Toward homogeneous Argo Index files

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

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

OBSOLETE 0.1 version. New link:

Argo_index_files_homogeneisation_V0.2.docx - Google Docs

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.

The suggestion from @RomainCancouet to have one index as listed below will ensure a homogeneous aspect:

- argo-one-prof-index.csv
- argo-one-meta-index.csv
- argo-one-traj-index.csv
- argo-one-tech-index.csv

As of today, there are only one tech and one meta, so there is nothing to add about them in this document. **On the other hand, there are several profile indexes and traj indexes.**

Two intertwined questions are related to a homogeneous list of fields:

- On which homogeneous fields do we agree?

- What are the size limit per file that we can afford and what about the access performance?

II - argo-one-prof-index.csv

There are currently six indexes that relates to profile or multiprofiles files:

Index file	Which input files?	Comment
argo_sprof_index.txt	{wmo}_Sprof.nc	Multiprofile file index
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

The multi-profile files content is quite different from the profile files. I suggest not including this type in the homogenisation of the argo-one-prof-index.csv and let it be as it is.

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

c) Homogeneous list of fields

New list

Would we like the argo-one-prof-index.csv to contain the concatenated information from the existing indexes related to profile files plus the information regarding the deep? If so, this would yield something like:

argo-one-prof-index.csv

file,**type,**date,latitude,longitude,ocean,profiler_type,institution,date_update,parameters,parameter_data_mode,**parameter_quality**,**ad_psal_adjustment_mean**,**ad_psal_adjustment_deviation**,**gdac_date_creation,**gdac_date_update,**n_levels**,**max_pression**

with type =

- 'C' for R/D core not deep
- 'CD' for core deep
- 'B' for BD/BR files
- 'S' for SD/SR files.
- Any other (BD, SD ?) ?

with max_pression = max(PRES) value with QC in 1,2,5,8 apart from 'B' files for which PRES_QC is not provided and max_pression could be set to max(PRES)

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

Let construct an example with a line for each "type"

argo-one-prof-index.csv:

file,type,date,latitude,longitude,ocean,profiler_type,institution,date_update,parameters,parameter_data_mode,parameter_quality,ad_psal_adjustment_mean,ad_psal_adjustment_deviation,gdac_date_creation,gdac_date_update,n_levels,max_pression

aoml/7901106/profiles/D7901106_043.nc,**C,**20230919132746,22.165**7**,-156.421**8**,P,846,AO,20230922130615,**PRES TEMP PSAL,DDD**,ABB,-0.005,0.000,20230922203741,20230922203741,496**,1599.1**

aoml/7901137/profiles/R7901137_010D.nc,CD,20240323150743,-51.8984,88.7911,I,874,AO,20240327020138,PRES TEMP PSAL,AAA,AAA,,,20240325104028,20240327024046,523,4005.2

aoml/7901108/profiles/SR7901108_002.nc,S,20240326060337,-38.6436,126.9879,I,846,AO,PRES TEMP PSAL DOXY CHLA BBP700 PH_IN_SITU_TOTAL NITRATE DOWN_IRRADIANCE380 DOWN_IRRADIANCE443 DOWN_IRRADIANCE490 DOWNWELLING_PAR,AAARAARRRRRR,AAAFAAFFAAAA,0.0000,0.0000,20240402095137,20240402095137,554,1599.8

aoml/1900722/profiles/BD1900722_001.nc,B,20061022021624,-40.316,73.389,I,846,AO,PRES TEMP_DOXY BPHASE_DOXY DOXY,RRRD,B,,,20120520122644,20200312153230,71,2000.0

Corresponding additional size

'C' and 'CD' data: 173-144 = **+ 29** ('C') / **+ 30** ('CD') characters per line with respect to detailed index (+2/3 for type, +2 for additional position digit, + 15 for parameters (most encountered PRES TEMP PSAL), + 4 for data_mode, -1 for profile_QC, + 7 for PRES_MAX)

'S' : 293 - 249 = **+ 44** characters per line with respect to detailed index (+2 for type, +2 for additional position digit, +14 for psal_adj mean and std, + 15 for date_creation, + 4/5 for N_LEVELS, + 7 for PRES_MAX)

'B' : 165 - 129 = **+36** characters per line with respect to index (+2 for type, +2 for additional position digit, +(**n_param**+1) for profile_QC, +2 for void psal_adj mean and std, + 15 for date_creation, + 4/5 for N_LEVELS, + 7 for PRES_MAX)

Each character is coded on 1 byte. 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
ar_index_global_prof.txt	2 954 990	274 482 107 Bytes (261.8 MB)	N/A	N/A	62 + 29	98 + 30	91.9 + 29
argo_profile_detailled_index.txt	2 954 944	417 535 602 Bytes (398.2 MB)	+ 29 x 2 954 944 = + 85 693 376 bytes	503 228 978 bytes (479.9MB)	99 + 29	218 + 30	140.3 + 29 = 169.3
argo_synthetic-profile_index.txt	308 333	46 345 633 Bytes (44.2 MB)	N/A	N/A	88 + 44	316 + 44	149.3 + 44
argo_synthetic-profile_detailled_index.txt	308 330	48 635 006 Bytes (46.4 MB)	+ 44 x 308330 = + 13 566 520 bytes	62 201 526 Bytes (59.3 MB)	93 + 44	336 + 44	156.7 + 44 = 200.7
argo_bio-profile_index.txt	309 541	87 952 947 Bytes (83.9 MB)*	+ 36 x 309541= + 11 143 476 bytes**	99 096 423 Bytes (95.5 MB)	91 + 36**	1338 + 36**	283.1 + 36** = 319.1
argo-one-prof-index.csv hypothetic new index	**3 572 815**	**554 123 555 Bytes (528.5 MB)**		**664 526 927 Bytes (633.7 MB)**	**62 + 29**	**1338 + 36**	**184.99**

*) 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)

Growing size extrapolated to One Argo

Profile files indexes are growing each day

Current growing rate:

	4th April		5th April		Growing Rate
Index file	Number of lines (without header lines)	Size	Number of lines (without header lines)	size	
ar_index_global_prof.txt	2 954 990	274 482 107 Bytes (261.8 MB)	2 955 502	274 529 933 Bytes	+ 512 lines/day + 47 826 Bytes/day
argo_profile_detailled_index.txt	2 954 944	417 535 602 Bytes (398.2 MB)	2 955 429	417 615 451 Bytes	+ 485 lines/day + 79 849 Bytes/day
argo_synthetic-profile_index.txt	308 333	46 345 633 Bytes (44.2 MB)	308 417	46 359 688 Bytes	+ 84 lines/day +14 055 Bytes/day
argo_synthetic-profile_detailled_index.txt	308 330	48 635 006 Bytes (46.4 MB)	308 399	48 646 955 Bytes	+ 69 lines/day + 11 949 Bytes/day
argo_bio-profile_index.txt	309 541	87 952 947 Bytes (83.9 MB)*	309 634	87 982 482 Bytes	+ 93 lines/day + 29 535 Bytes/day
Sum	**3 572 815**	**554 123 555 Bytes (528.5 MB)**	**3 573 462**	**554 244 888 Bytes**	**+ 121 333 Bytes/day** **+ 647 lines/day**

Extrapolated when OneArgo is operational:

One Argo target is 1000 BGC, 1250 deep and 2450 core only = 4700 operational in total.

OneArgo target will represent + 470 core/deep profiles per day and 100 BGC profiles per day (assuming a classical 10-day cycle) or else said: + 470+100+100= + 670 lines per day **= + 244 717 lines/year**.

Additional size = + 470 x 169.3 (for core/deep) + 100 x 200.7 (synthetic) + 100 x 319.1 (bio) per day = 131 551 bytes per day = **+ 48 049 002 bytes / year (+ 45.8 MB/year)**

Assuming OneArgo from today would mean argo-one-prof-index.csv reach **1 GB** (=1 073 741 824 bytes) in (1 073 741 824-664 526 927)/48 049 002 = **8.5 years = October 2032**

Assuming OneArgo from today would mean argo-one-prof-index.csv reach **5 million of lines** in (5 000 000 - 3 572 815)/ 244 717 = **5.8 years = January 2030**

File splitting strategy

Up to now, the file splitting strategy was to separate types as described here above. However, the index with a limiting size is the core one. 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 type, as of today ?
- 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.)

III - argo-one-traj-index.csv

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

b) Homogeneous list of fields

New list

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

with type =

- 'C' for R/D core not deep
- 'CD' for core deep
- 'B' for BD/BR files
- Any other (BD?) ?

Any other field to add ?

Let's construct an example with a line for each "type"

argo-one-prof-index.csv:

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

aoml/13857/13857_Rtraj.nc**,C**,6.931,0.008,-15.014,-33.808,845,AO,**PRES TEMP,**20210428200335

bodc/3901578/3901578_BRtraj.nc**,B**,,,,,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

Corresponding additional size

'C' and 'CD' data: **+ 17** ('C') / **+ 18** ('CD') characters per line with respect to traj index (+2/3 for type, + 15 for parameters (most encountered PRES TEMP PSAL))

'B' : **+2** characters per line with respect to Btraj index (+2 for type)

Each character is coded on 1 byte. 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 character per line (Actual + additional fields)		
					Min	Max	Ave
ar_index_global_traj.txt	20 498	1 710 671 Bytes (1.63 MB)	+17 x 20 498 = + 348 466 bytes	2 059 137 bytes (1.96 MB)	56 + 17	89 + 17	82.4 + 17 = 99.4
ar_index_global_Btraj.txt	103	20 848 Bytes (0.02 MB)	+ 2 x 103= + 206 bytes	21 054 bytes (0.02 MB)	89 + 2	593 + 2	196.5 + 2 = 198.5
argo-one-traj-index.csv hypothetic new index	**20 601**	**1 731 519 Bytes (1.65 MB)**		**2 080 191 Bytes (1.98 MB)**	**56+17**	**593 + 2**	**99.9**

Growing size extrapolated to One Argo

Argo trajectory indexes are growing each day

Current growing rate: (less significative on a 1-day analysis, let's update this in a few weeks)

	4th April		5th April		Growing Rate
Index file	Number of lines (without header lines)	size	Number of lines (without header lines)	size	
ar_index_global_traj.txt	20 498	1 710 671 Bytes (1.63 MB)	20 520	1 712 508 Bytes	+ 22 lines + 1 837 Bytes/day
ar_index_global_Btraj.txt	103	20 848 Bytes (0.02 MB)	103	20 848 Bytes	Not representative over 1 day
Sum	**20 601**	**1 731 519 Bytes (1.65 MB)**	**20 623**	**1 733 356 Bytes**	**+ 22 lines/day** **+ 1 837 Bytes/day**

Extrapolated when OneArgo is operational:

One Argo target is 1000 BGC, 1250 deep and 2450 core only = 4700 operational in total.

OneArgo target will represent + 490 new core-only floats per year (assuming a lifetime of 5 years) and + 562 new BGC and deep floats per year (assuming a lifetime of 4 years) , which means (490 + 562*2)= **+ 1 614 lines per year**.

Additional size = + 99.4 * (490+562) + 198.5*562 = 216 125.8 Bytes per year **(+0.2 MB per year**)

File splitting strategy

The file size shall never grow above a non-manageable size. There is no need to split this trajectory index.