Horizon 2020

Project title: AuTonomous intraLuminAl Surgery

Data Management Plan (DMP)

Deliverable number: D8.2

Version 1.1.0

Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 813782

Project Acronym: ATLAS

Project Full Title: AuTonomous intraLuminAl Surgery

Call: H2020-MSCA-ITN-2018

Grant Number: 813782

Project URL: https://atlas-itn.eu

Deliverable nature:	Report (R)
Dissemination level:	Public (PU)
Contractual Delivery Date:	September 30, 2019
Actual Delivery Date	October 11, 2019
Number of pages:	12
Keywords:	DMP, Data Management Plan
Authors:	Johan Philips, KUL
Peer review:	Gianni Borghesan, KUL Diego Dall'Alba, UNIVR Elena De Momi, POLIMI Emmanuel Vander Poorten, KUL Arianna Menciassi, SSSA Benoit Rosa, UNISTRA

Abstract

This document contains the *initial* Data Management Plan for *ATLAS*. It has been drafted using the *DMPonline tool* offered by KU Leuven to all its researcher (https://dmponline.kuleuven.be) and uses the Horizon 2020 template from the Digital Curation Centre (UK). This document answers questions on data generation and collection, storage, security and preservation as well as any ethical or legal issue that need to be addressed within *ATLAS* regarding data sharing.

The Data Management Plant will be updated every time concrete changes will be implemented with respect to the data management. Changes will be made available on-line, and provide in annex to periodic reports.

The structure of this document will follow the DMP template with 17 questions that were answered in the DMPonline tool. For completeness, the 17 addressed questions are summarised in the Appendix of this deliverable.

Contents

List of Tables

List of Acronyms

CI	Continous Integration
DICOM	Digital Imaging and Communications in Medicine
DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable and Reusable
NetCDF	Network Common Data Form
JSON	JavaScript Object Notation
ODF	Open Document Format
OWL	Web Ontology Language
RO	Research Objective
XML	eXtensible Markup Language

1. Data Summary

1.1. Purpose of data collection

Data collection activities (involving humans) in this project can be roughly categorised into four categories:

1. Data regarding the patient, gathered during a medical procedure.
2. Data regarding the surgeon, gathered during a medical procedure.
3. Data regarding a subject (*e.g.* surgeon, test person or researcher), gathered on experimental lab setups.
4. Non-personal and non-sensitive data gathered on experimental lab setups.

Data regarding patients targets mainly geometry, mechanical properties (*e.g.* deformation/force ratio), and camera images of lumens. Data processing will strive to obtain models – geometrical, mechanical – or to study autonomous systems. In addition, classifiers will be trained to segment and classify tissue.

Data regarding the surgeon contains data ranging from the description of the surgical workflow to instrument tracking. This data will be used for workflow and gesture modelling.

Examples of non-personal and non-sensitive data are sensor readings from experimental lab setups of a 3rd party (patient, surgeon or subject) present (*e.g.* user-driven navigation of a catheter in a mock-up) or without information or data of a 3rd party (*e.g.* autonomous navigation of a catheter in a mock-up).

The purpose of data-gathering efforts is to collect and share these data and their analysis in a straightforward way that stimulates data exchange within the consortium. Moreover, making these data available, increases credibility of scientific publications. The shared repository will eventually help with research continuity as it permits follow-up projects that build on these data. Publication of these data will also allow independent validation of results by fellow researchers from the community.

The challenges, that this data management plan addresses, are to:

- make this data FAIR, where re-use is done in line with standards and regulations (chapter 2),
- maintain this data (chapter 3, 4), and
- ensure that data is managed in line with ethical and legal standards and regulations (chapter 5).

1.2. Relation to *ATLAS* project's objectives

The data collection activities is functional to the fulfilment of the five Research Objectives (ROs) of the *ATLAS* project; For reference, the five ROs are listed here:

RO1: stretching limits of actuation to distributed, precisely controllable, compliant mechanisms;

RO2: to distributed sensing, featuring proprioception of the own complex shape with exteroception;

RO3: real-time reconstructed models of the complex geometry and episode in the surgical workflow;

RO4: distributed control over the interaction with the lumen and cognition to act in this fragile context;

RO5: based on a structured integration approach and a lean framework that supports seamless integration, and maximal exploitation of commonalities among clinical use cases.

Regarding data collection, ROs of the *ATLAS* project can be mapped as follows:

1. Instrument design - to define specifications and desired characteristics of novel medical instruments.
2. Autonomous navigation - to characterize the geometry and investigate which kind of information can be provided to a control system that provides autonomous navigation in lumens.
3. Classification - Training and test of expert systems that use camera images and other type of information to perform: *i)* classification of tissues, and *ii)* classification of surgeon gestures.
4. Instrument interfaces - by studying the surgeon(s) behaviour at kinematic level and more abstract level (surgical workflows, for example), to understand how to design control algorithms that allow for intuitive use and can easily blend in the current procedure.

These objectives will be updated as data collection will take place. Specific data gathering campaigns will be connected to one or more of these topics. If people are included in the data sets, informed consent forms will be prepared. These forms will clearly states whose of these objectives will be pursued.

1.3. Data types and formats

The data sets collected and shared within the consortium will be: *i)* images and 3d models of lumens (colon, urethra, etc), *ii)* endoscopic images, echos, MRI, *iii)* surgical flow analysis, *e.g.* description of how a procedure is carried out, and spacial information of instruments, via optical or electromagnetic tracking. Table 1.1 gives a more complete overview of data types that will be gathered. This table will be complemented as the project progresses.

Table 1.1: Data Types.

Data Type
CBCT (Cone beam computed tomography) 2D and 3D data
Pre-operative CT or MRI volumetric dataset
External camera images of patient's parts
Optical tracking data (for external devices poses)
Endoscopic camera images of patient or phantom
Optical Coherence Tomography images
Intra-operative ultrasound data from external probe
Intravascular ultrasound (IVUS) images
Electromagnetic tracking of instrument
FBG shape sensing data
Motor data (actuation, position, *etc.*)
Low level kinematic and sensor data from robotic controller
Virtual Reality multi-modality dataset (video, trajectories, shape sensing info, ecc...) acquired from simulations

All the used formats in *ATLAS* are either open formats such as the Open Document Format (ODF), tab or comma delimited format or plain text files or either readable via community or open source software libraries. The images are either medical images (Digital Imaging and Communications in Medicine (DICOM), processed or not) or (video) camera images (*e.g.* external cameras to track the surgeons movements). For the latter, open source codecs and standards will be used, while for the former open source libraries exist the view the data. The reason of using DICOM is that it is a well established standard within medical imaging and it allows for easy integration with several medical devices and infrastructures. Also, time series data from instrument tracking or sensor info, will be collected in either pandas (https://pandas.pydata.org/), rosbag (http://wiki.ros.org/rosbag), Network Common Data Form (NetCDF) (https://www.unidata.ucar.edu/software/netcdf) or plain CSV format with descriptive headers. All these formats have freely available (open source) readers and writers in various programming languages.

Other user experiments data consist of *i)* any type of sensory data in lab experiments (master interface encoders for a robot, instrument tracking, *etc.*) are also stored in open readable formats (pandas, rosbag, NetCDF, CSV or plain text, and *ii)* questionnaires filled in by surgeons to fill in, which will be stored in ODF.

The workflow analysis would typically involve a mix of attending procedures, transcripts of discussions with surgeons, and bibliography and these will be stored in ODF as well.

For all published files, version control will be in place to keep track of author, status, description, origin, and all relevant data regarding publication activities. This information will be stored in a separate metadata file.

1.4. Data reuse

The consortium is investigating the use of the *Cholec80 Dataset* [3], that consists of 80 videos of cholecystectomy surgeries performed by 13 surgeons. The data is made available from the University of Strasbourg[1].

In addition, LapOntoSPM [2] is considered as ontology for laparoscopic surgeries. This dataset might be extended with some research outcomes of *ATLAS*. As other open datasets are considered, they will be listed in Table 1.2.

Table 1.2: Data Reuse - Table will be updated with open or acquired datasets as project progresses.

Data set	Description	URL
Cholec80	Data set consisting of 80 videos of cholecystectomy surgeries performed by 13 surgeons.	http://camma.u-strasbg.fr/datasets
LapOntoSPM	Ontology for laparoscopic surgeries. It contains a Web Ontology Language (OWL)- that includes terminologies of instruments, anatomical parts, procedures, actors, *etc.* commonly encountered in the description of a surgical workflow.	https://link.springer.com/article/10.1007%2Fs11548-015-1222-1

1.5. Data size

The shared repository that will be compiled to realise the research objectives, will be hosted at the data center of KU Leuven. Some data may also be hosted at any partner institution in the consortium. The data repository size is estimated at 10 TB, which includes all raw data (time series from sensors, camera images and video recordings) as well as processed and structured data (documents, surveys, data analysis, results).

1.6. Data utility

Curated datasets will come out of this project and will be deposited at the project website and an open access data repository such as Zenodo. It is still to be decided which data repository will be used. The main target group are the consortium partners who are part of the project's network. The consortium tries to ensure that both the robotics research community as well as the surgical research community and medical professionals will benefit from these data sets. Furthermore, these datasets will be used by consortium partners as benchmarks in future research.

[1] see http://camma.u-strasbg.fr/datasets

2. FAIR Data

2.1. Findability

During the recordings of surgical procedures, metadata will be (1) automatically attached to all medical images (confined within the DICOM format) and (2) collected via a graphical user interface for the researcher or nurse. These metadata are uploaded to a project database with references to the raw data. Aside from DICOM, few standards are available, but an effort will be made to develop a common metadata structure for all consortium partners.

Curated datasets will be published on Zenodo which makes them openly accessible and discoverable. In addition, data will also be made available via the *ATLAS* website accompanied with the respective metadata. The data will be indexed using the EU Open Data Portal (http://data.europa.eu). All published data sets will receive a DOI that will be referred to in any scientific publication that made use of this data set.

2.2. Accessibility

Curated data sets will be uploaded in an open format (*e.g.* **CSV!** (**CSV!**), NetCDF, pandas) to a domain specific repository or Zenodo, under the widely-used permissive MIT license (https://opensource.org/licenses/MIT). This license has limited restriction on reuse, as it allows both non-commercial and commercial use. The metadata will be provided in a readable format (json, eXtensible Markup Language (XML) or **CSV!**). The software developed during the project will be hosted on the GitLab server owned by the department of Mechanical Engineering (KU Leuven). The consortium has the possibility to use automated GitLab Continous Integration (CI) pipelines to release parts of the development to the community. The CI pipelines ensure that the source code compiles and builds and that tests are ran before releasing it to a wider audience. Links to the software will be published on the project website and a mirror of each release might be made available via GitHub. Documentation and a user guide with examples will be published as an online tutorial via the website and will accompany any release.

As specified to the IPR policy of the consortium (defined in the CA), the collected data will be shared with the scientific community, when it does not concern sensitive information. Regarding clinical data, for confidentiality reasons, data will be anonymised before being shared even if such sharing takes place only with the consortium.

2.3. Interoperability

Since the project will collect many medical images, we will use that domain specific standard (DICOM) as a starting point for those data. Other data, such as time series, will be stored in open formats, following a domain specific standard, if any exists. If we see the need to use wider standards such as Dublin Core, we will provide proper mappings during the project. To ensure machine readable or actionable data sets, **CSV!**, JavaScript Object Notation (JSON) or XML will be used to format the meta data.

2.4. Re-usability

During the project we will curate and publish datasets and software that we assess as relevant and ready for release. This ensures fast visibility and re-use within the research community. All data and software will be licensed with MIT licenses as this permits both research and industry to re-use our data or software. The GitLab CI pipelines that will be developed will allow us to set up a data quality check inline. This helps our researchers to validate datasets, ensures code correctness and increases re-usability.

The published datasets will also be referenced in scientific publications, using their DOIs, and at conference presentations, to maximise re-use.

3. Cost & Allocation of Resources

3.1. Estimated cost

The shared repository that will be compiled to realise the research objectives, will be hosted at the data center of KU Leuven or any partner institution in the consortium. In view of the expected size of the database (10 TB), the estimated cost will be 1000 EUR to set up the shared repository and an annual fee of 200 EUR for storage maintenance and support per TB. Adequate budget is allocated to be in line with the local data preservation regulations at KU Leuven. Latter stipulates that data needs to be preserved at least 5 years after the end of the project, unless there are legal or ethical obligations to remove it earlier.

3.2. Data manager

Within this project, Johan Philips, PhD, research expert in reproducible science from KU Leuven, is responsible for implementing the DMP, reviewing the data capturing and metadata production process. He will safeguard the data quality, setup required infrastructure for storage and backup needs within *ATLAS* and facilitate data archiving and data sharing. He will also update the DMP any time conditions change and will make sure the DMP is effectively applied during the project.

4. Data Security

4.1. Data storage

Active data generated or collected during the project will be stored at the data center of KU Leuven or any of the consortium partners. Datasets collected in clinical studies will be anonymised before use in the project.

In line with ethical guidelines, where applicable, curated datasets will be published at the project website, on Zenodo or in domain-specific data repositories.

4.2. Data back-up

Within the consortium, datasets will be shared via private data repositories, that run at one or multiple beneficiaries (*e.g.* at the KU Leuven data center, which has automated backup and mirroring). The research data manager of this project, Johan Philips, will inform the consortium on how to safely upload datasets to such a shared data repository to ensure that all of the collected datasets are correctly stored and backed up. The research data manager will act as Data Protection Officer for the project, and fulfil the requirements stated in Deliverable D9.2.

4.3. Data preservation

As long as the informed consents for datasets originating from the clinical studies allow this, all data will be preserved on the university's central servers, (which have automatic back-up procedures) and this for a period of 5 years after the project ends.

Datasets originating from experimental measurements outside clinical studies (*e.g.* lab experiments), will be stored at one or multiple beneficiaries (*e.g.* KU Leuven's data center) as well.

Storage tiering will be set up to lower the maintenance costs. This means that datasets will move from faster, instant access (but more expensive) storage to slower, archive (cheaper) storage as the project progresses.

4.4. Data access

All datasets that will be used in this project will be anonymised. In the case this polocy would change during the course of the project, a secure environment for sensitive datasets will be set up, *e.g.* by using KU Leuven's data vault. Proper pseudonymization and anonymisation techniques will be used to secure any personal data.

5. Ethical & Legal Aspects

We will collect and reuse datasets from clinical studies done both at KU Leuven and other consortium partners. For all of these datasets the ethics approvals will have been received and for datasets that will be reused. Clearance for secondary use and proper consent forms will be checked. Based on this information, data preservation and sharing requirements as well as data security and safe storage are determined. All datasets will be anonymised before they are used in the project.

Details on Ethics, (ethical approvals, informed consents, personal and sensitive data gathering and management, *etc.*) are described in deliverables D9.1 to D9.5, due March 2020. Some of these deliverables describes actions that directly impact the data management activities, as reported in Table 5.1.

Table 5.1: WP9 deliverables related to the DMP

Del. N.	Covered Topics
D9.1	• Information on how observation of surgical procedures is carried on, and how to avoid that data gathering influences the medical procedure. • Information on how informed consents module are draft, and
D9.2	• Detailed information on what kind of data will be collected and how this relates to further processing of personal data. • The host institution must confirm that it has appointed a Data Protection Officer (DPO). The contact details of the DPO are made available to all subjects involved in the research. For host institutions not required to appoint a DPO under the GDPR, a detailed data protection policy for the project must be kept on file. • Description of the anonymisation/pseudonymisation techniques that will implemented • Detailed information on the procedures for data collection, storage, protection, retention, and destruction, and confirmation that they comply with national and EU legislation. • Detailed information on the informed consent procedures in regard to the collection, storage, and protection of personal data. • In case of further processing of previously collected personal data, relevant authorisations (secondary use) will be obtained.

6. KU Leuven context

KU Leuven's data policy demands retention of data for a period of at least 5 years after the end of a project, PhD or publication (barring 3rd party agreements). It may also be useful to point out that KU Leuven supports its researchers in the area of RDM (research data management) by the provision of a customized and free Data Management Plan (including guidance to KU Leuven legal requirements and policy guidelines) based on the templates provided by the Digital Curation Centre (UK), and a dedicated RDM support desk advising on data storage, metadata and preparing data for sharing. The university also continues to invest in affordable long-term storage and curation.

This deliverable has been drafted using the DMPonline tool (https://dmponline.kuleuven.be) that KU Leuven offers to its researchers, and the Horizon 2020 template provided by the Digital Curation Centre (UK).

A. Horizon 2020 template questions

Table A.1: Resume of the "*H2020 templates: Data management plan*", [1]

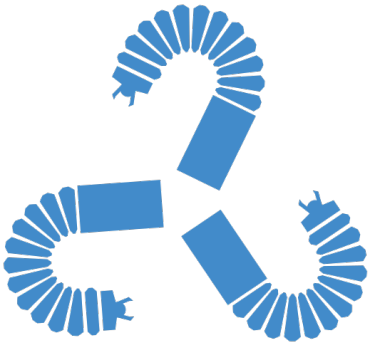
Sections	Questions
Data summary	**1** State the purpose of the data collection/generation and their origin. Relate this to the objectives of the project. **2** Specify the types and formats of data generated/collected **3** Specify if existing data is being re-used (if any) **4** State the expected size of the data (if known) **5** Outline the data utility: to whom will it be useful
FAIR data	**6** How will you make sure your data are FINDABLE by others? This question is twofold and deals with context (what metadata will be provided, according to what metadata standard) AND technical background (will you make use of persistent and unique identifiers *e.g.* DOI)? **7** How will you make your data openly ACCESSIBLE? If certain datasets cannot be shared explain why, clearly separating legal & contractual reasons from voluntary restrictions. **8** Assess the INTEROPERABILITY of your data: in how far is data exchange and reuse possible? **9** What steps will you take to allow maximal RE-USE of the data?
Costs & Allocation of resources	**10** Estimate the costs for making your data FAIR and describe how you intend to cover these costs. **11** Clearly identify responsibilities for data management in your project.
Data security	**12** Where will the data be stored? **13** Which back-up procedures are in place? **14** Where will the data be preserved for the longer term? (cf. section on data accessibility) **15** Describe the data security procedures and who has access to the data.
Ethical & legal aspects	**16** Are there any ethical or legal issues that can have an impact on data sharing?
KU Leuven context	**17** Does your institution have a RDM policy and does it offer RDM support?

Bibliography

[1] *H2020 templates: Data management plan v2.0 – 15.02.2018*. 2018. URL: https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf.

[2] Darko Katić et al. "LapOntoSPM: an ontology for laparoscopic surgeries and its application to surgical phase recognition". In: *International Journal of Computer Assisted Radiology and Surgery* 10.9 (Sept. 2015), pp. 1427–1434. ISSN: 1861-6429. DOI: 10.1007/s11548-015-1222-1. URL: https://doi.org/10.1007/s11548-015-1222-1.

[3] Andru P. Twinanda et al. *EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos*. 2016. arXiv: 1602.03012 [cs.CV].



The ATLAS project

lastly modified: October 11, 2019

ATLAS-D8.2-1.1.0

Horizon 2020