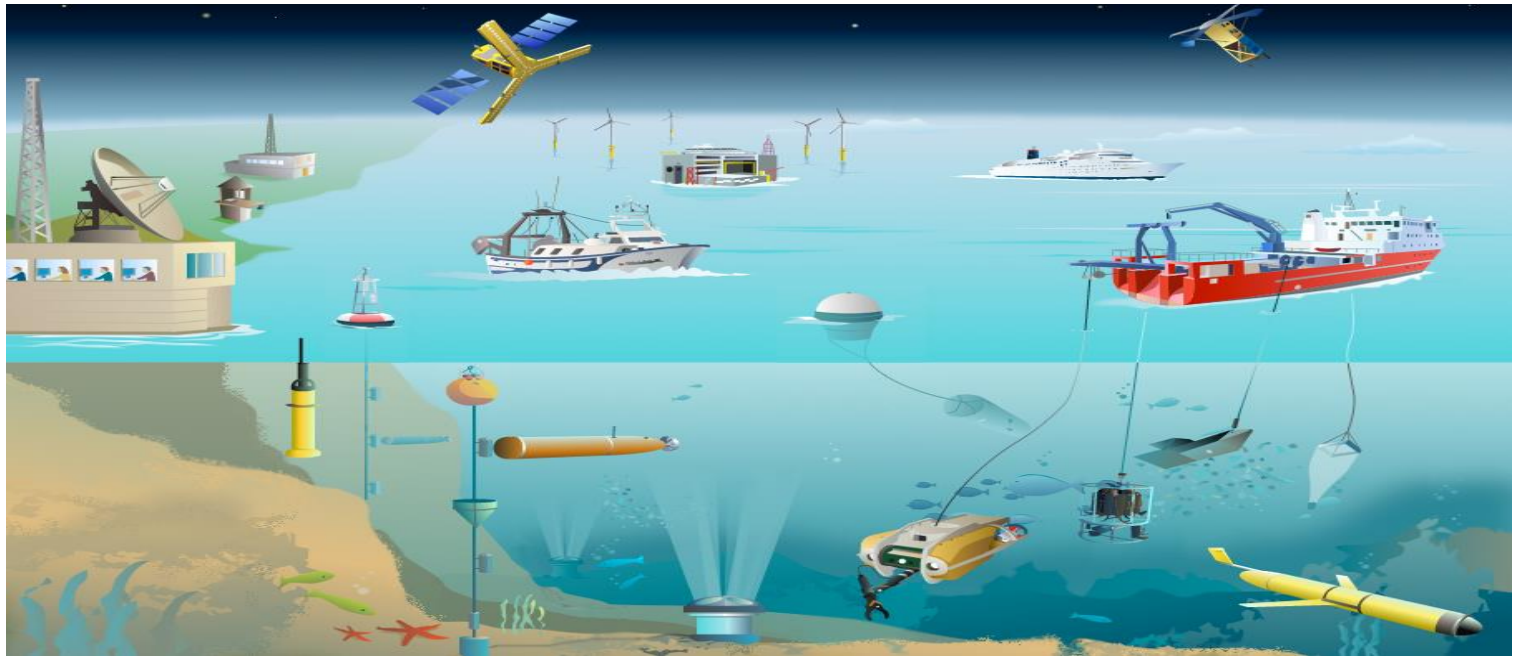
Foras na Mara
Marine Institute



Using Jupyter Notebooks as a Data Scientist “Work bench” for quality assurance of data processing flows and quality control of data series

Rob Thomas, Sarah Flynn, Will Meaney, Siobhan Moran, Ramona Carr, Adam Leadbetter

Foras na Mara
Marine Institute

OUR VISION

Our ocean wealth will be a key element of our economic recovery and sustainable growth, generating benefits for all our citizens, supported by coherent policy, planning and regulation, and managed in an integrated manner.

GOAL 1

THRIVING MARITIME ECONOMY

- Sustainable economic growth of our marine/ maritime sectors
- Increase the contribution to our national GDP
- Deliver a business-friendly yet robust governance, policy and planning framework

GOAL 2

HEALTHY ECOSYSTEMS

- Protect and conserve our rich marine biodiversity and ecosystems
- Manage our living and non-living resources in harmony with the ecosystem
- Implement and comply with environmental legislation

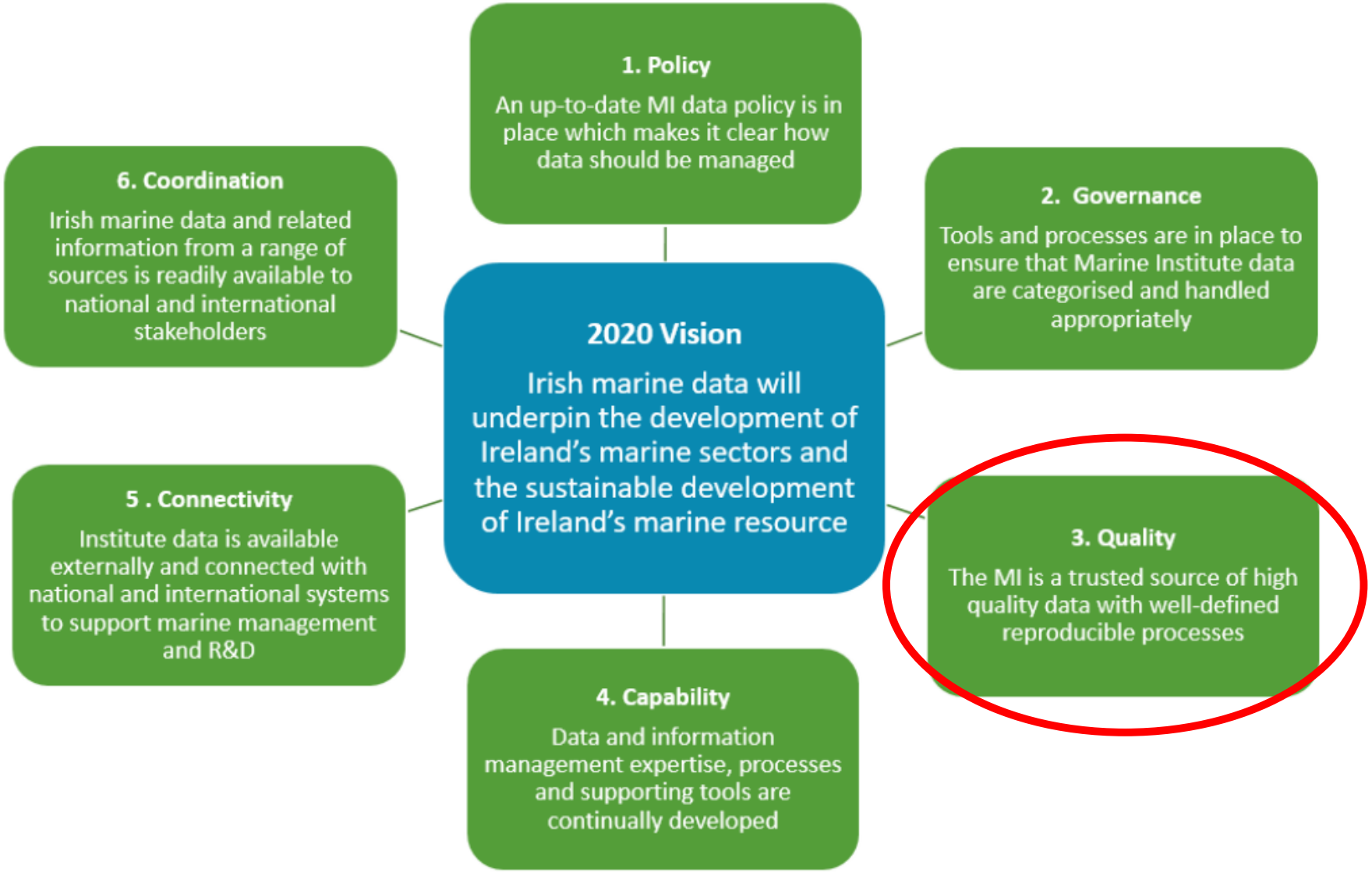
GOAL 3

ENGAGING WITH THE SEA

- Building on our maritime heritage, strengthen our maritime identity
- Increase our awareness of the value, opportunities and societal benefits
- Engagement and participation by all

Foras na Mara
Marine Institute

MI Data Strategy



Foras na Mara
Marine Institute

Why is quality important?

Decision making

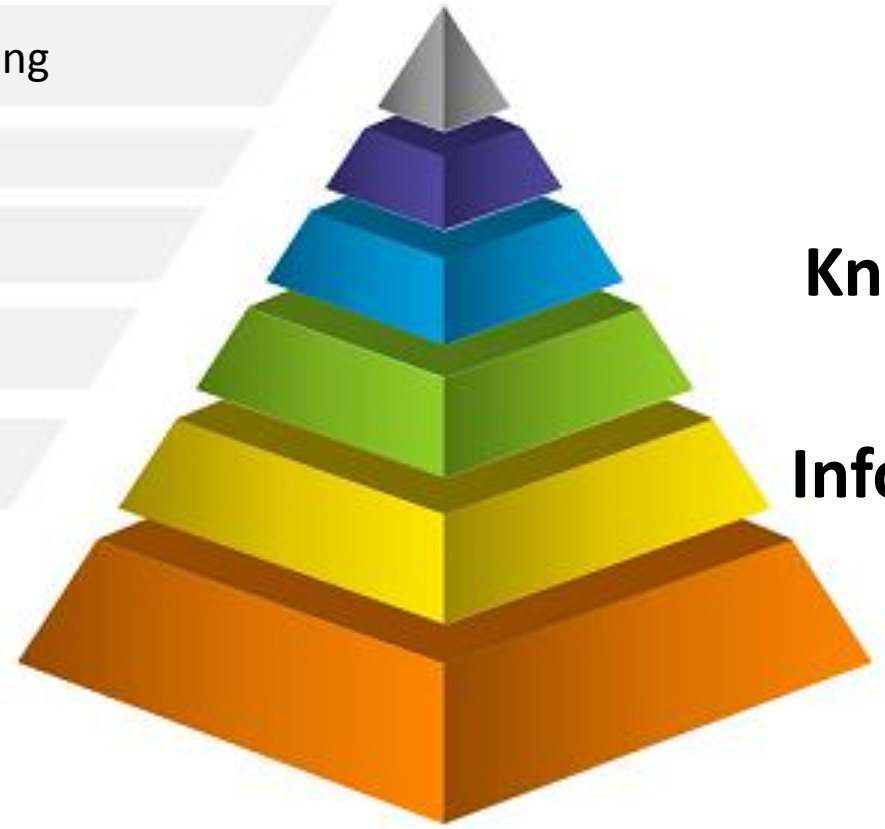
Synthesis

Analysis

Aggregation

Organisation

Collection

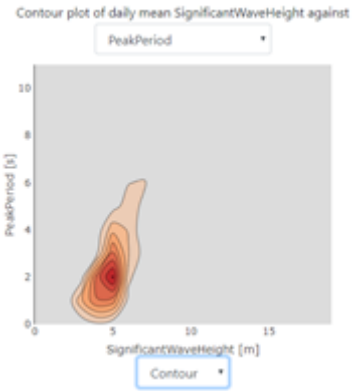


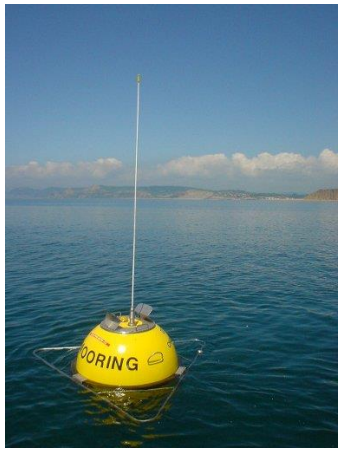
Policy

Knowledge

Information

Data



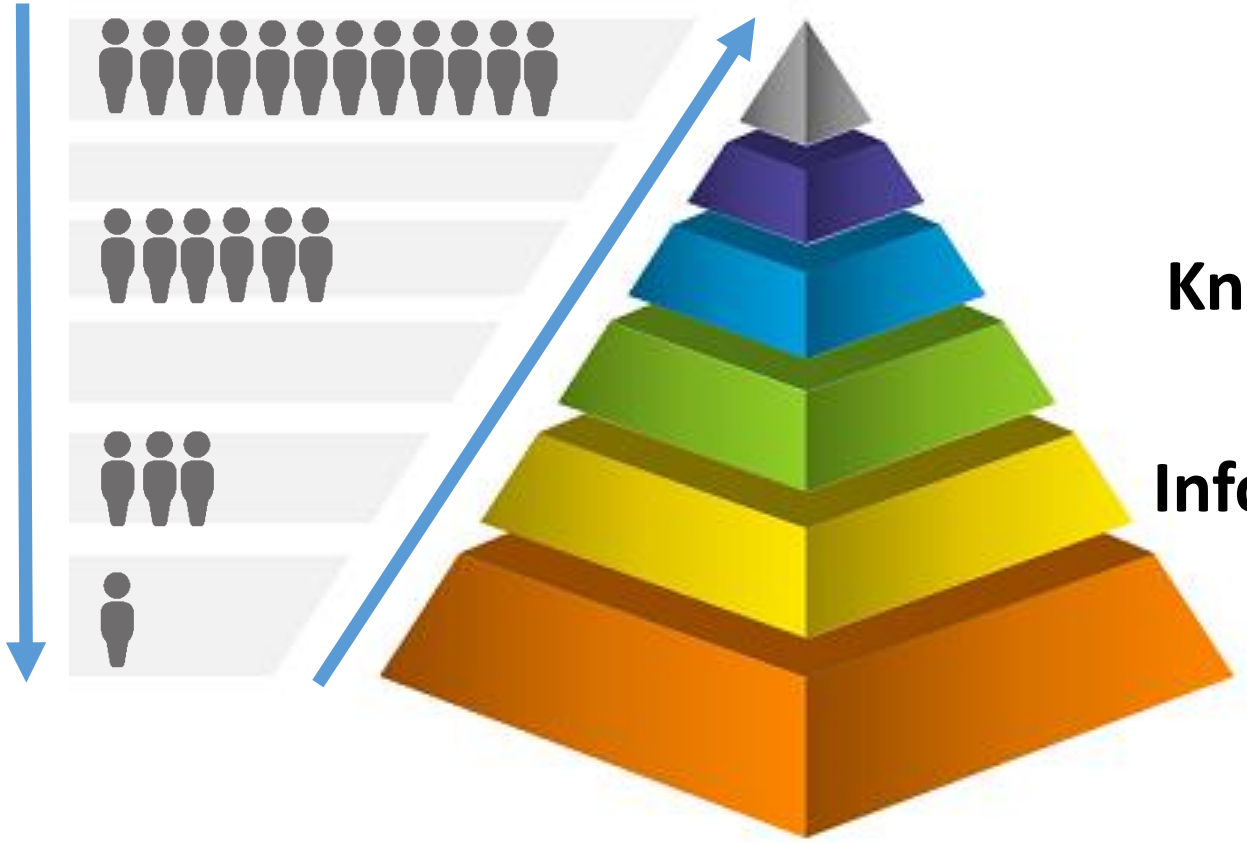


Foras na Mara
Marine Institute

Why is quality important?

Transparency

Trust

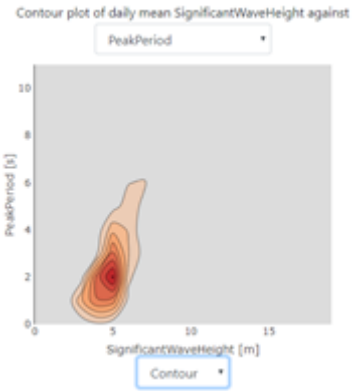


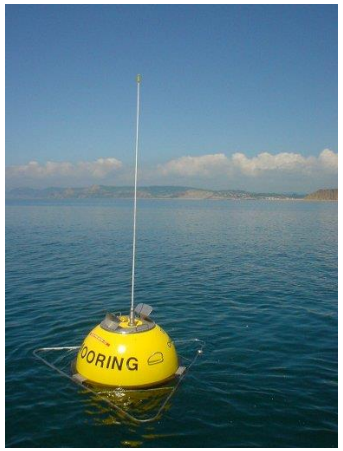
Policy

Knowledge

Information

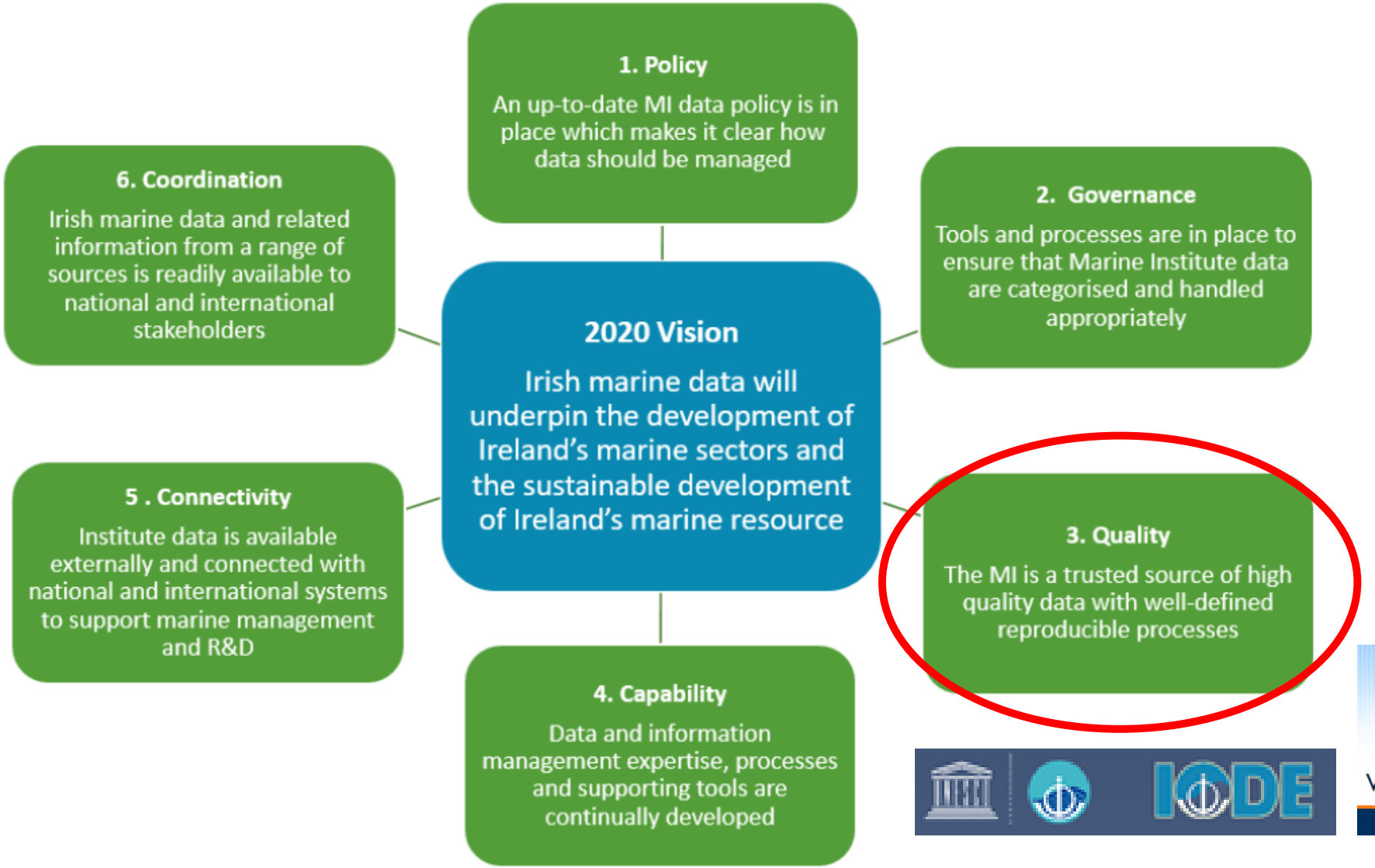
Data





Foras na Mara
Marine Institute

MI Data Strategy



European Maritime & Fisheries Fund

IODE

ICSU WORLD DATA SYSTEM

Quality Management Framework

- Data Technical Group created a QMF Working Group
- Data Co-ordinators and Stewards from:
  - Fisheries
  - Aquaculture
  - Marine Chemistry
  - Marine Environment
  - Oceanographic Sciences
  - Data management
- Corporate data out of scope
- IODE run course held September 2017 in Ostende, Belgium

QMF Principles & Benefits

1. Relationship management
2. Customer focus
3. Engagement of people
4. Leadership
5. Process approach
6. Continual Improvement
7. Evidence-based decision making

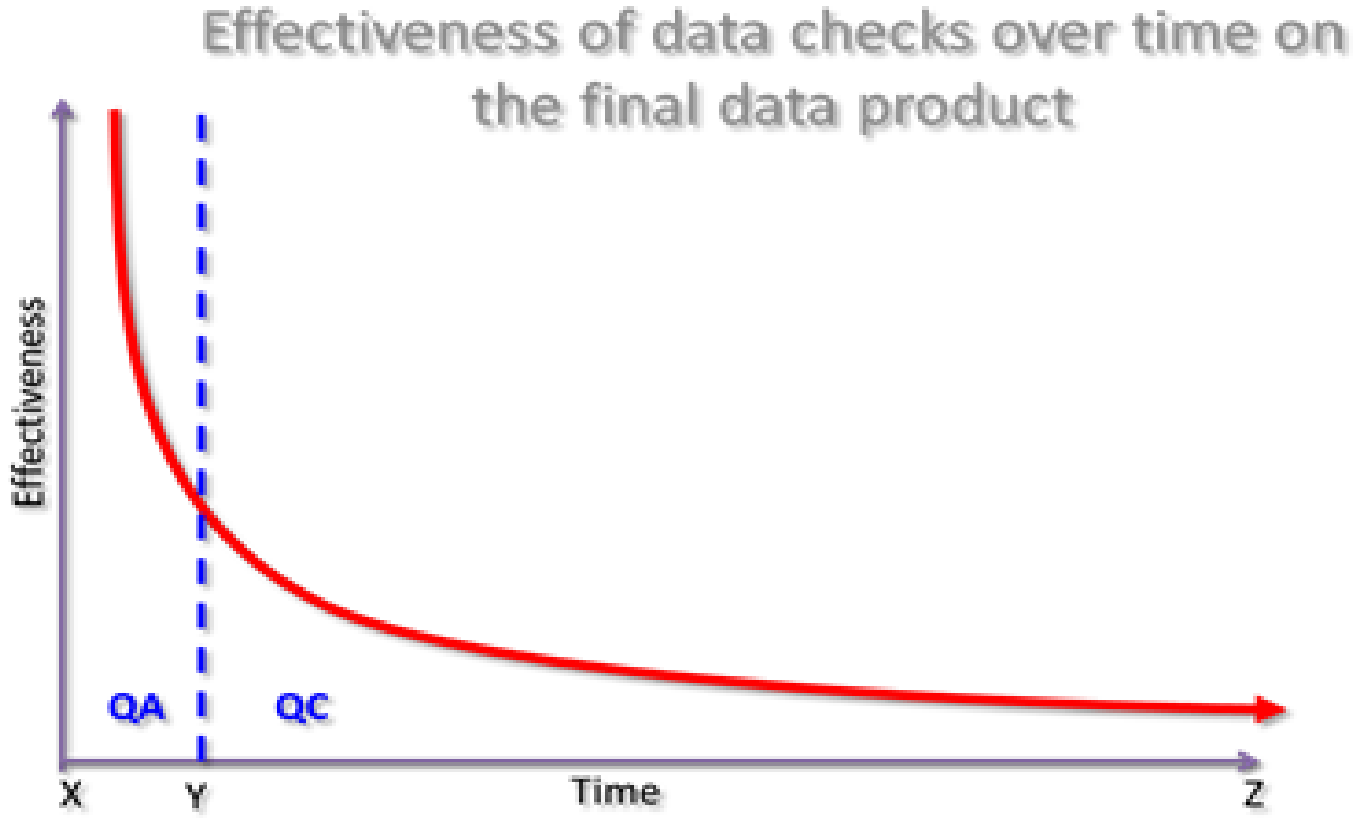
Assurance vs Control?

Quality Assurance

- Proactive
- Aim is preventing defects
- Focus on the process

Quality control

- Reactive
- Aim to detect and correct
- Doesn't consider the process



David Stokes (pers. comm.)

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

QA vs QC: a case study

16.8 hours per car
34 defects per 100 vehicles

All workers responsible for detecting defects during assembly process.

57 hours per car
78.7 defects per 100 vehicles

Separate QC team involved post assembly.

Womack et al. (1991)

Sutherland (2014)

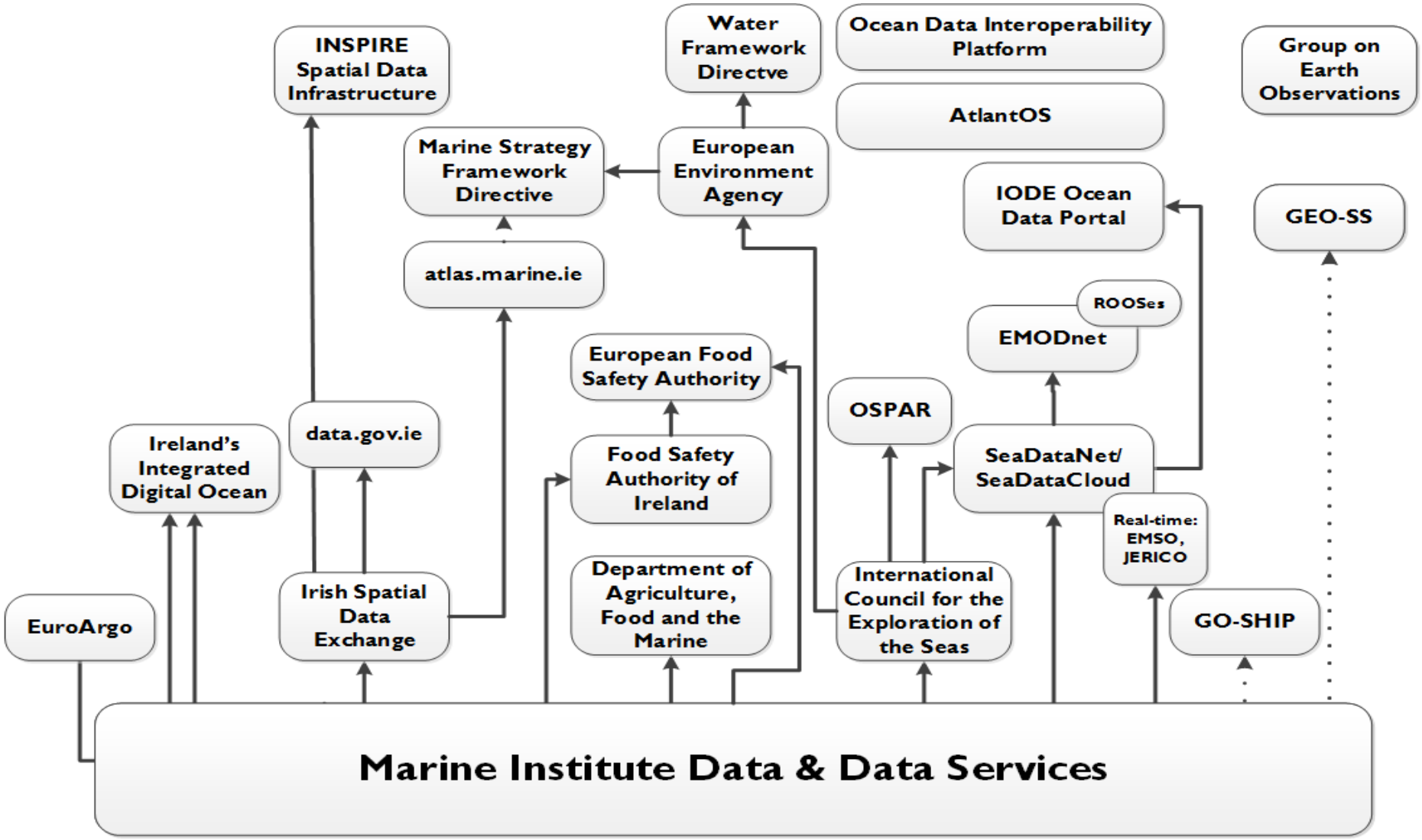
24x longer to fix an issue 3 weeks later

Today	3 weeks time
5 mins	2 hours
30 mins	~ 2 days
1 day	~ 1 month

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

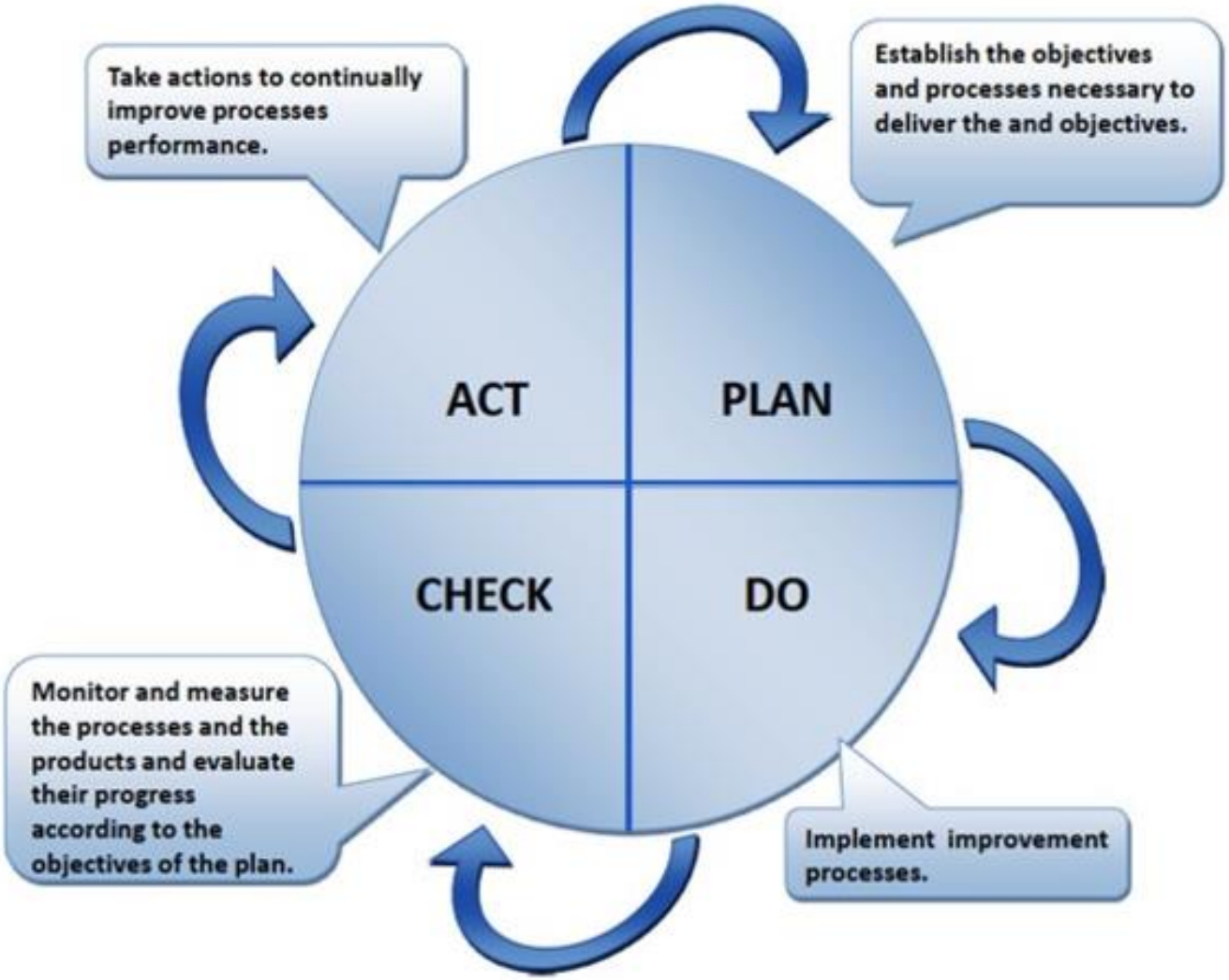
What journey do our data take?



Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

Quality Management Framework



Foras na Mara
Marine Institute

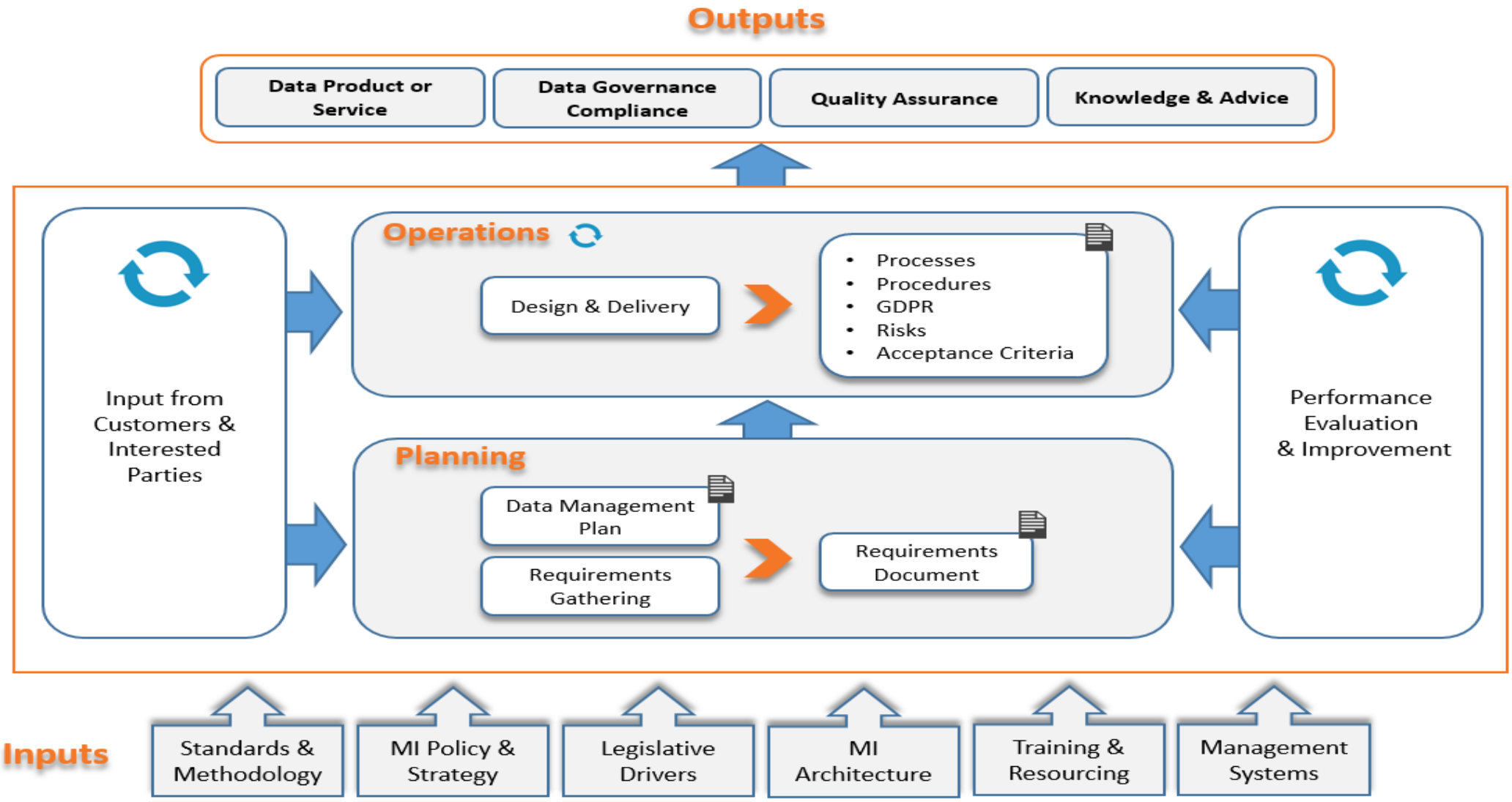
European Maritime & Fisheries Fund

MI QMF Manual

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

MI QMF Model



Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

MI QMF Implementation Pack

- Data Management Plan

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

MI QMF Implementation Pack

- Data Management Plan
- Requirements and User Acceptance Criteria

Foras na Mara
Marine Institute

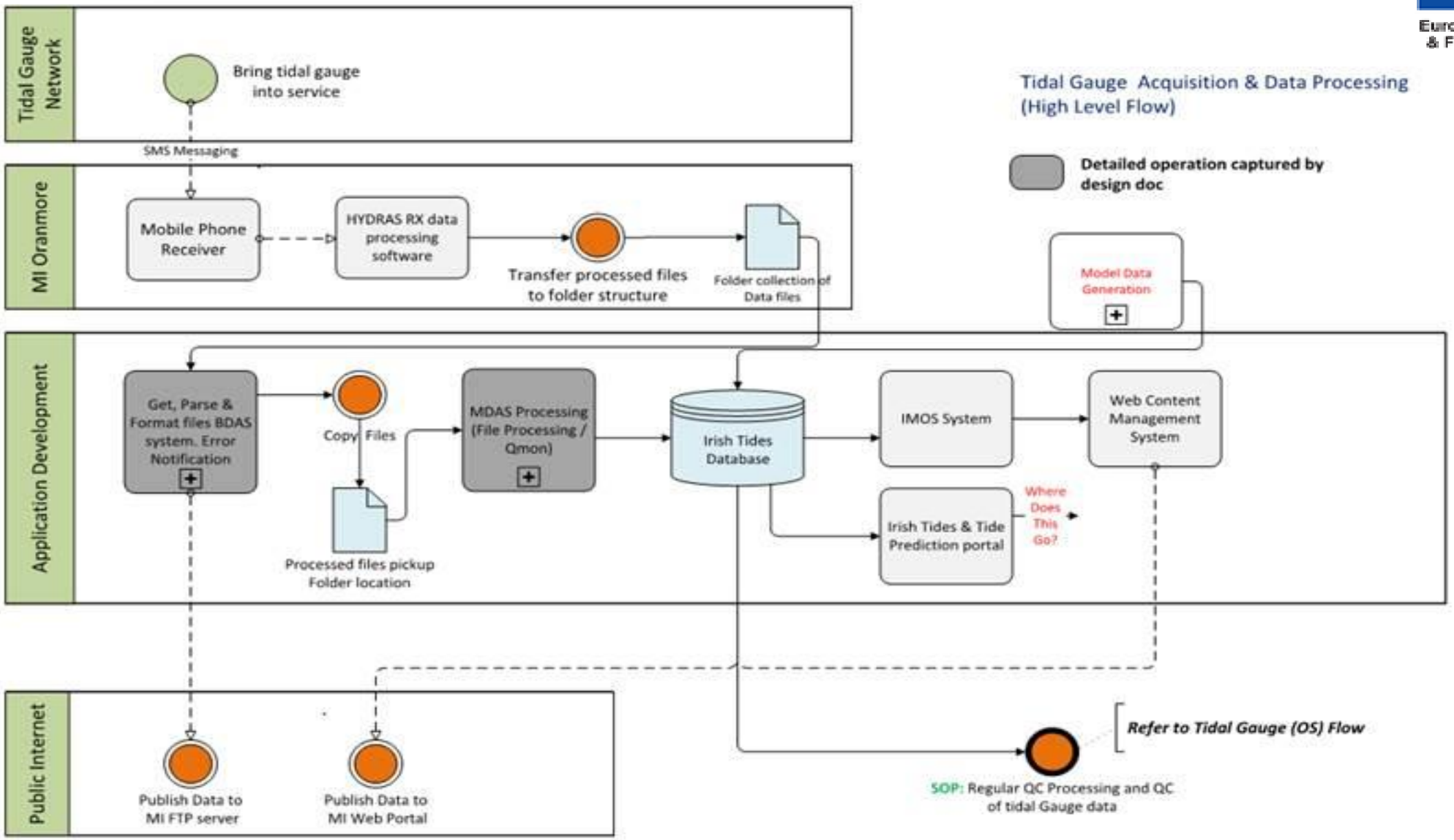
European Maritime & Fisheries Fund

MI QMF Implementation Pack

- Data Management Plan
- Requirements and User Acceptance Criteria
- Process Flow and associated information
  - ***What*** needs to happen to achieve the objectives
  - Flow diagrams preferable to pages of text.

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund



Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

MI QMF Implementation Pack

- Data Management Plan
- Requirements and User Acceptance Criteria
- Process Flow and associated information
  - ***What*** needs to happen to achieve the objectives
  - Flow diagrams preferable to pages of text.
- Standard Operating Procedures
  - ***How*** it will be done
  - Documentation should be active

Foras na Mara
Marine Institute

Data Requests

The following outlines the steps required to carry out the de
the online data request service page on the MI website

1. A data request is received by email from institute.n
   datarequests@marine.ie
2. The mail will contain a link to and details of the dat
   - Open https://www.marine.ie/Home/user and l
   - Click on the data request link in request email
   - Next check the detail of the request to see if th
     the online MI portals:
     - Marine Data Online – http://data.marine.ie
     - Marine Institute ERDAPP server - http://
     - Irelands Marine Atlas
     - Irelands Marine Renewable Energy Atlas
     - INFOMAR Web Map Viewer
   - If the data cannot be downloaded from any of
     needs to be sent to Data Steward and a licence
3. Click on 'Edit' tab on the top of the page. Change th
   name under 'Assigned To'.

4. Take note of the data request number in the forma
5. Open the Data Request Report Manager
   http://[redacted]/Report.aspx?It
   censeAgreement and use the request number to d

INSERT TABLE FROM DATA LICENCE

Please notify us if there are any constraints regarding publishing data as per conditions set out in Section "4. Publications" of this licence agreement.

Please note that all data delivered to the requestor must be delivered via the Data Request Service Team.
Also note if you are searching for previous emails they will be stored in the Data Request email folders and not your own.

Email the data requestor to acknowledge receipt of their data request through the MI Data Request Service. Attach the data licence agreement (pdf) for their signature. Use the following template:

From: datarequests@marine.ie

To: Internal Data Requestor

CC: datarequests@marine.ie

Subject: Marine Institute Data Request 2017/xxxx

The Marine Institute (MI) acknowledges receipt of your data request through the MI Data Request Service. We have taken your request details and shall pass them onto the relevant data steward in our organisation. We will be in contact with you in due course.

Our signed Marine Institute Data Licence Agreement is attached. We require this to be signed by you within the Acceptance of Agreement section above "For the Licensee" and returned to us before we release data to you.

Please note that if you intend publishing data supplied that the conditions set out in Section "4. Publications" of this licence agreement apply.

Note: The licence agreement needs to be signed by the data requester before the data delivered.

11. Once the data and signed licence agreement is received, the data is sent to the requester via data request: datarequests@marine.ie
    - If the data is too big to send via email then it can be sent via UFAnet File Sender.
12. Store a copy of the data and the signed licence agreement in.
13. Login to the data request page to close the data request and enter information to the following areas:
    - Status – Close
    - Area of activity – select as appropriate
    - Comments – use while data request is in progress to give any information that might be useful if another member of the team needs to pick up the request
    - Resolution – input what data was sent
    - Date drop down box: – When Licence Agreement was received
    - Date drop down box: When data was received
14. Each morning an email will be sent to the data request mailbox with a status of all requests which are still 'Open' or 'In Progress', to ensure we are tracking each data request.
    http://www.marine.ie/Home/open-data-requests
15. If a requester sends a query for data to your own personal mail, then you need to send them the online data request page to submit a data request via
    https://www.marine.ie/Home/marine-institute-request-digital-data

Wave rider data

- Monthly data is zipped from \\galwayfs03\WaveRider and transferred into \\galwayfs03\DataRequests\WaveRider
- The following documents to support the data are located on \\galwayfs03\DataRequests\WaveRider
  - DWR3_File_Details_v1
    - Galway Bay and West Wave – 2009 - 2012
    - All Belmulet data
  - DWR4_File_Details_v0
    - Galway Bay and West Wave – 2013 to date
- Data is sent through UFAnet or Filesender (depending on size)
  - The above documentation is provided to the data requester
  - Copy the link of where data is located and keep it in the data request ID folder, along with the signed license agreement.

European Maritime & Fisheries Fund

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

Jupyter Notebooks as SOPs

jupyter SCS_Underway_Overview Last Checkpoint: 10 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python [default]

Markdown

Marine Institute Foras na Mara

Ocean Science - Data Processing & QC notebooks

| Contents | Data Processing | Quality Control

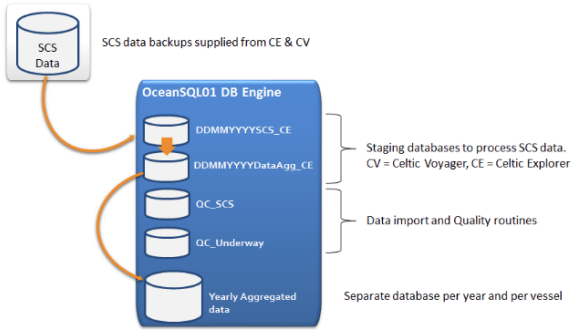
SCS Underway Data - Overview

This section captures the complete lifecycle of underway data for inclusion in the core Ocean Science datacentre. This process includes:

- Initial data collection.
- Extract Transform & Load (ETL) procedures.
- Quality controls.
- Final inclusion in the production Ocean Science datacentre.

Database Overview

The illustration below shows important databases used in the processing of Underway data. Each time a backup SCS database is received from a vessel, two new databases are created to process the data (SCS & DataAgg). Two additional databases are available to support data extraction and quality control stored procedures. Finally, the processed and aggregated data is loaded into a final yearly storage database.



jupyter SCS_Underway_QC Last Checkpoint: 04/26/2018 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python [conda root]

Markdown

Summary of Quality Check flags:

The following codes/flags are used to record the quality of each paramater under QC:

In [6]:
```
# Display all
df = pd.read_sql_query('''SELECT * FROM dbo.[tblflags]''',SCS_QC_engine)
df.head(20)
```

Out[6]:

	qc_flag_id	flag_description
0	0	No QC
1	1	Good
2	4	Failed physical limits
3	5	Failed local limits
4	9	Missing
5	13	Location falls on land
6	14	Location falls in port
7	15	Impossible location
8	16	Failed automated sounding check
9	17	Failed sea temperature check
10	18	Thermosalinograph fails comparison with Ti20
11	19	Failed visual sounding check
12	20	Failed air temperature check
13	21	Failed atmospheric pressure check
14	22	Failed salinity check
15	23	Failed fluoresence check
16	24	Failed visual check

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

MI QMF Implementation Pack

- Data Management Plan
- Requirements and User Acceptance Criteria
- Process Flow and associated information
  - ***What*** needs to happen to achieve the objectives
  - Flow diagrams preferable to pages of text.
- Standard Operating Procedures
  - ***How*** it will be done
  - Documentation should be active
- Performance Evaluation template

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

QMF - Additional Benefits

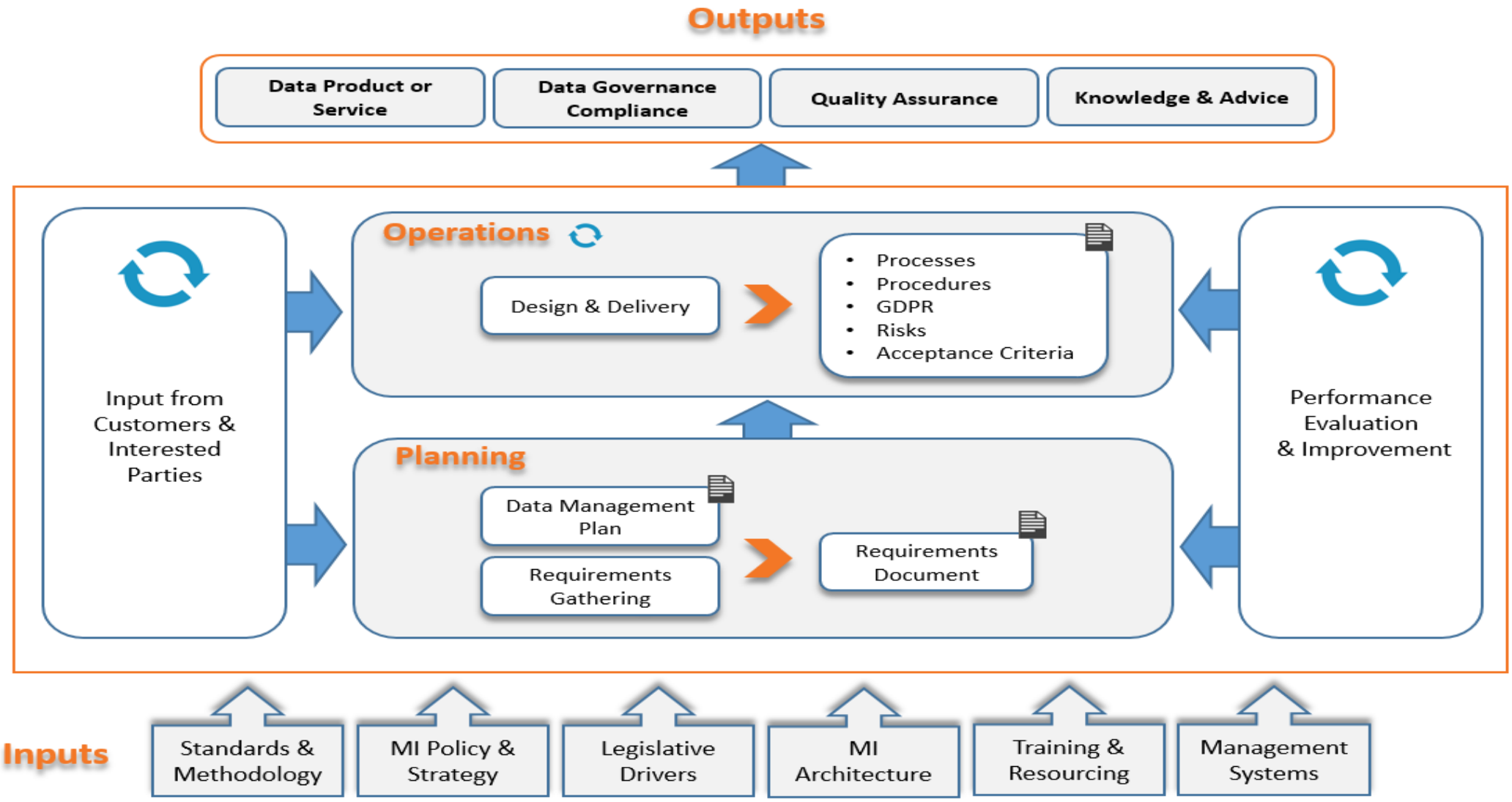
- Resources for new starters.
- Security around our digital data assets
- Understand our data protection obligations [GDPR].
- Business continuity and contingency planning.
- Transparency.
- Clear ownership and accountability for staff.
- IODE accredited NODC status…

submitted for review October 2018…

Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

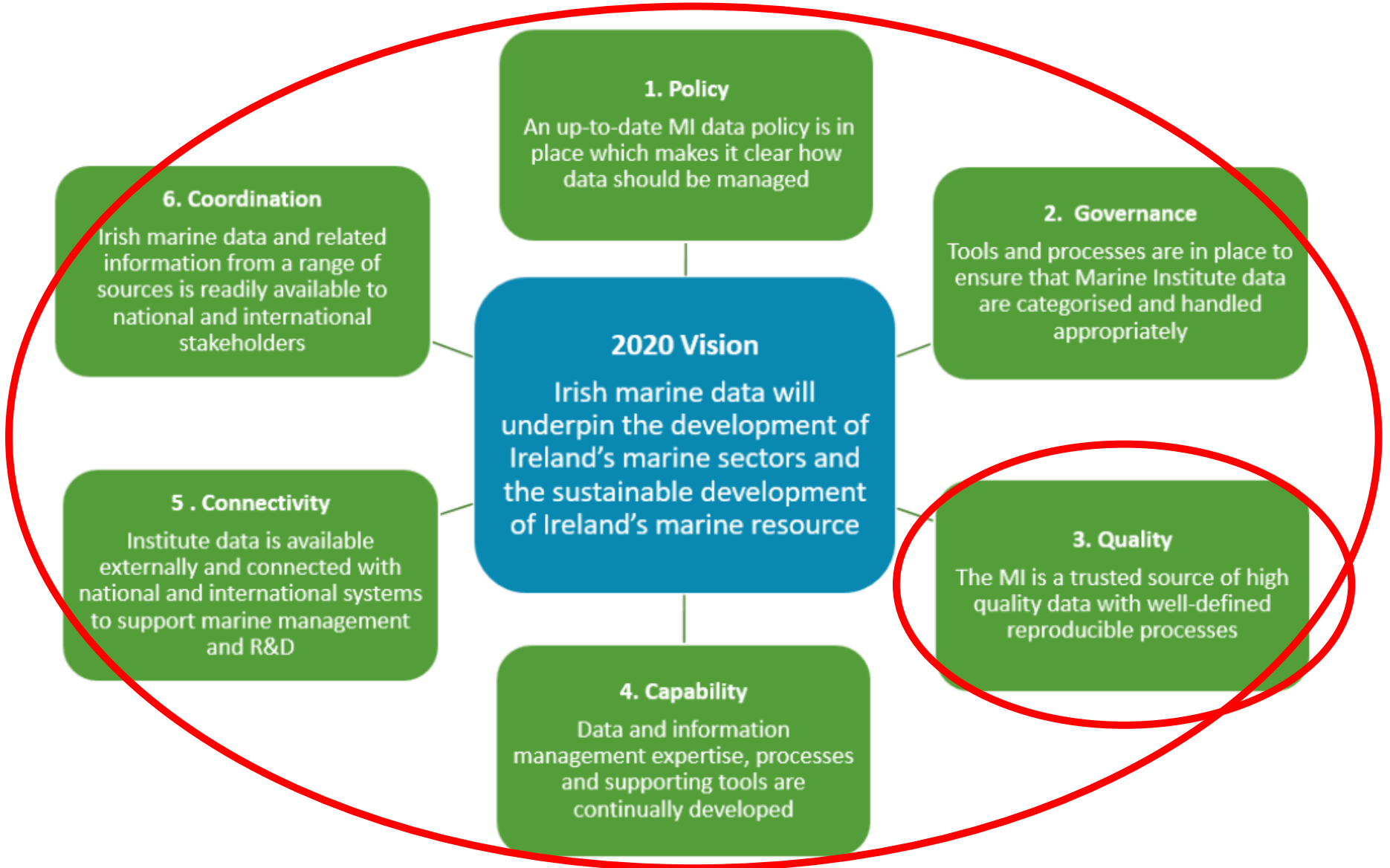
MI QMF Model



Foras na Mara
Marine Institute

European Maritime & Fisheries Fund

MI Data Strategy



Where next?

Foras na Mara
Marine Institute

Datacite DOIs for workflows

DataCite - International Data Citation

10.1	resourceTypeGeneral	1	The general type of a resource.	*Controlled List Values:* Audiovisual Collection DataPaper Dataset Event Image InteractiveResource Model PhysicalObject Service Software Sound Text[18] Workflow Other See Appendix for definitions and examples.

Foras na Mara
Marine Institute

Review and streamline processes

- Identify bottle necks
- Provide tools to help
  - Data stream visualisation screening tool for Galway Bay Cable Observatory
    - <200 lines of Python
    - Bokeh and Pandas toolboxes

GBCO CTD | GBCO fluorometer

Tide gauges | Wave bouys

Hide bad data

Time averaging: 0

x-axis

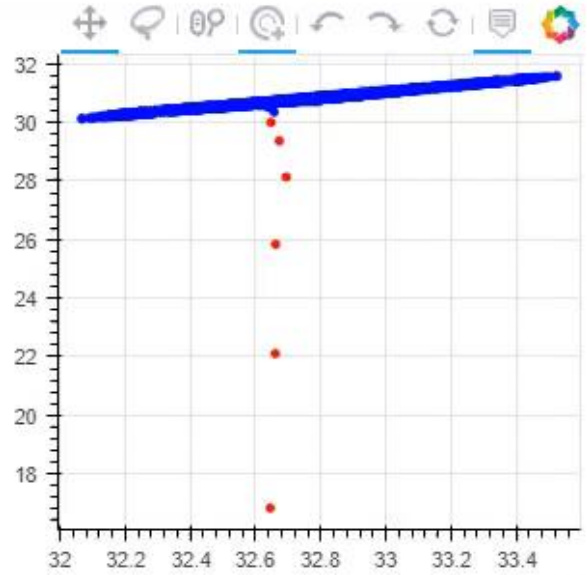
cond

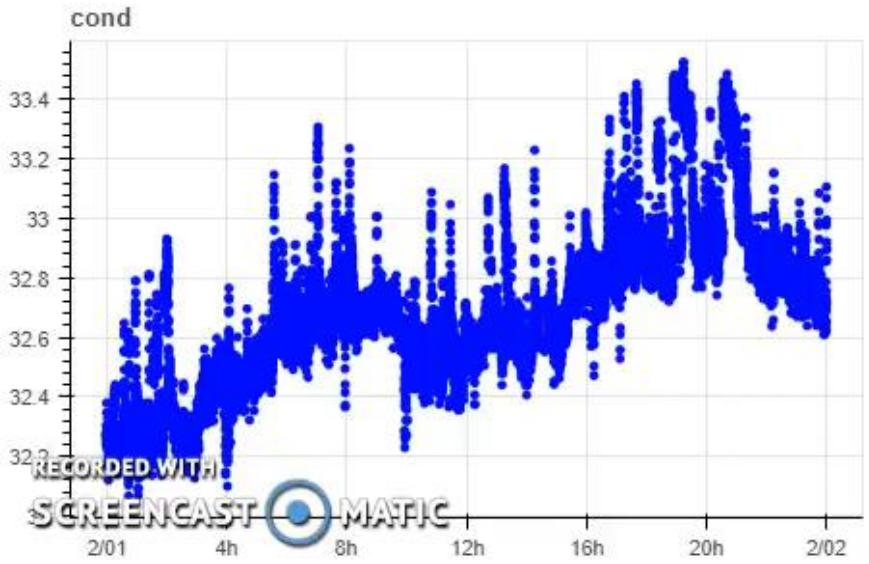
y-axis

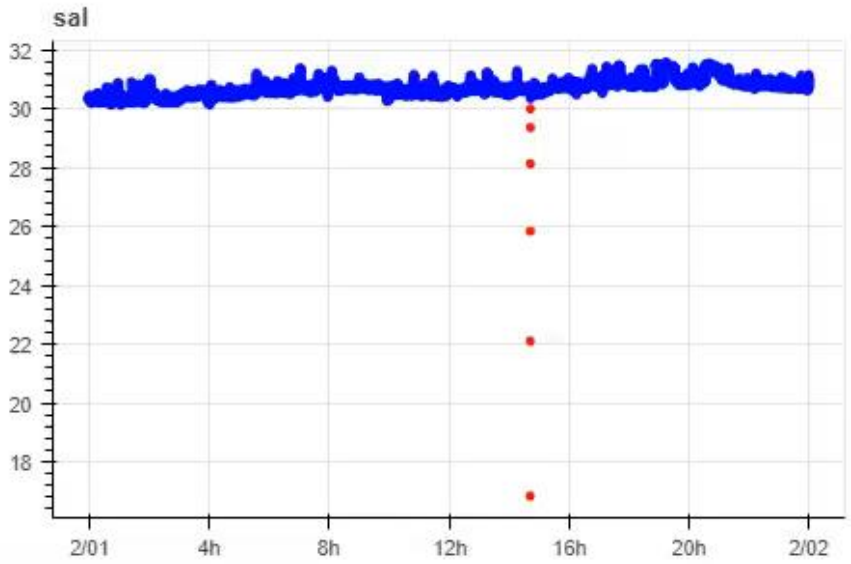
sal

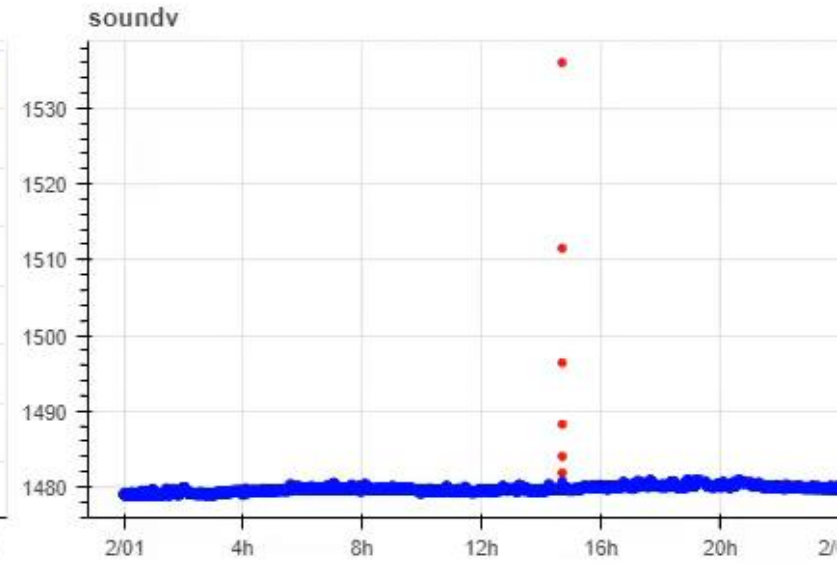
Flag selection

Save flags





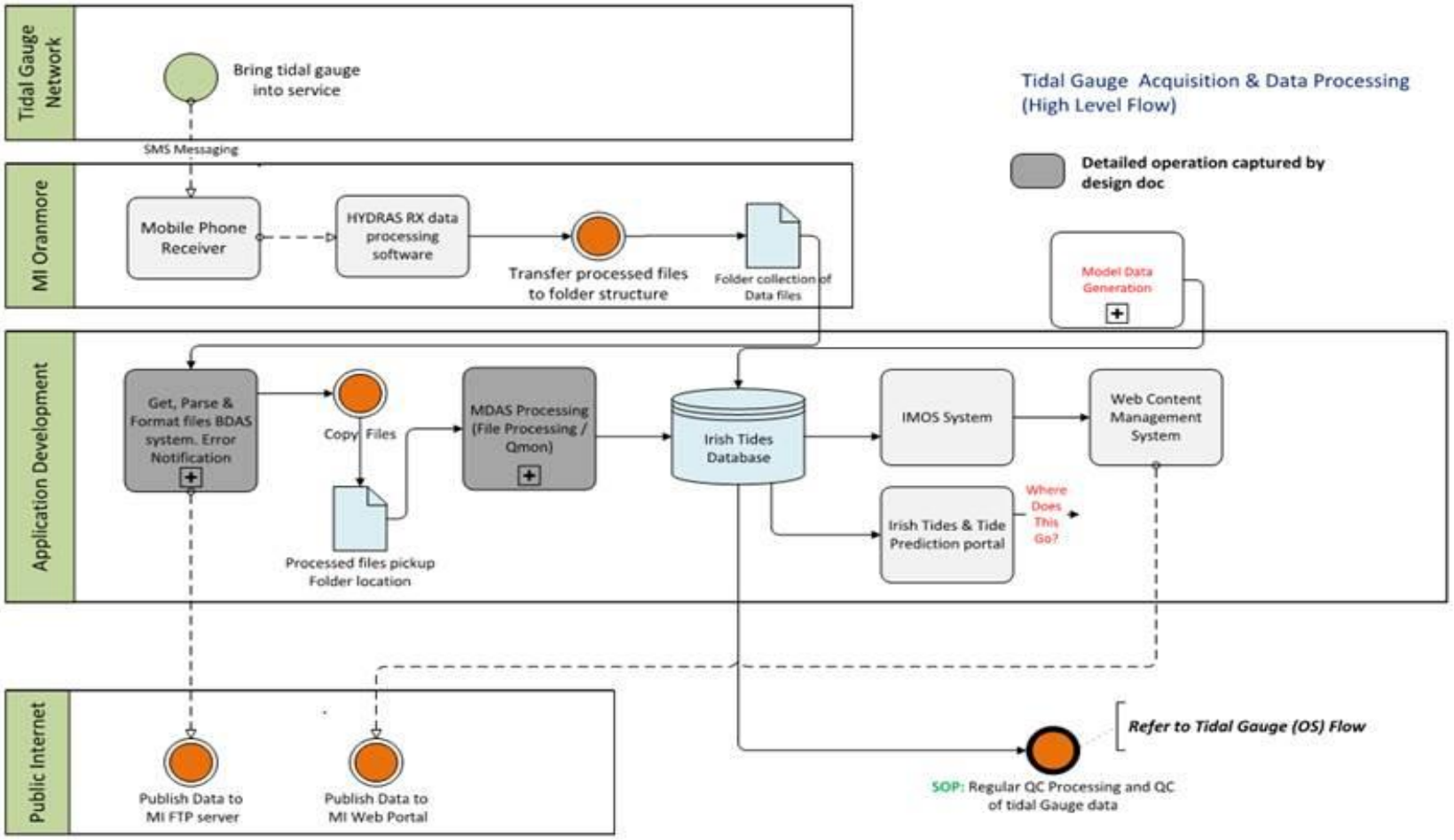


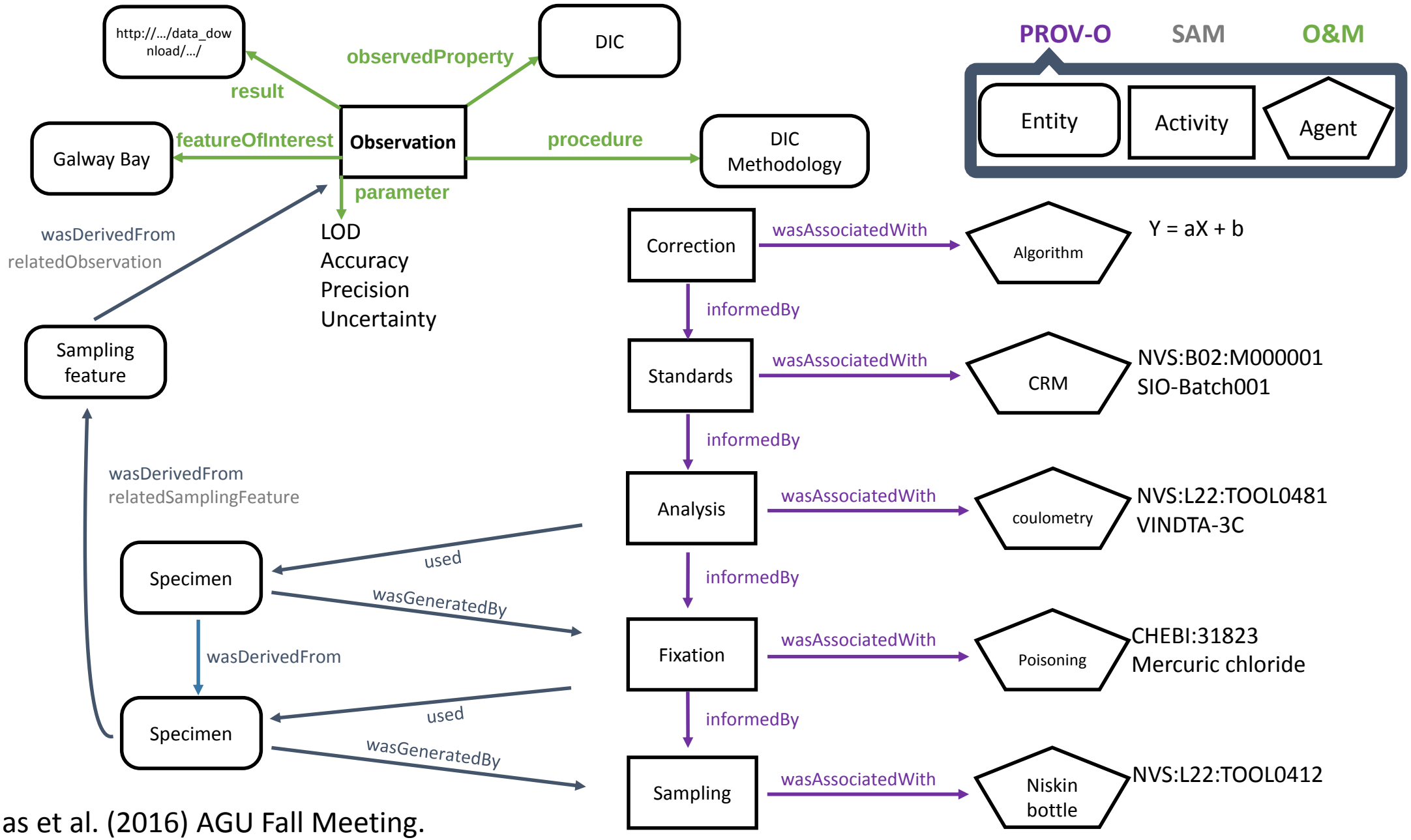


RECORDED WITH SCREENCAST-O-MATIC

Foras na Mara
Marine Institute

Enhanced metadata?





Thomas et al. (2016) AGU Fall Meeting.

Thank you to Greg and Claudia @ IODE