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

Rob Thomas, Marine Institute (Ireland), rob.thomas@marine.ie
Sarah Flynn, Marine Institute (Ireland), sarah.flynn@marine.ie
Will Meaney, Marine Institute (Ireland), will.meaney@marine.ie
Siobhan Moran, Marine Institute (Ireland), siobhan.moran@marine.ie
Ramona Carr, Marine Institute (Ireland), ramona.carr@marine.ie
Adam Leadbetter, Marine Institute (Ireland), adam.leadbetter@marine.ie

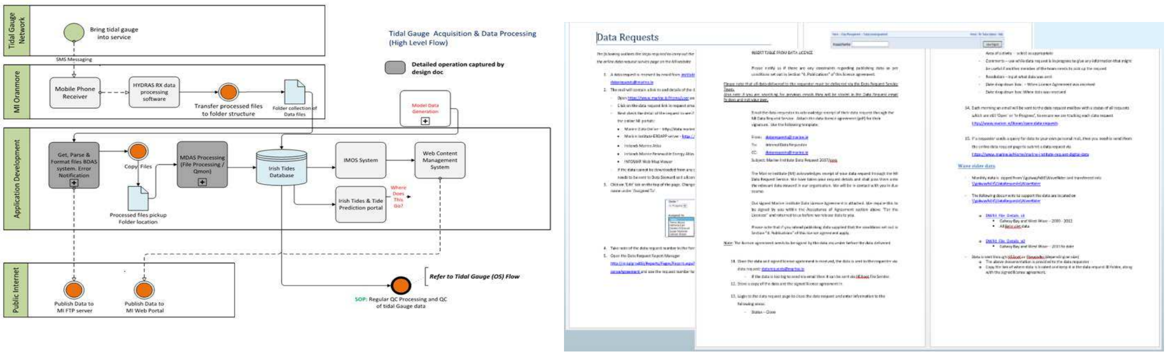
Abstract.

Marine data are increasingly being utilised for purposes or used by organisations beyond those intended at collection. In determining appropriate use, quality assurance is given greater importance to provide visibility of the processes used in generation, QC, curation and serving of data. This presentation describes the process of creating and implementing a Data Quality Framework in a cross-discipline organisation, showcases examples and benefits of using Jupyter notebooks as a data scientist workbench, also demonstrating the implementation of open source tools to provide an improved experience for scientists involved with data QC.

Background.

The Marine Institute (MI) has published its Strategy for 2018-2021 and the data it produces and manages are clearly identified as a “Strategic Enabler” for the Strategic Focus Areas. Supporting the role of data as a Strategic Enabler in the MI Strategy, the MI Data Strategy is being implemented with the vision that “Irish marine data will underpin the development of Ireland’s marine sectors and the sustainable development of Ireland’s marine resource”. Quality is one component in achieving this vision.

In order to meet the Quality element of the MI Data Strategy, a Working Group has been tasked with drafting and then implementing a Quality Management Framework for the environmental data collected by the MI. The IODE has recently introduced training to support data centres producing and implementing a Quality Framework. As a member of the IODE NODC network official recognition of the MI for its data management practices can be achieved by achieving accredited NODC status with the IODE. This then confers accreditation by the World Data System (WDS) through IODE’s membership of the WDS.


Tidal Gauge Network
MI Oranmore
Application Development
Public Internet
Bring tidal gauge into service
SMS Messaging
Mobile Phone Receiver
HYDRAS RX data processing software
Transfer processed files to folder structure
Folder collection of Data files
Tidal Gauge Acquisition & Data Processing (High Level Flow)
Detailed operation captured by design doc
Model Data Generation
Get, Parse & Format files BDAS system. Error Notification
Copy Files
Processed files pickup Folder location
MDAS Processing (File Processing / Qmon)
Irish Tides Database
IMOS System
Web Content Management System
Irish Tides & Tide Prediction portal
Where Does This Go?
Publish Data to MI FTP server
Publish Data to MI Web Portal
Refer to Tidal Gauge (OS) Flow
SOP: Regular QC Processing and QC of tidal Gauge data

Data Requests

INSERT TABLE FROM DATA LICENCE

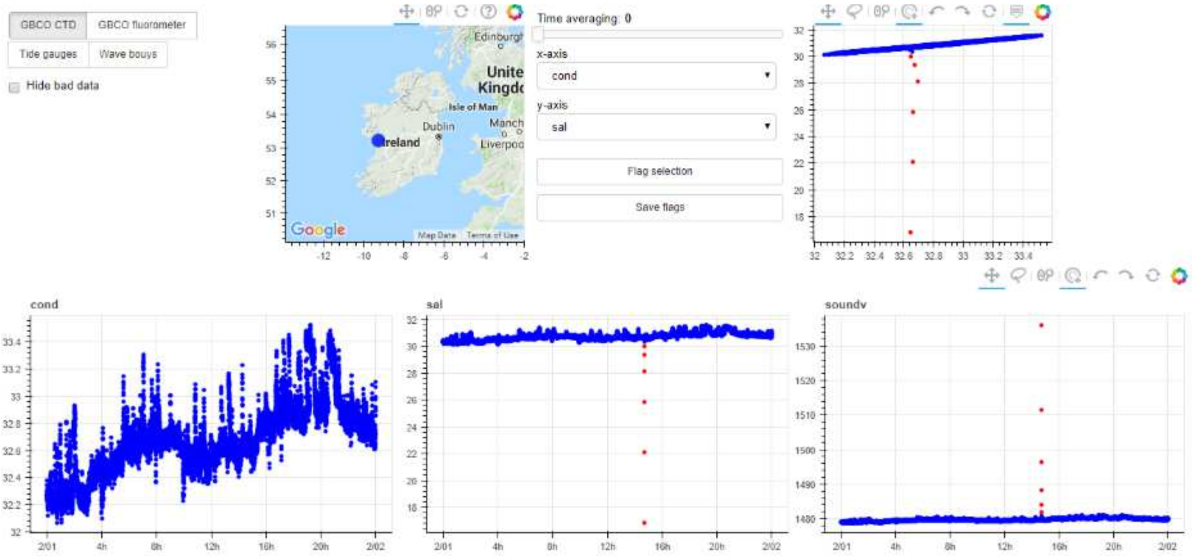
Wave rider data

Status – Close

Figure 1: Examples of a process flow documented in the Business Process Management Notation and documented SOP.

The MI has chosen to follow the IODE Quality Framework guidelines, with the structure following the ISO9001:2015 standard, and is documenting its processes through process flows to describe the high-level “what” a processes does and documenting standard operating procedures (SOPs) to describe “how” a process should be carried out.

For the MI Oceanographic Sciences section Jupyter Notebooks have become a key tool in the documentation of scripted workflows and are being used to implement a “Data Scientist Workbench” for the different data types handled by the section. The notebooks provide an interactive tool that allow users with limited scripting experience to work through a defined process or analysis with a consistent approach, while capturing the expert judgement decisions that need to be made on a case by case basis. These notebooks allow reproducibility with the flexibility to provide a mechanism to document different runs of a process (e.g. underway data processing documentation on a cruise by cruise basis). In this way they provide a great option when a process requires inputs that prevent the end-to-end process being automated.



□ Hide bad data

Time averaging: 0

x-axis

cond

y-axis

sal

Figure 2: Example of data screening GUI implemented using interactive visualisation toolbox.

By utilising open-source toolboxes available for use with Jupyter notebooks, data visualisation can be incorporated as a seamless part of the workflow and using interactive plotting toolboxes (e.g. Bokeh for Python) enables screening of data streams through a GUI. One advantage of such tools is that dynamic manipulation of high resolution data streams allows a scientist to see the impact of screening/flagging choices on binning of data by downstream users.

In providing clear processes, procedures and tools in support of a Quality Framework data this supports MI staff in meeting the Quality element of the MI Data Strategy and recognises the value of MI data as a strategic enabler in its future strategy. IODE accreditation also provides an external acknowledgement of the quality assurance for data collected and curated by the Marine Institute.