D6.1

DELIVERABLE 6.1

Title: "Data Management Plan"

PROJECT INFO	
Call/Topic	ICT-36-2020 Disruptive photonics technologies
Project title	COHERENT RAMAN IMAGING FOR THE MOLECULAR STUDY OF THE ORIGIN OF DISEASES
Project acronym	CRIMSON
Grant Agreement No.	101016923
Project website	http://crimson-project.eu/

DELIVERABLE INFO	
Deliverable Number	D6.1
Deliverable title	Data Management Plan
Nature	ORDP: Open Research Data Pilot
Work Package	WP6
Lead Beneficiary	3RDPLACE
Main Contributing Partner	3RDPLACE
Reviewers	POLIMI, FPM, INT, JUH
Dissemination level	PU - Public
Contractual delivery date	31 May 2021 (M6)
Actual delivery date	28 May 2021 (M6)
Version	1.0

History of changes

Version	Date	Comments	Main Authors
0.1	19/05/2021	First draft	Matteo Bregonzio (3rd Place), Leone De Marco (3rd Place), Manuela Bazzarelli (3rd Place), Dario Polli (POLIMI), Thomas Bocklitz (IPHT), Hervé Rigneault (IF-CNRS), Fabrizio Amarilli (FPM)
0.2	28/05/2021	Final draft	Dario Polli (POLIMI), Fabrizio Amarilli (FPM), Italia Bongarzone (INT), Michael Schmidt (IPHT)

Disclaimer

This document contains confidential information in the form of the CRIMSON project findings, work and products and its use is strictly regulated by the CRIMSON Consortium Agreement and by Contract no. 101016923.

Neither the CRIMSON Consortium nor any of its officers, employees or agents shall be responsible, liable in negligence, or otherwise howsoever in respect of any inaccuracy or omission herein.

The contents of this document are the sole responsibility of the CRIMSON consortium and can in no way be taken to reflect the views of the European Commission and the REA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016923.

PHOTONICS 21

PHOTONICS PUBLIC PRIVATE PARTNERSHIP

This project is an initiative of the Photonics Public Private Partnership. For further info, see: www.photonics21.org

Table of contents

EXECUTIVE SUMMARY

In this document, we present a first version of the Data Management Plan (DMP). In section 1, we specify that the core datasets produced within CRIMSON are vibrational spectra and hyperspectral images from cells and tissues, measured via spontaneous or coherent Raman spectroscopy (namely CARS and SRS). They will be accompanied by traditional biological methods such as fluorescence imaging, MALDI and histology. We will share in open access all data deemed useful for the scientific community, including data presented in scientific documents, unless the task leader can justify why data cannot be made openly accessible.

Following the "FAIR" principle (Findability, Accessibility, Interoperability, and Reuse of digital assets), we will make data findable and accessible through the multi-disciplinary open repository "ZENODO" maintained by CERN that created DOIs for each dataset. We will also link datasets through our project website and to the related scientific publications. Interoperability will be guaranteed using standard file formats such as ASCII text files. Access to data will be guaranteed using the "CC BY 4.0" licence, unless a more restrictive licence (e.g. the CC BY-NC 4.0) will be required in specific cases for commercial reasons.

Finally, in sections 3, 4 and 5 we discuss the allocation of resources and the issues related with data security, privacy and ethical aspects.

This document represents a work in progress and it will be updated over time based on project development. An updated version will be always available on Crimson Website at http://crimson-project.eu/.

LIST OF ABBREVIATIONS

The following table presents the acronyms used in the deliverable in alphabetical order.

Abbreviation	Description
DMP	Data Management Plan
FAIR	Findability, Accessibility, Interoperability, and Reuse of digital asset
DOI	Digital Object Identifier
WP	Work Package
EU	European Union
ISO	International Organization for Standardization
ECHR	European Convention on Human Rights
AI	Artificial Intelligence
CRS	Coherent Raman Spectroscopy
CARS	Coherent Anti-Stokes Raman Scattering
SRS	Stimulated Raman Scattering
SR	Spontaneous Raman

Data Management Plan

1 Data Summary

The data collection/generation process is of paramount importance for the success of the CRIMSON project. Briefly, the project is about implementing a new generation of bio-photonic imaging system coupled with state-of-the-art Artificial Intelligence (AI) powered software. Since the cornerstone of the project is a novel data-generating device (optical microscopes), data management is fundamental for driving and evaluating the progress in its implementation.

1.1 Which datasets?

The core dataset of the CRIMSON project will consist of:

(i) **Vibrational spectra:** signal-versus-frequency datasets, described by a two-column matrix, generally measured using spectrometers, often depicted using a simple two-dimensional plot.

(ii) **Hyperspectral images**: three-dimensional datasets, in which two dimensions represent the x-y coordinates of the image of a sample and the third dimension contains the aforementioned vibrational spectrum at each pixel (voxel) of the image, which can be used to identify the molecular components present at each position of the illuminated area.

These two datasets will be generated either using spontaneous Raman (SR) or Coherent Raman Spectroscopy (CRS) setups. In particular, CRS can be subdivided in Coherent Anti-Stokes Raman Scattering (CARS) and Stimulated Raman Scattering (SRS) techniques. Hyperspectral images are generated coupling a microscope (or an endoscope) to the spectrometer, so that a vibrational spectrum is measured at each sample position by raster scanning the sample or the laser with respect to each other. Spectroscopic maps will be invaluable to the bio-photonic imaging community while all relevant data generated from the wide variety of cancer samples employed will provide a wealth of information useful to advance cancer research.

Additionally, during the initial stages of the project, synthetic (ie. simulated) datasets will be produced in order to provide a head start to the development of the denoising AI powered module. Furthermore, in order to develop the classification chemometric AI module, it will be necessary to create reference datasets using several traditional biological methods such as:

- Western immunoblots
- Flow cytometry

- Molecular profiling
- Histology
- Immunohistochemistry
- Fluorescence microscopy
- MALDI imaging

All data will be newly generated specifically for the CRIMSON projects. In the initial stages, it will be possible that existing experimental data, already in possession of the involved institutions, will be used as needed.

In the initial phase of the project, data will be generated using the aforementioned techniques on commercial cultivated cell lines. For each cell line, data will be generated elucidating technical effects and biological differences in cell states. In a more advanced stage of the project, results will be validated in three case studies to investigate cellular mechanisms implicated in three diseases:

- Autophagy in liver cancer (case study 1)
- Immuno-oncology in head and neck cancer (case study 2)
- Senescence in thyroid cancer (case study 3)

For case study 1, the methods will be validated *in vitro* on liver 2D cell lines and/or 3D organoids in addition to non-alcoholic steatohepatitis (NASH) murine models to validate the methods *in vivo*. For practical, ethical and experimental reasons, only *ex vivo* samples (i.e. freshly explanted hepatocytes and hepatic slices) will be used in this project. For case study 2, the methods will be validated on 2D head and neck cancer cell cultures and co-cultures in addition to *ex vivo* head and neck cancer thin and thick slices. For case study 3, the methods will be validated *in vitro* on thyroid cancer cells in 2D, 3D cultures and co-cultures with other type of cells.

1.2 Expected data size

The expected data size will vary depending on the kind of data. As an example:

Vibrational spectra	~1-10 KiloBytes
Hyperspectral images	~1-10 MegaBytes (small images), up to ~250 MegaBytes (big images)
Fluorescence microscopy	depending on the image lateral resolution and image size, from a few MegaBytes up to a maximum of hundreds of MegaBytes

Histological images (H&E)	depending on the image lateral resolution and image size, from a few MegaBytes up to a maximum of 1 GigaByte
MALDI Imaging	from few MegaBytes to few GigaBytes

1.3 Which data to be shared?

We will share in open access all data deemed useful for the scientific community (see section 2.4), including data presented in scientific documents (reports, papers, conferences), unless the task leader can justify why data cannot be made openly accessible (e.g. for IP protection, protection of personal data, ethical or security issues, commercial exploitation purposes). If necessary, appropriate Intellectual Property Rights procedure (such as non-disclosure agreement) will be used.

2 FAIR data

2.1 Making data findable, including provisions for metadata

All relevant data, as described in section 1.3, will be made readily available on open-source platforms. As discussed in section 2.3, to enable the replication of experiments and to describe the shared data files, metadata files will be generated for each dataset and will include what, where, how, and when data was collected. Data will be findable on the specified platforms (see section 2.2) where keywords will be supplied in the manners allowed by the specific platforms to increase the findability of the datasets. Furthermore, a dedicated section within the CRIMSON webpage (www.crimson-project.eu) will provide a summary of each dataset and the relevant link for downloading the material.

All datasets will be organized according to naming conventions that will be specified in the next version of the DMP. Subsequent versions of the generated datasets will be easily identified by clear version numbers.

2.2 Making data openly accessible

To maximise the impact of CRIMSON research data, the results will be shared within and beyond the consortium. All the data and results will be shared with the scientific community and with other stakeholders through publications in scientific journals and presentations at conferences and events, as well as through open access data repositories.

In addition to local storage (the CRIMSON project datasets are first stored and organized in a database by the data owners on personal computers or on the institutional secure servers), public metadata and datasets will be made available

(data access policy unrestricted) to users through ZENODO (https://zenodo.org), the multi-disciplinary open repository maintained by CERN. Zenodo assigns a DOI (Digital Object Identifier) to all documents uploaded and allows the storage of metadata. Individual partner's institutional online repositories may also be used in addition to ZENODO to host and preserve data until the end of the project.

Relevant CRIMSON's metadata and dataset will be uploaded by the involved researchers to the ZENODO platform, compiling project-related information. This will enable automatic data extraction from the OpenAIRE platform, thus ensuring accessibility through a standard platform for Open Data access.

2.3 Making data interoperable

All data will be made available in "ASCII" format or similar, such as ".m" MATLAB or CSV. In this way, most files will be accessible with general use software such as Microsoft Word, Excel, Powerpoint and their open source alternatives such as the Libre Office suite, to allow as much as possible data exchange between researchers and institutions and to guarantee usage for years to come. As required, reference will be made to any software required to run it. Given the scope of this project it is anticipated that publicly available software will be used to store data. Barriers to access through interoperability issues are not anticipated.

The metadata format will follow the convention of the hosting research data repository. Metadata will be added as part of data collection process. Metadata files will be generated for each imaging dataset and will include what, where, how, and when data was collected. The main intention is to make metadata that allows the replication of experiments. Appropriate metadata file templates will be generated.

Partners will also ensure that CRIMSON data observes FAIR data principles under H2020 open-access policy.

As the project progresses and data is identified and collected, further information on making data interoperable will be outlined in subsequent versions of the DMP. In specific, information on data and metadata vocabularies, standards or methodology to follow to facilitate interoperability and whether the project uses standard vocabulary for all data types present to allow interdisciplinary interoperability.

Each partner involved in collection, elaboration, and production of scientific data will be responsible for the production of their own data and for the upload on the shared Zenodo repository. 3rdPlace, which is responsible for the DMP, will guarantee that data are shared, accessible, and reusable ensuring also the increase of data re-use through open or non-proprietary formats and licences.

2.4 Increase data re-use (through clarifying licences)

In order to ensure the widest reuse possible we will publish all public datasets under the CC BY 4.0 license where possible. For commercial reasons, the CRIMSON partners reserve the right to publish some datasets under the CC BY-NC 4.0 license on repositories hosted on proprietary servers. During the collection and production of data, the responsible partners will evaluate the need for a more restrictive sharing licence.

Additionally, there will be an effort to make the data publicly available following the publication of the related papers in a timely manner, according to catalogation, embargo, formatting and publishing times.

All data will be freely usable by third parties in accordance with the sharing platform terms and conditions. Specifically, ZENODO states that the retention period of the data corresponds to the lifetime of the repository that is, currently, the lifetime of the host laboratory CERN, which at the moment has an experimental programme defined for the next 20 years at least.

3 Allocation of resources

3rdPlace will be responsible for CRIMSON data management.

The costs for making data FAIR includes:

- Project website operation
- Data archiving at ZENODO: free of charge

Costs related to open access to research data in Horizon 2020 are eligible for reimbursement under the conditions defined in the H2020 Grant Agreement, in particular Article 6 and Article 6.2.D.3, but also other articles relevant for the cost category chosen. Costs will not be claimed by the CRIMSON partners retrospectively. The costs related to the description of data through metadata and to the publication are part of the project activities. Project beneficiaries will be responsible for applying for reimbursement for costs related to making data accessible to others beyond the consortium.

4 Data security

Every partner is responsible to ensure that the data are stored safely and securely and in full compliance with European Union data protection laws. All CRIMSON researchers commit to the highest standards of data security and protection in

order to preserve the personal rights and interests of study participants. They will adhere to the provisions set out in the:

- General data protection regulation (GDPR)
- Directive 2006/24/EC of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communication services or of public communications networks
- Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)

The following guidelines will be followed to ensure the security of the data:

- Store data in at least two separate locations to avoid loss of data.
- Encrypt data if it is deemed necessary by the participating researchers.
- Label files in a systematically structured way, in order to ensure the coherence of the final dataset.

All project deliverables will be stored and shared also in a folder (other cloud storage folders will be notified if necessary) restricted to the project consortium. As an initial step, only the consortium partners will have access to the repository where dataset and metadata are filed. Following scientific publications and articles, the datasets will be shared on ZENODO whose servers are managed according to the CERN Security Baseline for Servers.

5 Ethical aspects

CRIMSON project complies with the highest ethical principles, including the highest standards of research integrity and applicable international, EU, and national law, satisfying by design all relevant compliance requirements for each specific activity. The collected data will never include information raising ethical issues to ensure the optimal balance between the objectives of the research and the means by which the project partners go about achieving them. It will ensure respect for people and for human dignity in accordance with the European Convention on Human Rights (ECHR), fair distribution of burden and research benefits, while at the same time it will protect the value, rights and interests of all research stakeholders. The CRIMSON research will be conducted in accordance with the Declaration of Helsinki (2013) requirements and guidance provided in ISO 14155 (2012) and applicable local government regulations and Independent Ethics Committee policies and procedures. Additional cases that may arise will be

regulated, during the project lifetime, according to the Grant Agreement and the Consortium Agreement. Data will be formalized in structured databases for the purposes of elaborations to be carried out in the project.

Details on Ethics (ethical approvals, informed consents, personal and sensitive data gathering and management, etc.) are described in deliverables D4.1 and D7.4.

Deliverable	**Covered topics**
D4.1	Description of the experimental protocols for ethical committee approval
D7.4	POPD (Personal data protection) - Requirement No. 4

In particular, in Deliverable 7.4 we stated that human participants (only adults) might be involved in the research in the study case 2 of work package 4 (WP4) related to the study of the interaction between head and neck cancer and immune cells that occurs when the immune system attempts to target and attack cancer cells (see task T4.5). We will only use residual tumor material, taken out during the clinical routine surgery of the patients. Residual material is remaining material not used for diagnostic purposes. The participation at the study does not have any additional risks for the participating patients. The analysis of the residual material does not have any influence on the treatment of the patients. In particular, data will be **fully anonymized**. CRIMSON will not use or transfer any personal data. As the principles of data protection in the GDPR apply to information concerning an identified or identifiable natural person, the GDPR does not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. Therefore, anonymized data generated in CRIMSON is **excluded from GDPR regulation**.

To secure the **confidentiality, accuracy, and security of data** and data management in task T4.5, the following measures will be taken:

- Data obtained within the studies will be transmitted (as applicable) to partners within the consortium only after anonymization. Each study patient gets a sequential number (CRIMSON1, CRIMSON2, …). If the residual tumor material is split in several parts, sub-numbers are used (CRIMSON1.1, CRIMSON1.2 …). By this action, the individual is no longer identifiable and therefore the data is no longer personal data. A pseudonymization is not planned.