Gladstone Healthy Harbour Partnership DIMS

ISP009-2019-20: Provision of ongoing technical support for GHHP Data and Information Management System

Final Report

30th June 2020

Author: Aaron Smith, Eric Lawrey, Murray Logan, Marc Hammerton, Gael Lafond, Australian Institute of Marine Science

Project team: Eric Lawrey, Murray Logan, Marc Hammerton, Gael Lafond

Date: June 2020

Version: 3

Final report

The aim of the ISP-009 DIMS 2019-20 maintenance project was to ensure that the GHHP DIMS report card system was maintained and updated to allow the production of the GHHP Gladstone 2019 report card. This included a range of maintenance tasks, one-off tasks to enhance the system and general assistance with the GHHP website. The maintenance tasks were:

- To ensure that the servers kept running and were secure (with Operating System patches). This included updating the DIMS Pydio system (see Task 6 for more information) to the newest version.
- Finalising the integration of changes to the social and economic scripts into production.
- Priority bug fixes/changes were made to the system. This included fixing 2016 report card scores and trend-plots confidence bounds and removing the environmental scores 2014 from the trend plots.
- Assistance was provided to data providers, where necessary.

The project also included one-off tasks to further expand the capability of the report card scripts. These included:

- Changes to the water quality script regarding Limits of Reporting (LoR) and total vs dissolved metal values (Task 2 and Task 3).
- Integration of mangrove indicator by translating the algorithms into R as part of the environmental scripts (Task 4).
- Integration of two fish health scores into the fish and crab indicator (Task 5).
- Adjustments to new names for PCIMP sites and Sense of Place indicators (Task 8).

Details of each of these tasks is outlined below.

Background

The GHHP Data and Information System is a system for managing the generation of the GHHP Gladstone Report Card. It aims to fully automate the processing of raw or near raw data through to the final scores and grades of the Report Card. This automation helps ensure that the provenience of the data and subsequent processing is fully recorded making it possible to track back from the Report Card scores and grades to the original data.

The initial Gladstone Report Card System was developed from 2014 – 2016. During this time the system and the Gladstone report card indicator structure and logic was also developed. A pilot report card was released in 2014, followed by a more complete version in 2015. Each year there have been new indicators added to the report card and the DIMS Report Card System. In 2016 the DIMS Report Card System and associated scripts were complete and integrated, but not in time for the published report card. In 2017 the Gladstone report card was produced from the results generated by the DIMS – Report Card System for the first time.

The GHHP DIMS is now a maintenance project with only small aspects being adjusted each year, typically to include new indicators that have been developed.

The GHHP DIMS consists of off-the-shelf software components (Document Repository) and customs pieces of software (DIMS Report Card System, the analysis scripts and the reporting scripts). The structure of this system is shown in Figure 1.

Data Providers upload their data into the Report Card System, these are then processed by the analysis scripts to calculate the scores and grades from the raw data. These scores and grades are then converted to a series of intermediate reports (in Word format) to be reviewed by the GHHP science team and the Independent Science Panel (ISP). Once reviewed and any issues with the data are resolved, the final documents are sent to the communication team to be presented via the public website. In addition to this a Technical report is produced by the GHHP science team. To assist in this process the Report Card System generates a template of the Technical report (in Word format), with pre-populated tables and graphs. The GHHP science team then add synthesis text explaining the results and the report card process in detail.

While the *Content Management System* for the public website is hosted on the same server as the *DIMS – Report Card System* and the *Document Repository* the two sections (as indicated by the dashed line in Figure 1) are managed by different teams. AIMS manages the server itself, the *Document Repository*, the *DIMS Report Card System* and its associated scripts. The *Content Management System* is managed by a third party contracted by GHHP.

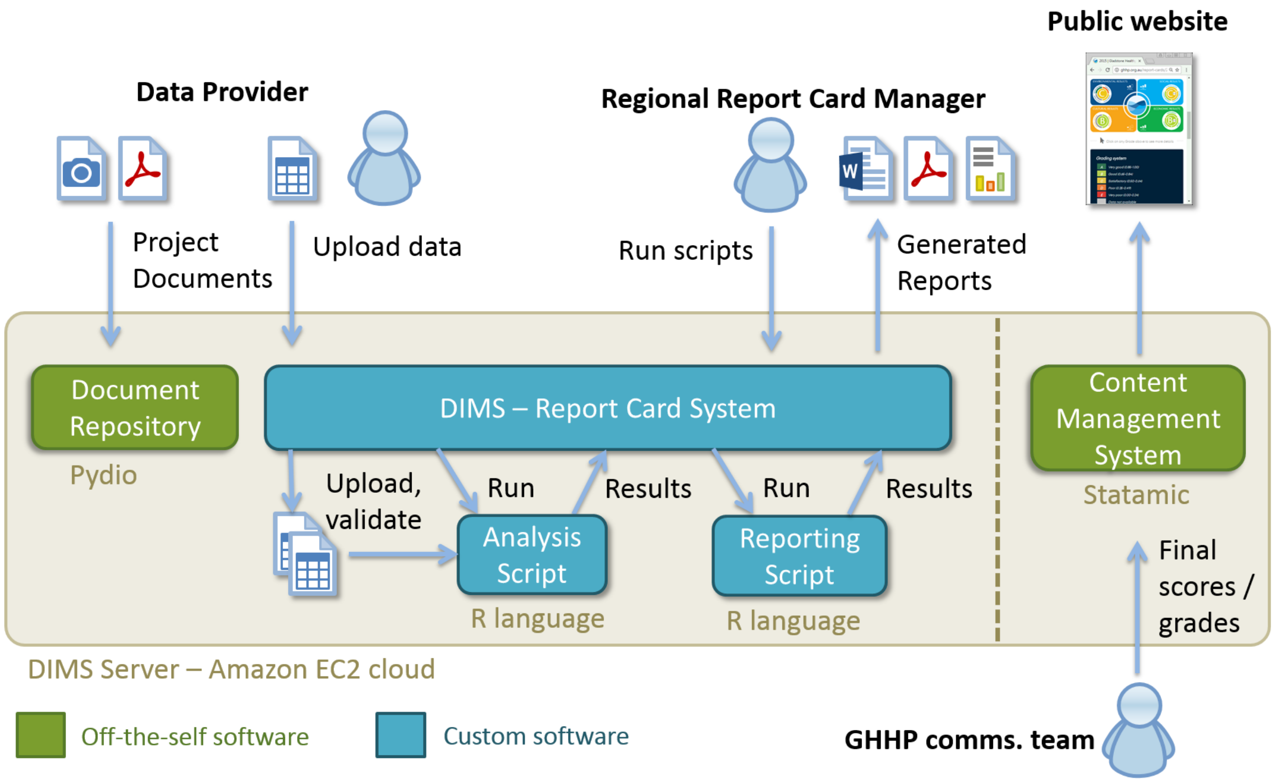


Figure 1. Structure of the GHHP Data and Information Management System (DIMS). Note: The Content Management System is run on the same server as the DIMS but is not part of the DIMS.

Routine tasks

Task 6: Server Administration

The aim of this task is to maintain a secure server for the GHHP report card products. This task may involve,

- *Ensuring that the DIMS server remains online,*
- *Ensuring that the DIMS server is backed up and security patches are routinely applied to the Operating System and other system software including Apache, Tomcat, Pydio, etc.*
- *Routine testing to ensure that updates and patches do not affect the performance and operation of the server*
- *Making improvements to the DIMS Report Card System necessary to improve its overall robustness (priority bug fixes)*
- *Any minor updates to the existing R scripts to work seamlessly with the newly added indicators. Following three minor updates are required for the 2018 report card. If any further minor updates are identified during the generation of report card scores and outputs these will be communicated to the project team.*

Any adjustments or bug fixes will be prioritised and those that fit in the small pool of time available (3 days) will be completed. If there is any remaining time, then it will be used to further improve the system documentation or alignment between the generated reports and the previous year's published Report Card technical report.

Routine Operating System Patches

Completed June 2020

Routine Operating System security patches were applied to the server on the Monday of each week. These ensure that all the packages on the server are up to date with the latest security patches. The automated backup system of the server (snapshots in Amazon EC2) was tested during the year by starting one of the backup images to ensure that they would run.

Pydio upgrade

Completed September 2019

Pydio is an open source web-based file sharing application. In the DIMS Pydio is used to manage user accounts and their permissions in the system. It is also used for storing unstructured data associated with the report card and documentation about the DIMS. As part of the server maintenance Pydio was upgraded from version 8.0.2 to the latest version 8.2.3. This upgrade also required adjustments to the DIM Report Card System to keep it aligned with changes to Pydios' user management Application Programming Interface (API).

Assisting with integrating changes in the social and economic scripts into production

Completed September 2019

Several adjustments to the social and economic scripts where submitted by Dr Jeremy De Valck as a list of changes in a MS Word document. We reviewed the changes from the document and integrated them into the production code. The changes were tested and pushed into production.

Fixing 2016 report card scores

Completed November 2019

The 2016 report card code in the DIMS is a reverse engineering of the scripts that were used to create the actual official report card. Unfortunately, this version of the scripts on the DIMS did not produce the scores for the indicator groups and above and this was not noticed until 2019. This results in a gap in the Trend Plots for the 2016 year. It does however generate all other scores.

To resolve this issue, we attempted to identify the bug in the 2016 code. We were not able to identify the bug in the available time and so it is still in this code base. This version of the environmental script is stochastic and so multiple runs of the script result in different scores (due to the bootstrap technique). As a result, even if the bug was fixed, we could not reproduce the published scores exactly.

The official generated scores.csv file from Nov 2016 was then tracked down and checked against the published technical report. The 56 entries in Table 1 of the 2016 technical report were verified against the scores file. This scores file is spliced in here as if the script generated it, allowing the scores and uncertainty estimates to be integrated into the DIMS state.

To put this in effect:

1. The original GHHP_write_scores.R file was copied to GHHP_write_scores_2016.R for any future reference.
2. The GHHP_write_scores.R file was modified to copy the 2016 scores file to the output scores location, as though the environmental scripts had generated it.
3. The 2016 scores file was added to the scripts directory, ready for copying.
4. The 2016 report card was unpublished and rerun to inject the official scores.
5. The 2016 report card was then republished, with this explanation added to the report card description.

Some scores were missing in the first run. The missing scores cannot be overwritten, therefore cannot exist in the system's state (the system ignores them while creating its state file).

A file named "input/hacked_scores.csv" was created, containing the missing scores which get combined to the script output on the first run of the script. The use of the term “hacked” was chosen specifically to signify that this is a workaround and is an accurate use of the term.

On the second run of the script, the outputs of the script are ignored and replaced with the scores file.

Fixing trend-plots confidence bounds

Completed November 2019

The trend plots script was using the variance to calculate the confidence bounds. This is a perfectly valid technique, but in the case of the environmental script, the output score file contains an upper and a lower value which more precisely represent the confidence level.

The system was modified to handle those new statistical variables and make them available to the trend plots. The state of previous report cards was also modified to contain the new variables, without re-running the script. Re-running the script would have resulted in different scores since the Environmental script is stochastic. Finally, the trend plots script was modified to use the confidence bounds variables when they are available, i.e. for the environmental script, otherwise it calculates them using the variance as it was doing before.

Remove Environmental 2014 scores from the trend plots

Completed November 2019

On the 26 Nov 2019, Mark Schultz notified us that the 2014 scores for the environmental script have been generated using incomplete water quality measures, therefore they should not be used in the trend plots.

Mark Schultz wrote:

- *Is it possible to remove the overall environmental scores for 2014 as in that year it was based solely on WQ?*
- *And is it possible to remove water quality for 2014 as this was based on an incomplete set of WQ measures?*

Since that is all there is for Environmental 2014, all environmental scores for 2014 were removed from the trend plots.

Task 7: Project management, training manual update and final report

This task requires documenting all work undertaken during this project and updating the DIMS user manual as necessary to reflect any updates to the system.

This task also includes the project management aspects of the project, coordinating with GHHP and team members.

The final report will provide a summary of each of the tasks completed over the life of the project.

Completed June 2020

There were no significant changes to the DIMS Report Card system code during this project and so there were no changes to the existing user and developer guides.

All changes to the server administration, such as Tomcat server updates and permission changes, were documented in the administration log file:

https://dims.ghhp.org.au/repo/ws-documentation-adminstration/Unix administration/Admin-log.txt

One-off tasks

Task 2: Changes to water quality score calculations.

Two changes to the water quality scoring script are required for 2019:

1. *Changes to the scripts are required to incorporate a new rule for Limits of Reporting (LOR) values.*

In all cases where water quality measures are below the LOR those values will be adjusted to 50% of the LOR.

LoR values will be requested from PCIMP by the GHHP and supplied to AIMS.

2. *Total vs dissolved (filtered) metal values.*

In all cases where total metal values are less than dissolved (filtered) metal values then the total value will be used.

Completed September 2019

The scripts were altered to reflect the above changes. In particular:

- when the total values are lower than the dissolved values (for WQ metals), use the total values
- LOR values are replaced by half LOR values (at level of raw data)
- Raw values lower than LOR are flagged
- Values are aggregated to Site/Date/Measure level
- Aggregated values lower than LOR are flagged. Note this does result in the potential for a flag when only one of multiple raw values were less than LOR.

Task 3: Change to water quality script to allow LoR values to be variable

LoR values are requested from PCIMP every year, in 2018 a number of LoR values changed over the course of the reporting year (Table 1). This is a result of different labs undertaking the analyse for these measures. As these types of changes are likely to occur again—flexibility in how the LoR values are coded in the DIMS will be required. The requirements of this task are to incorporate this flexibility into the DIMS and to provide a table in the GHHP short report which shows the LoR values (e.g. Table 1). All LoR levels also need to be updated in dot plots (blue line) for each survey period to show the correct LoR.

Measure	***2018 Limit of Reporting***
Chl-a	*0.02µg/L*
TN	*Aug (2017) 50 µg/L* *Nov (2017) 50 µg/L* *March (2018) 20 µg/L* *June (2018) 20 µg/L*
TP	*Aug (2017) 5 µg/L* *Nov (2017) 2 µg/L* *March (2018) 3 µg/L* *June (2018) 3 µg/L*
NOx	*2 µg/L*
Orthophosphate	*2 µg/L*
Sed-As	*0.5 mg/kg*
Sed-Cd	*0.5 mg/kg*
Sed-Cu	*0.5 mg/kg*
Sed-Pb	*0.5 mg/kg*
Sed-Hg	*0.5 mg/kg*
Sed-Zn	*0.5 mg/kg*
Sed-Ni	*0.5 mg/kg*

This task will be implemented by adding support for Limits of Reporting (LOR) data in the system. Measurements less than the LOR can be indicated using the '<' character, for example '<0.1' indicates that the value was less than the LOR of 0.1. The environmental script will then use this information to compile the set of LORs used in the year's data. If there is no data in the year below the LOR then a default LOR from the guidelines files will be used. The collated LORs will be reported in the generated Short Report file.

Completed September 2019

Both Task 2 and 3 require the application of a number of quality control rules to handle converting the raw PCIMP data into a form that is suitable for inclusion in the GHHP report card. These rules determine how limitations in measurements should be handled to ensure the report card captures as much of the available information that is available.

In previous years some of these QA/QC rules were being applied outside the DIMS by the PCIMP data provider, prior to importing them into the DIMS. The problem with this is that these rules are then not recorded in the DIMS and as new rules are developed it requires that the PCIMP data provider implement them, rather than have them captured in the DIMS.

To overcome these limitations the data ingestion from PCIMP was changed so that a rawer form of the data is provided to the DIMS. This then allows the QA/QC rules to be applied in the environmental scripts and for the results of these rules to be accurately reported on by these scripts. This approach also removes the manual step of needing to determine the Limits of Reporting (LOR) each year for each variable.

Changes to the PCIMP data ingest

To allow the QA/QC rules to be applied in the environmental script additional attributes were included in the uploaded PCIMP data. These included:

- Total metal values, to allow the implementation of Task 2
- Additional attributes to make each measurement unique ("REPLICATE","DEPTH"). Previously when there were multiple readings for a site they were averaged prior to upload into the DIMS. However since the DIMS will now handle Limits of Reporting and Total vs dissolved metals rules the handling of multiple readings per site is also necessary in the environmental script.
- The Limits of Reporting are now indicated in the uploaded data on each sample using a '<' character to indicate if the value is less than the Limit of Reporting. i.e. '<0.1' indicates that the value is smaller than 0.1 which is the LOR.

Changes to the DIMS Report Card System

To accommodate the handling of LOR values in the data such as '<0.1' an additional attribute type was added into the DIMS Report Card System software. Each attribute of the input files that needs validation during upload is indicated to the system in the dataset MANIFEST file. This file indicates what attributes to look for in the uploaded file, whether they are a number or string and what rules should apply to check the validity of the data. A new type was added to the system, called 'float_LOR', that indicates that inputs values may start with a '<' character.

The automatic dataset preview graphing in the DIMS Report Card System was upgraded to handle the new 'float_LOR' attribute type. Any values below the Limits of Reporting are now shown on these graphs as red dots. The scatter plots were also slightly improved by using transparency to show the density of points. Outliner values with only one or a few readings show in light grey, whereas values where there are many readings are shown in black.

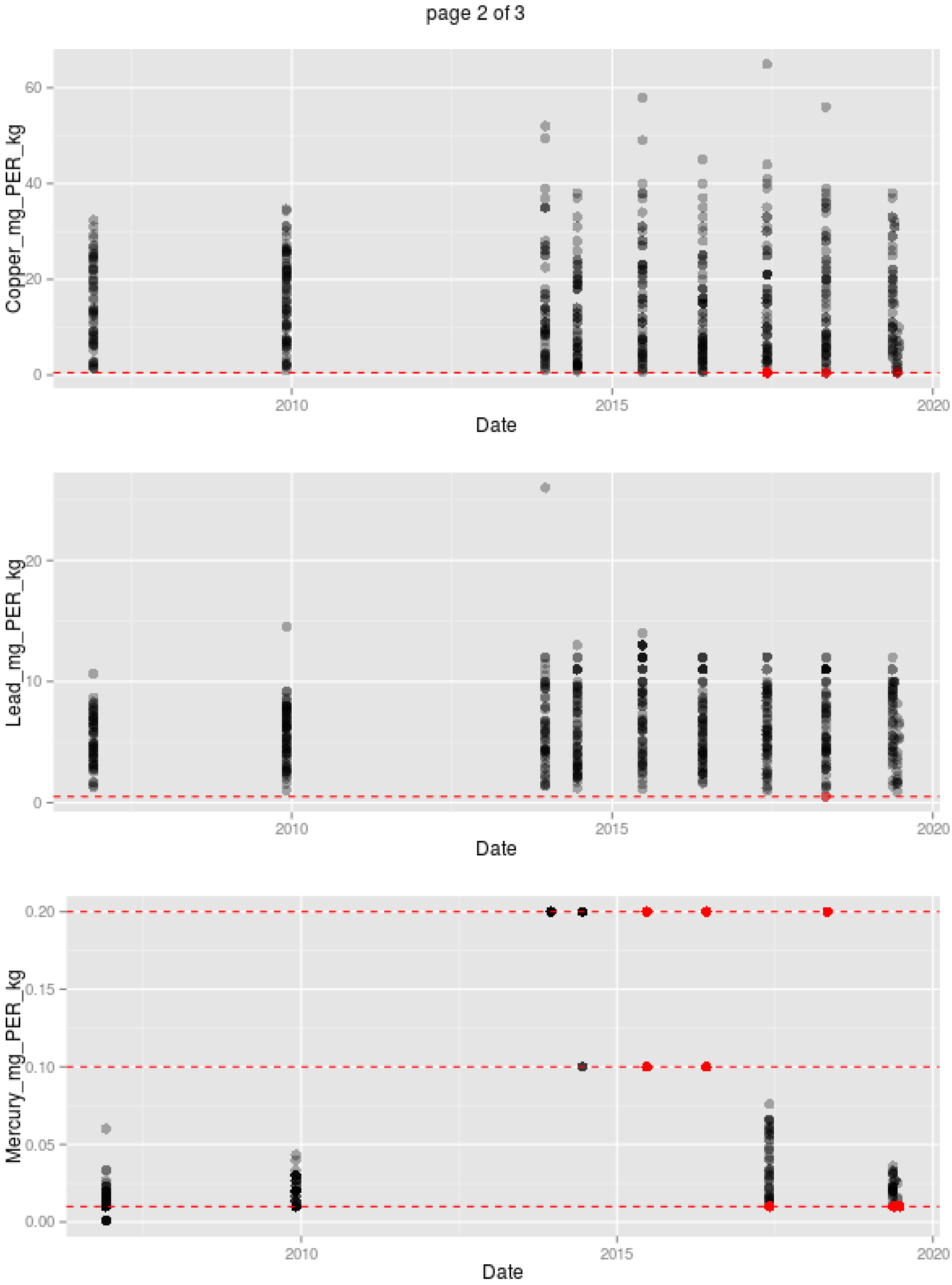


Figure 2. Example Scatter plot generated in the dataset validation for PCIMP in 2019 showing the Limits of Report (LOR) values detected in the uploaded data. Here we can see that for mercury the LOR varied over time resulting in 3 separate red LOR lines. The red dots indicate the individual readings were below the LOR.

Changes to the environmental script

The environmental scripts were altered to handle replicate PCIMP data, the application of QA/QC rules and the production of a summary table of the applied LOR values.

Task 4: Integration of Mangrove Indicator

Precomputed mangroves indicator scores were integrated into the environmental scripts in 2018. The 2019 task is to integrate the R scripts that calculate scores from the raw data. On completion of this task the DIMS should be able to accept 2019 mangrove raw data raw and generate report card grades and scores and associated report card outputs. The newly integrated mangrove calculation scripts should work seamlessly with the existing system and should be free of any technical errors. The script custodian for mangrove score calculation R scripts is Norm Duke at JCU/TropWATER.

This task involves working with Norm Duke to codify the indicator algorithms from the existing processes, based on Excel, into R script that is integrated with the DIMS environmental script. The full algorithms for the mangrove indicator are based on processing of satellite imagery using specialised software. The DIMS will not process the satellite imagery, but instead take in the information derived from the satellite imagery.

To ensure the timely development of the Report Card the R script development will occur after the production of the 2019 report card. This will allow time for the R scripts to incorporate changes to the mangrove indicator algorithms in 2019 which now include a comparison against the 2018 baseline.

Successful completion of this task is contingent with cooperation with the mangrove team.

Completed June 2020

The logic and calculations performed in spreadsheets provided by Norm Duke were translated into R code. There are clearly many calculations that are performed to obtain these spreadsheets in the first place. These were considered out of scope for the current code conversion project.

The R scripts import a set of input files (csv text files) based on the spreadsheets provided by Norm Duke. These scripts are:

- shoreline.csv (selected extract from GHHP2019_Shore_Score.xlsx:ERMP2019_Heli_Shore_Points)
- ndvi.condition.csv (selected extract from GHHP_NDVI_Condition_2019.xlsx:GHHP_NDVI_Condition_2019)
- ndvi.condition.5y.csv (selected extract from GHHP_NDVI_Condition_2019.xlsx:GHHP_NDVI_Condition_2014_2019)
- ndvi.water.mangrove.csv (selected extract from GHHP_NDVI_WCI_2019.xlsx:GHHP_NDVI_Water_Mangrove_2019)
- wci.csv (selected extract from GHHP_NDVI_WCI_2019:GHHP_WCI_2019_DATA)

The mangrove team were asked for feedback on the above. No feedback was received. We also highlighted several issues that were encountered throughout the translation process that could jeopardise the robustness of future index calculations. Unfortunately, the mangrove team did not provide comments or solutions to these issues.

To finalise the integration, the environmental script's MANIFEST file was adjusted to accept and validate the new files (shoreline.csv, ndvi.condition.csv, ndvi.condition.5y.csv, ndvi.water.mangrove.csv, wci.csv). It was decided to keep the ability of directly uploading the scores file (as in the previous report cards) until after the generation of the next report card which includes mangroves. This is to avoid possible delays as we have not received feedback from the mangrove team. The changes were tested and added to the production report card “2020 - Ready for data” on the production server.

Task 5: Integration of two fish health scores into the fish and crab indicator

In 2019 two new sub-indicators for fish health, ISP023A (Fish Health Assessment Index) and ISP023B (Visual Fish Condition Index) will be added to the Fish and Crab Indicator Group. A single harbor wide score will be provided by the contractors for each new project. ISP023A the fish health assessment index (HAI) is aggregated directly into the fish health indicator while with ISP023B the Visual Fish Condition indicators (VFC) are aggregated with body condition to provide a single score before this score is aggregated into the fish health indicator (Figure 3). The aggregated score from both projects will constitute the overall harbour score for the fish health indicator. This overall score will also be used as the zone score for fish health, for the calculation of the fish and crab indicator group score for all 13 GHHP environmental reporting zones and the calculation of the overall fish and crabs score.

This is a two-part task to facilitate the inclusion of these the new fish health indicator into the 2019 report card. These two parts are:

1. *The two fish health measures from ISP023 A & B will be added to the system and integrated into the Fish and Crabs indicator group. For 2019–20 the system will take in precomputed scores not raw data. These scores will be a single harbour wide score for the Health Assessment Index (ISP023A) and a single harbour wide score ISP023B which is comprised of Visual fish condition indicators and body condition. This will be done to allow the fish health results to be included in the 2019 report card prior to the availability of the R scripts that process the results from the raw data. A single score for the combined health indices projects will be calculated and applied to all 13 GHHP reporting zones.*

2. *The fish scripts that aggregate from the raw data, will be integrated into the environmental script. This will be done after the 2019 report card is complete and will be a component of the* ***next DIMS contract in 2020.***

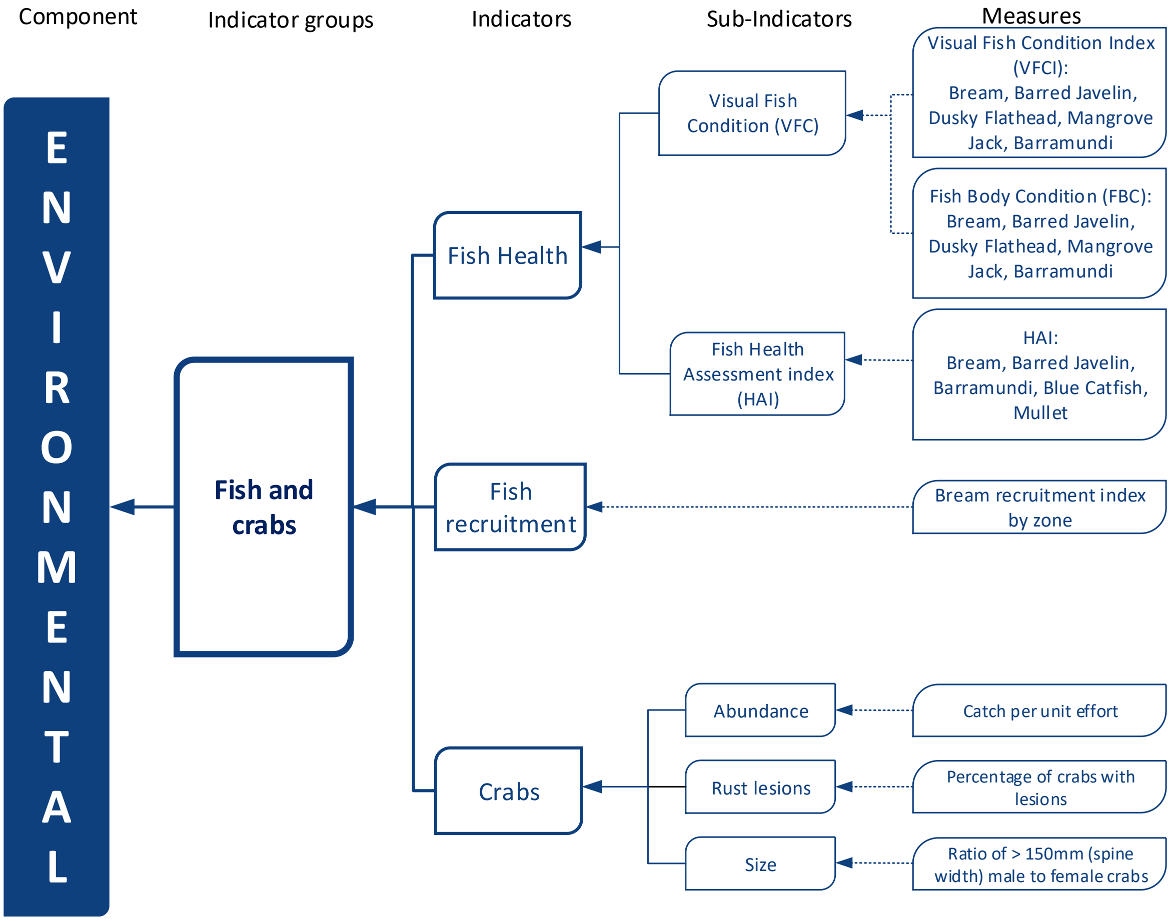

Component
Indicator groups
Indicators
Sub-Indicators
Measures
ENVIRONMENTAL
Fish and crabs
Fish Health
Fish recruitment
Crabs
Visual Fish Condition (VFC)
Fish Health Assessment index (HAI)
Abundance
Rust lesions
Size
Visual Fish Condition Index (VFCI): Bream, Barred Javelin, Dusky Flathead, Mangrove Jack, Barramundi
Fish Body Condition (FBC): Bream, Barred Javelin, Dusky Flathead, Mangrove Jack, Barramundi
HAI: Bream, Barred Javelin, Barramundi, Blue Catfish, Mullet
Bream recruitment index by zone
Catch per unit effort
Percentage of crabs with lesions
Ratio of > 150mm (spine width) male to female crabs

Figure 3. Aggregation of the fish and crabs indicator group. The Visual Fish Condition Index (VFC) is aggregated with body condition before being aggregated into the Fish Health indicator. Provided by Mark Schultz 18 September 2019

Completed October 2019

The environmental script was adjusted to incorporate the two additional fish health measures to the Fish and Crab indicator group. These two measures were Health Assessment Index (HAI) and Visual Fish Condition. At this stage, the input data for the two measures are the index scores. The resulting scores are then integrated into the Fish and Crabs Indicator.

Furthermore, a new MANIFEST file and example upload files were created for each dataset. These files are used to enable the upload and validation in the DIMS software system. The existing MANIFEST file for the fish recruitment input was adjusted to include the two new datasets as dependencies.

Task 8: Adjustments to new names for PCIMP sites and Sense of Place indicators

This task involves adjusting the names of the PCIMP sites in the DIMS to match those currently used by PCIMP. It also involves adjusting the names of the Sense of Place indicator names. This alignment will make it easier for the GHHP science team to perform quality checks on the data.

The PCIMP sites names in the input data acts as a pseudo IDs in the environmental script. Changing these input names requires that we add a process at the end of the environmental script to map these new site *names to the original site IDs. This ensures that the trend analysis plots can still track the measures through time. We must also adjust the input QAQC rules to accept the new site names.*

The trend plot script places all its generated plots in a directory structure matching the indicator tree of the report card. This tree structure is currently based on the indicator IDs. Since the IDs no longer match the display names (PCIMP sites and Sense of Place indicator names) we must adjust the script to use the label names for the directories rather that IDs as it currently does now.

Adjustments to the display names for the Sense of Place script:

- *Distinctiveness* → *Place attachment*
- *Continuity* → *Continuity*
- *Self-esteem* → *Pride in the region*
- *Self-efficacy* → *Well-being*
- *Attitudes to Gladstone Harbour* → *Appreciation of the harbour*
- *Values of Gladstone Harbour* → *Values*

Completed September 2019

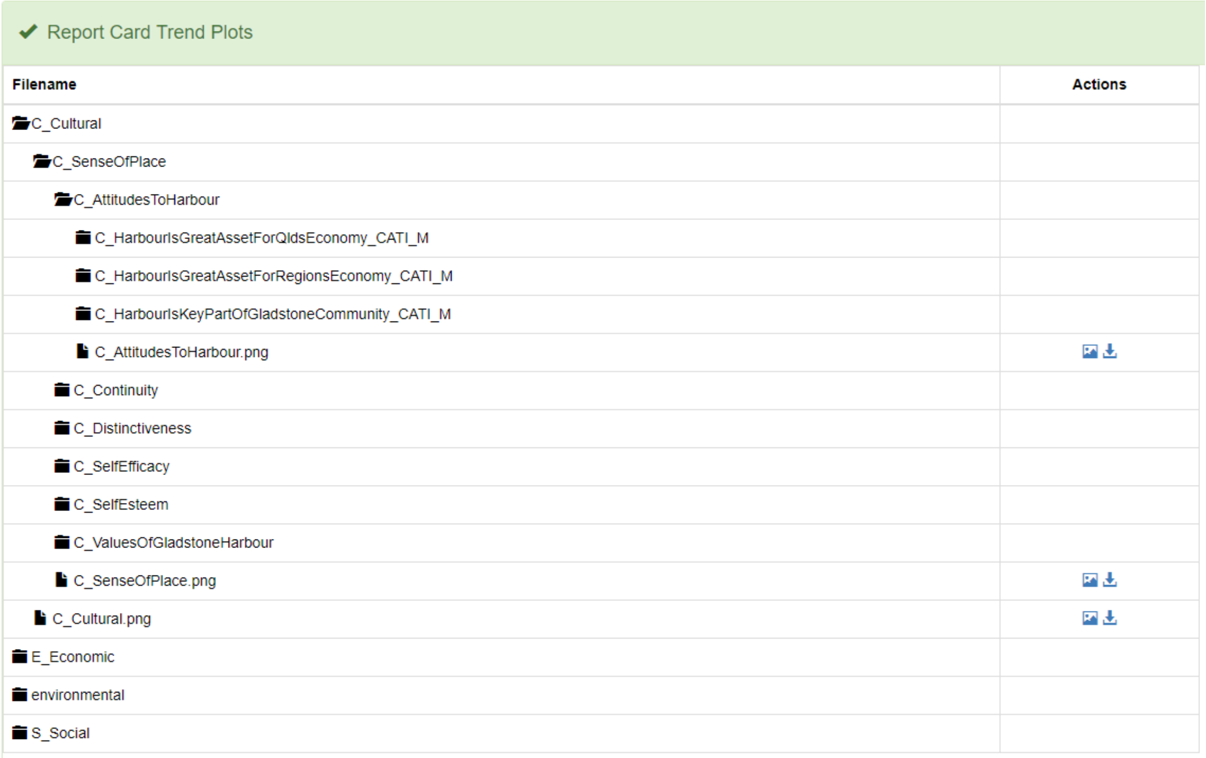
Adjusting the names of the Sense of Place indicator names

The names of the Sense of Place indicators have been adjusted in the labels file according to the list provided (see above).

Adjusting trend plot script

The trend plot script has been modified to use the names from the labels files instead of IDs. To cope with very long names a new override label file was introduced which holds a set of short names for some IDs. Furthermore, the script now also checks for long titles which do not fit on the chart and adds a line break.

Before:

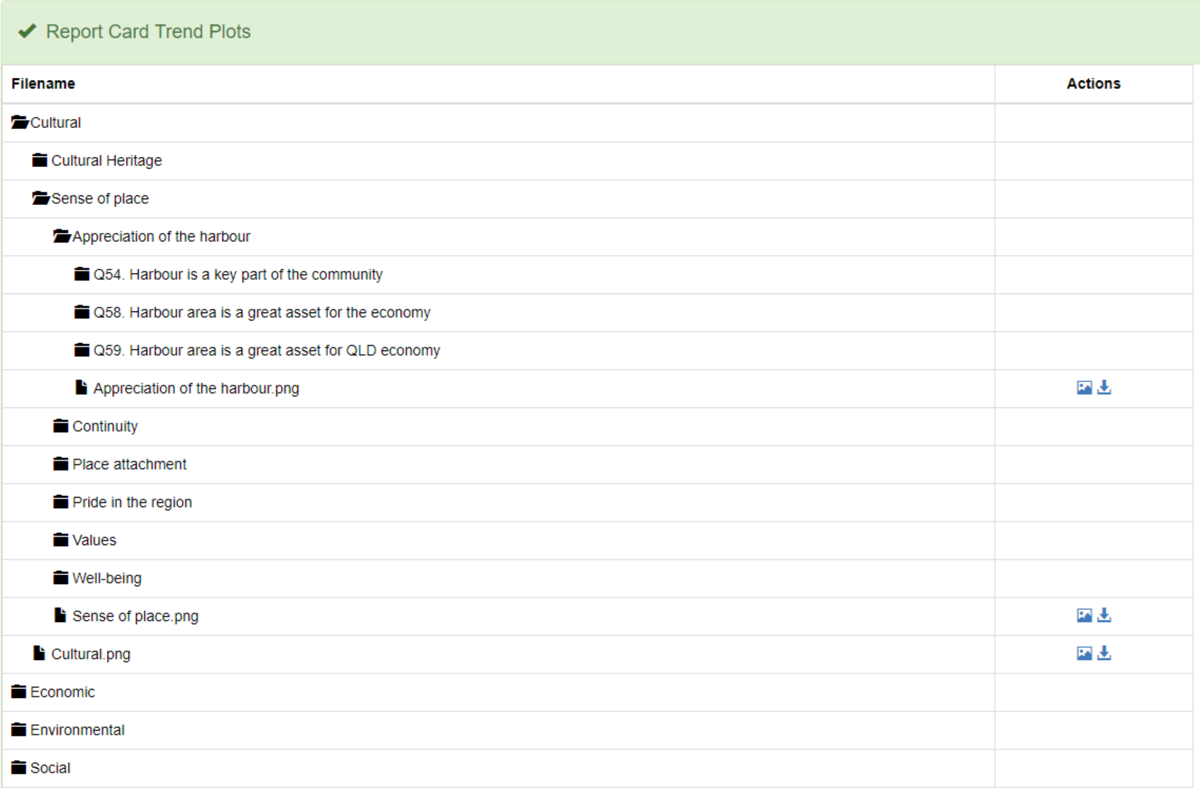


✔ Report Card Trend Plots

Filename	Actions
C_Cultural	
C_SenseOfPlace	
C_AttitudesToHarbour	
C_HarbourIsGreatAssetForQldsEconomy_CATI_M	
C_HarbourIsGreatAssetForRegionsEconomy_CATI_M	
C_HarbourIsKeyPartOfGladstoneCommunity_CATI_M	
C_AttitudesToHarbour.png	🖼 ⬇
C_Continuity	
C_Distinctiveness	
C_SelfEfficacy	
C_SelfEsteem	
C_ValuesOfGladstoneHarbour	
C_SenseOfPlace.png	🖼 ⬇
C_Cultural.png	🖼 ⬇
E_Economic	
environmental	
S_Social	

Folder and file names are based on IDs.

Now:

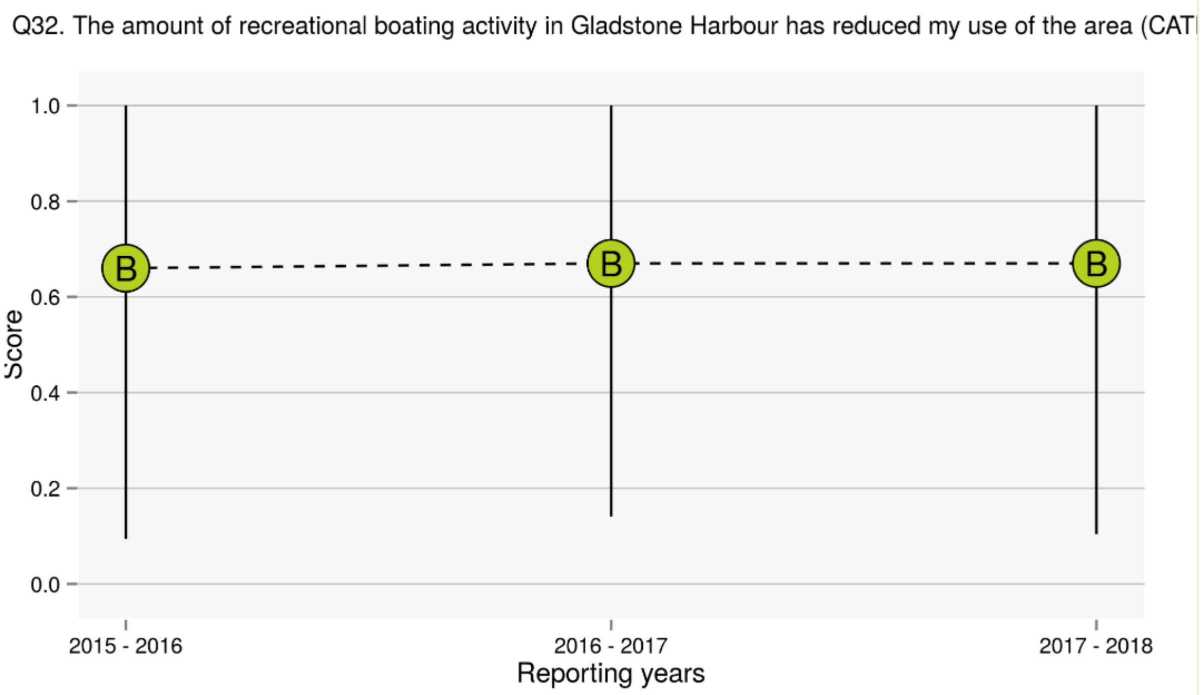


✔ Report Card Trend Plots

Filename	Actions
Cultural	
Cultural Heritage	
Sense of place	
Appreciation of the harbour	
Q54. Harbour is a key part of the community	
Q58. Harbour area is a great asset for the economy	
Q59. Harbour area is a great asset for QLD economy	
Appreciation of the harbour.png	🖼 ⬇
Continuity	
Place attachment	
Pride in the region	
Values	
Well-being	
Sense of place.png	🖼 ⬇
Cultural.png	🖼 ⬇
Economic	
Environmental	
Social	

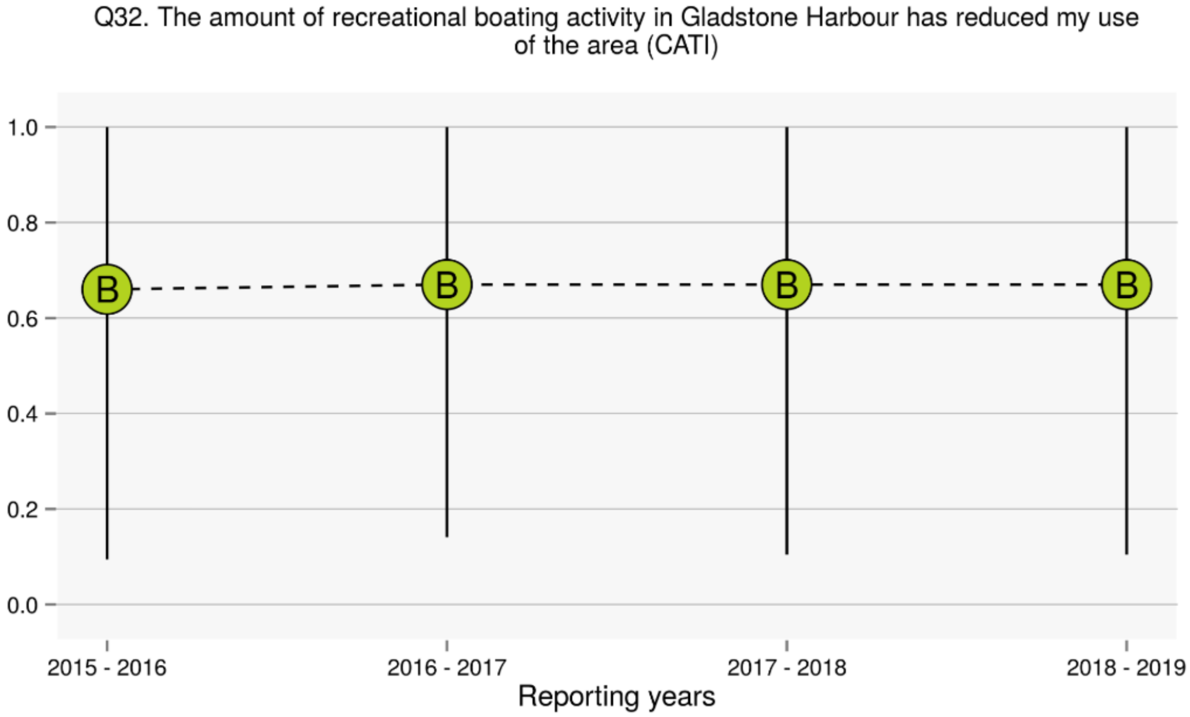
Folder and file names are based on names.

Before:



Very long titles would not fit on the chart.

After:



Very long titles now have a line break inserted to fit on the chart.