29/12/2019

Revision History

Version	**Edition**	**Author(s)**	**Date**
0	1	Simona Vatrano	02/12/2019
Comments:	Drafting of the document		
0	2	Manfredo Atzori	02/12/2019
Comments:	Document informal control		
0	3	Francesco Ciompi	12/12/2019
Comments:	First review		
0	4	Simona Vatrano	19/12/2019
Comments	Revision of document, adding information in Table 1		
1	0	Manfredo Atzori	29/12/2019
Comments	Final control		

Abstract

This document was written to describe the features of the first set of data, which was created following the consortium requirements and needs. The whole slide images and medical reports contained in this data set were selected from the over 600’000 AOEC and RADBOUDUMC cases and will be made available to the ExaMode consortium via the secure transfer module accomplished in Deliverable 7.4 “Secure data transfer module”.

Table of Contents

Index of Tables

List of abbreviations

CNN	Convolutional Neural Network
WSI	Whole Slide Images
NOMED	Systematized Nomenclature of Medicine
IHC	Immunoistochemistry

1 Introduction

In the ExaMode project all partners are working to obtain a machine learning system (also based on Convolutional Neural Networks, CNNs) able to link text and image information derived from a set of histopathology data, including diagnostic reports and Whole Slide Images (WSI). These data are provided by the hospital members of the consortium (AOEC and RADBOUDUMC), according to requirement previously established.

During the first year of the project, an example set of histopathology data was selected to start testing the WSI viewer under development in the consortium and to make manual annotations on digital images , as well as to test routines of feature extraction from text data (described respectively in Deliverable 3.1 “First set of cured publicly available multimodal and multimedia data” and Deliverable 5.3 “Example dataset”). These steps are very important to develop the weakly supervised knowledge extraction systems targeted within ExaMode WP2 “Semantic knowledge discovery and visualisation”, WP3 “Image content based knowledge discovery” and WP6 “Multimodal knowledge management”.

In order obtain models that can generalize well to different clinical settings, it is important to get as many images as possible to train the deep neural networks that are being developed in WP4 “Computational Pathology” and WP5 “Decision support and image enrichment”.

The annotations on the dataset described in Deliverable 5.3 “Example dataset” was the starting point to run computational tests, providing the consortium with examples of annotated data. The aim of the extended dataset described in this deliverable is instead to provide a first set of curated and available pathology data, including WSIs and text/clinical information, in order to train deep neural networks using the annotated data described in Deliverable 5.3 as well as the weakly supervised learning pipelines currently under development in WP2, WP3 and WP4.

This set of data corresponds to tissues and cancer types identified in Deliverable 5.1 “Priority list of tissue & cancer types”. As such, it is based on the data provided in Deliverable 3.1 “First set of cured publicly available multimodal and multimedia data” and on the guidelines provided in Deliverable 5.2 “List of the essential knowledge requirements for each tissue”. The identification of this first set of data with all correlated information has been decided based on consortium and project needs. In the document we describe in detail the dataset and the information provided with it.

The dataset described in this deliverable includes a total of 10’000 WSIs with corresponding text data from the pathology reports. The data were provided by AOEC and RADBOUDUMC. The dataset covers all the cases included in the priority list of tissue & cancer types provided in D5.1 “Priority list of tissue & cancer types” and it follows the indications provided in the list of the essential knowledge requirements for each tissue presented in D5.2 “List of the essential knowledge requirements for each tissue”.

2 First Dataset for ExaMode Consortium

In this deliverable we describe the features and characteristics of the first set of curated data available to the ExaMode consortium. In the following sections we describe the procedures for the identification, selection and transfer of the data.

2.1 Identification and selection of the first set of data for the consortium.

The identification and the construction of the dataset for the consortium is very important for training the computational models. Following the partners' suggestion and needs, the dataset consists of the clinical images, SNOMED codes and the diagnostic reports, as previously done for the first set of cured publicly available multimodal and multimedia data (Deliverable 3.1). In detail, for every selected case we provide:

- *WSI*: the images of the tissues, selected among the 600.000 slides digitized at the AOEC and Rabdoudumc Hospitals. The slides were collected following the specific characteristics and the disease classes previously identified and accepted by the consortium (Deliverable 5.1 "Priority list of tissue & cancer types" and Deliverable 5.2 "List of the essential knowledge requirements for each tissue"). According to the consortium requirements we planned to select a reasonable number of cases for each disease type with corresponding clinical data. All details are summarized in Table 1.
- *Clinical Report*: for each patient/case we provided also the corresponding pathological report which is available after clinical diagnosis. In this text data the consortium can find all the clinical and pathology information related to the WSI. This information is essential to extract medical knowledge from the data and link it to the visual content of the images. The pathology report is provided in the original languages (Italian and Dutch), in the form of an an excel file and/or, wherever possible, as pdf file.
- *SNOMED codess*: as thoroughly described in Deliverable 3.1 (first set of cured publicly available multimodal and multimedia data), for every diagnosis we provided also the corresponding SNOMED code (described in detail in Deliverable 7.3, "Instructions to understand the clinical reports"). The SNOMED codes reported are referred to as "T-topografy"-"D-disease/diagnosis"-"M-morphology" and "P-procedure", where every letter is followed by a code of at least 8 numbers [1-2]. These codes are univocal and identify a specific diagnostic workflow in Anatomic Pathology field. All data are provided in an excel table where each line corresponds to the "slide/case".

The dataset is characterized by high variability in terms of cases. Cases variability is one of the main requirements of the consortium, since it allows to train more robust and complete computational models. In detail, we selected several histological variants of Non-Small-Cell Lung Cancer, several benign and malignant variants of Colon Cancer and, for Coeliac Disease, every case was associated to the immunohistochemical stain.

Table 1: Summary of the first set of curated data.

	Colon Cancer	Uterine Cervix	Coeliac Disease	Lung Cancer
AOEC				
Number of cases	2000 cases	2000 cases	2000 cases	2000 cases
Magnification	40X	40X	40X	40X
Resolution	0.25 micron/px	0.25 micron/px	0.25 micron/px	0.25 micron/px
Staining	H&E	H&E	H&E-IHC	H&E-IHC
Scanner Used	Aperio AT2	Aperio AT2	Aperio AT2	Aperio AT2
NOMED Codes	T-M-P-D-codes	T-M-P-D-codes	T-M-P-D-codes	T-M-P-D-codes
Clinical reports	Available (pdf file)	Available (pdf file)	Available (pdf file)	Available (pdf file)
Rabdoudumc				
Number of cases	2000 cases	No Cases	No Cases	No Cases
Magnification	40X	/	/	/
Resolution	0.25 micron/px	/	/	/
Staining	H&E	/	/	/
Scanner Used	3DHISTECH P1000	/	/	/
NOMED Codes	T-M-P-D-codes	/	/	/
Clinical reports	Available (pdf file)	/	/	/

2.2 Transfer and Usage of First Dataset for the Consortium.

All the data will be transferred to SURFSARA BV from the server of each hospital using the Secure Transfer Module (Deliverable 7.4 “Secure data transfer module”). The creation of the secure transfer module allows to move data and corresponding information safely and easily from the Hospitals to the SURFSARA BV servers. The module is based on the AOEC software infrastructures and integrates AOEC and SURFSARA BV resources. The creation and description of the Secure Transfer Module is performed in more detail in Deliverable 7.4.

The creation of the curated dataset followed the consortium and ExaMode Privacy policy, thoroughly described in Deliverable 7.1 “General document for ethics and privacy regulations”. According to it, all WSI previously digitized for diagnostic purposes (AOEC) or obtained from archival data (RADBOUDUMC) which were selected for the creation of the ExaMode dataset, must undergo a procedure of full anonymization, which causes a complete loss of patient identifiable information. The identifiable data is removed from all data sets (Clinical Report) and replaced by an unidentifiable code. The entire anonymization process is performed using bioinformatics tools available to the consortium. Only the data managers of the local institution (respectively AOEC and RADBOUDUMC), who are allowed to replace the identifiable information with the new unidentifiable study code, have access to the initial non anonymized information, which is then destroyed before sharing data with researchers. Therefore, researchers involved in ExaMode project only have access to fully anonymized data. All types of data described in the previous section of this document is shared after data anonymization, where an unidentifiable study code links all the information. Only anonymized information is shared with partners in the consortium of the ExaMode project and in the form of publicly available data sets.

According to consortium requirements, AOEC annotates only few WSI following the instructions and annotation classes described in Deliverable 5.2 “List of the essential knowledge requirements for each tissue”. The rest of WSIs will be processed by the computational models using weakly supervised learning techniques.

3 Conclusion

In conclusion, Deliverable 7.5 “First set of data curated and available” was accomplished. This report described it and provided the information about the set of curated data that are now available to the ExaMode partners. All the data was selected following the requirements and needs of the consortium, to obtain an amount of examples of first collected and curated dataset, to train the computational models for the ExaMode multimodal and multimedia knowledge extraction purposes. This dataset can be considered as an excellent starting point which is expected to increase during ExaMode, since we should aim at a higher number of cases to achieve good performance of the algorithms. The dataset was created and is managed following the privacy policy described in Deliverable 7.1 “General document for ethics and privacy regulations”.

4 References

[1] *Ingenerf, J; Linder, R (2009). "Assessing applicability of ontological principles to different types of biomedical vocabularies". Methods of Information in Medicine.* ***48*** *(5): 459–467.*

[2] *Ruch, Patrick; Gobeill, Julien; Lovis, Christian; Geissbühler, Antoine (2008). "Automatic medical encoding with SNOMED categories". BMC Medical Informatics and Decision Making.* ***8****: S6*