

Digital preservation related aspects of the CoreTrustSeal certification: SOCIB's experience

Miquel Àngel Rujula,
Senior Software Engineer, on behalf of SOCIB's team
mrujula@socib.es

FAIRsFAIR webinar
November 4, 2021

www.socib.es

SUMMARY

- **Context**
- **Particularities of the SOCIB Data Repository**
- **The certification process: our experience**
- **Results & expectations**

Context

What is SOCIB?

It's a multi-platform ocean observing & forecasting system, from nearshore to open sea & from events to Climate Change

3 DRIVERS

- Science priorities
- Technology Development
- Society Needs

OPEN DATA PRINCIPLES

DATA ACCESS

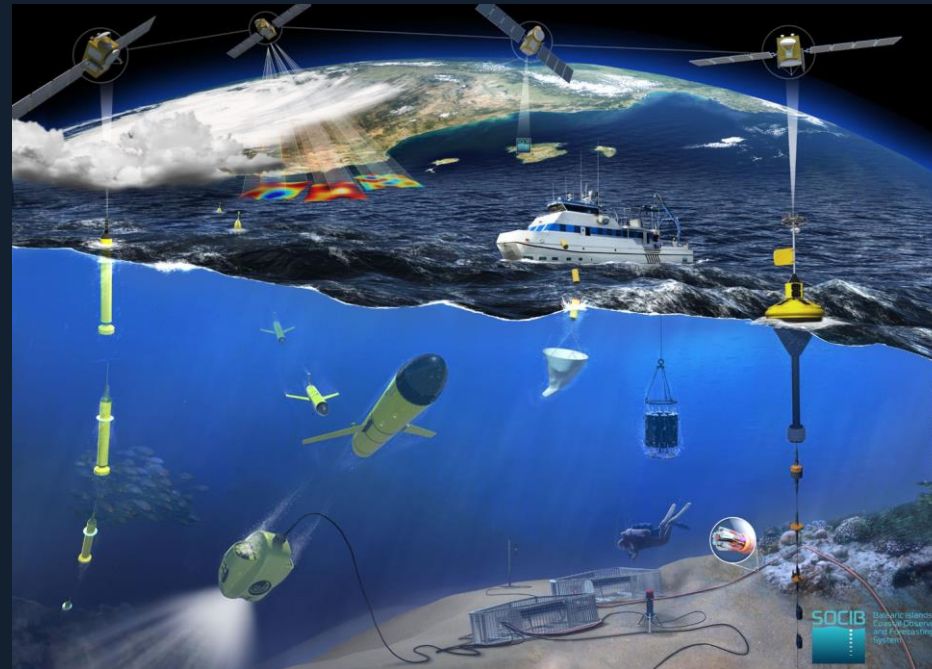
- Free/open data
- Endurance lines
- Competitive Open Access

COLLABORATIVE

- CSIC, IEO, UIB

EVALUATION

- Every 4 years



Since 2014, SOCIB is included in the Large Scale RI map of Spain

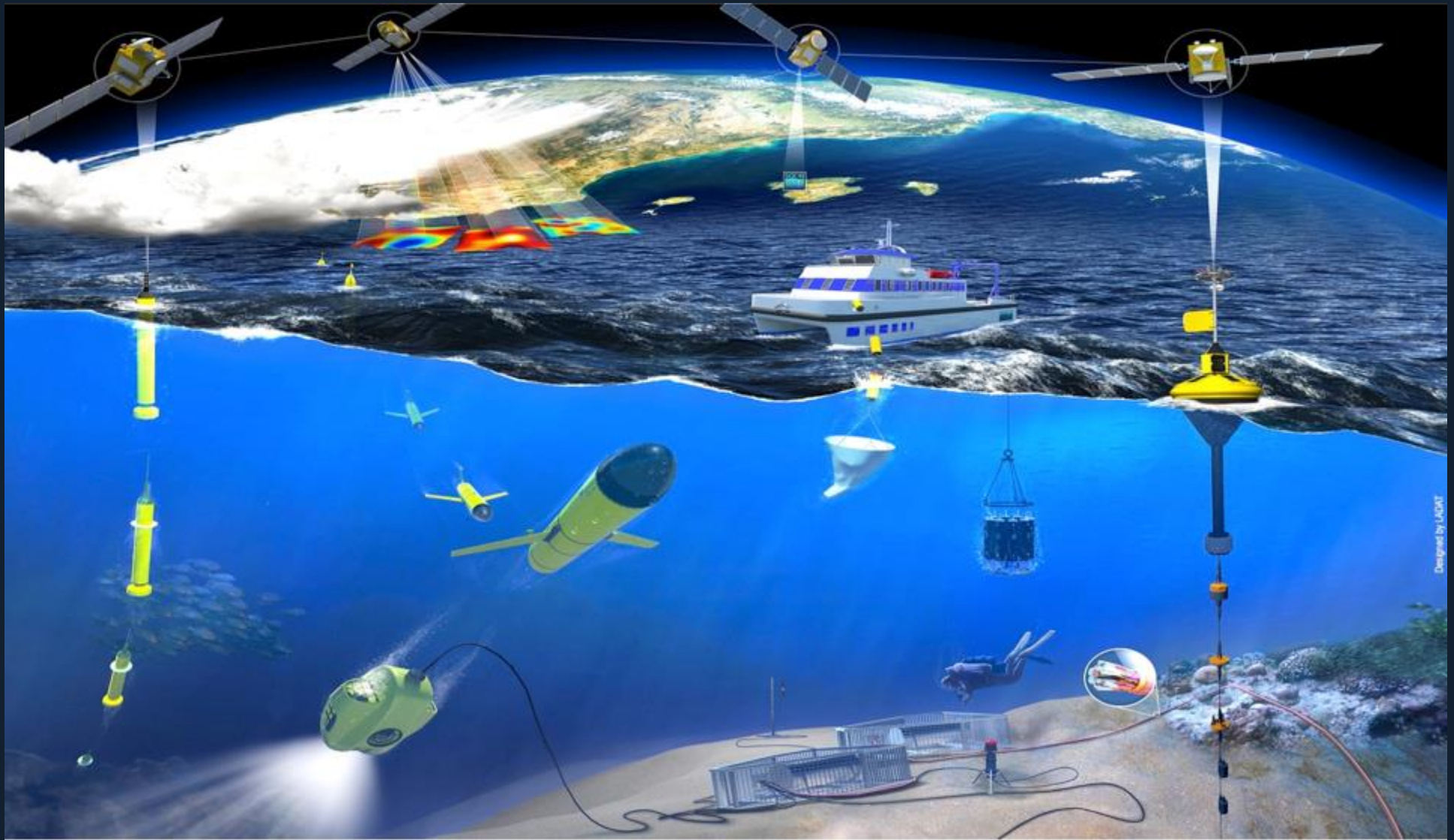
TIMELINE

- Proposal 2006 & approved in 2009
- Designed & built 2010-2013

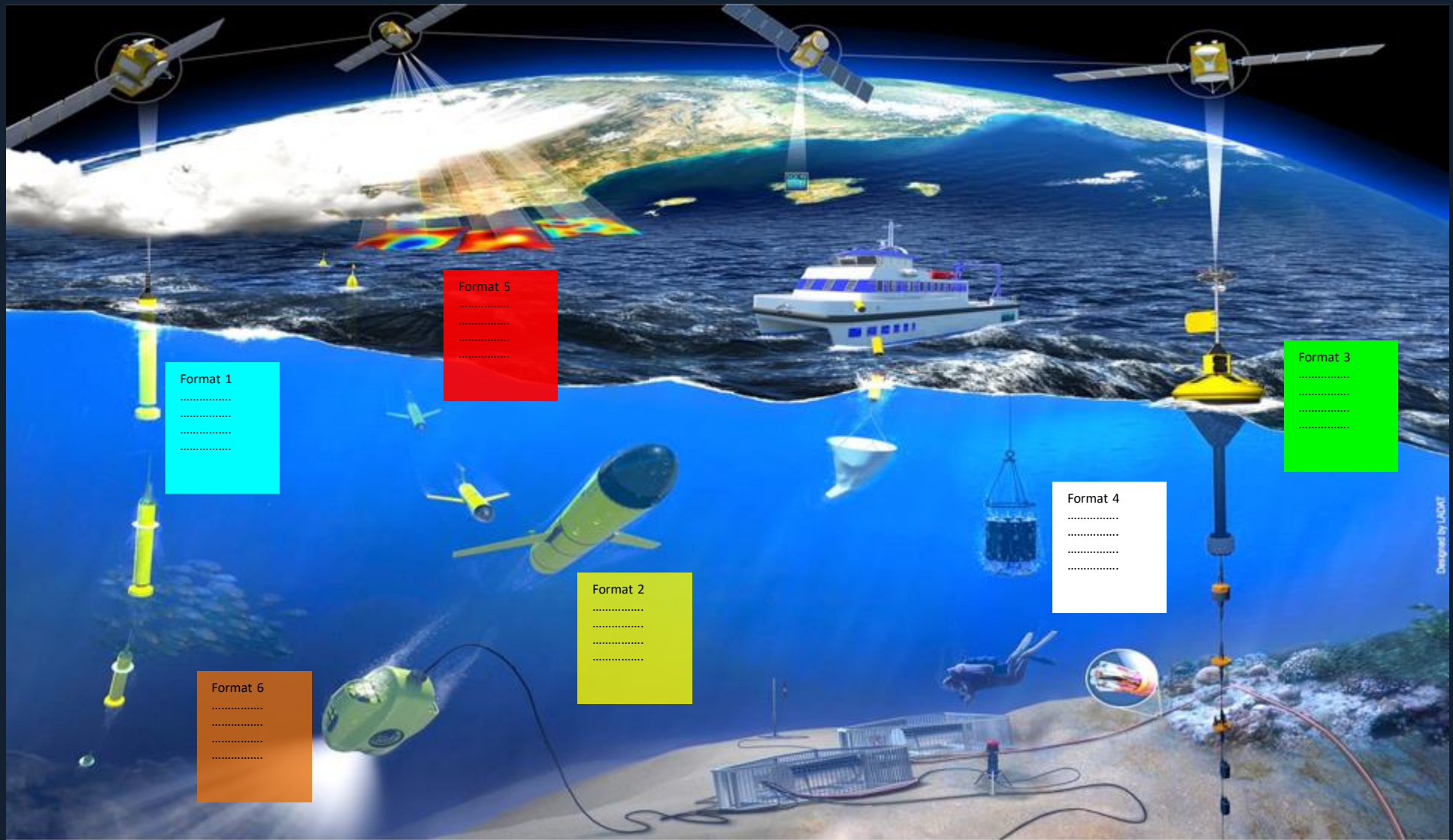
SCIENTIFIC PRODUCTION (KPIs):

- 168 papers, 2011 - 2020
- 16 EU projects, 2014 – 2020
- 7 contracts private sector
- 7 agreements public sector
- External funding > 5 M€
- Building trust and partnerships

Context: multi-platform ocean data



Context: multi-platform ocean data



Context: file formats

Challenge:

- Data from different instruments & sensors: raw data
- Instruments from different manufacturers...
- There isn't a standard file format!
- Scientists (users) need a common format to work with: **NetCDF**

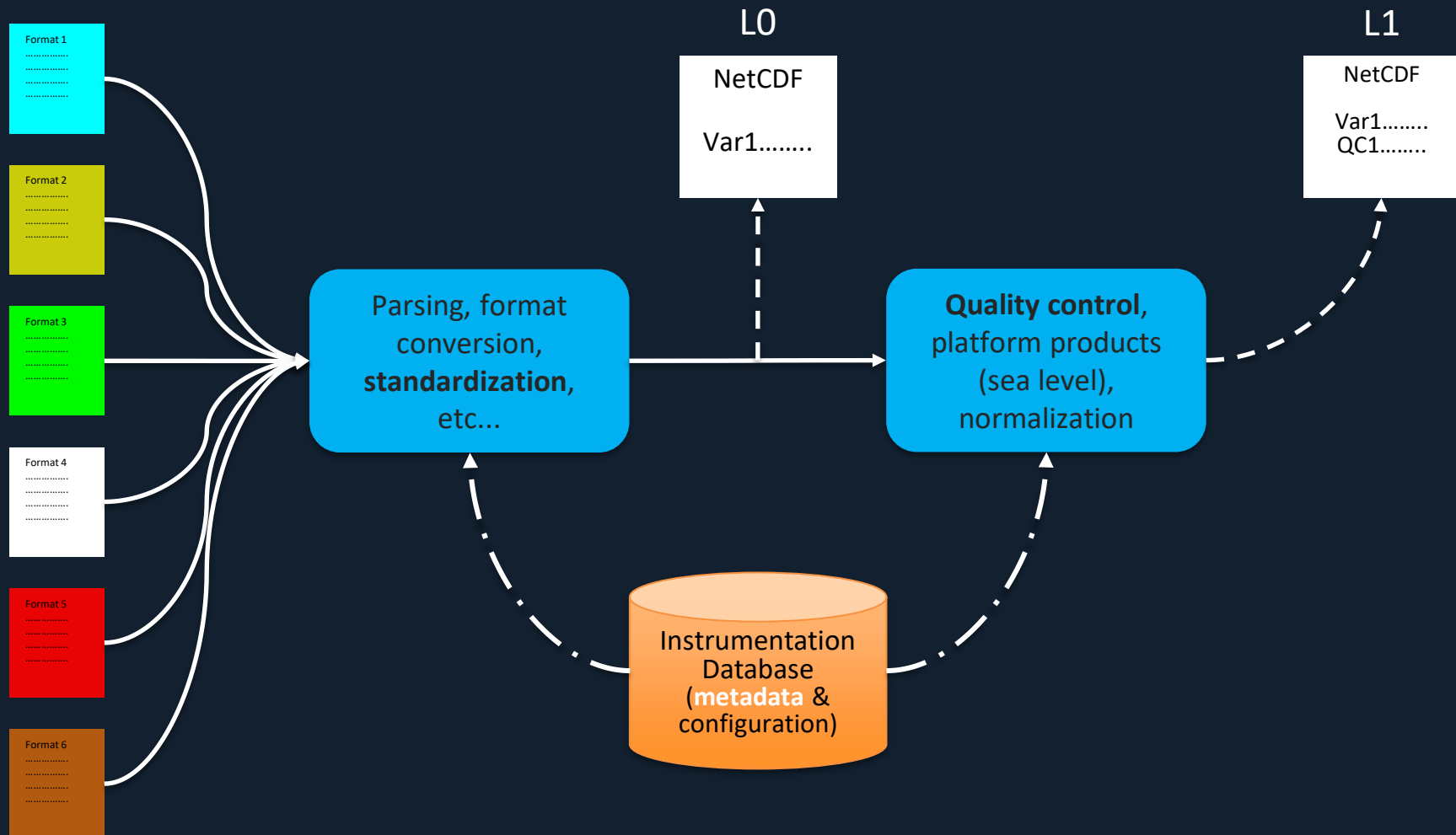
Barometer

```
"TOA5","PortoCristo","CR800","2083","CR800.Std.06","CPU:PORTOCRISTO.CR8","2
7063","Table1"
"TIMESTAMP","RECORD","ID","sampNum","WPRE","WTEM","APRE_Avg","ATEM_A
vg"
"TS","RN","","","PSI","C","mbar",""
","","Smp","Smp","Smp","Smp","Avg","Avg"
"2016-03-01 23:52:00",113649,2,3389767,16.3651,14.57,1022.032,16.76
"2016-03-01 23:52:30",113650,2,3389770,16.3897,14.56,1022.034,16.76
"2016-03-01 23:53:00",113651,2,3389773,16.4121,14.56,1022.033,16.76
"2016-03-01 23:53:30",113652,2,3389776,16.3755,14.56,1022.035,16.76
"2016-03-01 23:54:00",113653,2,3389779,16.3204,14.56,1022.033,16.76
"2016-03-01 23:54:30",113654,2,3389782,16.2765,14.56,1022.033,16.76
"2016-03-01 23:55:00",113655,2,3389785,16.2861,14.56,1022.035,16.76
"2016-03-01 23:55:30",113656,2,3389788,16.2945,14.56,1022.03,16.76
"2016-03-01 23:56:00",113657,2,3389791,16.278,14.56,1022.03,16.76
"2016-03-01 23:56:30",113658,2,3389794,16.3519,14.56,1022.033,16.76
"2016-03-01 23:57:00",113659,2,3389797,16.3809,14.56,1022.035,16.76
"2016-03-01 23:57:30",113660,2,3389800,16.3488,14.56,1022.035,16.76
"2016-03-01 23:58:00",113661,2,3389803,16.3511,14.56,1022.053,16.76
"2016-03-01 23:58:30",113662,2,3389806,16.3623,14.56,1022.041,16.75
"2016-03-01 23:59:00",113663,2,3389809,16.3968,14.56,1022.048,16.75
"2016-03-01 23:59:30",113664,2,3389812,16.3761,14.56,1022.04,16.74
"2016-03-02 00:00:00",113665,2,3389815,16.3328,14.56,1022.038,16.74
```

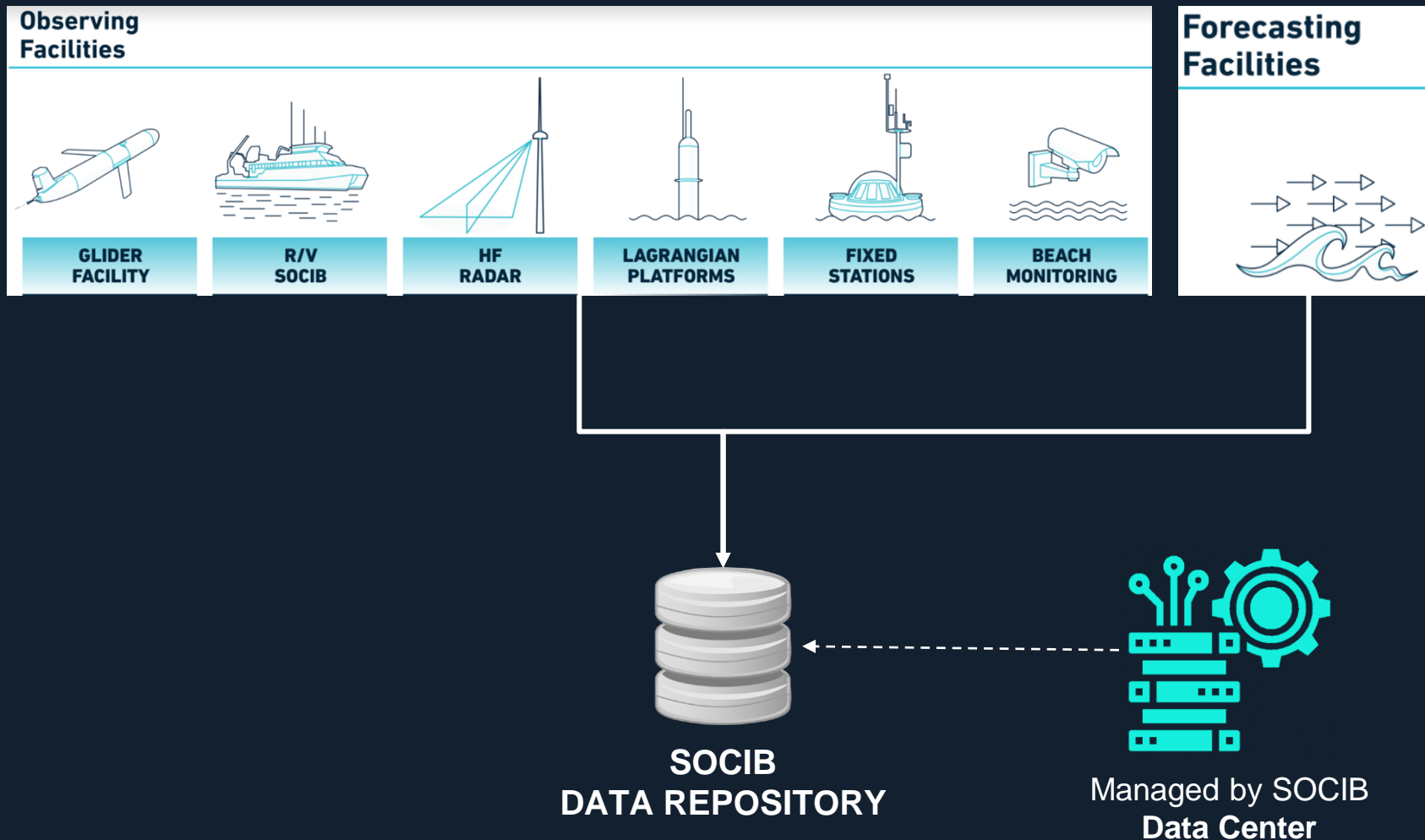
AWAC (current meter)

```
09 07 2011 15 00 00 000000000011100000 14.0 1540.3 323.8 0.6 1.8 18.662 26.61 0 154033 7
1 2.49 0.002 -0.029 0.008 124.0 131.0 127.0 0.029 176.05
2 4.50 -0.030 -0.022 -0.001 110.0 118.0 117.0 0.037 233.75
3 6.50 -0.035 -0.016 -0.004 103.0 107.0 108.0 0.038 245.43
4 8.50 -0.025 -0.026 -0.008 99.0 99.0 99.0 0.036 223.88
5 10.50 -0.016 -0.031 -0.006 91.0 92.0 92.0 0.035 207.30
6 12.51 -0.007 0.030 -0.012 82.0 86.0 82.0 0.031 346.87
7 14.51 -0.007 0.114 -0.011 91.0 88.0 96.0 0.114 356.49
09 07 2011 15 30 00 000000000011100000 14.1 1540.3 324.2 1.0 1.7 18.658 26.60 0 154033 7
1 2.49 0.017 -0.012 0.004 123.0 122.0 122.0 0.021 125.22
2 4.50 -0.002 -0.022 0.001 116.0 116.0 114.0 0.022 185.19
3 6.50 -0.013 -0.020 -0.001 101.0 101.0 106.0 0.024 213.02
4 8.50 -0.014 -0.011 -0.005 98.0 97.0 98.0 0.018 231.84
5 10.50 -0.009 -0.012 -0.004 92.0 90.0 90.0 0.015 216.87
6 12.51 -0.007 0.025 -0.009 86.0 85.0 85.0 0.026 344.36
7 14.51 0.016 0.102 -0.013 102.0 100.0 103.0 0.103 8.91
09 07 2011 16 00 00 000000000011100000 14.1 1540.2 324.0 0.8 1.7 18.667 26.57 0 154023 7
1 2.49 0.000 -0.008 0.006 117.0 123.0 116.0 0.008 180.00
2 4.50 -0.016 -0.002 0.004 107.0 111.0 108.0 0.016 262.87
3 6.50 -0.018 -0.014 -0.003 100.0 100.0 102.0 0.023 232.13
4 8.50 -0.029 -0.021 0.000 93.0 92.0 95.0 0.036 234.09
5 10.50 0.005 0.016 -0.007 90.0 91.0 88.0 0.017 17.35
6 12.51 0.021 0.031 -0.005 84.0 84.0 87.0 0.037 34.11
7 14.51 0.003 0.158 -0.006 94.0 92.0 105.0 0.158 1.09
09 07 2011 16 30 00 000000000011100000 14.1 1540.2 324.0 0.7 1.7 18.666 26.56 0 154023 7
1 2.49 -0.036 -0.004 -0.001 118.0 122.0 121.0 0.036 263.66
2 4.50 -0.010 -0.021 0.002 107.0 111.0 106.0 0.023 205.46
3 6.50 -0.023 -0.018 -0.003 102.0 101.0 98.0 0.029 231.95
4 8.50 -0.010 -0.002 0.000 98.0 97.0 97.0 0.010 258.69
5 10.50 -0.020 0.029 -0.007 90.0 93.0 93.0 0.035 325.41
6 12.51 0.000 0.044 -0.005 88.0 85.0 87.0 0.044 0.00
7 14.51 0.018 0.079 -0.012 102.0 93.0 110.0 0.081 12.84
```

Context: data ingestion



SOCIB



Particularities

- Most data is self-produced
- Topic-related (metocean data). Not a general archive
- Small percentage of data comes from third-party institutions/users
 - Competitive Open Access → Gliders fleet, Research Vessel
 - European & national projects → particular agreements
- Free open access
 - Few exceptions (restricted access):
 - to protect the environment of sensitive areas and/or of sensitive species
 - low level of curation
 - to protect data which may be related to Spanish strategic interests (e.g. Spanish Navy)
 - to give PhD. students and scientists time for publishing before wider distribution (“moratorium”)

CoreTrustSeal certification important for:

- Now → Competitive Access & European/national projects
- Future → To become a national ocean data repository

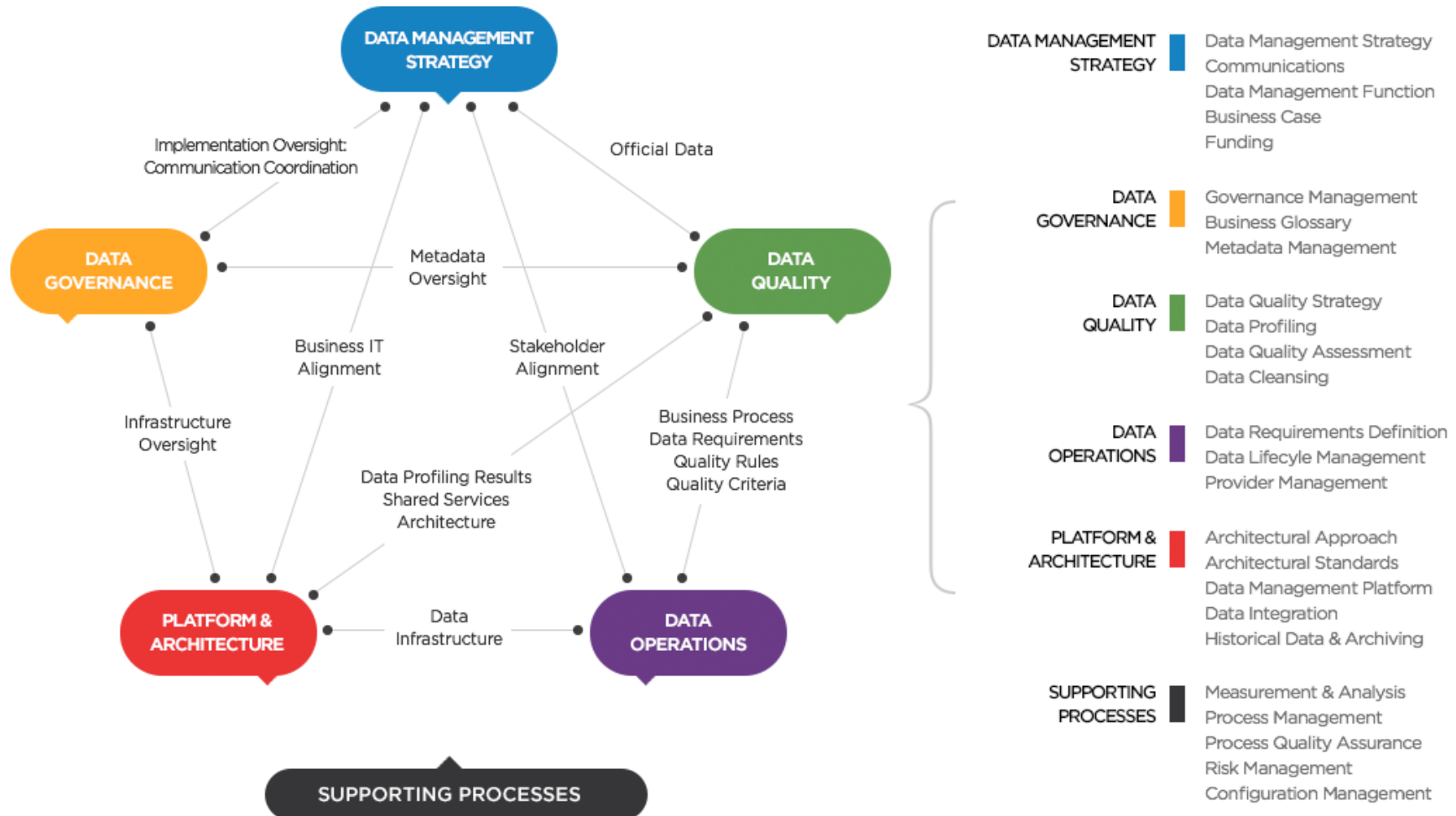
Before starting the certification process

- The data repository didn't exist as an entity
- Managed processes, but partially defined
- We didn't have a clear idea of what a preservation plan (and digital preservation in general) consisted of 😞

“What's preservation? Backups...!?” 🤔

Before starting the certification process

- Since 2018 working on Data Management Program, based on a [Data Management Maturity model](#) (CMMI Institute)



Process of gathering the information

- Working groups:
 - Involved staff: mainly Data Center, C&IT, direction
 - Requirements analysis (R0-R16)
 - Answering questions from the CTS guide
- Identification of gaps/weaknesses
 - Continuity of access (R3)
 - **Preservation plan (R10)**



Align new Strategic Plan 2021-2024 to fill the gaps

Preparing R10: Preservation Plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

- Identified preservation-related actions/processes → We were doing preservation
- **BUT**, barely documented in separate, internal documents
 - IT infrastructure, backups policy, etc
 - Data Management Plans
 - Quality Control procedures, NetCDF user manual....



We didn't have an overall **plan/strategy!**



Can we apply without a Preservation Plan???




**KEEP
CALM
AND
MAKE A
PLAN**

Preservation Plan

Let's do it!

- FAIRsFAIR workshop (February 2021)
 - Preservation Policy planning worksheet helped a lot



Preservation Policy Planning - Worksheet

This exercise has been developed to support selected repositories in the FAIRsFAIR repository support programme in getting CoreTrustSeal certified. It is part of a variety of resources prepared by FAIRsFAIR to support services and individuals in their FAIR data practices. Depending on your situation, it may take around 60 minutes to complete the exercise. The worksheet has been partially adapted from the Digital Preservation Coalition's (DPC) [policy toolkit](#) and the [Conversational CoreTrustSeal working paper](#). This worksheet is not endorsed by the CoreTrustSeal Board and will not guarantee repository certification.

Objectives

- To identify the status of relevant policy components to have a preservation policy in place.
- To structure and plan the writing process of the preservation policy.

Keep in mind that the construction of a preservation policy depends on the context of your repository. There is no one-size-fits-all solution, and you can adjust the content, format and style of your policy accordingly.

Instructions

- To start, indicate the status of each policy component in column 'Status'. Use 0 (doesn't exist), 1 (basic information exists), 2 (managed for this particular area) and 3 (defined and forms part of an integrated management and documentation set). Note that 'R' refers to the CoreTrustSeal assessment requirements. Although all of the below mentioned components are relevant when developing a preservation policy, **we recommend that you focus first on components #6 and #8** as these are some of the most challenging. You may also leave your comments or questions. We will collate them and address the most urgent ones during the upcoming workshop.
- Identify the people within your organisation who are subject experts of each policy component and a responsible person assigned to a specific component (you do not need to share this information with us).

Preservation Plan

How we faced the "blank paper"

Index/template based on DANS & UKDA —————→ ***Don't reinvent the wheel!***

1. **Purpose** → R0, R1
2. **Scope** → R0
3. **Requirements**
 1. **SOCIB'S requirements**
 2. **Legal and regulatory framework** → R1
4. **Roles and responsibilities** → R5
5. **Model (OAIS)**
 1. **Pre-ingest function**
 2. **Ingest function** → R0, R8
 3. **Archival storage function**
 4. **Data management function** → R8
 5. **Access function** → R13
 6. **Administration function**
6. **Preservation planning and strategy**
 1. **Preservation strategy overview** → R3, R5
 2. **Integrity measures** → R8
 3. **Monitoring, review and feedback** → R0, R6
7. **IT Architecture** → R15
8. **Security** → R16
9. **Co-operation**
10. **Funding and Resource Planning** → R5
11. **Appendix: Definition of Terms**

Preservation Plan

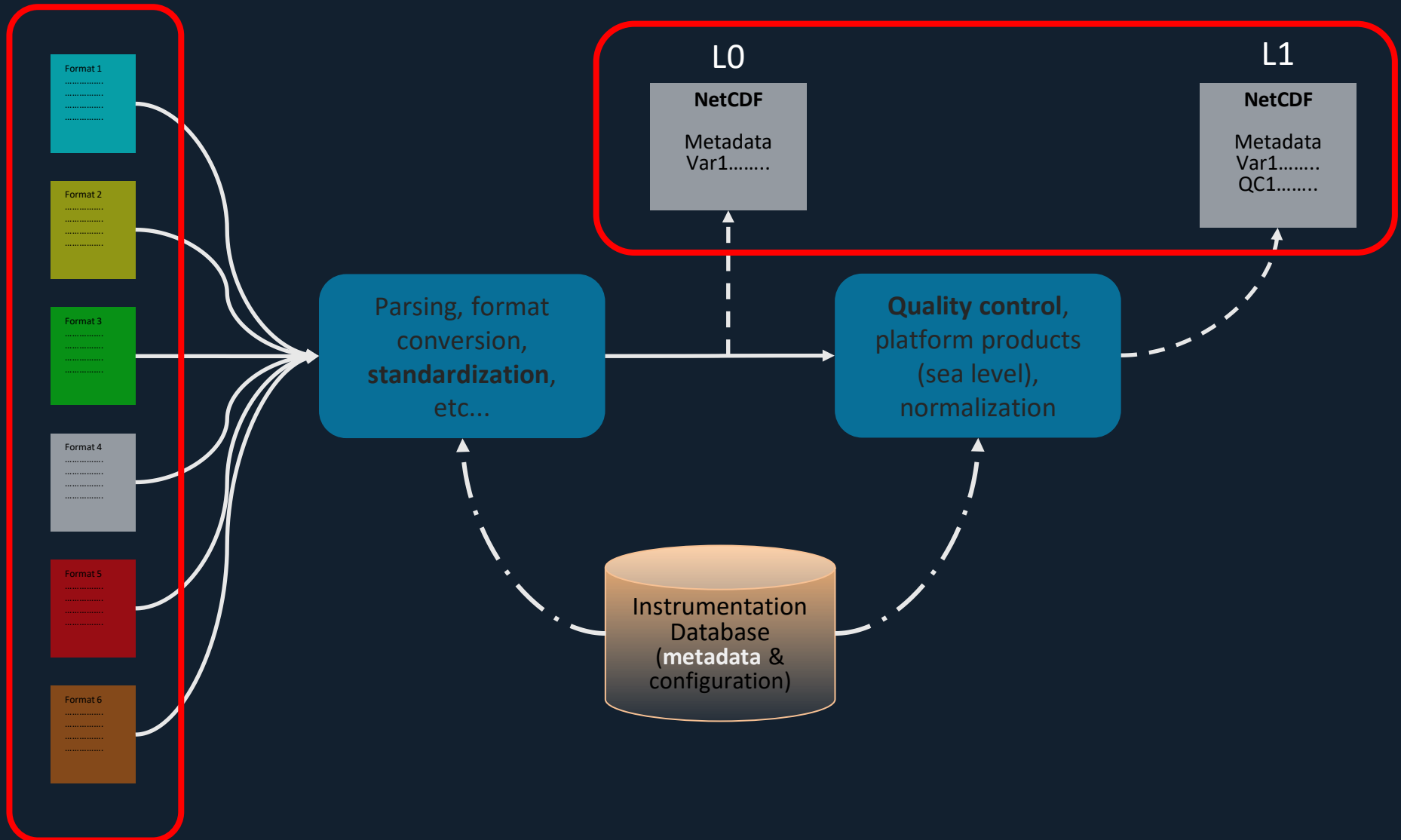
Things to take into consideration

- Information from other CTS requirements can (and has to) be used to write the Preservation Plan → consistency
- Practical day-to-day implementation is out of the scope of the document → generally described in separate, more detailed docs
- Use OAIS concepts (not necessary, but it helps):
 - Functions: pre-ingestion, ingestion, archival, etc...
 - Information packages & preferred formats. What are your...
 - SIP (Submission Information Package)
 - AIP (Archival Information Package)
 - DIP (Dissemination Information Package)
 - Do you preserve all SIPs, AIPs and/or DIPs?
 - Are AIPs accessible?

Preservation plan: identifying SIPs, AIPs & DIPs

SIPs & AIPs (raw files)

DIPs & AIPs (NetCDF files)



Digital preservation & CoreTrustSeal

What is expected from the applicants?

- preservation is not just backups, but also: clearly documented processes, a policy, a model, etc...
- clear idea of the status of your preservation plan, whether you have it or not. It's fine to say:
*“we don't have a preservation plan as such, but we do these preservation-related actions and we have to write them down in a document... We expect to have a **preservation plan** in x months...”*
- it's better to create the plan following a reference model, such as [OAIS](#)
- provide evidence!

Results

- Boosted up internal documentation
- Helped us detect weaknesses (& strengths!)
- SOCIB Data Repository has become an entity
 - Landing page: www.socib.es/data
 - All information in one place available for users: licenses, data policy, preservation plan, ethics/confidentiality, etc...

Expectations



- Obtain CTS certification! (2nd attempt soon)
- National impact
 - Alignment with DIGITAL.CSIC: they already have the CTS certification
 - Become a repository of reference in marine data

- International impact
 - Improve interoperability
 - Become an [IODE Associate Data Unit](#)
 - Regular member of the World Data System
 - Best Practices: contributions in Data Management



Thank you!

Investigamos el mar, compartimos futuro



www.socib.es



@socib_icts



@ICTSSOCIB



ICTS SOCIB