Session 4: Future for SOCAT: Data Quality, Management and Products,
Data2Flux Workshop, Unesco, Paris
12 and 13 September 2011

Report for SOCAT website by Dorothee Bakker and Letitia Barbero (23/11/2012)
(Any mistakes/inconsistencies are unintentional and have been made by DB).

Session 4: Early breakout group on SOCAT automation

Lunch breakout session on Monday 12/09/2011

Chair: Dorothee Bakker; Participants: Stephen Jones, Heather Koyuk, Steven Hankin, Benjamin Pfeil, Alex Kozyr, Yukihiro Nojiri, Denis, Pierrot, Rik Wanninkhof (part of session), Chris Sabine, Ute Schuster, Nicolas Metzl, David Hydes, ...

Followed by a smaller breakout group afterwards

Participants: Benjamin Pfeil, Heather Koyuk, Steven Hankin, Dorothee Bakker, Alex Kozyr

The participants agreed that streamlining and automating SOCAT is essential for prompt, regular, future SOCAT releases, e.g. every 1 to 2 years after the release of SOCAT version 2. In the two breakout groups we discussed how to:

- Streamline information exchange in the SOCAT global group;
- Automate submission of metadata;
- Automate submission of data;
- Automate initial quality control;

Below follows a summary of these discussions. Rik Wanninkhof, Chris Sabine and others indicate that SOCAT should remain a delayed mode data product with quality control and regular releases, e.g. at 1 to 2 year intervals from version 3 onwards.

Any strategy for data and metadata submission to SOCAT should also allow making these original (meta-)data public, either upon (meta-)data submission or latest upon the release of the SOCAT version containing these data, depending on the wishes of the data PI.

The marine CO_2 community agreed on how to report surface water CO_2 data and metadata (http://www.ioccp.org/FinalRpts/IOCCP_WS2Summary.pdf, http://cdiac.ornl.gov/oceans/submit.html) at the Tsukuba meeting in January 2004. However, few PIs have strictly followed these criteria. In addition, some formats for data submission (date, time, file format e.g. text or CSV, need for a header) have not been fully defined. Metadata formats have been less well defined than data formats.

It would be extremely useful to have PIs submit data & metadata for SOCAT in the recommended Tsukuba format. It is suggested to provide a carrot for PIs who submit data and metadata in the recommended Tsukuba format, in the form of a reduction in the chances of error during data ingestion, (future) instant feedback on data quality and an increase in the likelihood that the data will be included in the next SOCAT release (as the process of inclusion is

more automated and less labour intensive for the SOCAT data manager). At the same time we do not want to deter PIs from data submission, thus other data and metadata formats will be accepted too.

Some PIs have a contractual obligation to submit data and metadata to a national data centre. Thus, it would be beneficial if software for automated data and metadata submission can be implemented at data centres around the world without any licence-related issues.

Metadata submission should be fully standardized by e.g. web-based pull down menus. Ideally submission of metadata is as easy as possible and a PI should be able to access information of previous metadata submissions and submit metadata for multiple cruises easily.

Ideally data are run during submission across simple software which recognizes what data are in which columns with input from the PI. Bob Key (on Tuesday in session 4) suggests that using exact column headers instead of an exact column order during data submission would make it easier to ingest submitted data into SOCAT. The software would ideally be web-based or alternatively downloadable for the PI to run on his/her computer. This software could create an Expocode for the data file, either on the basis of the date and time of the first and last data point or on the date and time the ship left and returned to port.

The LAS might be able to provide tools for initial QC upon data submission.

Session 4: Future for SOCAT: Data Quality, Management and Products (13/11/2012)
Chairs: D.C.E. Bakker and B. Pfeil, Rapporteur: L. Barbero

Lessons learnt from SOCAT version 1.5

Discussion on the authors of the two SOCAT ESSD papers:

Position A) List as authors those who have actively contributed to SOCAT version 1.5 (e.g. all quality controllers (QCers), regional group leaders, LAS developers, data managers), around 45 authors. Acknowledge data PIs in a table or in the acknowledgements.

Reasons:

R. Wanninkhof: Authors hopefully have to submit data anyway.

C. Sabine: How fair is it to cite a data PI who submitted one of the 1851 cruises in SOCAT at the same level as those who have QCed the data or prepared the software etc.?

C. Sabine: Each individual dataset is published on its own; SOCAT is a synthesis product, different from and going beyond each cruise, so the data PIs can choose to publish their own dataset and get credit for it by writing an individual paper for the data submission.

Position B) Include everyone as authors, notably those who have actively contributed to SOCAT version 1.5 and data PIs, around 100 authors.

Reasons:

Authorship will encourage authors to submit more data in the future, because they benefit from SOCAT releases.

A.J.Watson: Other large communities such as physics or astronomy simply list everybody in major papers (up to hundreds of authors) which also benefits for junior investigators. In many countries the number of citations is very important for the advancement of scientific careers. The creation of large datasets has been delayed, because individual scientists were reluctant to provide their data for various reasons (difficulty in putting the data together, no or inadequate citation. In future data PIs may choose not to submit their data, unless they are authors on the ESSD articles. The citing system may be not fair but is used and is important for the careers of scientists. Maybe it should have been made clearer that you may not be cited unless you contributed actively to SOCAT.

P.S. D.C.E. Bakker: After the session the lead authors on the ESSD articles decided to include as authors all those who have actively contributed to SOCAT, as well as data PIs.

Other comments:

B. Pfeil: Often only the PIs providing the data are cited, but not the PhD students who have collected or worked on the data.

R. Feely suggests sending around a list of co-authors on the ESSD articles so that if someone is found to be missing he/she can added. Reply D.C.E. Bakker: This will be done with initial circulation of the list to the regional group leaders.

B. Pfeil: Every cruise in SOCAT has an Expocode, a doi-number and metadata on e.g. who provided the data and other relevant information. Cruise data files with recalculated fCO_2 in a uniform format are available at http://www.pangaea.de/. Every data point in SOCAT has a link it to its cruise of origin via the doi-number.

C. Sabine: The SOCAT website has a clear data policy on how SOCAT should be cited (http://www.socat.info/SOCAT_data_policy_public _release_v2.htm).
B. Pfeil: Some scientists submitted data to SOCAT, but their data were suspended from version 1.5 for a variety of reasons. If these scientists, are not acknowledged in the ESSD articles they might be discouraged from submitting further data.
D. Pierrot suggests updating the SOCAT data set awaiting secondary quality control (QC) on the LAS regularly (e.g. monthly, quarterly) such that QC can be carried out continuously by SOCAT QC-ers.

Needs and priorities for future SOCAT

Operational – Automation of data submission & initial QC

B. Key: Using exact column headers instead of exact column order during data submission would make it easier to ingest submitted data into SOCAT.
R. Wanninkhof: SOCAT version 1.5 was a painful process, so automation of SOCAT will enthuse people into submitting data for future SOCAT releases. Having your data in a global data product such as SOCAT presents enormous advantages for data PIs, such as the utilization of the LAS (Live Access Server) for QC of new data, creation of maps for reports. The LAS is a powerful tool and will motivate data PIs to submit data for future SOCAT releases.
B. Pfeil: In January 2004, the community agreed on how to report data and metadata (http://www.ioccp.org/FinalRpts/IOCCP_WS2Summary.pdf, http://cdiac.ornl.gov/oceans/submit.html). We realize now that this has not been properly implemented. We should make sure that we stick to these 2004 protocols in future. Proper reporting of data and metadata avoids having to get back to the PI for information on historic cruises.
R. Wanninkhof: Create a crystal clear data format cookbook for data submission. Data PIs must be made aware that SOCAT is a continuing effort with regular releases or they will stop submitting data.
B. Pfeil: PIs should continuously submit data for inclusion in future SOCAT releases.
B. Pfeil: The data in SOCAT version 1.5 were at CDIAC, public websites and data holdings around the world. In some cases cruise data were updated at a different location to that of the initial data submission. Submission of all data to one site will make it easier to keep track of updates (http://cdiac.ornl.gov/oceans/submit.html).
D. Bakker: The priority is to speed up the SOCAT process by reducing the effort (man power) required.
C. Sabine: Who is responsible for the next SOCAT release to figure out what happened to cruises that were suspended? B. Pfeil will go through the list of suspended cruises to see why they were suspended and to get new data from the PI. Most PIs are already aware of the cruises that were suspended. S. Hankin suggests automating the process of informing the PI of the suspension of their cruise during SOCAT QC.

User Requirements? Frequency of release? Age of data?

C. Sabine, C. Sweeney, B. Pfeil and D.C.E. Bakker: What criteria should be set for the submission deadline for future SOCAT releases? We might consider a SOCAT release every 1 to 2 years after the release of SOCAT version 2. A submission deadline could be set 6 months after the end of the year to be included in the SOCAT release, e.g. a 30 June 2014 deadline for the submission of data from 2013. This would make it easy for PIs to remember the deadline and would encourage prompt data submission. One problem remains different release requirements in different countries. However, data PIs can generally submit data before they are contractually required to do so. Ultimately all data would be included in a specific SOCAT release that have passed secondary QC by a certain date.

User Requirements? Additional parameters or products?

C. Sweeney suggests producing something similar to a Globalview-type assimilation product, a simple average or low level analysis. Such products would entice scientists to contribute data to SOCAT and to use SOCAT.

C. Sabine: The Bulletin of the American Meteorological Society does a state of the climate publication every year and it could include the yearly SOCAT data that has been added in the latest release. Alternatively annual SOCAT releases might be made for the Global Carbon Project.

N. Metzl, R. Feely, S. Hankin, B. Key: An initial target in SOCAT was to produce maps at a decadal scale. This can now be done for the 1980's and 1990's. Such maps can be created as a community product of SOCAT or each scientist can carry out their own research. SOCAT provides gridded products without interpolation in time or space. Thus, the quality of the interpolations and extrapolations is critical (by e.g. multiple linear regression, neural networks, etc). It would be useful if these interpolations were shared online and compared. Users could be asked to send interpolation products derived from SOCAT, so they can be archived and shared with the community. Since the data is public, authors can only be asked, but not required to submit their new data products. The LAS could host mapping products by different authors and provide a platform for sharing ideas on how to improve them. Such a platform would be a way for community interaction and research activities. A suggestion is made for a workshop planned around an intercomparison between SOCAT and various products derived from it, perhaps a topic for a joint SOLAS/IMBER carbon (SIC) working group 1 workshop?

N. Metzl suggests asking scientists to inform SOCAT (submit@socat.info) of publications using SOCAT.

P.S. D.C.E. Bakker: Some SOCAT users would appreciate additional parameters in SOCAT, e.g. the atmospheric mixing ratio of CO_2, dissolved nutrient and oxygen concentrations, dissolved inorganic carbon, total alkalinity.

Areas with data gaps? New target regions?

B. Pfeil: Many coastal data have not been submitted for inclusion in SOCAT.

B. Pfeil: SOCAT has no Chinese data, possibly as a result of Chinese government policy on data sharing.

R. Wanninkhof: We can act as SOCAT ambassadors and help obtain data from PIs or reluctant governments by showing them the benefits of data sharing. A publicly available product like SOCAT helps in this direction.
X. XXXX: An effort needs to be made to reach out to the coastal group. Simone Alin, Burke Hales and Joe Salisbury might contribute to such an effort.

Management, e.g. Global and regional groups? Formalisation of the decision process?

B. Pfeil: Things will speed up, if the global group decides on most major issues.
R. Feely, D.C.E. Bakker: After some discussion it is suggested to post summaries of meetings and decisions by the SOCAT global group on a password protected website with access for all SOCAT contributors.
R. Feely suggests having a password restricted area for online dialogue, so people can contribute with ease. S. Hankin suggests that the LAS can host such a password protected discussion board, as well as summaries of meetings and decisions by the SOCAT global group.
P. Monteiro suggests formal arrangements for SOCAT in view of the political context of climate change.
D.C.E. Bakker: Making data public and creating synthesis products such as SOCAT is an excellent strategy for meeting public demands for more openness from the scientific community.

Future funding needs and opportunities for SOCAT.

D.C.E. Bakker: Please, inform the SOCAT global group of funding opportunities
R. Wanninkhof: It is important to prepare a budgetary vision of the funding requirements for future SOCAT and how these are currently met. Such a budgetary vision would also be extremely useful for program managers.
R. Feely suggests adding a webpage with logos thanking all the funding agencies. C. Sabine and B. Pfeil inform that the current preliminary version will be updated to acknowledge all contributors and funding agencies with their logos (http://www.socat.info/credits.html).
N. Metzl reminds the audience of the list of possible papers to be published from SOCAT, as discussed in the June 2009 Atlantic and Southern Ocean regional group meeting in Norwich (http://www.ioccp.org/FinalRpts/WR222_eo.pdf), and asks if this option is still being considered.

Action items from session 4 and its early breakout session

SOCAT version 1.5 (public release on 14 September 2011):

1. All those who have actively contributed to SOCAT version 1.5, as well as data PIs (~100 scientists) will be included as co-authors on the two SOCAT ESSD articles. A list of co-authors will be sent around for checking (Benjamin Pfeil, Dorothee Bakker, Ute Schuster, others?);
2. Update preliminary SOCAT webpage to acknowledge all contributors and funding agencies with their logos (http://www.socat.info/credits.html) (Benjamin Pfeil, Dorothee Bakker, others);
3. Create a password restricted website on the LAS for online dialogue and summaries of meetings of SOCAT global group (Steven Hankin, Heather Koyuk);

4. Ask scientists to inform SOCAT (submit@socat.info) of publications using SOCAT (Dorothee Bakker, Benjamin Pfeil);
5. Actively promote publication of SOCAT science (Chair of SIC WG 1, Dorothee Bakker, others);
6. Discuss an intercomparison of interpolation products based on SOCAT, e.g. a workshop. Ask authors of interpolation products to submit these to SOCAT for posting on the LAS (Chair of SIC WG 1, Steve Hankin, Dorothee Bakker);

Future SOCAT Releases:

Vision for Future SOCAT Releases

7. Start work on SOCAT version 2 shortly. A 31 December 2011 deadline for data submission to SOCAT version 2 is being discussed (Dorothee Bakker, Benjamin Pfeil, all);
8. Aim for regular SOCAT releases, e.g. every 1 to 2 years from SOCAT version 3 onwards. Set a clear submission deadline for data to be included in future releases (all);
9. Prepare a budgetary vision of the funding requirements for future SOCAT and how these are currently met (Dorothee Bakker, Chris Sabine, Are Olsen, others);

Data submission and data policy

10. Data PIs to continuously submit data for future SOCAT releases (http://cdiac.ornl.gov/oceans/submit.html);
11. Data PIs are urged to submit data and metadata following the recommended Tsukuba format (http://www.ioccp.org/FinalRpts/IOCCP_WS2Summary.pdf, http://cdiac.ornl.gov/oceans/submit.html). Data PIs are urged to adopt the more strict formats that meet automation requirements, once these have been defined;
12. Inform data PIs of cruises suspended from SOCAT 1.5 on the reasons why the cruises were suspended such that the PIs will hopefully submit revised data to SOCAT (Benjamin Pfeil);
13. Reach out to data PIs whose data might be included in SOCAT. Highlight the advantages of having data in SOCAT (Maciej Telszewski, all);
14. Reach out to the coastal group (Maciej Telszewski, Dorothee Bakker to contact Alberto Borges and coastal data PIs);
15. Develop a clear SOCAT data policy, including guidance for submission of data via submit@socat.info and co-authorship on future technical SOCAT publications (Dorothee Bakker, Chris Sabine, Benjamin Pfeil, Are Olsen, to consult widely);

Streamlining SOCAT

16. Frequent Skype meetings for SOCAT global group or the data managers and LAS staff to discuss progress and streamline the work. Meetings might take place e.g. every first Tuesday of the month at 18:00 UK time (19:00 Bergen/Paris, 9:00 Seattle). (Dorothee Bakker to prepare the agenda, Steven Hankin, Benjamin Pfeil, Alex Kozyr, Chris Sabine, Are Olsen, Maciej Telszewski, with the addition of Heather Koyuk);
17. Annual 3-day meeting of the SOCAT data managers and LAS staff for a thorough and efficient discussion of all SOCAT database-related issues, including automation (Heather

Koyuk, Steven Hankin, Alex Kozyr, Benjamin Pfeil and possibly 1-2 scientists, as required);

Speed up secondary QC on the LAS (for version 2, if possible)

18. Implement regular updates of the SOCAT data set awaiting secondary QC on the LAS (e.g. monthly, quarterly) (Steven Hankin, Heather Koyuk, Benjamin Pfeil);
19. Provide regular updates of ZIP files of data files awaiting secondary QC, such that files can be run through Matlab (suggestion by Ute Schuster) (Heather Koyuk, Steven Hankin, Benjamin Pfeil);

SOCAT automation (complete for version 3)

20. Automate SOCAT, notably submission of data and metadata, and initial quality control (Steven Hankin, Benjamin Pfeil, Alex Kozyr, Heather Koyuk, others);
21. Draft a strategy for automating SOCAT, notably on automating data and metadata submission and initial data quality control. Circulate for consultation (Steven Hankin, Benjamin Pfeil, Alex Kozyr, Heather Koyuk, others);
22. Define a strict and unambiguous metadata format, based on the Tsukuba format, such that it meets automation requirements. Create a crystal clear cookbook for submission of metadata (Alex Kozyr to lead, Heather Koyuk, Steven Hankin, Benjamin Pfeil, others);
23. Develop software for automated metadata submission. Ideally this should be freeware or very common software, such that it can easily be used by data centres around the world (web-based pull down menus?). In addition a PI should be able to use earlier metadata input for subsequent data submissions and be able to easily submit multiple data files (Alex Kozyr, Steve Hankin, others);
24. Define a strict and unambiguous data file format, based on the Tsukuba format, such that it meets automation requirements. Create an example data file. Create a crystal clear cookbook for data submission (Benjamin Pfeil, Alex Kozyr, Heather Koyuk, Steven Hankin, others);
25. Identify or develop software for automated data submission, which recognizes the contents of columns and/or exact data headers. Ideally this should be freeware or very common software, such that it can easily be used by data centres around the world (Heather Koyuk, Steven Hankin, Benjamin Pfeil, others);
26. Publicize request to submit data and metadata in a clearly defined, recommended format. Ideally this is done, once formats for metadata and data have been defined to meet automation requirements. Use e.g. IOCCP, SOCAT.info, socat.googlegroup, CDIAC, CarboChange etc., while emphasing the advantages of data submission in these formats (Benjamin Pfeil, Dorothee Bakker, Alex Kozyr, Maciej Telszewski, ...);
27. Automate the process of informing PIs of the suspension of their cruises during SOCAT QC (Steve Hankin, Heather Koyuk, Benjamin Pfeil);
28. The LAS has tools for automating QC. Such tools might be made available to data PI prior to data submission and become part of online data submission (Steven Hankin, Heather Koyuk);

SOCAT based products and additional parameters

29. Consult SOCAT community and SOCAT users on the addition of extra parameters to future versions of SOCAT (Dorothee Bakker, SOCAT global group);
30. Consider creation of SOCAT based products (Chris Sabine, others);

Resources needed:

Annual 3-day meeting for SOCAT data managers and LAS staff to move things forward, streamline the work, apply lessons from SOCAT version 1.5 and discuss automation (first in early 2012) (Heather Koyuk, Steven Hankin, Alex Kozyr, Benjamin Pfeil and possibly 1-2 scientists, as required).