Metrological evaluation and testing of robots in international competitions

Deliverable title	D4.1: RAMI Evaluation plan
Deliverable lead	Francisco Javier Pérez Grau
Related task(s)	T4.1 RAMI Competition definition
Author(s)	Gabriele Ferri Fausto Ferreira Francisco Javier Pérez Grau
Dissemination Level	Confidential, only for members of the consortium/public
Related work package	WP4: Inspection and maintenance
Submission date	30th June, 2020
Grant Agreement #	871252
Start date of project	1st January, 2020
Duration	36 months
Abstract	The evaluation plan includes the definition of scenarios, episode, task, and functional benchmarks and their associated execution and assessment procedures. It takes into account scientific, economic, ethical, legal, and safety aspects, as well as inputs from stakeholders.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Versioning and Contribution History

Version	Date	Modified by	Modification reasons
V0	26th June 2020	Francisco Javier Pérez Grau	First version
V1	29th June 2020	Francisco Javier Pérez Grau	Integrated comments from Rémi Régnier's review
V2	30th June 2020	Francisco Javier Pérez Grau	Integrated comments from Anne Kalouguine
V3	29th April 2021	Francisco Javier Pérez Grau	Updates after dry-run campaign execution
V4	25th June 2021	Francisco Javier Pérez Grau	Updates before 1st cascade campaign execution

List of Abbreviations and Acronyms

Abbreviation/Acronym	Meaning
ATEX	Appareils destinés à être utilisés en ATmosphères EXplosives
AUV	Autonomous Underwater Vehicle
CATEC	Advanced Center for Aerospace Technologies
CMRE	Centre for Maritime Research and Experimentation
DVL	Doppler Velocity Log
FBM	Functionality Benchmark
GNSS	Global Navigation Satellite System
I-AUV	Inspection AUV
I&M	Inspection and Maintenance
MTOW	Maximum Take-Off Weight
OPI	Object of Potential Interest
O&G	Oil and Gas
RAMI	Robotics for Asset Maintenance and Inspection
RMSE	Root-Mean-Square-Error
ROV	Remotely Operated Vehicle
SAUC-E	Student Autonomous Underwater Vehicle Challenge – Europe
TBM	Task Benchmark
UAV	Unmanned Aerial Vehicle
VTOL	Vertical Take-Off and Landing

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Executive summary

The current document is the Evaluation Plan for RAMI competition, the robotics competition of METRICS project focused on the inspection and maintenance domain. This evaluation plan is based on METRICS methodology, and comprises the procedure, benchmarks and evaluation metrics for the competition, both the field and the cascade campaigns.

This evaluation plan includes the definition of scenarios, episodes, task and functionality benchmarks and their associated execution and assessment procedures. It considers scientific, economic, ethical, legal, and safety aspects, as well as inputs from stakeholders.

After considering scientific aspects and inputs from relevant stakeholders, different use cases have been identified as potentially having a major impact on the I&M sector, while keeping the competition attractiveness for participants, and focusing on aerial and underwater robots.

1 Introduction to RAMI competition

RAMI (Robotics for Asset Maintenance and Inspection) is one of the four challenge-led robotic competitions of METRICS project. RAMI competition aims at addressing Inspection and Maintenance (I&M) tasks achieved by aerial and underwater robots, offering the possibility of increasing the spatial/temporal resolution of the inspection process, improving the operation persistency and the quality of the acquired data. At the same time, these robotic domains have the potential to reduce the operational costs and to increase the safety of workers, especially in dangerous areas, such as explosive atmosphere (ATEX) environments, or works at height.
However, to tackle the different challenges of the I&M sector, and increase the added value of using robots, it is key to increase their autonomy level. A high degree of autonomy is especially required when a direct link with an operator cannot be guaranteed, or when it is required to perform inspection tasks in a repetitive way. Autonomous decisions can also increase the robot mission performance and guarantee robot survival in hostile or cluttered environments, where it is difficult to teleoperate robots safely.
Moreover, it has been identified that the most promising applications in the I&M sector require the use of aerial and underwater robots due to the risks and costs associated to work at height or underwater inspection performed by human operators. Therefore, the RAMI competition will focus on these two types of robots, to push the state of the art in terms of autonomy navigation and performance. RAMI will focus on the Oil&Gas and renewable energy sectors, both off-shore and on-shore facilities. In these domains, commercial robots used for I&M are usually teleoperated such as the Remotely Operated Vehicles in these industries, or drones for visual inspection of large infrastructures in refineries. RAMI addresses this need by increasing, assessing and evaluating the robot autonomy in I&M tasks.
Therefore, the evaluation process of RAMI competitions will mainly involve tasks related to autonomous navigation and data acquisition for I&M purposes. Since aerial and underwater domains are very different, both domains will be evaluated separately in two different tracks. CATEC, the Advanced Center for Aerospace Technologies, will organize the aerial domain, while CMRE, the Centre for Maritime Research and Experimentation, will oversee the underwater domain. Both will conduct the competition evaluations in realistic environments of interest for I&M end-users.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

1.1 Competition Outline

RAMI competition, as well as the other competitions in METRICS project, will be organized as a series of campaigns. There are two types of competitions:

- A Field Evaluation Campaign, or field competition, in which the physical devices are tested in realistic environments (i.e., physical test-beds).
- A Cascade Evaluation Campaign, or cascade competition, in which software is tested on data generated during the field competition.

All the competitions will be designed in a similar spirit: the first year is a dry-run that allows validating the evaluation procedure. After this, a competition of each type will be organized once a year for the two remaining years of METRICS project.

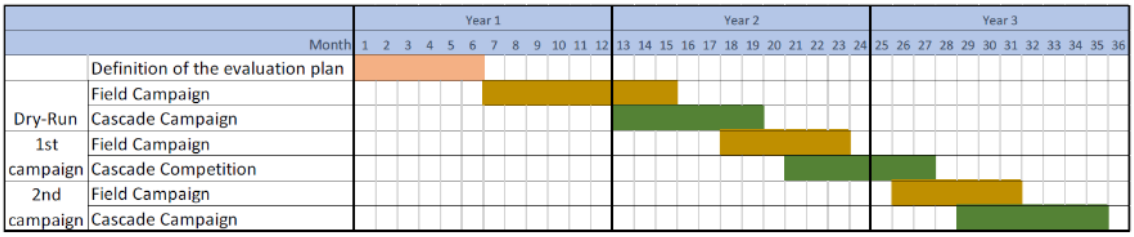

Figure 1. RAMI competition outline

In the case of RAMI competition, the dry-run phase during the first year of METRICS project will not involve external participants. The evaluation protocol and evaluation tools will be tested on CATEC and CMRE aerial and underwater robots.

1.2 Document Outline

This document is structured as follows, according to the common evaluation framework that will be applied during METRICS competitions, covered in the document "D2.1 Methodological Framework".
Section 2 details the tasks that are part of RAMI competition and will be evaluated across the different competition campaigns, and the motivation that made them suitable for METRICS.
Section 3 describes the testing scenarios that will host the campaigns of RAMI competition, along with explanations about their benefits and how to overcome their limitations.
Section 4 introduces the metrics and scoring mechanisms that will be used to assess the performance of the teams attending the competitions.
Since RAMI competition has two different tracks, one for aerial robots and another one for underwater robots, each of these sections will have two subsections detailing the information that is specific to each track.
Finally, Section 5 highlights how this evaluation plan follows METRICS methodology.

2 Evaluation tasks

Evaluating the overall performance of a robotic system while performing a task is interesting for assessing the global behaviour of the application, but neither does it allow the evaluation of the contribution of each component, nor does it put in evidence which components are limiting the system performance. On the other side, the good performance of each element in a set of components does not necessarily mean that a robot built with such components will perform well: system-level integration has, in fact, a deep influence on this, which is not investigated at all by component-level

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

benchmarking. For these reasons, METRICS comprises two categories of benchmarks for evaluating the performance of teams:

- **Functionality Benchmarks (FBMs)** focus on specific capabilities required for the target application. A Functionality is conventionally identified by researchers as a self-contained unit of capability, which is too low-level to be useful on its own to reach a goal (e.g., self-localization, crucial to most applications, but aimless on its own). A Functionality can be provided by a single component or by a set of components, and usually involves both hardware and software. An FBM is a benchmark that investigates the performance of a robot component when executing a given functionality. A Functionality is as independent as possible of the other functionalities of the system, to control it as the sole dependent variable in the evaluation.
- **Task Benchmarks (TBMs)** combine multiple functionalities for the execution of complex activities. A Task is an activity of a robot system that, when performed, accomplishes a goal that is considered useful on its own. A task always requires multiple functionalities to be performed (e.g., finding and fetching an object, which involves functionalities such as self-localization, mapping, navigation, obstacle avoidance, perception, object classification/identification, grasping). A TBM is a benchmark that investigates the performance of a robot system when executing a given task. TBMs are designed by focusing on the goal of the task, without constraining how such goal is reached.

The selection of FBMs and TBMs included in RAMI competitions is still under discussion, given the current interactions with relevant stakeholders from the I&M sector. For each proposed benchmark, a brief description is provided along with the reasons behind their choice.

2.1 Underwater domain

The goal of the underwater field campaigns within RAMI competitions is to advance the state of the art in so-called Inspection Autonomous Underwater Vehicles (I-AUVs). While over the past years AUVs have become more and more advanced and autonomous in a myriad of tasks, the degree of precision required in I&M tasks is still a challenge for most AUVs. Several technological issues are still to be fully solved making popular the use of I-AUVs in this kind of tasks.

For these reasons, today Remotely Operated Vehicles (ROVs) are mostly used in Oil&Gas industry. To operate this kind of robots a pilot is needed. The control station is connected via a cable to the robot, and the robot is teleoperated to accomplish different tasks, such as inspection and manipulation. However, a new generation of I-AUVs would open novel opportunities in the underwater inspection and maintenance scenario. Tasks of monitoring, survey and manipulation could be achieved at a longer distance (there is no more the need of a cable connecting the robot with a control station), in a more persistent way and, above all, reducing the workload for pilots.

However, to reach a mature technology for the use of I-AUVs several barriers are still present. One of the strongest barriers to the proliferation of I-AUVs in industries such as O&G is the lack of precise navigation and localization underwater. This is a common problem that affects all AUVs but to a less extent ROVs used in O&G industry. The other main barrier is free floating manipulation. Maintenance tasks are among the hardest to be tackled in underwater robotics. ROVs have been used for decades for this kind of tasks but I-AUVs are much more recent and less widespread. Tackling this challenge is extremely relevant for the industry.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

In METRICS we propose evaluation tasks that address these barriers, including autonomous navigation, inspection, advanced perception mapping and manipulation. The criteria to evaluate these tasks will be based on the accuracy of navigation, object detection and recognition (for inspection), quality/coverage of mapping, and the autonomy degree, completeness and robustness of the manipulation.

The tasks were chosen taking into account different stakeholders including typical I&M industries such as O&G. Indeed, on the one hand, we will evaluate mapping and object recognition that are generic functionalities needed across the whole spectrum of Inspection tasks. On the other hand, pipeline inspection and the intervention tasks chosen address directly two issues found in O&G industry. We will have a task concerning valve closing which is currently done in a similar way by ROVs. Similarly, touching the pipe and keeping the contact is another task required in O&G that we will evaluate in the field campaign. During the project lifespan, we also plan to engage the industry to help us determine the best objects to be tested.

The general evaluation scenario requests the robots to reach, inspect and map the operation area where Objects of Potential Interest (OPIs) are deployed. OPIs are of different nature: submersed buoys (of different dimensions, colors and with some identification signs such as numbers), pipes of various lengths, pipeline assembly structures and several objects of different colors and shapes. Then the robots have to intervene in the environment, closing/opening valves, staying in touch with a pipe for its inspection and have to perform pick and place with some objects in dedicated areas. The evaluation tasks are divided in a modular way into FBMs and TBMs as per METRICS methodology and are detailed below.

The evaluation constraints and requirements for robots will be set to guarantee comparability between different performances, but at the same time they must not prevent technological creativity. All robots participating in the competition will be allowed to have the same type of sensors and potential advantages will be eliminated by checking the list of requirements and limitations as part of the qualification process. Results should be provided to the evaluators in standard formats easily processed by common tools used in robotics. A precise list of formats accepted and instructions for data production will be given to the teams in advance of the evaluation campaigns. This will help the comparability of results.

In the following, a description of all FBMs and TBMs to be executed at the field campaigns.

FBMs:

- **FBM1-M:** mapping the area: This FBM has the objective of assessing the capability of the robot to map an area and to localise the encountered OPIs. The robots start the benchmark at a location inside the area to inspect, to limit the influence of the navigation error accumulated along a possible transit. The quality of the produced map will be assessed. The area coverage will be considered as well as the accuracy in the localization of OPIs in a geo-referenced frame. Both real-time and post-processed results will be used in the performance assessment;
- **FBM2-M**: object recognition: The second evaluation assesses the capability of the robot to classify known objects underwater. The objects, different in shape and color, are shown to the robot before the mission (e.g., a circular red panel with a number printed). Then some objects are positioned underwater and shown to the robot. The robot is positioned at different distances and at different angles of view from the objects. Results will assess the capability of the robot to classify the detected object in the different tested configurations. Both real-time and post-processed results will be used in the performance assessment. Metrics that can be used are F-measure, precision and recall so as to assess the quality of the classification. In the

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

future, O&G experts can suggest us on how to modify the shapes for an evaluation more appealing for the industry end-users;

- **FBM3-M:** shape recognition: FBM3-M addresses the capability of the robot to estimate the position and the shape of an unknown (or roughly known) object. The robots are expected to analyse some objects positioned underwater at different distances and at different angles of view. The dimensions/shapes of the object (e.g., a trapezoidal sign with a marker printed) are not (perfectly) known a priori. The robot has to determine the shape, dimensions and position relative to a given reference frame point, according to the provided OPIs. Both real-time and post-processed results will be used in the performance assessment. The accuracy and precision in the determination of the object dimension/position are evaluated.

TBMs:

- **TBM1-M:** pipeline area inspection:
  - In the first TBM, a robot has to reach the area of inspection through autonomous navigation. The objective of the underwater navigation task is to assess the ability of the robot to estimate its position and orientation during the I&M operations. Underwater navigation is still an open research topic since underwater GNSS methods are not available and a combination of proprioceptive (e.g., Inertial Measurement Units) and exteroceptive sensors (e.g., Doppler Velocity Logs) has to be used. Passive beacons will be added to facilitate the analysis. Obstacles (underwater buoys) will be placed along the robot navigation path to assess its obstacle avoidance capabilities and reactive behaviour.
  - Once in the area, the robot has to map it (some OPIs will be deployed in the area) and inspect the pipeline structure identifying a leak (a marker on a pipe) reporting the dimension and the position of the leak;
- **TBM2-M:** intervention on the pipeline structure:
  - In this TBM, the robot has to approach the pipeline assembly structure and to intervene in the area. Objects placed in proximity of the pipeline assembly structure have to be classified and their dimensions/shape/position computed.
  - The robot has to localise the valves and close/open them (in a supervised or autonomous way).
  - Then the robot has to touch a pipe and stay in touch with it despite possible environmental disturbances (e.g., waves). This is of interest in the gas and oil industry since this action is usually conducted to detect via electric sensors if a leak is present along the pipe.
  - Finally, the robot has to grab a pole from a console (in a supervised or autonomous way), to remove it from the structure and to place it in another location of the console (pick and place);
- **TBM3-M**: complete I&M mission: In the final TBM, the robot has to perform the tasks of TBM1-M and TBM2-M.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

2.2 Aerial domain

Due to ageing, environmental factors, changes in use, damages caused by human/natural factors, inadequate or poor maintenance and deferred repairs, Oil&Gas and renewable energy infrastructure is progressively deteriorating, urgently needing inspection, assessment and repair work. Currently, inspection activities are primarily done through visual observations by inspectors. It relies upon the inspector having access to low accessible areas via specific equipment (ladders, rigging and scaffolds) or vehicular lifts (man lifts or bucket trucks). This is potentially dangerous for the inspectors. Robotics, and more specifically aerial robotics, can provide the required elements for in-depth inspection.

The evaluation process of the aerial domain track will be mainly focused on addressing the following autonomy-based functions and inspection tasks:

- Precise autonomous navigation without GNSS.
- Automatic detection of defects using advanced AI algorithms.
- Performing punctual inspections in difficult access areas.
- Obtaining images from the same location in a repetitive way.

During I&M activities for Oil&Gas, renewable energy or civil infrastructure sectors, GNSS coverage is often poor when operating in cluttered environments, or for example below bridges. Moreover, inspections could also be necessary indoors. The safe introduction of autonomous aerial robots for these activities needs a higher level of maturity in the currently existing navigation approaches. Such approaches must rely only on sensors carried by the aerial platform, and not on existing infrastructure or additional sensors deployed in the environment, to provide a truly scalable solution.

The automatic detection of defects is important for inspectors when they face considerable amounts of data to review, which can be a tedious task. Having a system that provides the inspector with only the images that contain defects for simple review is a very powerful tool to simplify their job and allow for faster processing times.

Punctual inspections from specific locations are common operations when dealing with I&M tasks. For example, there are particular known points where infrastructure is subject to more stress, and therefore are interesting areas to inspect. Furthermore, being able to inspect that same location repetitively, at different time intervals, is even more valuable to assess the evolution of the infrastructure. All this also needs to include safe navigation capabilities without GNSS.

The aerial competitions of RAMI are designed to be more focused on inspection activities than maintenance activities. The main reason is that the maturity level of the required functionalities is slightly higher in the former case. Robotic aerial manipulation is an active research topic, but we consider that introducing maintenance tasks, most likely involving aerial manipulation, would increase the difficulty and raise the entry level requirement of research teams.

The specific FBMs and TBMs for the aerial domain track of RAMI competitions are the listed below.

FBMs:

- **FBM1-A:** precise navigation without GNSS: This FBM assesses the ability of the aerial robots to navigate precisely in a safe way, without the use of GNSS or magnetometers, due to the complex magnetic environment in Oil&Gas infrastructures. The execution of the FBM consists in an aerial robot performing specific trajectories, and the evaluation will be based on the comparison of a precise positioning system with respect to the results of the navigation system of the aerial robot. The specific trajectories will be the same for every competing team, thus guaranteeing fair comparisons.
- **FBM2-A:** automatic detection of defects: The objective is to assess the ability of a technological solution to automatically detect defects, such as corrosion on pipes, while in operation using advanced AI algorithms. The assessment will be based on an offline analysis of the images

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

obtained and processed by the aerial robot. Several defects will be artificially placed along the scenario, and the aerial robot will need to inspect the area of interest in order to collect images. After the offline analysis of the images dataset, the results of the automatic system will be compared with the ground truth, estimated by human experts.

TBMs:

- **TBM1-A:** punctual inspection in difficult access areas:
  - This TBM is focused on safely reaching a specific area with the aerial robot in an autonomous way, in a cluttered environment and without GNSS. In this task, the aerial robot has to estimate its pose with onboard sensors and be able to navigate while maintaining a minimum distance to obstacles. The obstacles will emulate the typical infrastructure in a refinery (pipes, vessels, auxiliary structures, etc.).
  - This Task Benchmark is a combination of the two previous Functionality Benchmarks, so the aerial robot can now emulate a full inspection mission as it would be carried out in a real environment.
  - The area of interest to be inspected will be provided to the system. Once the aerial robot reaches that area, it must take a picture and associate it with the specific location. After this has been done, the aerial robot must return safely to the starting location. Robot navigation is required to be fully autonomous during the whole trial.
  - There can be different areas of inspection with similar complexity regarding their accessibility; this will be the independent variable. The evaluation of this TBM consists of different achievements that need to be checked during the execution of the task.
- **TBM2-A:** repetitive inspection:
  - The evaluation will be focused on the capability of the aerial robot to precisely navigate to the same goal position and be able to take images of the same location repetitively.
  - The area of interest to be repetitively inspected will always be the same in the scenario, and the acquired images must include such area.
  - The capability of acquiring images of the area of interest will be used to compare the performance of the systems.

3 Testing settings

As stated in the previous section, RAMI competition will be organized in two different tracks: underwater and aerial. Underwater robots competitions will take place in the seawater basin of CMRE (La Spezia, Italy), while aerial robots competitions will be held in the indoor testbed of CATEC (Seville, Spain).

3.1 Underwater domain

The physical testbed used by CMRE to run the competitions is CMRE's sea basin. This protected harbour provides excellent conditions for real-life challenges. It has been used for marine robotics competitions in the past 10 years and has proven a challenging yet accessible testbed (see Figure 1). The seawater basin is 50 x 50 m with a depth of 4-5 meters. Both student-based competitions (Student Autonomous Underwater Vehicle Challenge- Europe/SAUC-E) and more advanced competitions such

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

as euRathlon 2014, European Robotics League Emergency 2018 and 2019 took place in this testbed. The latter ones were dedicated to benchmarking and thus, this testbed has proved being suited for metrological evaluation.

Figure 2. View of the CMRE water basin.

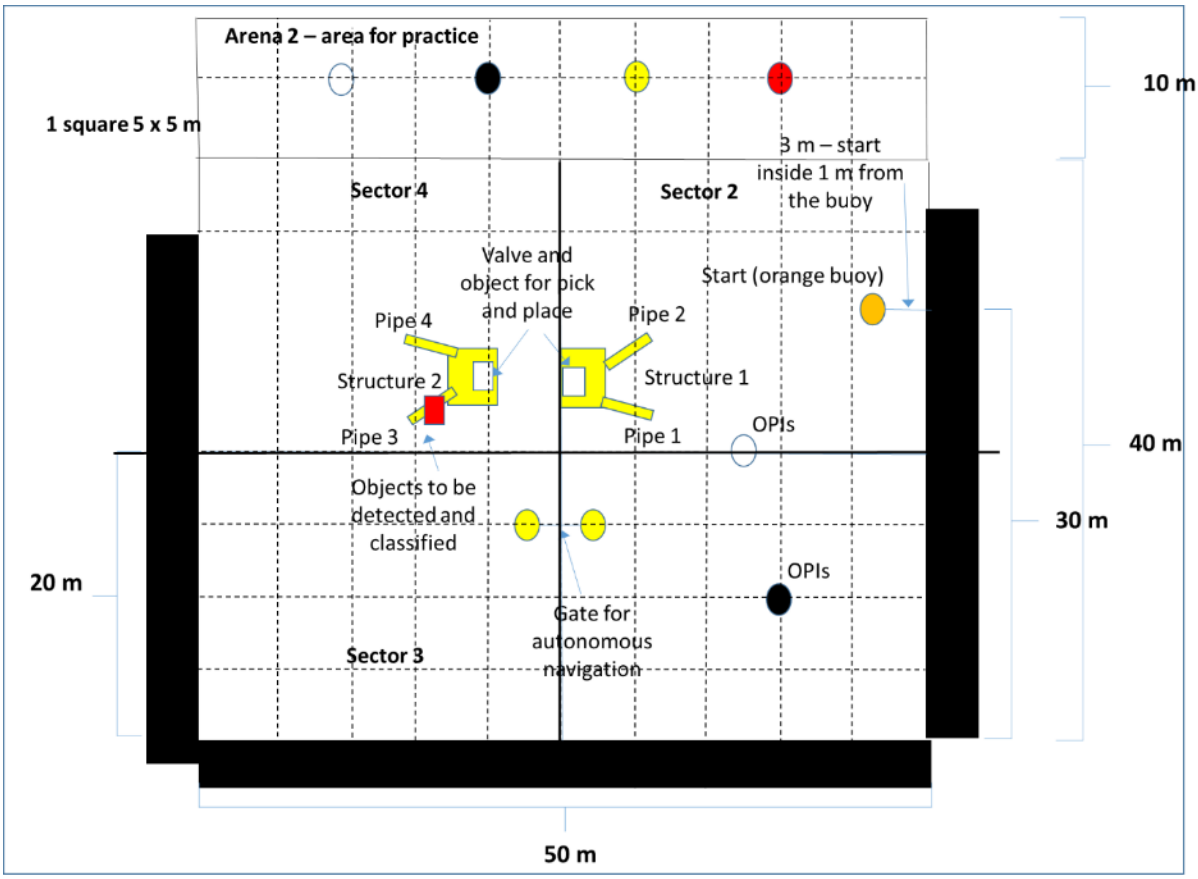

Figure 3. Indicative arena configuration for the METRICS evaluation campaigns. An area of 40 m x 50 m is considered for the evaluation arena. A second area, Area 2, will be prepared for allowing the teams with OPIs for practicing.

The same arena is available to all teams so reproducibility of evaluation can be achieved for what regards the tasks to be completed. However, being an outdoor scenario and an open sea testbed, environmental conditions can change and are not controllable. Wind, waves, tides and luminosity/cloud cover can vary throughout the day/competition.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Being a protected harbour and with the competition planned for summer months, waves are typically small and calm and do not vary significantly except in case of rare inclement weather (including strong winds). The tide difference is around 0.5 meters and it is known so that factor can be easily taken into account when planning teams' timeslots. Luminosity/cloud cover is partially predictable but can change rapidly. However, it can also be taken into account into the planning/scheduling.
One way of controlling the influence factors is to have all teams experiencing the different conditions through the week of competitions. For instance, all teams are given timeslots in the morning in the first day and in the afternoon in the second day. This addresses both luminosity and tide changes as similar tides occur at similar times over two consecutive days and the luminosity in the morning is typically very different than in the afternoon.
Another possible way is to measure these parameters (tides, luminosity) and attribute an "influence factor" for each of the factors to be multiplied by the scores of the teams. This could be hard to implement as, while for luminosity and waves, it is easy to understand what are good conditions or bad conditions; for the tides, it is hard to say if it is better high tide or low tide. For this reason, we are not considering this option at the moment.
Finally, another possible option to explore is to define an acceptable range of conditions for which the expected performance is similar and exclude from the evaluation observations taken out of this range. For instance, bad weather conditions (to be defined precisely) invalidate any results. Extreme cloud cover is also outside the range of validity for luminosity. This should be done for each influencing factor.
For the accessibility of the testbed, considering the past years' experience, there are not many limitations preventing teams to attend the evaluation and bring their robot to test. There are security hurdles for nationals coming from non-NATO countries (or partner countries). This constrains teams to apply and send personal details in advance of the evaluation campaign to allow enough time for the access control procedure. Teams with small budgets have participated in the past and teams with different robot sizes can participate. There is a limitation on the size/weight of the vehicle that prevents big/heavy vehicles to participate. However, this is mostly for safety of the vehicles and takes into account the dimension of the testbed (50 x 50 m) and the transportation to the site. Teams can test their vehicles in swimming pools prior to attending the evaluation, both at their labs and in the competition area. Not everything can be tested in a small pool, but there are currently several simulators dedicated to underwater robotics (e.g., UWSim) that can be used to mitigate this issue.
There will be a qualification procedure for the teams wishing to participate. This will be done to select the teams (limited number of spots available) and to make sure vehicles are safe and ready to participate. The qualification procedure before the evaluation consists of two phases. In the first phase, teams send basic information regarding their team and vehicle. The information must show that teams' robots fulfil the list of requirements to enter the competition (mostly safety requirements). This serves to do a pre-selection and estimate how many teams are willing to participate. In the second phase, teams need to submit a paper/technical report and a video. The paper must describe the vehicle and the approach to the evaluation campaign. The video must show the robot doing basic manoeuvres in a pool. Based on the assessment of the paper and video, the teams are selected to the competition. Teams that do not provide enough proof of readiness will be invited to participate first in the cascade campaign and apply for the field campaign the following year. The qualification procedure will be completed at the testbed location with a safety check before the start of the evaluation campaign. No vehicle that did not pass the safety check will be allowed to participate.
Regarding the participation of judges in the competitions, a qualification process will also be made. For scientific profiles, we will take into account aspects of their experience in the underwater robotics research field, while for industrial profiles, we will consider factors such as the relevance and use of robotic solutions in their companies' I&M activities.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

For the cascade campaigns, the datasets to be acquired will consist of navigation data (raw – e.g., Doppler Velocity Log (DVL) for speeds and Inertial Measurement Unit (IMU) for angular velocity and accelerations and pressometer output for pression, and processed – i.e., speed and robot position-latitude, longitude and depth), optical camera images and forward looking sonar. Images and sonar images of the objects to be detected and mapped will be provided together with navigation data. Object detection/recognition and classification (including shape and dimensions) will be evaluated in the cascade campaigns based on data from FBM2-M and FBM3-M.

3.2 Aerial domain

CATEC counts with an aerial robotics indoor testbed that can be used to develop and test different algorithms and technologies applied to multiple aerial platforms. The tests can be conducted in a 15 x 15 x 5 m volume, although due to safety and the need to keep track of the aerial robot's pose during the whole run with the position tracking system, the operative area is 10 m x 10 m x 3.5 m size.
This testbed has an indoor localization system based on 20 VICON cameras, that only needs the installation of passive markers on the aerial vehicle and/or along the scenario. This system is able to provide, in real-time, the position and orientation of the aerial vehicle with millimetre precision; the sample rate of data acquisition is 100 Hz. This aerial robotics indoor testbed has been extensively used in previous European projects such as ARCAS and AEROARMS, for the development of aerial robotics technologies for I&M applications. Thanks to this, CATEC counts with mock-ups of infrastructures that can be used to recreate different scenarios inside this testbed.

Figure 4. Indoor testbed at CATEC for the field campaigns

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Figure 5. Obstacles and points of interest.

Figure 6. Pipe rack used as obstacle (left). Structure used as point of interest with defects (right).

In order not to restrict the technological solutions of the teams, the design of their aerial platforms will only be restricted due to general basic safety reasons. Only Vertical Take-Off and Landing (VTOL) and battery-powered platforms will be allowed to participate in RAMI competitions. According to CATEC's indoor testbed size, the platform's diagonal wheelbase must be smaller than 1000 mm and the maximum take-off weight (MTOW) must be smaller than 5 kg.

Being an indoor testbed, external factors usually affecting outdoor competitions such as extreme temperatures, strong winds, illumination changes due to clouds, possibility of rain, presence of vegetation or birds are not applicable. Therefore, there is guarantee that competitors are evaluated in the same conditions, in the same scenario.

The indoor testbed is covered by tarps to guarantee a priori ignorance for the teams, in a way that they cannot see how the scenario is mounted until they access the competition area.

The competing teams may find different pictures of defects along the execution of TBM1-A and TBM2-A. The defects contained in these pictures mimic cracks in real surfaces. It should be noted that there

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

may be pictures that do not contain any defects. The following figures show some examples of pictures used for this purpose.

Figure 7. Example of images with defects.

Figure 8. Example of images without defects.

Competition days will be subdivided into time slots, and these slots will be assigned to the different competing teams in a random manner. Teams will be allowed to switch time slots between each other if they agree.

Similarly to the underwater domain, there will be a qualification procedure for the teams willing to participate, before granting a team access to the indoor testbed, there is a qualification procedure for RAMI aerial domain competitions. This will be done to select the teams (limited number of spots available) and to make sure the aerial robots are safe and ready to participate. The qualification procedure before the evaluation consists of two phases. In the first phase, teams send basic information regarding their team and aerial platform. The information must show that teams' robots fulfil the list of requirements to enter the competition (mostly safety requirements). This serves to do a pre-selection and estimate how many teams are willing to participate. In the second phase, teams need to submit a paper/technical report, and a set of videos showing specific manoeuvres of the aerial platform. The paper must describe the vehicle and the approach to the evaluation campaign. The videos must show the robot doing basic manoeuvres in manual mode, to assess their performance and also the pilot's ability. Based on the assessment of the paper and videos, the teams are selected to the

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

competition. Teams that do not provide enough proof of readiness will be invited to participate first in the cascade campaign and apply for the field campaign the following year. The qualification procedure will be completed at the testbed location with a safety check during the first days of testing/practice. Platforms that did not pass the safety check will not be allowed to participate.
Regarding the participation of judges in the competitions, a qualification process will also be made. For scientific profiles, we will take into account aspects of their experience in the underwater robotics research field, while for industrial profiles, we will consider factors such as the relevance and use of robotic solutions in their companies' I&M activities.
For the cascade campaigns, the datasets to be acquired will consist of navigation data (raw position and orientation from the testbed's motion capture system), and optical camera RGB and depth images. Images of the objects of interest (i.e., defects) to be detected and mapped will be provided together with navigation data. Object detection/recognition and classification will be evaluated in the cascade campaigns based on data from FBM2-A.

4 Metrics and scoring

Details concerning rules, procedures, as well as scoring and benchmarking methods, are common to all Task Benchmarks (TBMs).
There will be mandatory pre-competition safety-checks described in Section 5 of this Rulebook. Only teams that successfully pass the safety checks will be able to participate in the competition.
Evaluation of the performance of a robot according to TBMs is based on performance equivalence classes. The criterion defining the performance equivalence class of robots is based on the concept of tasks required achievements. The ranking of the robot within each equivalence class is obtained by looking at the performance criteria. In particular:

- The performance of any robot belonging to performance class N is considered as better than the performance of any robot belonging to performance class M whenever M < N.
- Considering two robots belonging to the same class, then a penalization criterion (penalties are defined according to task performance criteria) is used and the performance of the one that received less penalization is considered as better.
- If the two robots received the same amount of penalization, the performance of the one that finished the task more quickly is considered as the best (unless not being able to reach a given achievement within a given time is explicitly considered as a penalty).

Performance equivalence classes and in-class ranking of the robots are determined according to three sets:

- A set A of achievements, i.e. things that should happen (what the robot is expected to do).
- A set PB of penalised behaviours, i.e. robot behaviours that are penalised if they happen (e.g., manual intervention).
- A set DB of disqualifying behaviours, i.e. robot behaviours that absolutely must not happen.

Scoring is implemented with the following 3-step sorting algorithm:

- If one or more of the elements of set DB occur during task execution, the robot gets disqualified (i.e., assigned to the lowest possible performance class, called class 0), and no further scoring procedures are performed.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

- Performance equivalence class X is assigned to the robot, where X corresponds to the number of achievements in set A that have been accomplished.
- Whenever an element of set PB occurs, a penalization is assigned to the robot (without changing its performance class).

One key property of this scoring system is that a robot that executes the required task completely will always be placed into a higher performance class than a robot that executes the task partially. Moreover, the penalties do not make a robot change class (also in the case of incomplete task).
For the specific Functionality Benchmarks, each domain has its own methods for scoring, detailed in the following subsections.

4.1 Underwater domain

The metrics used to measure the teams' performance in FBMs and TBMs will be inspired by previous projects such as SciRoc, RockEU2 and euRathlon. We will use a similar strategy of having different achievements as in SciRoc. We will also use weights, autonomy classes and penalties to handle more complex achievements that cannot be simply measured as success/failure and count as much as simpler achievements. For instance, mapping can be measured similarly to euRathlon, where the coverage of the map was estimated (between 0 and 100%) and it weighted 4 times more than a simple achievement. Similarly, where semi-autonomous mode is allowed (for TBM2-M), different autonomy classes will be used and autonomous mode will be considered above semi-autonomous.
The metrics will be as objective as possible and subjectivity will be limited to tasks where it is not possible to have a quantitative measure. Where a measure depends on the interpretation of the referees, the consensus among all evaluators measuring that achievement will be required.
Functionalities can be grouped in three types of metrics:

- Functionalities assessed by binary metrics (1/0), e.g. reached a waypoint, passed a gate, etc.
- Functionalities assessed by completion metrics (from 0 to 100%), e.g. valve closing, and mapping.
- Functionalities assessed by other quantitative metrics such as precision, recall, F-measure, RMSE, e.g. object recognition.

Non-functional metrics such as power consumption or total distance navigated are not considered at this point. However, time may be considered and a bonus depending on the total run time can be used in the TBMs.

4.2 Aerial domain

TBMs will be evaluated using binary metrics, as explained at the beginning of this section.
TBM1-A: Punctual inspection in difficult access areas
To evaluate the TBM1-A, there is a list of achievements based on the criterion explained before. Reaching the waypoint precisely, staying in the waypoint for a determined time slot, avoiding the obstacles safely and capturing the images of the defects are considered positively, among other achievements. After the three days of competition, the best run of each team is used to evaluate the TBM1-A. TBM1-A is evaluated as follows:

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Set A1: TBM1-A

- WP1:
  - A1.1) An aerial robot reaches WP1 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 2 seconds.
  - A1.2) An aerial robot reaches WP1 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.3) An aerial robot reaches WP1 with the correct orientation within 0.2m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.4) An aerial robot reaches WP1 with the correct orientation within 3 minutes.
  - A1.5) An aerial robot reaches WP1 with the correct orientation within 1 minute.
  - A1.6) An aerial robot reaches WP1 with the correct orientation within 30 seconds.

- WP2:
  - A1.7) An aerial robot reaches WP2 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 2 seconds.
  - A1.8) An aerial robot reaches WP2 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.9) An aerial robot reaches WP2 with the correct orientation within 0.2m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.10) An aerial robot reaches WP2 with the correct orientation within 3 minutes.
  - A1.11) An aerial robot reaches WP2 with the correct orientation within 1 minute.
  - A1.12) An aerial robot reaches WP2 with the correct orientation within 30 seconds.
- WP3:
  - A1.13) An aerial robot reaches WP3 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 2 seconds.
  - A1.14) An aerial robot reaches WP3 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.15) An aerial robot reaches WP3 with the correct orientation within 0.2m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.16) An aerial robot reaches WP3 with the correct orientation within 3 minutes.
  - A1.17) An aerial robot reaches WP3 with the correct orientation within 1 minute.
  - A1.18) An aerial robot reaches WP3 with the correct orientation within 30 seconds.
- WP4:
  - A1.19) An aerial robot reaches WP4 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 2 seconds.
  - A1.20) An aerial robot reaches WP4 with the correct orientation within 0.5m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.21) An aerial robot reaches WP4 with the correct orientation within 0.2m radius avoiding obstacles along the route staying for 10 seconds.
  - A1.22) An aerial robot reaches WP4 with the correct orientation within 3 minutes.
  - A1.23) An aerial robot reaches WP4 with the correct orientation within 1 minute.
  - A1.24) An aerial robot reaches WP4 with the correct orientation within 30 seconds.

Set A2: General

- A2.1) The aerial robot returns to the take-off/landing area once all the tasks have been done.
- A2.2) The aerial robot takes off autonomously.
- A2.3) The aerial robot lands autonomously.
- A2.4) The aerial robot transmits live position and images/video to the control station during the run.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

- A2.5) The aerial robot reaches all the waypoints sequentially.
- A2.6) The aerial robot detects the defects online and it is visualized in the control station.

The set **PB of penalised behaviour** for this task are:

- The robot needs manual intervention. A maximum of one intervention is permitted.
  Note: when the maximum number of interventions reaches the allowed maximum, the run is considered terminated and a new run is restarted.

Additional penalised behaviours may be identified and added to this list if deemed necessary.

The set **DB of disqualifying behaviours** for this task are:
- A robot damages the competition arena (including obstacles).
- A robot does not conform to safety regulations for the competition.
- The robot leaves the flight volumes defined by the organisation.
- The team does not provide the data after the required time.

Additional disqualifying behaviours may be identified and added to this list if deemed necessary. These sets will be completed in later rule revisions.
The output provided by the teams is a set of files that must be saved in a USB stick given to the teams before the test. The USB stick must be formatted with NTFS file system and all the files should be saved in a folder with the name of the team.

For this task, teams must provide the following data.
Navigation Data: this must be a file in .txt format that must satisfy the following requirements:
- Contain all the navigation events from the start of the flight until the end of the flight. These events must be: CURRENT, to indicate that the current target waypoint has changed, and REACHED, to indicate that the target waypoint has been reached.
- Each event must be written on a different line and in text format.
- Each line must contain the instant of time of the event, its type (CURRENT or REACHED), the waypoint identifier, as well as the estimated position (x, y, z) and orientation (x, y, z, w) of the robot.
- The fields of each line must be separated by a blank space.

Here is an example of a single line:
 # timestamp type identifier tx ty tz qx qy qz qw
1614061255095597525 REACHED 1 0.034987638631 0.0174963661549 0.127783673399 0.000749064780156 0.00480720140327 -0.627320874946 0.778745689923

Image Data: this must be a folder containing the set of images captured during the execution of TBM1-A in .jpg format and a file in .txt format that must satisfy the following requirements:
- Each event must be written on a different line and in text format.
- Each line must contain the time stamp of the capture of each image, the name of the image file, as well as the estimated position (x, y, z) and orientation (x, y, z, w) of the robot.
- The fields of each line must be separated by a blank space.

Here is an example of a single line:
 # timestamp image_name tx ty tz qx qy qz qw

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

1614061255095597525 frame0005.jpg 0.034987638631 0.0174963661549 0.127783673399 0.000749064780156 0.00480720140327 -0.627320874946 0.778745689923

Pose Data: this must be a file in .txt format that must satisfy the following requirements:

- Contain all the estimated positioning measurements from the start of the flight until the end of the flight at a rate of minimum 1 Hz*****(at least 10 Hz recommended).
- Each measure must be written on a different line and in text format.
- Each line must contain the instant of time of the measurement as well as the estimated position (x, y, z) and orientation (x, y, z, w) of the robot.
- The fields of each line must be separated by a blank space.

Here is an example of a single line:
#timestamp tx ty tz qx qy qz qw
1403636580.013555527 0.0125827899 -0.0015615102 -0.0401530091 -0.0513115190 0.8092916900 0.0008562779 0.5851609600

***Note:** If pose data .txt file contains poses at less than 1 Hz this run would not be considered.*

TBM2-A: Repetitive inspection
To evaluate the TBM2-A, there is a list of achievements based on the criterion explained previously Reaching the waypoint precisely and capturing the images of the defects are considered positively, among other achievements. After the three days of competition, the best run of each team is used to evaluate the TBM2-A. TBM2-A is evaluated as follows:

Set A1: TBM2-A

- A1.1) An aerial robot reaches the goal position (x, y, z, yaw) within 0.5m radius for the first time and stays for 2 seconds.
- A1.2) An aerial robot takes picture in goal position for the first time.
- A1.3) An aerial robot lands in the landing area after reaching the goal position for the first time.
- A1.4) An aerial robot reaches the goal position (x, y, z, yaw) within 0.5m radius for the first time and stays for 2 seconds for the second time.
- A1.5) An aerial robot takes picture in goal position for the second time.
- A1.6) An aerial robot lands in the landing area after reaching the goal position for the second time.
- A1.7) An aerial robot reaches the goal position (x, y, z, yaw) within 0.5m radius for the first time and stays for 2 seconds for the third time.
- A1.8) An aerial robot takes picture in goal position for the third time.
- A1.9) An aerial robot lands in the landing area after reaching the goal position for the third time.
- A1.10) Image 1 includes the defect completely.
- A1.11) Image 2 includes the defect completely.
- A1.12) Image 3 includes the defect completely.
- A1.13) Two out of three images have been captured in the same position (x, y, z, yaw) with an error less than 0.2m.
- A1.14) Three out of three images have been captured in the same position (x, y, z, yaw) with an error less than 0.2m.

Set A2: General

- A2.1) The aerial robot returns to the take-off/landing area once all the tasks have been done.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

- A2.2) The aerial robot takes off autonomously.
- A2.3) The aerial robot lands autonomously.
- A2.4) The aerial robot transmits live position and images/video to the control station during the run.
- A2.5) The aerial robot reaches the goal position three times.
- A2.6) The defect is detected offline in image 1.
- A2.7) The defect is detected offline in image 2.
- A2.8) The defect is detected offline in image 3.
- A2.9) The aerial robot detects the defects online and it is visualized in the control station.

The set **PB of penalised behaviour** for this task are:

- The robot needs manual intervention. A maximum of one intervention is permitted.
  Note: when the maximum number of interventions reaches the allowed maximum, the run is considered terminated and a new run is restarted.

Additional penalised behaviours may be identified and added to this list if deemed necessary.
The set **DB of disqualifying behaviours** for this task are:

- A robot damages the competition arena (including obstacles).
- A robot does not conform to safety regulations for the competition.
- The robot leaves the flight volumes defined by the organisation.
- The team does not provide the data after the required time.

Additional disqualifying behaviours may be identified and added to this list if deemed necessary. These sets will be completed in later rule revisions.

The output provided by the teams is a set of files that must be saved in a USB stick given to the teams before the test. The USB stick must be formatted with NTFS file system and all the files should be saved in a folder with the name of the team.

For this task, the teams must provide the same data requested in TBM1-A: **Navigation, Image and Pose Data.**

For the FBMs of the aerial domain, the scoring will be as follows.
FBM1-A: Precise navigation without GNSS
As stated before, the execution of the FBM consists in an aerial robot performing specific trajectories, and the evaluation will be based on the comparison of a precise positioning system with respect to the results of the navigation system of the aerial robot.
The metric for assigning scores for the competing teams will be based on the Root Mean Squared Error (RMSE) of the provided trajectory with respect to the ground truth trajectory that the aerial robot followed.
To evaluate this FBM, the data of the position estimated by the aerial robot during the execution of TBM1-A is used (**Pose Data**). For each TBM1-A runtime slot, the teams must provide a .txt file of the best flight.
The trajectory evaluation has been performed with a tool called RPG Trajectory Evaluation (https://github.com/uzh-rpg/rpg_trajectory_evaluation). This tool provides the Root-Mean-Square Error (RMSE) in rotation and translation.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

If the team had not provided the data or if the file had not met the indicated requirements, the score that it would have obtained in this FBM would have been the maximum possible (RMSE = ∞).
RMSE is calculated in each section, being a section the trajectory performed between one waypoint and the following, and the average of all the sections will be used to compare the performance between teams.
On the other hand, if the team has not finished the trajectory (i.e., the team has finished all the sections but the last one), their RMSE score is multiplied by a factor α which depends on the percentage of waypoints reached. In case of a tie in the translation's RMSE, rotation's RMSE is used to break the tie.
In the Cascade Campaign, no sections are considered, so the RMSE is calculated in the whole trajectory.

FBM2-A: Automatic detection of defects
After the offline analysis of the images dataset, the results of the automatic system will be compared with the ground truth, estimated by human experts. Precision, recall and F-measure metrics will be used to assess the performance of each team.
Non-functional metrics such as power consumption or total distance navigated can also be considered. Energy consumption could be evaluated through normalized calculations based on the platform's battery level. Thanks to the testbed's motion capture system, the total distance that the aerial platform traversed can be precisely measured, as well as the time it took to complete the trajectory.

To evaluate this FBM, the data of the images captured by the aerial robot during the execution of TBM1-A are used (**Image Data**). For each TBM1-A runtime slot, the teams must provide their best flight results.

Since the images are taken by the teams and are unlabelled, it is the responsibility of the referees to manually label them before the evaluation. Once it is done, the competing team must execute their detection algorithm over the **Image Data** files with the supervision of the referee. The following output data must be provided:

Detection Data: this must be a plane text file (.txt) satisfying:

- Each detection must be written on a different line and in text format.
- Each line must contain the name of the image file as well as the coordinates of the detection on the image plane in pixels as following: top-left corner (x_min, y_min) and bottom-right corner (x_max, y_max).
- The fields of each line must be separated by a blank space.

Here is an example of a single line:
 # image_name left top right bottom
frame0005.jpg 213 144 315 172

The *.txt* described along with the images are evaluated using a tool which associates the labels created by the referee with the detections on the images. The steps performed by this tool to obtain the evaluation metrics are:

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

- Read a line from file.
- Open the image with Image name (e.g., frame0064.jpg) and its label.
- Calculate the Intersection Over Union of the detection (IoU) with every label in the image.
- If the IoU is greater than a given threshold (0.4) this detection is a True Positive and the label is marked as detected.
- Else if none of the labels satisfies the IoU criteria, the detection is marked as a False Positive and no label is marked as detected.
- Once every detection has been evaluated, re-iterate over the input images set and mark every undetected label as False Negative.

The IoU is a metric calculated with 2 bounding boxes to measure the overlapping area proportion between both. The value is calculated dividing the overlapping region by the combined region.

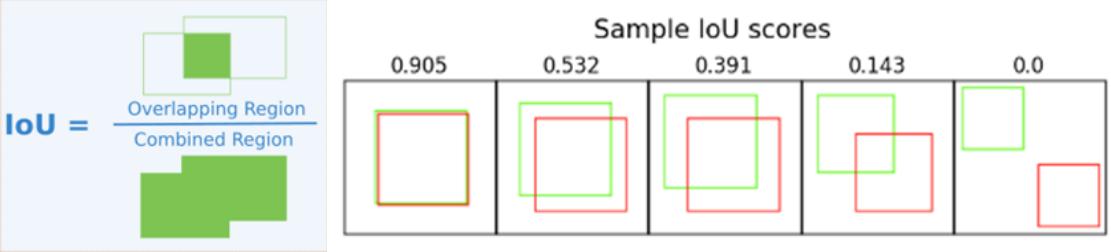

Figure 9. Intersection over Union formula (left) and examples (right).

The evaluation criteria used for this FBM is the Critical Success Index (CSI) or Threat Score (TS), which stands for the equation:

$$CSI = \frac{TP}{TP + FN + FP}$$

In the formula above the *TP* stands for True Positives, which are those defects correctly detected; *FN* for False Negatives, those cracks being undetected; and *FP* for False Positives, which are those detections with no defect associated.

5 METRICS common methodology

The following table covers a self-declaration that the evaluation plan of RAMI competition is compliant with the METRICS common methodology.

Topic	Taken into account	Detail
Organization of the evaluation		
The first occurrence of the competition is a dry-run	Yes	The dry-run will take place during the first season, as described in section 1.1.
The evaluation plan is formalized	Yes	The evaluation plan is presented in this document.
Evaluation tasks		
Each evaluation task is relevant for industry	Yes	They are relevant because they have been selected after extensive interactions with

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

		industrial partners from previous research projects. Nevertheless, the tasks can be refined along the project, and new tasks could be proposed.
The dependent and independent variable of each evaluation are identified	Yes	They are described in section 2 of the evaluation plan.
The evaluation is modular (FBM+TBM)	Yes	The FBMs and TBMs are described in section 2.
The constraints are adapted to the objective of the evaluation	Yes	The constraints on technology selection are exclusively based on safety of the competition, according to the testing settings described in section 3.
Testing environments		
Repeatability and reproducibility of the observations are maximized	Yes	The factors that could affect repeatability and reproducibility of observations are considered in section 3.
The accessibility of the test beds is maximized	Yes	Accessibility of the testing areas is covered in section 3.
A qualification procedure is defined and implemented	Yes	The qualification procedure, both prior to the competition, and during the competition itself, is described in section 3.
Scoring		
Measurements and estimations are clearly identified	Yes	The scoring and metrics to be used throughout RAMI competitions is described in section 4.
Subjectivity is addressed in an appropriate way	Yes	The achievements to be checked and metrics to be assessed are designed to minimize the influence of referees.
Metrics are properly designed	Yes	The chosen metrics are quantitative and fair assessments of the performance of teams.

Reference section

Annexes

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252