Toward homogeneous Argo Index files

Contributors: D.Dobler, T.Carval, M.Scanderbeg

Version history:

0.1 : initial version

0.2: include initial comments from Thierry Carval and Megan Scanderbeg

https://github.com/OneArgo/ADMT/issues/16

https://github.com/OneArgo/ADMT/issues/3

I - Introduction

During ADMT-24, the question was raised to create a deep index with only deep floats and additional information about max pressure reached. Given the number of indexes and the starting variety of fields, it was asked to think toward homogenous index files, limiting the parsing costs for users.

As a personal note (D.Dobler), when exploiting prof_index for the evolution of the DMQC status tool, I had to develop two parsers, mainly because data mode and parameters quality information are filled in 2 different ways. Additionally, I was missing PRES_PROFILE_QC in the core profile index.

As of today, there is only one tech and one meta. On the other hand, there are several profile indexes and traj indexes.

Several intertwined questions are related:

- On which homogeneous fields do we agree?

- What is the size limit per file that we can afford and what about the access performance?
- How do we build the new index files ?

For the latter question, we reckon we should start from a unique super-index file, which would include fields needed for all indexes, whatever the type of file (single core profile, single B profile, single S profile, multi-profile, meta, traj or tech file). This super index would be a reference file from which all the specific index files will be built. This super index could be used for big data purposes.

Section II to V recall the actual fields for each index. Section VI is a proposal for an homogeneous field list and some rules depending on the initial file type. Section VII is an estimate of file size and discusses file splitting strategy. Section VIII presents how one-argo-{type}-index.csv could be built from the super index.

II - Existing profile files associated indexes

There are currently six indexes that relate to profile or multiprofiles files.

Index file	Which input files?	Comment
argo_sprof_index.txt	{wmo}_Sprof.nc	Multiprofile file index for BGC floats
ar_index_global_prof.txt	Profiles/{R|D}_{wmo}_{cycle}{|D}.nc	
argo_profile_detailled_index.txt	Profiles/{R|D}_{wmo}_{cycle}{|D}.nc	+ quality + salinity adjustment + date_creation + n_levels compared to ar_index_global_prof.txt
argo_bio-profile_index.txt	Profiles/{BR|BD}_{wmo}_{cycle}{|D}.nc	Same fields as argo_synthetic-profile_index.txt
argo_synthetic-profile_index.txt	Profiles/{SR|SD}_{wmo}_{cycle}{|D}.nc	+ parameters + data_mode compared to ar_index_global_prof.txt
argo_synthetic-profile_detailled_index.txt	Profiles/{SR|SD}_{wmo}_{cycle}{|D}.nc	+ quality compared to argo_synthetic-profile_index.txt

a) argo_sprof_index.txt

file,profiler_type,institution,parameters,date_update

aoml/1901379/1901379_Sprof.nc,846,AO,PRES TEMP PSAL DOXY NITRATE,20221224115754

b) Current list of fields for the other five profile files indexes and examples:

ar_index_global_prof.txt

file,date,latitude,longitude,ocean,profiler_type,institution,date_update

aoml/7901106/profiles/D7901106_043.nc,20230919132746,22.166,-156.422,P,846,AO,20230922130615

argo_profile_detailled_index.txt

file,date,latitude,longitude,ocean,profiler_type,institution,date_update,profile_temp_qc,profile_psal_qc,profile_doxy_qc,ad_psal_adjustment_mean,ad_psal_adjustment_deviation,gdac_date_creation,gdac_date_update,n_levels

aoml/7901106/profiles/D7901106_043.nc,20230919132746,22.166,-156.422,P,846,AO,20230922130615,B,B,,-0.005,0.000,20230922203741,20230922203741,496

argo_synthetic-profile_index.txt

file,date,latitude,longitude,ocean,profiler_type,institution,parameters,parameter_data_mode,date_update

aoml/7901108/profiles/SR7901108_002.nc,20240326060337,-38.644,126.988,I,846,AO,PRES TEMP PSAL DOXY CHLA BBP700 PH_IN_SITU_TOTAL NITRATE DOWN_IRRADIANCE380 DOWN_IRRADIANCE443 DOWN_IRRADIANCE490 DOWNWELLING_PAR,AAARAARRRRRR,20240402095137

argo_synthetic-profile_detailled_index.txt

file,date,latitude,longitude,ocean,profiler_type,institution,parameters,parameter_data_mode,parameter_quality,date_update

aoml/7901108/profiles/SR7901108_002.nc,20240326060337,-38.644,126.988,I,846,AO,PRES TEMP PSAL DOXY CHLA BBP700 PH_IN_SITU_TOTAL NITRATE DOWN_IRRADIANCE380 DOWN_IRRADIANCE443 DOWN_IRRADIANCE490 DOWNWELLING_PAR,AAARAARRRRRR,AAAFAAFFAAAA,20240402095137

argo_bio-profile_index.txt

file,date,latitude,longitude,ocean,profiler_type,institution,parameters,parameter_data_mode,date_update

aoml/1900722/profiles/BD1900722_001.nc,20061022021624,-40.316,73.389,I,846,AO,PRES TEMP_DOXY BPHASE_DOXY DOXY,RRRD,20200312153230

III - Existing meta files associated index

There is currently one index that is related to the metadata file.

a. Actual list of fields

ar_index_global_meta.txt

file,profiler_type,institution,date_update

aoml/13857/13857_meta.nc,845,AO,20181011200014

IV - Existing traj files associated indexes

There are currently two indexes that are related to trajectory data:

Index file	Which input files?	Comment
ar_index_global_traj.txt	{wmo}_{R|D}traj.nc	
argo_bio-traj_index.txt	{wmo}_B{R|D}traj.nc	+ parameters

a - Current list of fields

ar_index_global_traj.txt

file,latitude_max,latitude_min,longitude_max,longitude_min,profiler_type,institution,date_update

aoml/13857/13857_Rtraj.nc,6.931,0.008,-15.014,-33.808,845,AO,20210428200335

argo_bio-traj_index.txt

file,latitude_max,latitude_min,longitude_max,longitude_min,profiler_type,institution,parameters,date_update

bodc/3901578/3901578_BRtraj.nc,,,,,836,BO,PRES C1PHASE_DOXY C2PHASE_DOXY TEMP_DOXY DOXY RAW_DOWNWELLING_IRRADIANCE380 RAW_DOWNWELLING_IRRADIANCE412 RAW_DOWNWELLING_IRRADIANCE490 RAW_DOWNWELLING_PAR DOWN_IRRADIANCE380 DOWN_IRRADIANCE412 DOWN_IRRADIANCE490 DOWNWELLING_PAR VRS_PH PH_IN_SITU_FREE PH_IN_SITU_TOTAL FLUORESCENCE_CHLA BETA_BACKSCATTERING700 FLUORESCENCE_CDOM CHLA BBP700 CDOM TEMP_NITRATE TEMP_SPECTROPHOTOMETER_NITRATE HUMIDITY_NITRATE UV_INTENSITY_DARK_NITRATE UV_INTENSITY_DARK_NITRATE_STD FIT_ERROR_NITRATE UV_INTENSITY_NITRATE NITRATE PPOX_DOXY,20240110013415

V - Existing tech files associated index

a) Actual list of fields

ar_index_global_tech.txt

file,institution,date_update

aoml/13857/13857_tech.nc,AO,20210428200335

VI - Homogeneous list of fields

New list

argo-one-prof-index.csv

file,**type,**date,latitude,longitude,**date_min,date_max,lat_min,lat_max,lon_min,lon_max,**ocean,profiler_type,institution,parameters,parameter_data_mode,**parameter_quality**,**ad_psal_adjustment_mean,ad_psal_adjustment_deviation**,**date_creation,date_update,**gdac_date_update,**n_levels,max_pressure,ice_detection**

with type =

- 'PC' for R/D core
- 'PB' for BD/BR files
- 'PS' for SD/SR files
- ‘MPC’ for multiprofile files for ‘core’
- ‘MPS’ for multiprofile files for ‘synthetic’
- ‘M’ for metadata file
- ‘TE’ for tech file
- ‘TR’ for traj file

A few complementary considerations:

DEEP: when the float is a deep float (from profiler_type indication) then: type = concatenation of ‘type when not deep’ and ‘D’. For instance, the deep core profile files type would be ‘PCD’.

POSITION: with latitude and longitude displaying 4 digits in the decimal part when available (request from a user in POKaPOK).

ICE DETECTION: Additional suggestion at the ADMT meeting was to consider adding a flag for whether the float detected ice. It could be achieved by retrieving the same information as the one used to display the parameter TECH_FLAG_IceDetection_NUMBER in the FleetMonitoring (e.g. with 6903258), when it exists. A void field would mean the parameter does not exist, a zero would mean ice was never detected. As the information is within the technical file, it makes sense to add it in the lines associated with 'TE' type.

DATES: date_update and **date_creation** refer to the associated fields inside the netCDF file, gdac_date_update refers to the last file change as stated by the stat command/modify date.

Empty fields and specific information with respect to "type":

The following table sums up the fields that would remain empty as not applicable to the file type:

File Type	Empty fields	Specific information
PC, PS	**date_min,date_max,lat_min,lat_max,lon_min,lon_max, ice_detection**	max_pressure = max(PRES where PRES_QC in (1,2,5,8))
PB	**date_min,date_max,lat_min,lat_max,lon_min,lon_max, ad_psal_adjustment_mean,ad_psal_adjustment_deviation,ice_detection**	max_pressure = max(PRES)
MPC, MPS	date,latitude,longitude,ocean,parameter_data_mode,**parameter_quality,ad_psal_adjustment_mean,ad_psal_adjustment_deviation,n_levels, ice_detection**	max_pressure = max (PRES where PRES_QC in (1,2,5,8))
M	date,latitude,longitude,**date_min,date_max,lat_min,lat_max,lon_min,lon_max,**ocean,parameter_data_mode,**parameter_quality,ad_psal_a**	Shall we use date,latitude,longitude to mention launch_dat, launch_lat, launch_lon ?

	djustment_mean,ad_psal_adjustment_deviation,**n_levels,max_pressure, ice_detection**	
TE	date,latitude,longitude,**date_min,date_max,lat_min,lat_max,lon_min,lon_max,**ocean,parameter_data_mode,**parameter_quality**,**ad_psal_adjustment_mean,ad_psal_adjustment_deviation**,**n_levels,max_pressure**	ice_detection = 1 if at least once the relevant technical parameter (TECH_FLAG_IceDetection_NUMBER for nke floats) mentions that ice was detected. ice_detection = 0 if the relevant technical parameter was found and never indicated that ice was detected. ice_detection =’’ if the relevant technical parameter was not found.
TR	date,latitude,longitude,ocean,parameter_data_mode,**parameter_quality**,**ad_psal_adjustment_mean,ad_psal_adjustment_deviation**,**n_levels, ice_detection**	parameter is TRAJECTORY_PARAMETERS field content max_pressure = max(PRES where PRES_QC in (1,2,5,8))

Let construct a few examples with each "type"

aoml/7901106/profiles/D7901106_043.nc,20230919132746,22.166,-156.422,P,846,AO,20230922130615,B,B,,-0.005,0.000,20230922203741,20230922203741,496

Type	example	Additional size
‘PC’	aoml/7901106/profiles/D7901106_043.nc,**PC,**20230919132746,22.165**7**,-156.421**8,,,,,,,**P,846,AO,**PRES TEMP PSAL,DDD**,ABB,-0.005,0.000,20230922203741,20230922130615,20230922203741,496**,1599.1,**	+ 37 bytes compared to argo_profile_detailled_index.txt

‘PCD’	aoml/7901137/profiles/R7901137_010D.nc,**PCD**,20240323150743,-51.898**4**,88.791**1,,,,,,,**I,874,AO,**PRES TEMP PSAL,AAA**,AAA,,,20240325104028,20240327020138,20240327024046,523,**4005.2,**	+ 38 bytes compared to argo_profile_detailled_index.txt
‘PB’	aoml/1900722/profiles/BD1900722_001.nc,**PB,**20061022021624,-40.316**0**,73.389**0,,,,,,,**I,846,AO,PRES TEMP_DOXY BPHASE_DOXY DOXY,RRRD, **B,,,20120520122644**,20200312153230,**20200725074645**,**71,2000.0,**	+59 bytes compared to argo_bio-profile_index.txt
‘PS’	aoml/7901108/profiles/SR7901108_012.nc**,PS**,20240707010139,-40.723**0**,126.251**8,,,,,,,**I,846,AO,PRES TEMP PSAL DOXY CHLA BBP700 PH_IN_SITU_TOTAL NITRATE DOWN_IRRADIANCE380 DOWN_IRRADIANCE443 DOWN_IRRADIANCE490 DOWNWELLING_PAR,AAAAAAAARRRR,AAAAAABAAAAA,**,,20240709020639**,20240709020639,**20240709020639**,**554,1699.3,**	+55 bytes compared to argo_synthetic-profile_detailled_index.txt
‘MPS’	aoml/1901379/1901379_Sprof.nc,**MPS,,,,20091106152133,20131105033522,21.853,24.912,-170.12,-157.979,**846,AO,PRES TEMP PSAL DOXY NITRATE,**,,,,20240629213017**,20240629213017,**20240629213019**,,**1298.1**,	+111 bytes compared to argo_sprof_index.txt
‘M’	aoml/7901137/7901137_meta.nc,**M,,,,,,,,,**874,AO,**TEMP PSAL PRES,,,,,20240710215609**,20240710215609,**20240710224155,,**	+62 bytes compared to ar_index_global_meta.txt
‘TE’	aoml/1901379/1901379_tech.nc,**TE,,,,,,,,,,**AO**,,,,,,,,20210428220054**,20210428220054,**20210428224520,,,**	‘+53 bytes compared to ar_index_global_tech.txt
‘TE’ with ice detected	coriolis/6903258/6903258_tech.nc**,TE,,,,,,,,,,**IF**,,,,,,,,20230927090759**,20240627232909,**20240628003621,,,1**	‘+54 bytes compared to ar_index_global_tech.txt

'TR'	bodc/3901578/3901578_BRtraj.nc**,TR**,,,,**20230316203450,20240109112930,,,,,**836,BO,PRES C1PHASE_DOXY C2PHASE_DOXY TEMP_DOXY DOXY RAW_DOWNWELLING_IRRADIANCE380 RAW_DOWNWELLING_IRRADIANCE412 RAW_DOWNWELLING_IRRADIANCE490 RAW_DOWNWELLING_PAR DOWN_IRRADIANCE380 DOWN_IRRADIANCE412 DOWN_IRRADIANCE490 DOWNWELLING_PAR VRS_PH PH_IN_SITU_FREE PH_IN_SITU_TOTAL FLUORESCENCE_CHLA BETA_BACKSCATTERING700 FLUORESCENCE_CDOM CHLA BBP700 CDOM TEMP_NITRATE TEMP_SPECTROPHOTOMETER_NITRATE HUMIDITY_NITRATE UV_INTENSITY_DARK_NITRATE UV_INTENSITY_DARK_NITRATE_STD FIT_ERROR_NITRATE UV_INTENSITY_NITRATE NITRATE PPOX_DOXY**,,,,,20230519163111**,20240110013415,**20240126221144,,2020.7**,	'+ 79 bytes compared to argo_bio-traj_index.txt

Additional size:

Here are the various sizes as of 4th April 2024, with estimated increase due to additional fields:

Index file	Number of lines (without header lines)	Actual size	Additional fields	New size	Number of characters per line (Actual + additional fields)		
					Min	Max	Ave
argo_profile_detailled_index.txt	2 954 944	417 535 602 Bytes	+ 37 x 2 954 944 = + 109 332 928 bytes	526 868 530 Bytes	99 + 37	218 + 38	140.3 + 37 = 177.3
argo_synthetic-profile_detailled_index.txt	308 330	48 635 006 Bytes	+ 55 x 308330 = + 16 958 150 bytes	65 593 156 Bytes	93 + 55	336 + 55	156.7 + 55 = 211.7
argo_bio-profile_index.txt	309 541	87 952 947 Bytes*	+ 59 x 309541= + 18 262 919 bytes**	106 215 866 Bytes	91 + 59**	1338 + 59**	283.1 + 59** = 342.1

OneArgoIndex-prof.csv hypothetic new index	**3 572 815**			**698 677 552 Bytes**	**99+37**	**1338 + 59**	**194.99**
ar_index_global_traj.txt	20 498	1 710 671 Bytes	+79 x 20 498 = + 1 619 342 bytes	3 330 013 bytes	56 + 79	89 + 79	82.4 + 79 = 99.4
ar_index_global_Btraj.txt	103	20 848 Bytes	+ 79 x 103= + 8 137 bytes	28 985 bytes	89 + 79	593 + 79	196.5 + 79 = 198.5
OneArgoIndex-traj.csv hypothetic new index	**20 601**	**1 731 519 Bytes**		**3 358 998 Bytes**	**56+79**	**593 + 79**	**99.9**
TO BE CONTINUED with S files, tech and metadata files							

*) despite one-less field, the size of argo_bio-profile_index.txt is much greater than argo_synthetic-profile_detailled_index.txt because the intermediate parameters are also listed in there, whereas they are not in the synthetic index.

**) it depends on the number of parameters, low value provided (3 parameters)

VII - File Splitting Strategy for Super index

The question is: what size should we not over cross in order for most users to be able to transfer (eased when zipped) and to load (Matlab, Python) with minimal file manipulation?

- Shall we split by size ?
- Shall we split by number of lines ?
- How do we define the limits if we decide to split by size or number of lines ? (500 MB ? 1 GB ? more ? 5 million lines ? 10 ? more/less? Etc.)

The number of line criterion may be the easiest one. To be discussed.

VIII - Super index subsetting for specific information

subsetting file	criteria
OneArgoIndex-profMulti	type in (MPC,MPS,MPCD,MPSD)
OneArgoIndex-profCore	type in (PC,PCD)
OneArgoIndex-profBGC	type in (PS,PSD)
OneArgoIndex-profDeep	type in (PCD,PBD)
OneArgoIndex-traj	type in (TR)
OneArgoIndex-tech	type in (TE)

OneArgoIndex-meta	type in (M)