“D7.7”

First set of annotated digital pathology data

Version: 1.0

Last Update: 30/06/20

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 AI	SIRMA	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:	12/06/2020 30/06/2020

Revision History

Version	Edition	Author(s)	Date
0	1	Simona Vatrano	19/06/2020
Comments:	Creation of first draft with provisional data		
0	2	Francesco Ciompi, Manfredo Atzori	22/06/2020
Comments:	Providing comments on version 0.1		
0	3	Simona Vatrano	23/06/2020
Comments:	Corrections and reply to comments on version 0.2		
0	4	Simona Vatrano, Francesco Ciompi, Manfredo Atzori	30/06/2020
Comments:	Final revision of the document		
1	0	Manfredo Atzori, Francesco Ciompi, Simona Vatrano	30/06/2020
Comments:	Final check before submission		

Executive summary

The first set of annotated Whole Slide Images (WSI) is made available to the consortium. Such data are selected from the AOEC data and are expected to include at least 100 annotated WSI in the final annotated dataset. The requirements for making manual annotations were defined by AOEC and MICROSCOPEIT together and were made available to the consortium, in Deliverable 5.2 "List of the essential knowledge requirements for each tissue" and Deliverable 5.3 "Example datasets". Data annotation was performed mainly using the Virtum software developed by MICROSCOPEIT, while few annotations were still made with ASAP, developed by RADBOUDUMC.

Despite several hospitals (such as AOEC, responsible of this deliverable) were required to strongly focus on managing the Covid-19 outbreak, AOEC managed to achieve this deliverable in time, also thanks to the good collaboration with RADBOUDUMC.

Table of Contents

Table of Figures

Index of Tables

List of abbreviations

WSI	Whole Slide Image,
CNN	Convolutional Neural Network (CNN)
ROI	Region of Interest
DNN	Deep Neural Networks

1 Introduction

The main objective of the ExaMode project is to develop machine learning models for weakly supervised knowledge discovery of exa-scale heterogeneous data in highly specific domains, such as histopathology in the era of Digital Pathology. In order to achieve this aim, we use Deep Learning algorithms such as Convolutional Neural Networks to extract knowledge and value from medical data, such as WSI, which are the first source of medical data in Histopathology. For this purpose, in parallel to unsupervised approaches, we can feed the machine with Regions of Interest (ROI) extracted from the WSI and associate these ROIs to specific Classes of Annotations, similar to how a senior pathologist teaches a young pathologist to identify a disease on a WSI [1-2].

As for the human brain, the “quantity” of images acquired during the training period is proportional to the “quality” of diagnosis, the same for the machine: the more annotated ROIs on a WSI you give to the computer model, the more accurate the computer prediction should be. Detection of pathologic areas is based on ROIs delineation from background (not clinically relevant) tissue. A precise ROIs detection and delineation by experts is a tedious and time-consuming process, but a useful step to produce reliable information for development of computer methods. This valuable information, can be exploited in a weakly-supervised learning approach to initialize the learning procedure, or to complement the large amount of unlabelled, or weakly-labelled data. The aim is to build decision support tools that are resilient to data acquired from different sources, different staining and cutting protocols and different scanners.

The objective of ExaMode is to reduce human interaction in training models, by reducing (and possibly completely remove) data annotation requirements while using semi-supervised approaches. The advent of clinical digital pathology and the concomitant increase in the number of histopathology images derived from public sources has made digital pathology an excellent candidate for the application of deep learning based classification models. The possible creation of weakly supervised approaches to extract, link and retrieve multimodal information from highly heterogeneous and unstructured medical data is the main objective of ExaMode project. Nevertheless, having annotated regions of interest is fundamental for ExaMode in order to test the developed models, to develop models that are based on the combination of strong and weak approaches, and to allow the companies involved in the project (MICROSCOPEIT and ONTOTEXT) to start working on products as soon as possible.

2 Annotation Procedure

The aim of this Deliverable was to build and make the first set of annotated WSI data available to the consortium, containing manually annotated regions of interest of diseases identified by the consortium partners as of interest for the project.

Training machine learning algorithms requires data annotations from physicians, which are difficult to obtain. Unfortunately, the 2020 year has started with a global health problem, the Covid-19 pandemic, which has stressed not only the health systems worldwide, but also the economy and society, which underwent a devastating period, with different difficulties still present worldwide. Italy, such as other European countries, was completely locked down and all medical specialists (including the pathologists) were enrolled to face off the pandemic problem. Since the annotation requires the employment of an expert Pathologist and considering what happened, the *Annotation Procedure* started late. Nevertheless, the consortium managed to achieve the deliverable releasing 100 and 50 cases from medical partners AOEC and RADBOUDUMC, respectively. Among the four diseases included in the ExaMode project (D5.5 "Priority List of tissue and cancer types"), we focused the attention on Colon Cancer only, to achieve the best results without dispersion of energy and optimization of sources and data.

2.1 Selected Cases and Annotation Workflow

The cases selected for the first annotated dataset were extracted from the "first set of data curated and available" (D.7.5). The cases were selected in order to obtain a good representation of all possible annotation classes, following the specific characteristics identified and accepted by the consortium (D.5.2 "List of essential knowledge requirement for each tissue"). The annotation procedure was carried out by two different experts from Italy and the Netherlands respectively. Among the 2000 and 2500 cases from AOEC and RADBOUDUMC Hospitals respectively, we selected a subset of 200 WSI to be annotated:

- 100 cases from *AOEC Hospital* were identified with a relatively balanced distribution of diagnosis and subsequent annotation classes. The cases were selected based on Pathology Report (final diagnosis), also available with the WSI data. The dataset images derived from two different scanners, Aperio AT2 (Leica) and 3DHISTECH P1000 (ThermoFisher Scientific), improving the variability in terms of images and resolution. All selected WSI were

H&E without IHC slides associated. The main characteristics of the selected cases are summarized in Table 1.

- 50 cases from *RADBOUMC Hospital* were identified from the mentioned dataset with a good distribution of diagnosis and subsequent annotation classes. The cases were selected based on Pathology Report (final diagnosis), also available with the WSI data. The dataset images derived from one scanner, 3DHISTECH P1000 (ThermoFisher Scientific), and also in this dataset all selected WSI were H&E only. The main characteristics of the selected cases are summarized in Table 1.

Table 1: Summary of selected cases for the first set of Annotated dataset.

Features of Selected Cases	AOEC Hospital	RADBOUDUMC Hospital
Number of WSI (n.)	100	50
Diagnosis (%):		
Cancer	36 (35%)	45 (90%)
High Grade Dysplasia	32 (30%)	3 (6%)
Low Grade Dysplasia	0 (0%)*	1 (2%)
Hyperplastic Polyp	32 (35%)	1 (2%)
Type of Staining	All H&E	All H&E
Scanner Used:		
Aperio AT2	60 (60%)	0
3DHISTECH P1000	40 (40%)	50

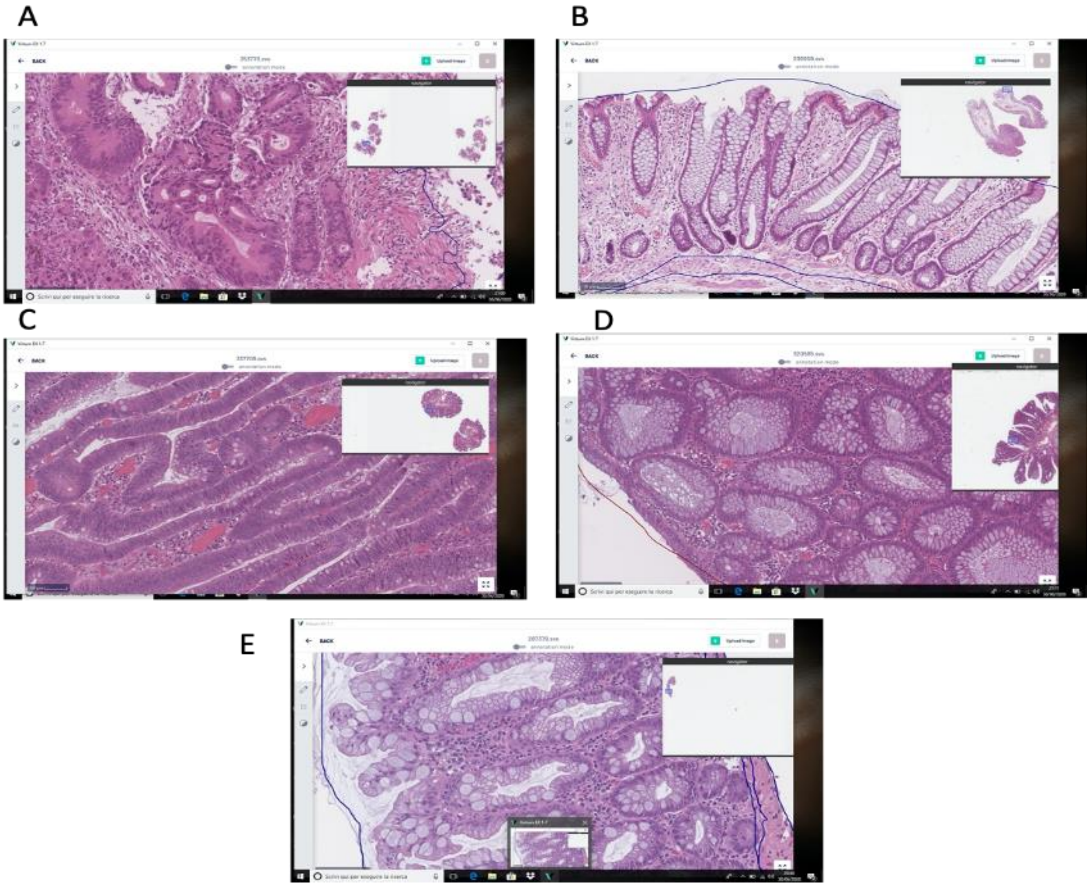
*This diagnostic category was not selected according to the consortium. NA: not available - the distribution of annotations among the Radboud cases is under evaluation by the expert pathologist.

According to consortium requirements, the annotation procedure was performed using the MICROSCOPE-IT annotation software, VIRTUM P1 (v1.7). MICROSCOPEIT Product Prototype 1 called Virtum Album is a Cloud-based software to manage histopathological slide collections. As a result of discussion, research and analysis performed within Task 5.1 and Task 5.2, some general and disease-specific features were defined for this prototype and described within Deliverable 5.4 ("Specifications of products MICROSCOPEIT-P1, MICROSCOPEIT-P2 and MICROSCOPEIT-P3"). At RADBOUDUMC, some cases were initially annotated using the in-house developed ASAP viewer. Successively, functionalities to import ASAP annotations into Virtum were developed by MICROSCOPEIT, which allowed to continue and finalize the annotation procedure using Virtum.

All selected cases were uploaded on *"Virtum Desktop"* and annotated following the instructions within the User Manual available for the consortium and well reported in the Deliverable 5.5 ("Prototype of product 1- VIRTUM-ALBUM (MICROSCOPEIT-P1)"). All the annotations were carried out using the annotation classes and features reported in the Deliverable 5.2 ("List of the essential knowledge requirements for each tissue") and considering the requirements for the different pathology decision support tools within the Deliverable 7.2 ("Decision support tool requirements report").

An example of WSIs selected for the creation of this first annotated dataset is reported in Figure 1, for each class of diagnosis considered. These images refer to the WSI before the annotation procedure.

Figure 1. Examples of cases selected for the annotation procedure from VIRTUM Desktop. (a) cancer; (b) No Cancer; (c) High Grade Dysplasia; (d) Low Grade Dysplasia and (e) Hyperplastic Polyp.



2.2 Results from the First Set of Annotated Cases

A total of 2977 annotations were performed by the two groups, working on selected WSI. The annotations were performed with the latest version of Virtum P1 software which is easier to use than the previous one. The software was able to manage all files from the different scanners and the multiple annotations tools available were easy to use. During the work and following a discussion between the pathologists and the consortium, one new class was added to allow a more complete annotation of the WSI, increasing the information requested to train the machine learning procedures. The new class added is “Normal Mucosa”.

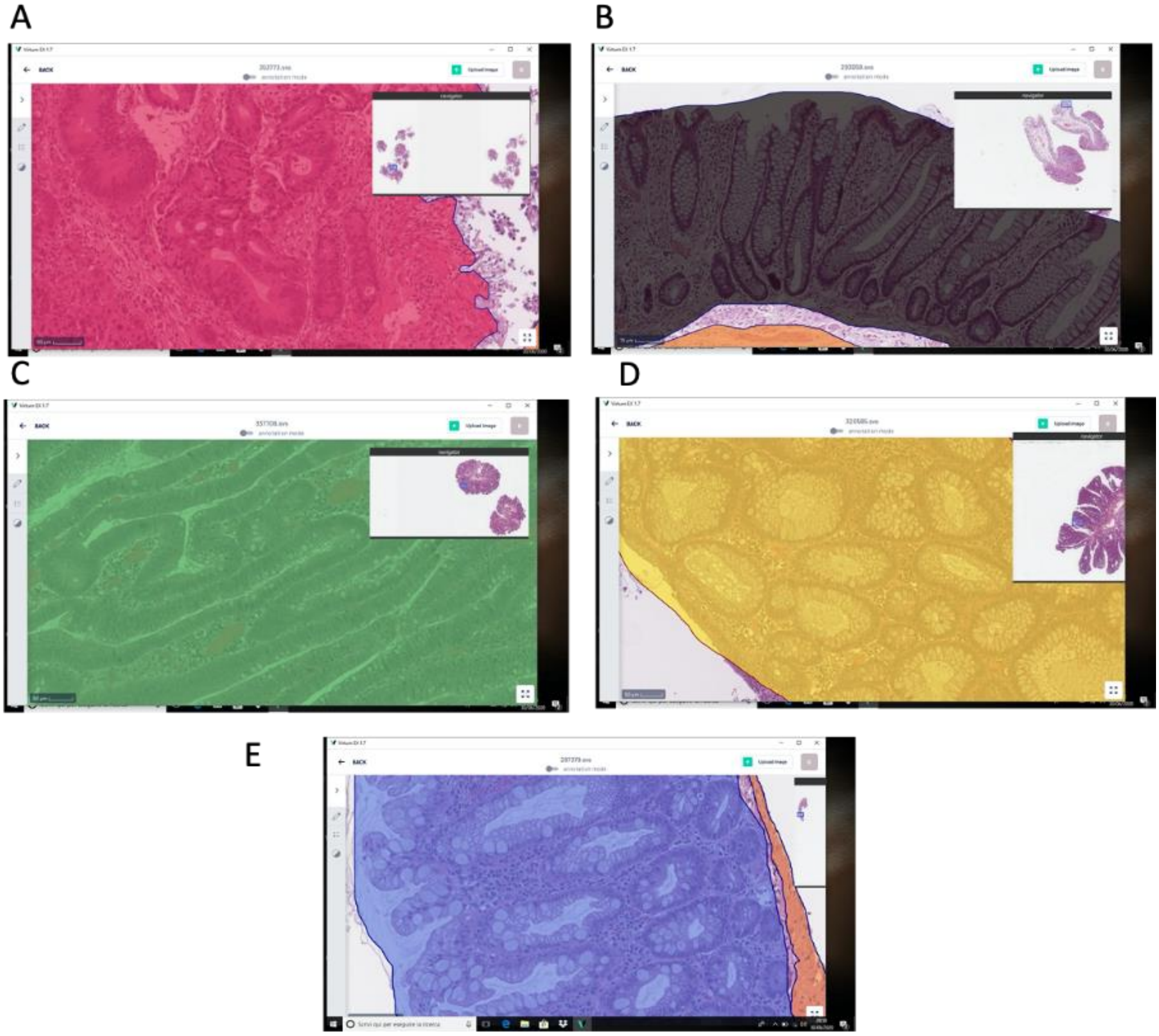
An example of annotations performed is illustrated in Figure 2, using the same cases reported in Figure 1. In detail, considering the different classes of annotation used, the number of annotations were:

- *AOEC Hospital*. Among the 100 cases annotated the distribution of annotations (tot. 2977) were for the different classes: 1417 (48%) Colon Cancer; 756 (25%) non informative; 301 (10%) Normal Mucosa; 304 (10%) High Grade Dysplasia; 104 (3,5%) Low Grade Dysplasia and 95 Hyperplastic Polyp (3,5%). An average of 30 annotations were performed for each case.
- *RADBOUDUMC Hospital*. Among the 50 cases annotated the distribution of annotations (tot. 352) was: 85 (24%) Colon Cancer; 113 (32%) High Grade Dysplasia; 76 (22%) Low Grade Dysplasia; 78 (22%) Healthy Glands; 121 (35%) Non-informative.

To underline the importance and the need to have a good annotation phase, a brief comment has been requested to one of the two pathologists regarding the annotation procedure:

- Pathologist 1 (GB): “During the annotation work of the selected cases, several critical issues emerged, to improve the annotation procedure. Together with the partners we are going to consider “intramucosal carcinoma” as cancer and to note the “hyperplastic mucosa” as normal. Considering the VIRTUM software, I suggest improving the focus closely to also allow macro-annotations using low magnification (using the pen tool)”.

Figure 2. Examples of annotation classes obtained after analysis of selected cases for the annotation procedure from VIRTUM Desktop. (a) cancer; (b) No Cancer; (c) High Grade Dysplasia; (d) Low Grade Dysplasia and (e) Hyperplastic Polyp.



3 Conclusion

In this Deliverable we reported the first set of annotated digital pathology data. Despite the Covid-19 outbreak, according to the consortium requirements, we were capable to achieve the deliverable and selected a total of 150 cases from the AOEC and RADBOUDUMC data to perform the annotations. Two different persons from the two hospitals worked to obtain this dataset, using the MICROSCOPEIT Product Prototype 1 called Virtum Album and, at the beginning, the software ASAP developed by RADBOUDUMC. Two main improvements were done with respect to Deliverable 5.2 "List of the essential knowledge requirements for each tissue" and Deliverable 5.3 "Example datasets". First, the class "Normal mucosa" was added to the annotations, which was not included before. Second, the cases were identified in order to have a more homogeneous distribution of classes. Moreover to implement the annotation we considered: *(i)* “intramucosal carcinoma” as cancer; *(ii)* “hyperplastic mucosa” as normal and *(iii)* to exclude from annotation the “stromal counterpart” where possible.

This first dataset of annotated data is very important for the following phases of the project. Even if the main objective of ExaMode is to reduce human interaction in training models, by reducing (and maybe completely removing) data annotation requirements in favour of semi-supervised approaches, data annotations are fundamental to test the developed models, to develop models that are based on the combination of strong and weak approaches, and to allow the companies involved in the project (MICROSCOPEIT and ONTOTEXT) to start working on products as soon as possible. In the era of Digital Pathology, the development of Neural Network able to manage, analyze and produce healthcare information starting from the WSI will be helpful for the pathologist and a new great opportunity for the industries.

4 References

[1] Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. Sci Rep. 2017;7:46450. Published 2017 Apr 18. doi:10.1038/srep46450.

[2] Ciresan, D., Giusti, A., Gambardella, L. & Schmidhuber, J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013, vol. 8150 of Lecture Notes in Computer Science 411–418 (Springer Berlin Heidelberg, 2013).