

## 4.1 Final publishable summary report

### 4.1.1 Executive Summary

VISCERAL defined and executed a targeted benchmark framework to speed up progress towards the objective of automated anatomy identification and pathology identification in 3D (MRI, CT) radiology images. The following objectives were met in carrying out the project:

1. Create an evaluation infrastructure and software based partly on existing tools to allow the benchmarks to be carried out efficiently and effectively, but that also allows continuous evaluation to take place beyond the benchmarks;
2. Innovate through the use of a cloud infrastructure for the evaluation of algorithms on huge amounts of data;
3. Run two benchmarks and two workshops at which competition results will be discussed;
4. Fuse a large number of automated entries to create a very large silver corpus;
5. Create a small but sufficiently large gold corpus by manual annotation of the radiology images (used to evaluate the quality of the evaluation by the silver corpus);
6. Release the radiology data and the silver and gold corpora as research collections at the end of the project for a continuing impact.

The main results of the VISCERAL project include the framework for managing benchmarks and managing manual annotations; the work on implementing efficient evaluation metrics and selecting the best evaluation metrics for a task; the organisation of three Benchmark series: Anatomy, Detection and Retrieval; and the creation of large amounts of ground truth in the form of gold and silver corpora.

VISCERAL produced scientific results of interest in the following areas, leading to 17 scientific publications in these areas:

- Medical imaging: the benchmarks and their results are of interest to research groups working in the medical imaging area;
- Evaluation infrastructure: the cloud-based evaluation infrastructure is of interest to research groups working, amongst others, in the areas of Big Data analysis, eScience and Information Retrieval;
- Radiology: the outcomes of the Benchmarks and their potential impact on radiology workflows.

The results of the VISCERAL project have the potential to lead to impact in a number of areas. The launch of the Evaluation-as-a-Service initiative (<http://eaas.cc>) in the VISCERAL project will lead to a continuation of work on the ideas of cloud-based evaluation and privacy-preserving evaluation, making impact in the area of open innovation in data science. The continuing availability of high quality, professionally annotated 3D radiology image data will overcome many of the challenges of getting access to such data currently faced in medical image analysis research.

Two of the open source software packages released have the potential to impact the way in which medical image analysis research is done. Wide adoption of the very efficient metric calculation software would lead to fewer uncertainties in publications about which definitions and implementations of segmentation

comparison metrics have been used, leading to higher reproducibility of results. The Manual Annotation Ticketing System allows manual annotation and quality control to be very efficiently coordinated, leading to a decrease in the time needed to do manual annotation.

#### 4.1.2 Project Context and Objectives

Diagnostic decision-making (using images and other clinical data) is still very much an art for many physicians in their practices today due to a limited availability of quantitative tools and measurements. Traditionally, decision making has involved using evidence provided by the patient's data coupled with a physician's a priori knowledge and experience of a limited number of similar cases. With advances in electronic patient record systems, a large number of pre-diagnosed patient data sets are now becoming available. These data sets are often multimodal consisting of images (x-ray, CT, MRI), videos and other time series, and textual data (free text diagnostic reports and structured clinical data). Analysing these multimodal sources for disease-specific information across patients can reveal important similarities between patients and hence their underlying diseases and potential treatments. Researchers are now beginning to use techniques of content-based retrieval to search for disease-specific information in visual data to find supporting evidence for a disease or to automatically learn associations of symptoms and visual abnormalities with diseases.<sup>2</sup>

With ever more sophisticated imaging techniques, the number of images acquired per day is exploding. There are however many challenges in processing 3D (MRI, CT) and 4D (MRI with a time component) radiology images that have currently not been solved. Particular challenges remain in the areas of anatomy identification that can be a first step to all further analysis, and pathology identification that requires very localized data analysis as regions of interest are most often very small. Semi-supervised learning based on radiology reports is also of interest as fully manual annotation is extremely expensive and does not scale well if the goal is to deal with realistically sized data collections. Most current research projects work on small collections and often, a large part of the budget of research projects is spent on creating local data collections to validate algorithms and afterwards these collections cannot be shared with other researchers.

Creating local small collections means the same effort in terms of ethics approval and similar paperwork, so creating much larger collections that can be shared is more efficient. The same is true for annotation, as manual annotation is expensive and creating this for a larger number of researchers is also much more efficient.

VISCERAL defined and executed a targeted benchmark framework to speed up progress towards the objective of automated anatomy identification and pathology identification in 3D (MRI, CT) radiology images. The following objectives were met in carrying out the project:

1. Create an evaluation infrastructure and software based partly on existing tools to allow the benchmarks to be carried out efficiently and effectively, but that also allows continuous evaluation to take place beyond the benchmarks;
2. Innovate through the use of a cloud infrastructure for the evaluation of algorithms on huge amounts of data;

---

<sup>2</sup> Henning Müller, Jayashree Kalpathy-Cramer, Barbara Caputo, Tanveer Syeda-Mahmood, Fei Wang, Overview of the First Workshop on Medical Content-Based Retrieval for Clinical Decision Support at MICCAI 2009, Medical Content-Based Retrieval for Clinical Decision Support, Volume 5853 of the series Lecture Notes in Computer Science, pp 1–17

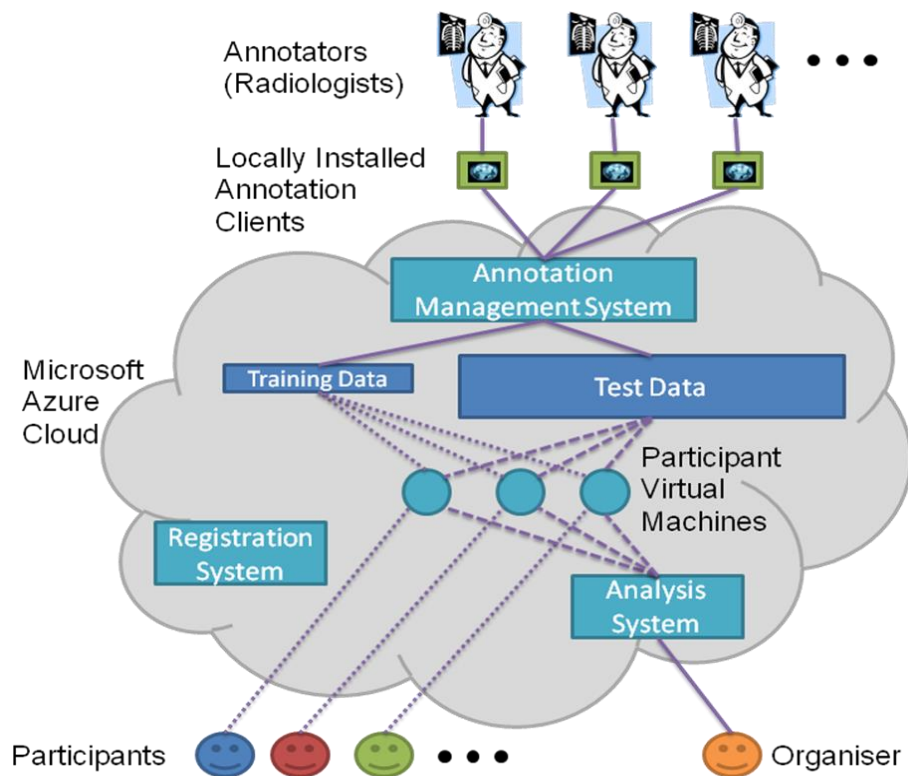
3. Run two benchmarks and two workshops at which competition results will be discussed;
4. Fuse a large number of automated entries to create a very large silver corpus;
5. Create a small but sufficiently large gold corpus by manual annotation of the radiology images (used to evaluate the quality of the evaluation by the silver corpus);
6. Release the radiology data and the silver and gold corpora as research collections at the end of the project for a continuing impact.

### 4.1.3 Main Results

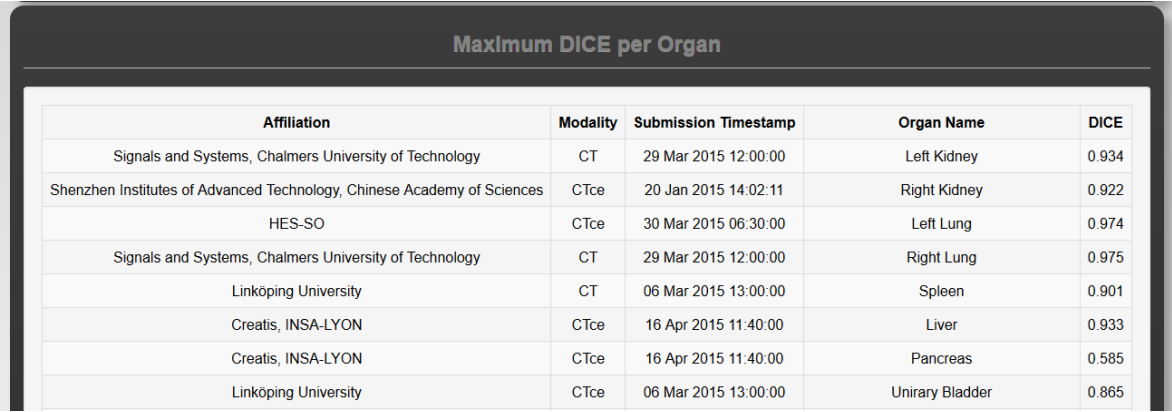
The main results of the VISCERAL project are described in this section. They include the framework for managing benchmarks and managing manual annotations; the work on implementing efficient evaluation metrics and selecting the best evaluation metrics for a task; the organisation of three Benchmark series: Anatomy, Detection and Retrieval; and the creation of large amounts of ground truth in the form of gold and silver corpora.

#### Benchmark and Annotation Framework

A benchmarking and image annotation framework was created. This framework manages the Benchmark participants and their access to Virtual Machines and Data, as well as the annotation of the images by radiologists. The framework runs on the Microsoft Azure Cloud, and is shown in the diagram below:



The training data is placed on the cloud. During the training phase of the benchmark (dotted lines above), participants register using the Registration System, and get assigned a Virtual Machine (VM) in the cloud with access to the training data. By the end of the training phase, participants must have installed software completing the provided segmentation task in their VM. During the testing phase (dashed lines above), the organisers take over the VMs and, using the Analysis System, run the installed software on the test data in order to evaluate how well the participant programs perform. In the second year of the project, the testing phase was automated to a large extent. It is now possible that when a participant submits a VM, metrics are calculated without intervention of the organisers and shown to the participant. Participants can decide which results should be published on a publicly visible Leaderboard, a screenshot of which follows:



Affiliation	Modality	Submission Timestamp	Organ Name	DICE
Signals and Systems, Chalmers University of Technology	CT	29 Mar 2015 12:00:00	Left Kidney	0.934
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences	CTce	20 Jan 2015 14:02:11	Right Kidney	0.922
HES-SO	CTce	30 Mar 2015 06:30:00	Left Lung	0.974
Signals and Systems, Chalmers University of Technology	CT	29 Mar 2015 12:00:00	Right Lung	0.975
Linköping University	CT	06 Mar 2015 13:00:00	Spleen	0.901
Creatis, INSA-LYON	CTce	16 Apr 2015 11:40:00	Liver	0.933
Creatis, INSA-LYON	CTce	16 Apr 2015 11:40:00	Pancreas	0.585
Linköping University	CTce	06 Mar 2015 13:00:00	Urinary Bladder	0.865

The Annotation Management System manages the manual annotation of the radiology images. The aim of the system is to assign the manual annotation resources, given the assumption that there is a large amount of data, and that it is possible to only annotate part of it with given annotation resources. Therefore, the system aims at using these resources optimally, by ranking the volumes and organs to identify those for which annotation would mean the maximum information gain, while at the same time performing quality checks of the annotations. The system generates annotation “tickets” for specific annotators that specify the volume and anatomical structures to be annotated, and accepts the upload of finished annotations.

## Evaluation Metrics

In order to evaluate the segmentations generated by the participants, it is necessary to compare them objectively to the manually created ground truth. There are many ways in which the similarity between two segmentations can be measured, and at least 22 metrics have each been used in more than one paper in the medical segmentation literature. We implemented these 22 metrics in the EvaluateSegmentation software, which is available as open source on GitHub,<sup>3</sup> and can read all image formats (2D and 3D) supported by the ITK Toolkit.<sup>4</sup> The software is specifically optimised to be efficient and scalable, and hence can be used to compare segmentations on full body volumes.

<sup>3</sup> <https://github.com/codalab/EvaluateSegmentation>

<sup>4</sup> <http://www.itk.org>

## Benchmarks

The VISCERAL project has organised a series of Benchmarks described below:

*Anatomy Benchmarks:* A set of medical imaging data in which organs are manually annotated is provided to the participants. The data contains segmentations of several different anatomical structures in different image modalities, e.g. CT and MRI. Participants in the Anatomy Benchmarks have the task of submitting software that automatically segments the organs for which manual segmentations are provided, and this software is tested on images which the participants have not seen. Three rounds of the Anatomy Benchmark have been organised, and the Anatomy Benchmark is continuing beyond the end of the VISCERAL project.

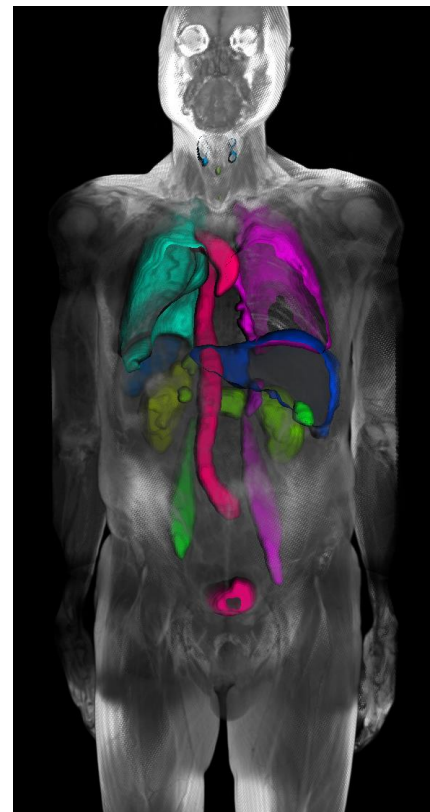
*Detection Benchmark:* A set of medical imaging data that contains various lesions manually annotated in anatomical regions such as the bones, liver, brain, lung, or lymph nodes is distributed to participants. Participants have the task of submitting software that will automatically detect these lesions. The software is tested on detecting lesions on images that the participants have not seen. The Benchmark data and ground truth continue to be available beyond the end of the VISCERAL project as the Detection2 Benchmark.

*Retrieval Benchmark:* One of the challenges of medical information retrieval is similar case retrieval in the medical domain based on multimodal data, where cases refer to data about specific patients (used in an anonymised form), such as medical records, radiology images and radiology reports or cases described in the literature or teaching files. The Retrieval Benchmark simulates the following scenario: a medical professional is assessing a query case in a clinical setting, e.g., a CT volume, and is searching for cases that are relevant in this assessment. The participants in the Benchmark have the task of developing software that finds clinically-relevant (related or useful for differential diagnosis) cases given a query case (imaging data only or imaging and text data), but not necessarily the final diagnosis. The Benchmark data and relevance assessments continue to be available beyond the end of the VISCERAL project as the Retrieval2 Benchmark.

## Gold Corpora

The VISCERAL project produced a large corpus of manually annotated radiology images, called the gold corpus. An innovative manual annotation coordination system was created, based on the idea of tickets, to ensure that the manual annotation was carried out as efficiently as possible. The gold corpus was subjected to an extensive quality control process, and is therefore small but of high quality. Annotation in VISCERAL served as the basis for all three Benchmarks. For each Benchmark, training data was distributed to the participants and testing data was kept for the evaluation.

For the Anatomy challenge series, volumes from 120 patients were manually segmented by the end of VISCERAL by radiologists, where the radiologists trace out the extent of each organ. The following organs were manually segmented, as shown in the image on the right: left/right kidney, spleen, liver, left/right lung, urinary bladder, rectus abdominis muscle, 1st lumbar vertebra, pancreas, left/right psoas major muscle, gallbladder, sternum, aorta, trachea, and left/right

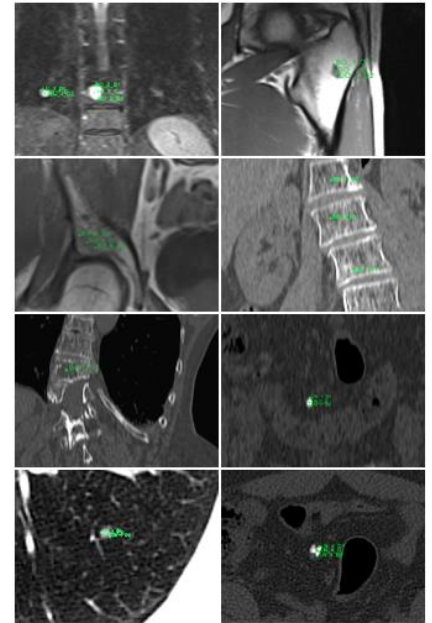


adrenal gland. The radiologists also manually marked landmarks in the volumes, where the landmarks include: lateral end of clavícula, crista iliaca, symphysis below, trochanter major, trochanter minor, tip of aortic arch, trachea bifurcation, aortic bifurcation, and crista iliaca.

For the Detection Benchmark, overall 1609 lesions were manually annotated in 100 volumes of two different modalities, in five different anatomical regions selected by radiologists: brain, lung, liver, bones, and lymph nodes. Examples of the manual annotation of lesions are shown on the right.

For the Retrieval Benchmark, we collected more than 10.000 medical image volumes, and selected about 2.000 for the challenge, and extracted terms describing pathologies and anatomical regions from the corresponding radiology reports.

The use of these images for the purpose of automated segmentation and landmark detection was cleared by the relevant ethics boards. The capability to provide the images once on the cloud instead of needing to distribute the images to the participants was a key aspect in getting the approvals of the ethics boards.



## Silver Corpus

In VISCERAL, we collected and annotated clinically relevant data, and organized benchmarks during which participants could train on large amounts of data, and evaluated algorithms on test data. In addition to the gold corpus of expert annotated imaging data described in the previous section, this offers the possibility to generate a far larger silver corpus of data, which is annotated by the collective ensemble of participant algorithms. Even though this silver corpus annotation is less accurate than expert annotations, the fusion of participant algorithm results is more accurate than individual algorithms and offers a basis for large-scale learning. It was shown by experiments that the accuracy of a silver corpus annotation obtained by label fusion of participant algorithms is higher than the accuracy of individual participant annotations. Furthermore, this accuracy can be improved by injecting multi-atlas label fusion estimates of annotations based on the gold-corpus annotated data set.

In effect, the silver corpus is large and diverse, but not of the same annotation quality as the gold corpus. The final silver corpus of VISCERAL Anatomy Benchmarks contains 264 volumes of four modalities (CT, CTce, MRT1, MRT1cefs), containing 4193 organ segmentations and 9516 landmark annotations.

### 4.1.4 Potential Impact

The launch of the Evaluation-as-a-Service initiative (<http://eaas.cc>) in the VISCERAL project will lead to a continuation of work on the ideas of cloud-based evaluation and privacy-preserving evaluation, making impact in the area of open innovation in data science.

The continuing availability of high quality, professionally annotated 3D radiology image data will overcome many of the challenges of getting access to such data currently faced in medical image analysis research.

Two of the open source software packages released have the potential to impact the way in which medical image analysis research is done. Wide adoption of the very efficient metric calculation software would lead to fewer uncertainties in publications about which definitions and implementations of segmentation comparison metrics have been used, leading to higher reproducibility of results. The Manual Annotation Ticketing System allows manual annotation and quality control to be very efficiently coordinated, leading to a decrease in the time needed to do manual annotation.

#### 4.1.5 Dissemination

VISCERAL produced scientific results of interest in the following areas, leading to 17 scientific publications in these areas:

- Medical imaging: the benchmarks and their results are of interest to research groups working in the medical imaging area;
- Evaluation infrastructure: the cloud-based evaluation infrastructure is of interest to research groups working, amongst others, in the areas of Big Data analysis, eScience and Information Retrieval;
- Radiology: the outcomes of the Benchmarks and their potential impact on radiology workflows.

A number of workshops were organised by the VISCERAL project, which are described below. Over 80 people attended the *MICCAI Workshop on Medical Computer Vision: Algorithms for Big Data* (bigMCV - <https://sites.google.com/site/miccaimcv2014/>) on the 18<sup>th</sup> of September 2014, where the results of the VISCERAL Anatomy2 Benchmark were presented in a special session. After an overview presentation of the Anatomy2 Benchmark by the organisers, five participants presented their segmentation approaches. The proceedings are published as Springer LNCS 8848. The following photo shows the workshop:



Over 20 people attended the *Multimedia Retrieval in the Medical Domain (MRMD) workshop*, held on the 29<sup>th</sup> of March in conjunction with the European Conference on Information Retrieval (ECIR) in Vienna, Austria (<http://www.visceral.eu/workshops/mrmd-2015/>). The programme consisted of two invited speakers (Camille Kurtz of the University Paris Descartes and Eldad Elnekave of Zebra Medical Vision). Papers on multimodal medical information retrieval that passed a review process were also presented, as were papers presenting the results of participants of the VISCERAL Retrieval Benchmark. The proceedings of the

workshop will be published in the next months as Springer LNCS 9059. The following photo shows the workshop in progress:



Participants of the Anatomy3 Benchmark gathered at a challenge workshop held at 2015 IEEE International Symposium on Biomedical Imaging (ISBI) on the 16th of April 2015 in New York, USA (<http://www.visceral.eu/workshops/anatomy-grand-challenge-workshop>). Ten people took part in the workshop. The proceedings of the workshop, including an overview paper, are published in the CEUR online proceedings (<http://ceur-ws.org/Vol-1390/>). The following photo shows the workshop attendees:



An interim VISCERAL Anatomy2 workshop, framed as the VISCERAL Organ Segmentation and Landmark Detection Challenge, was held at 2014 IEEE International Symposium on Biomedical Imaging on May 1<sup>st</sup>, 2014 in Beijing, China. In summary, the goal of our ISBI challenge session was three-fold: i) the submitting groups presenting their techniques, ii) reporting the results of the evaluation, iii) publicizing our benchmark series widely. With over 50 participants, the challenge session was a success; and we have received very positive feedback from participants both at the session and afterwards via email. The submitted written contributions have been published as online CEUR proceedings (<http://ceur-ws.org/Vol-1194/>). The following photo shows session attendees:





Twelve experts in Evaluation Infrastructure were invited to an *Evaluation-as-a-Service (EaaS)* workshop held in Sierre, Switzerland on the 5th to 6th of March 2015. Two days of intensive discussion led to a summary paper published in the SIGIR Forum, and plans for a more detailed white paper. The workshop attendees started the Evaluation-as-a-Service Initiative (<http://eaas.cc>), which will continue to develop and promote the EaaS approach in multiple areas. The workshop participants are shown in the following photo:



An article about VISCERAL appeared in the *research\*eu results* magazine N° 38 / December 2014 / January 2015 - <http://www.visceral.eu/assets/assets/VISCERAL-ZZAC14010ENN-002.pdf>.

## 4.1.6 Exploitation

The exploitable results created in the VISCERAL project are described in this section.

### Software

- *Evaluation Infrastructure*: Open source software for the registration and administration system for the benchmarks, as well as the automated evaluation capability. The software is described in VISCERAL Deliverable D1.4. The code is available for download from GitHub: <https://github.com/Visceral-Project/registration-system>. There is also available for download a Virtual Machine (VirtualBox) that is equipped with everything necessary to continue further the development of the registration system, which can simplify further use of the code. This can be used with any cloud system that supports VirtualBox.
- *Evaluation Metrics*: EvaluateSegmentation, the software for efficiently calculating twenty evaluation metrics comparing two segmentations, as well as landmark detection metrics, is available on Github as part of the CodaLab project: <https://github.com/codalab/EvaluateSegmentation>. It is documented extensively in Section 5 of VISCERAL Deliverable D5.5.
- *Gold Corpus Ticketing System*: The gold corpus ticketing system software used to organise efficient manual annotation by multiple annotators is available as open source software on GitHub: <https://github.com/Visceral-Project/annotationTicketingFramework>. The package includes installation guidelines, documentation on input and output parameters for each Matlab function, Matlab tutorial scripts, and workflow overview descriptions.
- *Silver Corpus Merging Framework*: The silver corpus merging framework used to merge multiple participant segmentations is available as open source software on GitHub: <https://github.com/Visceral-Project/silverCorpusFramework>. The package includes installation guidelines, documentation on input and output parameters for each Matlab function, Matlab tutorial scripts, and workflow overview descriptions.

### Documentation

All public deliverables from the VISCERAL project are available for download on the VISCERAL website (<http://www.visceral.eu/resources/deliverables/>). These include the results of the Benchmarks (D4.3, D4.4, D5.3, D5.4), the evaluation methodologies and tutorial material (D5.5). This material is currently being edited to appear in an open access book to be published by Springer.

### Data

The data created in VISCERAL consists of the anonymised radiology images and corresponding anonymised radiology reports, the manual segmentations of the gold corpus, and the silver corpus created by merging participant submissions to the Anatomy Benchmarks. This data continues to be available in the Anatomy3, Detection2 and Retrieval2 Benchmarks that are running beyond the end of the VISCERAL project.

#### **4.1.5 Project website**

<http://visceral.eu>