DEEPHEALTH

D8.7 Open Research Data Pilot

Project ref. no.	**H2020-ICT-11-2018-2019 GA No. 825111**
Project title	Deep-Learning and HPC to Boost Biomedical Applications for Health
Duration of the project	1-01-2019 – 31-12-2021 (36 months)
WP/Task:	WP8/ T8.4
Dissemination level:	PUBLIC
Document due Date:	30/06/2019 (M6)
Actual date of delivery	30/08/2019 (M8)
Leader of this deliverable	STELAR
Author (s)	Matthias Pocs
Version	v1.0

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 82511

Document history

Version	Date	Document history/approvals
0.1	27/03/2019	First draft contents
0.2	12/04/2019	Partner contributions
0.3	03/06/2019	Contribution by WP leader
0.4	11/06/2019	Addition of guidance, proposal for ORD
0.5	21/06/2019	Partner contributions
0.6	24/06/2019	Consolidation of contributions
0.7	28/06/2019	Reviewed
0.8	28/06/2019	Revised draft
1.0	28/06/2019	Submitted to EC
0.9	03/07/2019	Reviewed
0.91	26/07/2019	Revised draft
0.92	26/08/2019	Partner contributions
0.93	27/08/2019	Consolidation of contributions
0.94	28/08/2019	Finalised
1.0	30/08/2019	Submitted to EC

DISCLAIMER

This document reflects only the author's views and the European Community is not responsible for any use that may be made of the information it contains.

Copyright

Table of contents

Executive summary

This document extracts the information related to the DeepHealth project’s open research data from the Data Management Plan (D8.4), in particular, section 2.2 on data openness.

1 Data Summary

This section summarises the details of the collection/generation of data that is made open in the DeepHealth project. Four main categories of data will be managed in the project:

1. **Software. The code and its documentation** of the DeepHealth toolkit, that includes the European Distributed Deep Learning Library (EDDLL), the European Computer Vision Library (ECVL), and a front-end to easily use the functionalities of the libraries,
2. **Documents and dissemination material.** Scientific publications, newsletters, website contents, etc.
3. **Medical datasets from use cases and from public repositories.** The medical data, which are mainly images but also includes other clinical data in some of the use cases.
4. **Operational data and metadata.** (This is a tentative item)

As this document describes the first iteration of the DMP, the focus on these data categories represent the current state of knowledge. It is therefore possible that new categories of data appear in subsequent iterations of the DMP (e.g. code for the different platforms).

As described in the Grant Agreement, the code of the DeepHealth toolkit and its documentation will be available in public and certified repositories, as it is the case of GitHub[1], in order to make it easy for anyone interested to access the code and its documentation. The code of the libraries and the front-end will be open source and free software.

Scientific publications will be open access according to the rules of Horizon 2020. Other kind of publications as they are press releases, website contents and newsletters are open access for obvious reasons. No information that could jeopardize the proprietary software of industrial partners or their internal development lines will be published in any kind of publication.

Concerning to medical data, three of the 14 use case datasets are open access and can be shared. The remaining 11 data sets have restricted access in accordance with any data sharing agreements (non-disclosure agreements). The medical datasets are subject to the GDPR and other protection rules with respect to the privacy of personal data.

Trained predictive models are defined by the topology of the designed neural network and the estimated values of the weights. The weights are simply real values that are dropped every time a new estimation is ready to be used, i.e. predictive models are replaced periodically. The process of training predictive models is done frequently as new labelled/annotated data is available. So this kind of data is not necessary to be maintained, it can be regenerated from scratch when necessary. In fact, the training procedure starts from scratch, i.e. from a random initialization of the weights. For all these reasons, predictive models are not considered as a data category in this Data Management Plan.

[1] https://github.com/

2 FAIR data

This section describes the application of the FAIR data principles in the DeepHealth project to the data that is made open.

2.1 Making data findable, including provisions for metadata

In the DeepHealth project, the properties of each of the various open datasets to be used to promote findability of the data (associated metadata, unique persistent identifiers, naming conventions, etc.) are specified in the table attached as Annex to this document.

2.2 Making data openly accessible

Numerous datasets of the DeepHealth project will be made open. This subsection will outline which datasets and how and where they will be made publicly available.

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

The following data created within the project will be made openly available:

- DeepHealth toolkit
- Public deliverables
- Scientific publications (through Green or Gold Open Access)
- Dissemination and Communication material (including website, newsletters, leaflet, press releases, general publications, etc.).

The DeepHealth associated software developed in the project will be open source and free of charge. The code and documentation of the DeepHealth toolkit (EDDLL, ECVL and front-end) will be publicly accessible in the GitHub repository.

The website contents and the newsletters will be public and available in the website. Public deliverables of the project will also be available in the website of the project. Concerning (scientific) publications, prior notification will be covered by the consortium agreement to help identify confidential information and to keep draft publications closed if needed.

Some of the medical datasets used in the 14 Use Cases are already publicly available (see previous section). However, most of the medical datasets will be mostly kept private for legal reasons of protection of personal data. This is in line with the consortium agreement and its clauses governing the non-disclosure of information (especially due to legal reasons of privacy and data protection).

Use case	UC1 – Migraine & Seizures prediction (partner WINGS)
Dataset	Migraine & Seizures datasets
Private or public	Private
Reasons (if private)	Legal protection of personal data. Informed consent and any other process related to GDPR will be followed from WINGS members participating in DeepHealth.
Means of availability (if public)	NA
Use case	UC2 – UnitoPath: Classification of whole-slide histological images of colorectal biopsy samples (partner UNITO)
Dataset	UNITOPath dataset
Private or public	Private
Reasons (if private)	Legal protection of personal data

Means of availability (if public)	NA
Use case	UC3 – Brain (partner UNITO)
Dataset	UNITOBrain dataset
Private or public	Private
Reasons (if private)	Legal protection of personal data
Means of availability (if public)	NA
Use case	UC4 – Chest (partner CDSS)
Dataset	Chest dataset
Private or public	Private
Reasons (if private)	Legal protection of personal data. CT scans to be anonymised before moving the data from the medical premises to the ODH storage space.
Means of availability (if public)	NA
Use case	UC5 – Deep Image Annotation (partner UNITO)
Dataset	Automatic annotation of image-based medical examinations with text in natural language
Private or public	Public
Reasons (if private)	NA
Means of availability (if public)	Publicly available on the Internet
Use Case	UC6 – PROMORT: A vertical application to support digital pathology in the context of prostate tumor diagnosis (partners Karolinska Institutet and CRS4)
Dataset	Promort dataset
Private or public	Private
Reasons (if private)	Legal protection of personal data. UC6 on PF6 will only use pseudonymised images. Images will be transferred to CRS4 under tight contractual control.
Means of availability (if public)	NA
Use case	UC7 – Major Depression (partner OVGU)
Dataset	MRI based functional and strutural imaging
Private or public	Private
Reasons (if private)	Legal protection of personal data. Conditions by university’s IT and data protection departments and ethics board. Data sharing will be subject to NDA forbidding any exploitation or commercialisation of the health data.
Means of availability (if public)	NA
Use case	UC7 – Major Depression (partner OVGU)
Dataset	Prospective data monitoring depressive symptoms and adherence to treatment
Private or public	Private
Reasons (if private)	Legal protection of personal data. During pilots, the users will be asked for their informed consent for their data to be collected, stored and processed in WINGS infrastructure (MigraineNet platform) and

	for their data to be accessed and processed by both the WINGS and OVGU teams upon patients consent.
Means of availability (if public)	NA
Use case	UC8 – Dementia (partner OVGU)
Dataset	Retrospective data
Private or public	Private
Reasons (if private)	Legal protection of personal data. Private infrastructure to be determined. Where data sharing is needed (in connection with PF4 PIAF), an NDA will be signed for data sharing and Each Party expressly agrees that any exploitation or commercialization of the health data that may be shared between the Parties in the frame of the discussions covered by this non-disclosure agreement is strictly forbidden.
Means of availability (if public)	NA
Use case	UC9 – Study of structural changes in lumbar spine pathology (partner FISABIO)
Dataset	Lumbar spine pathology dataset
Private or public	Private before the hackathon
Reasons (if private)	Legal protection of personal data. File care file not to be shared for privacy reasons. Anonymised datasets will be shared with CEA (PF3 ExpressIF), UNITO (PF5 Open-DeepHealth) and EVR (PF7 Lumen) during the training and test phases according to FISABIO directives.
Means of availability (if public)	NA
Use case	UC10 – Predictive and Population Model for Alzheimer's Disease (AD) using Structural Neuroimaging (partner FISABIO)
Dataset	Alzheimer’s Disease dataset
Private or public	Private before the hackathon
Reasons (if private)	Legal protection of personal data. File care file not to be shared for privacy reasons. Anonymised datasets will be shared with UNITO (PF5 Open-DeepHealth) and EVR (PF7 Lumen) during the training and test phases according to FISABIO directives.
Means of availability (if public)	NA
Use case	UC11 – Image Analysis and prediction for Urology (partners SIVECO and SCHTB)
Dataset	Urology retrospective / prospective data
Private or public	Private
Reasons (if private)	Legal protection of personal data. No connection is being made to PHILIPS or any third party.
Means of availability (if public)	NA
Use case	UC12 – Skin Cancer (partner UNIMORE)
Dataset	Skin cancer melanoma detection dataset
Private or public	Public
Reasons (if private)	NA

Means of availability (if public)	Public archive available online Internal database will be organised and made public within the ethical requirements.
Use case	UC13 – Epileptic Seizures Detection (partners EPFL and CHUV)
Dataset	Physionet CHB-MIT EEG Database
Private or public	Public
Reasons (if private)	NA
Means of availability (if public)	Public on the Internet
Use case	UC13 – Epileptic Seizures Detection (partners EPFL and CHUV)
Dataset	CHUV epilepsy EEG database
Private or public	Private
Reasons (if private)	Legal protection of personal data. Due to the restriction of the data to be maintained either in the CHUV or in the EPFL premises, during this period, MigraineNet mechanisms will be set-up within the EPFL infrastructure (upon the necessary NDAs).
Means of availability (if public)	NA
Use case	UC14 – Multiple Sclerosis (partners EPFL and CHUV)
Dataset	MICCAI 2008 MS lesion segmentation challenge dataset & ISBI 2015 longitudinal MS lesion segmentation challenge database
Private or public	Public
Reasons (if private)	NA
Means of availability (if public)	Public on the Internet
Use case	UC14 – Multiple Sclerosis (partners EPFL and CHUV)
Dataset	CHUV MRI dataset from MS patients
Private or public	Private
Reasons (if private)	Legal protection of personal data. Deployment of the Open Innovation platform is done inside the EPFL / CHUV premise. No connection is being made to PHILIPS or any third party.
Means of availability (if public)	NA

How will the data be made accessible (e.g. by deposition in a repository)?

Software, publications and communication material will be made accessible by publishing them on the Internet on appropriate repositories for use by any third party (see later the particular chosen repositories).

According to the GDPR all private medical datasets need NDA signature and security measures to be put in place prior to making accessible any data. This process regularly involves the organisation's internal IT/data protection officer. Some (university) data owners also need to seek approval by an internal ethics committee.

Concerning the public medical datasets, the datasets have been made accessible by deposition in an online repository. Some of them require online registration (login credentials) for access.

What methods or software tools are needed to access the data?

For data that is going to be made public, no specific methods or software tools are needed but only to access to the specific repositories on-line.

For medical datasets (not public), methods and tools will be defined for each use case. Usually, data in datasets is stored according to standard formats, so they exist open-source and free-software tools for both retrieving the samples (images and other clinical data) from tables in a database, or for loading the contents of files (it is typical to have the images of scans as single files in a filesystem).

Is documentation about the software needed to access the data included?

In this project no specific software for databases is developed, therefore, it does not apply.

Some functions for reading/writing images from/to files using different formats are provided as part of the ECVL, so the documentation of such functions is included in the documentation of the ECVL.

Functions for loading/saving deep neural networks using the ONNX format are provided as part of the EDDLL, so the documentation of such functions is included in the documentation of the EDDLL.

Is it possible to include the relevant software (e.g. in open source code)?

The source code of the DeepHealth toolkit will be publicly available in GitHub under the terms of open-source and free software.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

- Code and documentation of the three components of the DeepHealth toolkit in GitHub, with open access.
- Website contents and newsletters in the website of the project in open access mode.
- Public deliverables will also be in the website of the project in open access mode.
- Scientific publications in the respective website of the journal or conference publisher.
- Zenodo will be used for scientific publications and other public material such as public deliverables and communication material as default. This repository is recommended by the European Commission. This does not apply for website contents. However, the public deliverables and the newsletters will be in the website of the project, that will play the role of repository for this kind of data.
- The location of data deposition is listed below for the medical datasets. Note that the public medical datasets of UC 5 (Deep Image Annotation), UC 12 (Skin Cancer*)*, UC 13 (Epileptic Seizures Detection) and UC 14 (Multiple Sclerosis) were made public by third parties and are managed by third parties:

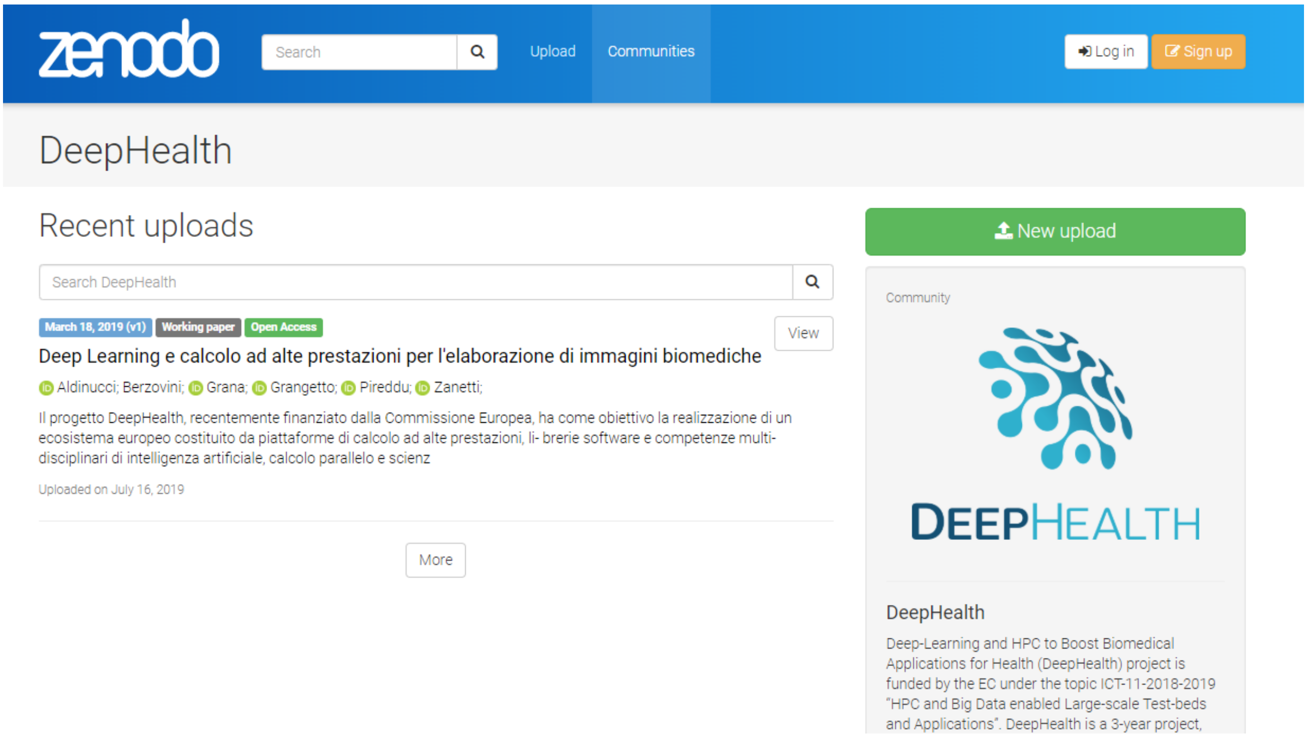
Use case	UC5 – Deep Image Annotation (partner UNITO)
Dataset	Automatic annotation of image-based medical examinations with text in natural language
Location of deposition	"Indiana University Chest X-Rays": https://openi.nlm.nih.gov/gridquery?q=&it=xg&coll=cxr PEIR Radiology, Chest: http://peir.path.uab.edu/library/index.php?/category/111
Use case	UC12 – Skin Cancer (partner UNIMORE)
Dataset	Skin cancer melanoma detection dataset
Location of deposition	Public archive: ISIC 2017 (login required): https://challenge.kitware.com/#challenge/583f126bcad3a51cc66c8d9a, Public archive: ISIC 2018 (login required): https://challenge2018.isic-archive.com/participate/
Use case	UC13 – Epileptic Seizures Detection (partners EPFL and CHUV)
Dataset	Physionet CHB-MIT EEG Database

Location of deposition	Physionet servers: https://www.physionet.org/pn6/chbmit/
Use case	UC14 – Multiple Sclerosis (partners EPFL and CHUV)
Dataset	MICCAI 2008 MS lesion segmentation challenge dataset & ISBI 2015 longitudinal MS lesion segmentation challenge database
Location of deposition	NITRC servers: MICCAI 2008: https://www.nitrc.org/frs/?group_id=745 Smart Stats servers: ISBI 2015: https://smart-stats-tools.org/lesion-challenge

Have you explored appropriate arrangements with the identified repository?

Up to now, partners responsible for the different software elements have already created the GitHub repositories.

Concerning Zenodo, a *DeepHealth community*[2] has already been created (see the following Figure), indicating the link to the EC as funding agency and the GA number. As public material is available the Community will be filled in with new contents.



DeepHealth

Recent uploads

Search DeepHealth

March 18, 2019 (v1) Working paper Open Access

Deep Learning e calcolo ad alte prestazioni per l'elaborazione di immagini biomediche

Aldinucci; Berzovini; Grana; Grangetto; Pireddu; Zanetti;

Il progetto DeepHealth, recentemente finanziato dalla Commissione Europea, ha come obiettivo la realizzazione di un ecosistema europeo costituito da piattaforme di calcolo ad alte prestazioni, li- brerie software e competenze multi-disciplinari di intelligenza artificiale, calcolo parallelo e scienz

Uploaded on July 16, 2019

DeepHealth

Deep-Learning and HPC to Boost Biomedical Applications for Health (DeepHealth) project is funded by the EC under the topic ICT-11-2018-2019 "HPC and Big Data enabled Large-scale Test-beds and Applications". DeepHealth is a 3-year project,

Figure 1. DeepHealth community at Zenodo.

If there are restrictions on use, how will access be provided?

Only applicable to private datasets, which will be accessible to other partners in the project. In theses cases access will be managed by concluding NDAs for bilateral data sharing. Further details on which partners will conclude NDAs and the templates used by those partners are given in D1.6.

Is there a need for a data access committee?

No, for scientific publications, public deliverables, newsletters, website contents, code and documentation.

[2] https://zenodo.org/communities/deephealth/

Yes, for datasets containing medical data (images and other clinical data), except for the public ones with open access, the public ones with restricted access need a data access committee.

Are there well described conditions for access (i.e. a machine readable license)?

Code and documentation will be available in open access mode at GitHub. No conditions for access are needed.

Public deliverables, website contents and newsletters will be available in open access mode. No conditions for access are needed.

Scientific publications will be available in the website of the publisher and in Zenodo (unless it is not possible for publisher limitations). Access conditions will be clearly described both in the published and in Zenodo, where the license of the publication will be specified as part of the data Zenodo requests for each uploaded file.

The public medical datasets may be accessed online by any user. Servers of NITRC, Smart Stats, Indiana University Chest X-Rays, PEIR Radiology (Chest), ISIC 2017, ISIC 2018 and *Physionet* (see sources above) provide open access mode. Some of them (ISIC 2017 & 2018) require prior registration for login.

Private medical datasets are intended to be kept confidential. Access require the conclusion of NDAs, taking of security measures and, if applicable, ethical approvals.

For each of the medical datasets, the table above lists the respective access conditions.

How will the identity of the person accessing the data be ascertained?

For software, as well as publications, no specific user identification will be needed to access the data. For medical data, the person must be logged in the website of the repository to have access to datasets. Both for public repositories and for private datasets with access only for people of the institution that is the data holder.

2.3 Making data interoperable

In the DeepHealth project, the properties of each of the various open datasets to be used to promote interoperability of the data (data types, data formats, whether existing or new data, data origin, expected data size) are specified in the table attached as Annex to this document.

2.4 Increase data re-use

In the DeepHealth project, the properties of each of the various open datasets to be used to promote re-use of the data (limitation of the necessary data embargo, third-party use, length of time for long-term preservation of data) are specified in the table attached as Annex to this document.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

In the DeepHealth project, the software libraries will be open source and published together with the public deliverables of the project.

For (scientific) publications, the maximum necessary data embargo period is set to six months.

Concerning the medical Use Cases, most of the datasets will be kept private for protection of (sensitive) personal data. The medical datasets that are public are already published, online addresses are listed per public dataset in the Annex.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

In order to develop the pilot test beds proposed in the DeepHealth project, different datasets will be used that can be classified in the following four groups:

1. Open software to be developed within the project that will be publicly available free of charge.
2. Public datasets available in open access mode not generated by any of the partners of the project.
3. Private datasets described in D1.1 and mentioned in the Section 1 of this document, that will not be available for reuse by third parties. These datasets are restricted because they contain personal data and the process of anonymisation cannot be guaranteed to be 100% irreversible. Therefore, these datasets cannot be made public nor shared among partners of the DeepHealth project according to the GDPR and other Personal Data Protection laws. In some of the cases, the software developed will be installed on the premises of the use case provider for running the training algorithms and evaluating the designed models.
4. Public datasets, also described in D1.1 and mentioned in the Section 1 of this document. These datasets will be shared among partners of the DeepHealth project to run experiments and evaluate the designed predictive models. Some of them are already public, and other ones will be made public during this project. At the end of the project these datasets will be publicly available for being reused by third-parties.

It is intended that all the public datasets used and/or generated in the DeepHealth project will be allocated in public data repositories, in open access mode but after registering and applying to use them. Registering the users that download a dataset is the best way for having information about its use.

How long is it intended that the data remains re-usable?

All the public data used and/or generated in the DeepHealth project will be available in the long term in order to ensure it is reused by any interested person or institution. Both the code of the DeepHealth toolkit and the public datasets will be available for reuse, and the documents (public deliverables, newsletters and scientific publications) will be available to be consulted.

Annex (Open Research Data Table)

Dataset name	Brief description	Data Category	Data types	Metadata	Data formats	Naming structure	Type of Persistent identifier	New data or data reuse	Origin of data	Expected data size	Owner	To whom outside the project data might be useful	Publicly available?	Location of deposition/Storage for public access	Length of time for deposition (duration of preservation)	Preservation responsible
EDDLL	European Distributed Deep Learning Library	Software	Software library	NA	source code (plain text)	No specific naming, but common guidelines for coding will be followed. "Google C++ Style" could be an option.	Commits' ID on the git repository	New	Created through WP2, WP3 and WP5 activities	Around 100MB incl. code and documentation.	UPV and other WP2 partners?	- IT companies - Data scientits in the health domain - software developers - machine-learning experts	Public	The source code and the associated documentation will be publicly available in GitHub	Long-term preservation	TBD but in principle UPV will be involved
ECVL	European Computer Vision Library	Software	Software library	NA	source code	The ECVL source code will follow the "Google C++ Style"	Commits' ID on the git repository	Both	NEW: created through WP3. RE-USE: public available data [https://opencv.org, https://github.com/opencv/opencv, https://www.wxwidgets.org, https://github.com/wxWidgets, https://github.com/sprinfall/dcm, https://imagej.net/ImageJ2, https://github.com/imagej/imagej, and other]	10MB - 1GB	UNIMORE and other WP3 partners?	- IT companies - Data scientits in the health domain - software developers - machine-learning experts	Public	The source code is and will be stored on GitHub. The documentation is stored on a public UNIMORE server now, but the final host must be defined (TBD).	Long-term preservation	https://github.com/deephealthproject/ECVL
DeepHealth toolkit front-end	Framework to exploit ECVL and EDDLL for processing images or other data type.	Software	Web-based GUI and/or desktop app (TBD)	NA	source code	TBD	Commits' ID on the git repository	New	Created through WP2 and WP3 activities	TBD	UNIMORE and other WP2 & WP3 partners?	Any company (large/SME) and any institution (hospital, ministry, agency, etc.) to exploit Hybrid and Heterogeneous HPC and Big Data clusters for processing images or any other data type - software developers - expert users	Public	The source code will be stored on GitHub. The documentation's final host must be defined (TBD).	Long-term preservation	https://github.com/deephealthproject/front-end.git
COMPSs distributed programming framework	Framework to distribute the computation of the inference and/or training processes included in the DeepHealth libraries on the underlying computing infrastructure	Software	NA	NA	NA	NA	NA	New	Exsiting COMPSs distributed programming framework	NA	BSC	NA	Public	NA	Long-term preservation	BSC
COMPSs	Programming development framework for distributed applications	Software	Software library	NA	source code	NA	Commits' ID on the git COMPSs repository	Re-use	http://compss.bsc.es/gitlab/compss/framework	~100MB	BSC	scientific communities, SW developers targeting distributed HPC/cloud-based infrastructures	Public	Github	Long-term preservation	BSC
UC5: Deep Image Annotation	Automatic annotation of image-based medical examinations with text in natural language.	Medical dataset	Medical radiological images and associated textual reports	Public dataset 1: MeSH terms	• Images: png • Reports: plain text	Same as original datasets	NA	DATA-REUSE	Public dataset 1: "Indiana University Chest X-Rays", https://openi.nlm.nih.gov/gridquery?q=&it=xg&coll=cxr Public dataset 2: PEIR Radiology, Chest, http://peir.path.uab.edu/library/index.php?/category/111	Dataset 1: 7471 images, 3955 reports Dataset 2: 102 images, 102 reports Size (for both): 2TB	Dataset 1: OpenI, US National Library of Medicine Dataset 2: PEIR Digital Library	Scientific communities	Public	Public on the internet	2 years	UNITO
UC12: Skin cancer melanoma	Skin cancer melanoma detection	Medical dataset	Dermoscopic and Confocal images	Most images are paired with the following metadata: binary segmentation masks, approximate age, general anatomic site, type of diagnosis, sex.	JPEG, DICOM	N/A	N/A	Previous data	Publicly available [ISIC 2017 (login required): https://challenge.kitware.com/#challenge/583f126bcad3a51cc66c8d9a, ISIC 2018 (login required): https://challenge2018.isic-archive.com/participate/] Dermatology hospital of Modena	10 GB - 1 TB	ISIC for the public data, UNIMORE for the other	Scientific community	Public/ UNIMORE	Public/UNIMORE repository	Long-term preservation	UNIMORE
UC13: Epileptic seizure detection	Physionet CHB-MIT EEG Database	Medical dataset	Scalp EEG records	Time stamps of epileptic seizures	EDF for signals, MIT-Annot for seizure annotations	N/A	URI	Public dataset	Publicly available data [https://www.physionet.org/pn6/chbmit/]	~45GB	Public dataset managed by Physionet	Scientific community	Public	Physionet Servers	Long-term preservation	N/A
UC14: Multiple sclerosis patients	MICCAI 2008 MS lesion segmentation challenge dataset	Medical dataset	T1-weighted (T1w), T2-weighted (T2w), and FLAIR MRIs	Ground truth for segmentation	NRRD	N/A	URI	Public dataset	Publicly available data [https://www.nitrc.org/frs/?group_id=745]	~10GB	Public dataset managed by NITRC	Scientific community	Public	NITRC Servers	N/A	N/A

Dataset name	Brief description	Data Category	Data types	Metadata	Data formats	Naming structure	Type of Persistent identifier	New data or data reuse	Origin of data	Expected data size	Owner	To whom outside the project data might be useful	Publicly available?	Location of deposition/Storage for public access	Length of time for deposition (duration of preservation)	Preservation responsible
UC14: Multiple sclerosis patients	ISBI 2015 longitudinal MS lesion segmentation challenge database	Medical dataset	T1w, T2w, proton density-weighted (PDw), and FLAIR MRIs	Ground truth for segmentation	N/A	N/A	URI	Public dataset	Publicly available data after registration [https://smart-stats-tools.org/lesion-challenge]	~3GB	Public dataset managed by Smart-Stats	Scientific community	Public	Smart-Stats Servers	N/A	N/A
Public deliverables	Public Deliverables from the project	Documents	Documents and reports	Metadata as provided by documents. Tags associated in public repositories (project identifier, author, etc.)	PDF	As specified in Project Handbook	DOI assigned in Zenodo	New	As a result of project activities (all WPs)	<2GB	Producer of the deliverable	Scientific community/ goverment/ healthcare proffessionals	Public	Project repository/ Zenodo/project website	2 YEARS	Coordinator/ Zenodo
Scientific publications	Scientific publications created disseminating project results	Documents	Scientific papers	Key worrds and document associated data	PDF	As required by the publisher	DOI assigned /Zenodo DOI	New	As a result of project activities (all WPs)	<2GB	Producer(s) of the paper	Scientific community/ goverment/ healthcare proffessionals	Public	Project repository/Zenodo/ project website/Open Access Journal website	Long-term	Zenodo/ OA Journals
Other public publications	Articles, posts in non-scientific media	Documents	Documents and reports	Key words	PDF	As required by the publisher	NA	New	As a result of WP7 activities	<2GB	Producer(s) of the content	General society, websites, press, consumers, healthcare professionals, etc	Public	project website, project SM channels	NA	NA
Project communication material	leaflet, flyers, brochures, posters, project presentation, newsletter	Dissemination material	Document	Metadata as provided by documents. Tags associated in public repositories (project identifier, author, etc.)	PDF	As defined in project handbook	DOI assigned by Zenodo	New	As a result of WP7 activities	<2GB	Consortium , producer of the content	General society, websites, press, consumers, healthcare professionals, etc	Public	Website, repository	2 YEARS	everis, WINGS
Public datasets for library development	Datasets needed for iterative testing during library developments	Research data and metadata	TBD	NA	TBD	NA	NA	Re-use	Public datasets to be identified	TBD	Depending of each dataset	NA	Public	Each dataset original repository and partner premises	Not needed	NA