www.visceral.eu

Data format definition focusing on Competition 1

Deliverable number	*D2.2.1*
Dissemination level	*Public*
Delivery date	*28 January 2013*
Status	*Final V1.0*
Author(s)	*Tomas Salas, Georg Langs, Henning Mueller, Bjoern Menze*

SEVENTH FRAMEWORK PROGRAMME

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Grant Agreement Number: 318068

Executive Summary

VISCERAL will provide a very large data set of medical images which will be used for an image retrieval benchmark and the automated annotation of these images.

These data will come mostly from electronic health records, and have been collected to provide health care.

Original data will have to go through a series of transformations in order to address legal issues and also to ensure that conforms to the needs of the benchmarking process.

This deliverable describes the format conventions for the collection, storage, and distribution of data in the VISCERAL project with a focus on competition 1. Data encompasses image data, and various information regarding annotations of regions, and landmarks in the imaging data. The deliverable provides a detailed description of these conventions, and how they should be implemented in the VISCERAL project.

To keep data management overhead at a minimum, the conventions are fixed at the beginning of the project, and it is planned to keep these conventions throughout the project lifetime. The conventions are formulated in a way that allows appending additional information later in the project, if relevant annotation aspects arise later. In any case newer conventions should be backwards compatible, so that existing pipelines can stay fixed.

Table of Contents

List of Abbreviations

MRI	Magnetic Resonance Imaging
CT	Computed Tomography
NIFTI	Neuroimaging Informatics Technology Initiative
DICOM	Digital Imaging and Communications in Medicine

1 Introduction

This deliverable contains the specifications of the data formats used for data distribution, annotation, and algorithm results. The two main categories of annotations and results are

1. **Region segmentations**. These encompass regions that correspond to anatomical structures (e.g., liver), or sub-parts. The regions are volumetric entities, in volume data.
2. **Landmarks**. These encompass points in the 3D space of the volumetric data. They can be either anatomical landmarks (e.g., distal point of the radius), or landmarks that are identified computationally across the population (e.g., a sampling of a bone surface, with corresponding points across the population)

In this deliverable we describe the format we will use for the competition preparation and competition focusing on competition 1.

2 Format of the image data

The first competition will be primarily dealing with multi-modal 3D imaging data (Magnetic Resonance Imaging, Computed Tomography). The standard format in which these images are saved in the clinical context is DICOM. There is a host of alternative formats that focus on specific applications. A widely used and accepted format that significantly reduces the file number in case of large 3D data is NIFTI[1]. For these reasons DICOM and NIFTI were chosen as the image data distribution formats both for annotation, and for the competitions.

Previous to their distribution, original DICOM objects and images will have to be processed in order to meet specific project requirements including:

- Ensure data privacy
- Provide information needed to perform annotations, both manually and automated (e.g., patient’s age)
- Remove, when necessary, information that may cause biases in the process of benchmarking (e.g., certain original annotation or measurements)

On an interim basis, and more detailed discussion pending, transformations performed on original DICOM objects will adhere to the following specifications:

- Some basic DICOM objects will have to be always present within the data set

SOP Class Name	SOP Class UID
CT Image Storage	1.2.840.10008.5.1.4.1.1.2
Enhanced CT Image Storage	1.2.840.10008.5.1.4.1.1.2.1
MR Image Storage	1.2.840.10008.5.1.4.1.1.4
Enhanced MR Image Storage	1.2.840.10008.5.1.4.1.1.4.1

[1] ***nifti***.*nimh.nih.gov*

- Additional DICOM objects may be removed, if present within the original imaging study, from the data set

SOP Class Name	SOP Class UID
Secondary Capture Image	1.2.840.10008.5.1.4.1.1.7
Grayscale Softcopy Presentation State	1.2.840.10008.5.1.4.1.1.11.1
Color Softcopy Presentation State	1.2.840.10008.5.1.4.1.1.11.2
Pseudo-Color Softcopy Presentation State	1.2.840.10008.5.1.4.1.1.11.3
Blending Softcopy Presentation State	1.2.840.10008.5.1.4.1.1.11.4
XA/XRF Grayscale Softcopy Presentation State	1.2.840.10008.5.1.4.1.1.11.5
Raw Data	1.2.840.10008.5.1.4.1.1.66
Spatial Registration	1.2.840.10008.5.1.4.1.1.66.1
Spatial Fiducials	1.2.840.10008.5.1.4.1.1.66.2
Deformable Spatial Registration	1.2.840.10008.5.1.4.1.1.66.3
Segmentation	1.2.840.10008.5.1.4.1.1.66.4
Surface Segmentation	1.2.840.10008.5.1.4.1.1.66.5
Basic Text SR	1.2.840.10008.5.1.4.1.1.88.11
Enhanced SR	1.2.840.10008.5.1.4.1.1.88.22
Comprehensive SR	1.2.840.10008.5.1.4.1.1.88.33
Procedure Log	1.2.840.10008.5.1.4.1.1.88.40
Key Object Selection	1.2.840.10008.5.1.4.1.1.88.59
Chest CAD SR	1.2.840.10008.5.1.4.1.1.88.65
X-Ray Radiation Dose SR	1.2.840.10008.5.1.4.1.1.88.67
Colon CAD SR	1.2.840.10008.5.1.4.1.1.88.69
Implantation Plan SR Document	1.2.840.10008.5.1.4.1.1.88.70
Encapsulated PDF	1.2.840.10008.5.1.4.1.1.104.1
Encapsulated CDA	1.2.840.10008.5.1.4.1.1.104.2

While some of these objects could be removed from origin, and therefore will not be part of VISCERAL data sets, others could be removed for specific purposes, e.g., create different data sets for manual annotation and benchmarking.

- Modifications to the DICOM header will be performed in order to remove personal information and to ensure that the data set fits the intended use. These modifications could include:
    - Removal of personal information or replacement of this information by dummy values, depending on the DICOM Value Representation rules for each. Personal information should be understood here as information that can lead to identify or make identifiable a person and will therefore include both identifiers and quasi identifiers.

- Standardization of text values that may be needed according to the intended use of the data set (e.g. 'Study Description').
- Removal of information that could interfere with the benchmarking process (e.g., 'Scan Options', 'Protocol Name').
- Removal of proprietary tags.

All these transformations performed over DICOM objects must respect DICOM Information Object Definitions or, in general, any specification affecting construction, transmission, retrieval, display and post processing of DICOM instances.

Quality Assurance processes will have to be designed and completed prior to data disclosure in order to ensure conformity of those derived DICOM objects to the standard and fulfilment of the data protection laws.

DICOM objects will have to be disclosed either uncompressed or using lossless compression methods.

For both data formats the voxel dimension has to be provided. That is, the size of a voxel along x-, y-, and z-axis has to be given.

All imaging data has to contain (either natively or in a corresponding text file, in any case unified across the entire data set) information regarding voxel dimensions.

3 Format of the annotations

There are two categories of annotations that are either provided by the annotators, or by the algorithms. Regions are labels that are assigned to a set of voxels (e.g., all voxels in the liver). Landmarks are coordinates of specific points in the anatomy (e.g., distal end of the radius). Both annotations can carry additional information regarding the *uncertainty*. For segmentations this is typically a continuous value between 0 and 1, where 0 means guaranteed absence, and 1 guaranteed presence of a label. For landmarks uncertainty can be expressed by an estimated covariance matrix of the position estimates.

3.1 Regions and Segmentations

Regions will be annotated as voxel labels in a volume corresponding to an actual medical imaging volume (e.g., CT volume). The format for the volume annotations is NIFTI, together with a csv file that holds the assignment between organ or structure annotated, and label in the NIFTI volume. This is consistent with well-established conventions such as those implemented in 3Dslicer[1].

If uncertainty is defined for labels, labels are stored in separate volumes, and uncertainty for volume annotations is expressed numerically by assigning each voxel in the respective annotation volume a value in the interval [0,1]. There, 0 is assigned to voxels for which the structure (e.g., liver) is not present. 1 is assigned to voxels where the structure is present, and for all other voxels the value corresponds to the probability estimated of the presence of the particular structure.

The naming conventions for annotation files are described in Sec.4.

Name of the structure (consistent with RadLex)	Label

[1] http://www.slicer.org/

Liver	32

Table 1: Label information for volume labels.

3.2 Landmarks

Landmarks will be stored as csv files with columns holding the ID that identifies a specific landmark together with the coordinates in the coordinate system of the specific volume. The coordinate system will be unified across all volumes, with the origin in [0,0,0] voxel space. Units of the landmark coordinates should be mm, and the voxel dimensions of each volume have to be provided in mm (see Sec.2). If defined the uncertainty of the landmarks is given by the coefficients of the covariance matrix estimate.

ID of the landmark	x-coordinate	y-coordinate	z-coordinate	Uncertainty
321	134,5	423,5	262,5	[c11,c12,c13,c21,c22,c23,c31,c32,c33]

Table 2: Landmark information.

3.3 Identifying annotations across images and subjects

The main tokens of identification across the data are

1. ID of a landmark
2. ID of a region
3. ID of a subject
4. ID of a particular data (e.g., a specific MRI volume, if there are many volumes acquired for a single subject)

These identities are reflected in the filenames, and in the csv files, that hold information regarding region labels, and landmark IDs.

4 Filename conventions

Filenames of the data will be chosen to convey critical information. However, for all data plain text files will be provided that explain naming conventions, data characteristics, and relevant correspondences across the files. In case of additional Meta information, there is a parallel txt file to each nii file, holding this information.

Filenames for imaging data will be formatted

```
subjectXX_acquisitionZZ[_modalityYY].nii
```

For example

```
subject21_1[_modalityCT].nii
```

Filenames of region annotations

```
subjectXX_acquisitionZZ[_modalityYY] _regionannotation_annotatorXX.nii
```

For example

```
subject21_1_modalityCT_regionannotation_3.nii
```

If uncertainties are assigned to region annotations the annotations of individual structures are stored in individual files

```
subjectXX_acquisitionZZ[_modalityYY]_regionannotation_labelXX_annotatorXX.nii
```

For example

```
subject21_1[_modalityCT]_regionannotation_label32_3.nii
```

Filenames for landmark annotations

```
subjectXX_acquisitionZZ[_modalityYY]_lmannotation_annotatorXX.nii
```

For example

```
subject21_1¢_modalityCT]_lmannotation_3.nii
```

5 Conclusion

This deliverable describes the format conventions for image data, and annotations (both originating from annotators, and algorithms) to be used in the VISCERAL project. The formatting is essential, since it impacts the distribution of data, the processing pipelines for preparing, analysing, and annotating data, and lastly the information provided to the participants. For this reason the format conventions are fixed at the beginning of the project.

The nature of the project makes it possible that additional aspects of the data, and annotations have to be added during the project lifetime, if their relevance becomes apparent. In this case the conventions should be appended, and the overall conventions should be only changed in a way, that allows for backward compatibility. That is, any processing pipeline built for the current convention has to be operable for any future versions of the conventions during the project.

6 References

[1] Lenzerini, M. (2002): Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, New York, 233–246.

[2] Sheth A 1998, Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, in Interoperating Geographic Information Systems, M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds) Kluwer Publishers.