"Bringing together Research and Industry for the Development of Glider Environmental Services"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 635359

DELIVERABLE D3.2

"Standardization of data/metadata for gliders supporting marine science and blue economy"

ABSTRACT

This document describes the recommended standards and/or best practices for managing marine observations ("data") and supplementary information ("metadata") collected from gliders. Information is based on glider models commercially available: what work flow is used at the manufacturer, operator, scientist, and end-user levels and how that compares with the recommendations of the partner report D3.3 "Interface standards for applications of deep and ultra-deep glider." In that report, a proposed interface between data/metadata and data users was described, based on Open Geospatial Consortium standards and current practices. In this report, suggestions on how that interface can and should be implemented are described. Since this work is progress on state of the art, it concerns not only the D and UD Explorer, but gliders in general.

DOCUMENT TYPE	Deliverable
DOCUMENT NAME:	BRIDGES_D3.2_Data_Formatting_Standards
VERSION:	vfinal
DATE:	27-10-2016
STATUS:	S0
DISSEMINATION LEVEL:	PU

Authors, Reviewers				
Author(s):	Daniel Hayes, Matthes Reike, Simon Jirka, Laurent Beguery, Justin Buck			
Affiliation(s):	CSCS, 52N, 52N, ALSEAMAR, NOC (NERC BODC)			
Further Authors:				
Peer Reviewers:	Michael Field, Pierre Testor			
Review Approval:	Approved	Yes	Rejected (to be improved as indicated below)	No
Remarks / Improvements:	No remarks.			

Version History			
Version:	Date:	Comments, changes, Status:	Person(s) / Organisation Short Name:
v0.1	10/08/2016	Draft version	Daniel Hayes / CSCS
v1.1	12/10/2016	First review	Daniel Hayes / CSCS
vFINAL	27/10/2016	Document ready for submission	Daniel Hayes / CSCS

Version Numbering	
v0.x	draft before peer-review approval
v1.x	After the first review
v2.x	After the second review
vfinal	Deliverable ready to be submitted!

Status / Dissemination Level			
Status		Dissemination Level	
S0	Approved/Released/Ready to be submitted	**PU**	Public
S1	Reviewed	**CO**	Confidential, restricted under conditions set out in the Grant Agreement
S2	Pending for review		
S3	Draft for comments	**CI**	Classified, information as referred to in Commission Decision 2001/844/EC.
S4	Under preparation		

TABLE OF CONTENTS

CONTENT

1 Introduction

1.1 Purpose of document

This document is the final deliverable for Task 3.2 Standardization of data formatting and is named “Standardization of data/metadata for gliders supporting marine science and blue economy”. Its purpose is **to review existing standards to format environmental data** in collaboration with industry and scientific standards groups. From this review, a **standardization solution will be proposed** that will strike a balance between hardware capabilities, transmission costs, and user needs while still meeting the needs of stakeholders in marine science and the blue economy. The document will also **outline an implementation plan** for a proposed standardization protocol.

1.2 Purpose of work package WP3 and Task 3.2

The aim of WP3 is to find or develop current best practices for platform manufacturing, data formatting and interface capabilities. A systemic approach and standardization are required to produce a reliable and flexible glider design and thus allow ongoing system upgrade, enhancement and technology insertion and ultimately commercialization of the platform. This particular task (T3.2) is meant to carry out the review of data and metadata formatting and recommend an implementation plan, as described above, as well as take an active role in the ongoing discussion of data standardization (e.g., the ISO/Open Geospatial Consortium or ISO/OGC standards). The other tasks deal with hardware standards for manufacturers (T3.1) and interface standards for ensuring data visibility, storage, and provision to a wide range of stakeholders (T3.3). It is clear that the latter task is closely tied to the present task, since ultimately, the interfaces of T3.3 will serve the data and metadata discussed here to operators, scientists, industry, government, and the wider public.

1.3 Related work packages and projects

Task 3.2 is directly related to work package 5, since the information about and collected by the sensor packages (metadata and data, respectively) described there will have to populate the data structures in the format proposed here. In the ideal case, smart sensors will be used, so that sensor metadata are queried and supplied using standard protocols such as PUCK and SensorML, respectively. In fact, this is one of the recommendations of this report to WP5 via T5.1, as will be discussed later. This task is also related to the overall system development in WP4, since the sensor metadata will be merged with the observed data, along with platform and processing metadata in the final glider prototypes. The populated data structures will then be served via the system described in T3.3.

SenseOCEAN, a collaborative project funded by the European Union 7th Framework Programme under grant agreement No. 614141, aims to “provide a quantum leap in the ability to measure crucial biogeochemical parameters” (http://www.senseocean.eu/). BRIDGES partners (NERC/BODC) are involved in the standardisation of sensor metadata to enable ‘plug and play’ sensor integration and SenseOCEAN will implement the OGC’s Sensor Web Enablement (SWE) standards.

NexOS, also a collaborative project funded by the European Commission 7th Framework Programme, http://www.nexosproject.eu/, aims to "develop new cost-effective, innovative and compact integrated multifunctional sensor systems." Stated objective of WP4 is "to enable interoperable Web access to marine sensors. This will facilitate a rapid integration of useful sensor data into standard open data portals." Some BRIDGES partners (ALSEAMAR, 52N) are involved in evaluating the PUCK protocol, which has been found to be very useful in managing sensor hardware (maintenance and calibration procedures) and the data sets it produces (keeping metadata and data connected).

AtlantOS, a large scale EU Horizon 2020 research and innovation project, https://www.atlantos-h2020.eu/, aims "to deliver an advanced framework for the development of an integrated Atlantic Ocean Observing System that goes beyond the state-of–the-art, and leaves a legacy of sustainability after the life of the project." Part of that effort involves data harmonization because of the multi-platform, multi-parameter data sets expected. In particular, according to Deliverable 7.1 [1], "the work of AtlantOS WP7 is dedicated to improve the data management and interoperability among the observation networks and Integrators involved in AtlanOS." A vocabulary matrix for AtlantOS Essential Ocean Variables (EOVs) was built and validated by partner BODC.

In GROOM, http://www.groom-fp7.eu/, an FP7 research infrastructure design study for a European glider infrastructure which is now completed, the observation file standard was specified. It is known as the "EGO 1.2" format or Everyone's Gliding Observatory 1.2 [2, 3]. It will be recommended in this report to use this format.

1.4 Scope of Task

With community input, this report summarizes data and metadata formats used by the data originators, providers, managers, and user communities. Then conclusions are drawn about future developments that are likely to contribute to the widely-accepted adoption of standard formats. Ultimately, the objective is to improve glider data and metadata discovery and download services by defining and promoting best practices for data collection, processing, and sharing. In particular, it is important to promote standards and best practices that are scalable and can evolve without becoming obsolete or cumbersome, as sensors and processing chains used by gliders increase in type and number. Therefore, we must **ensure key metadata and technical data from sensors are never lost,** as well as secure efficient **data processing, archival, and seamless delivery.** This requires **standardization**. Our working hypothesis is that EGO netcdf format (or perhaps NetCDF Climate and Forecast (CF) Metadata Conventions: http://www.cfconventions.org/documents/) must **be integrated into standard discovery and download services** for glider data.

The information presented here comes from a range of sources. First-hand knowledge as glider operators and principal investigators has provided insight into issues and possible improvements in the current work flow. Other operators and/or scientists have expressed their views at the 7th EGO Conference in September 2016. Glider developers (ALSEAMAR, Kongsberg/University of Washington, Scripps Institute of Oceanography) have provided feedback as to how they handle the standardization issue and their views on future developments. Regional (EuroGOOS Glider Task Team and front office) and global (nascent OceanGliders steering team in JCOMM-Observations Program Area: http://www.jcomm.info/index.php?option=com_content&view=article&id=21&Itemid=38) have provided some feedback as to potential ways forward and possible pitfalls. And of course, data management specialists have commented on the current state of the art and views on the way forward (British Oceanographic Data Center). Finally, other project consortia (such as AtlantOS, SenseOcean, NexOS, GROOM, EGO, JERICO, PERSEUS, SeaDataNet) have significant results relating to data and metadata standardization, however not all have been approached at the time of writing. In the future, feedback could be collected from end-users such as the oil and gas or deep sea mining companies, or their regulatory agencies.

2 WP3 Review of existing standards

In this section the existing standards currently in use are reviewed. A description of the relevant actors is followed by their particular needs regarding data handling and standardization, followed by the recommendations for standardization across the board.

2.1 Actors and existing best practice and workflows

In this section, the various groups involved in handling glider data are identified and their roles are described. How their roles are fulfilled regarding data standards and formatting is also described. A graphical view of the relationships between the actors is shown in Fig. 1. A summary of the groups and their practices is given in Table 1.

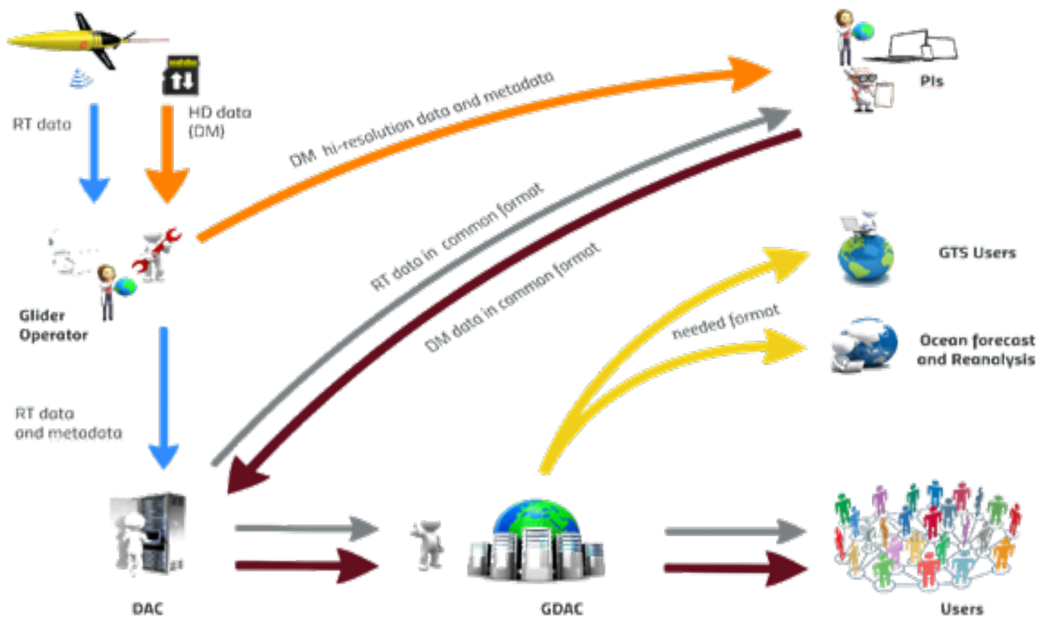

Figure 1.Data management stakeholders and workflows.

2.1.1 Hardware Developers

This group designs and builds glider platforms and/or sensors. Of course the main application of data standards for this group is to produce new and better platforms (i.e. this project's Deep and Ultradeep SeaExplorer) that can properly support a variety of sensors, both presently available and those to be available for years to come. This should be done in a scalable, flexible way so that future data can be collected and provided to the user along with historical data without any extra effort or lost information or quality. From another perspective, sensor manufacturers may want to design sensors which can be integrated or used by a large range of platforms with no extra development, which would be attractive to platform manufacturers because of the reduced development costs. If the data collected can easily arrive at a wide user group with known quality and processing, it is a further benefit of adopting data formatting standards.

In this project, feedback has been collected from the developer of the currently produced SeaExplorer glider and the MiniFluo hydrocarbon sensor (ALSEAMAR) and from the University of Washington who develops Seaglider with commercial partner Kongsberg Underwater Technology Inc. Through activity in developing and/or choosing sensor payloads (WP5), this consortium has also learned of the range of practices regarding data formatting by industrial developers. It was found that sensor manufacturers we have worked with in this project rarely provide data or metadata streams in a standard format. Typically, the user must have a separate user manual to interpret signals or query the sensor for its characteristics. These protocols are typically developed in house and are sent over RS-232 digital interface, and sometimes signal output is sent as an analog voltage output. This situation is a clear disadvantage since measurements collected by such a device are separated from the information about conditions under which they were taken without consistent user effort (e.g. manual data entry into an in-house database).

From a platform perspective, the manufacturers use in-house formats at the glider level, but provide more complete netcdf files to the operator after processing on land (basestation). Some care is needed, since this post-processing is also prone to user error, because of the need to vigilantly ensure that the metadata contained in an ascii text file are attached to the data correctly (*sg_calib_constants.m* on the basestation for some Seaglider configuration and calibration parameters and two files on the SeaExplorer glider: *sea.cfg* for all configuration parameters and *sea.msn* for default navigation behaviour). Please see Appendix A. This is the main disadvantage of this workflow: if a configuration file is lost or modified without proper record keeping, before the production of an acceptable netcdf file, data can become useless. It should be noted that ALSEAMAR's efforts to maintain a configuration file on the glider memory should aid in tracking the history of each glider and reduce the likelihood of lost metadata. In both cases, however, the configuration files are not of a standard format that can easily be integrated into a globally accepted workflow.

In both cases, once the merge of sensor and platform engineering and scientific information occurs on the basestation, all relevant data and metadata are contained in netcdf files provided to the operators, and thereafter could be kept in all derived products. This is a major achievement for the glider platform. Netcdf flavours vary, but should be CF (Climate-Forecast) compliant. These files have the advantage that extensive metadata can be included in a human- and machine-readable header, and that the data structure is described in the file itself. This is an accepted level of netcdf standard by all major data centers, although often a more refined version tailored for a particular platform or need are used (EGO netcdf for gliders, OceanSITES netcdf for fixed point platforms, and others). The tools used are based on in-house software written in Python or Matlab. It should be noted that Python is a freely available programming language that supports netcdf file operations, while Matlab supports the same netcdf functionality but is not free.

2.1.2 Operators

This group could be considered data providers, as it is the one preparing, deploying, and maintaining the equipment for the particular missions required by the end user (such as commercial client, principle investigator/scientist, or government agency). The technical know-how must translate to usable data sets: configuring the sensors and gliders in the desired way, while recording this configuration for future attachment to the data sets, is the primary reason operators should be concerned with data formatting. On one side, they are generally bound to the manufacturers' provided formats and tools, but on the other they are expected to produce properly documented data sets for real time use, and for future delayed-mode processing and archival. For the latter, conversion toolboxes have been developed at some operation centers. Some scientists/PIs are also operators, so advanced tools to carry out quality control are often included at this part of the process. This step regarding delayed-mode data is discussed further in the next section. In some cases, a local data management team also has developed tools for discovering and downloading that group's data sets.

Examples of such operator level toolboxes include the oceano_data_toolbox [4], the Spanish toolbox developed at SOCIB [5], the University of East Anglia's "The UEA Seaglider Toolbox" [6]. A full description of each is beyond the scope of the present report, but common elements can be identified: 1. They aim at handling multiple glider models, 2. They convert manufacturer's format to netcdf format (matlab in the case of UEA), and 3. They are freely available. There is a possibility that they are or will be used for QC processing, since most of them are driven by a data scientist or principal investigator driving the development to aid in their research. Obviously there is some duplication of effort, but certainly these toolboxes are fit for purpose for the labs that support them. In Australia, there is a toolbox for sensor level only, the Australian IMOS Matlab toolbox [7] as well as matlab codes (pers. Comm.) for converting to IMOS CF format [8].

2.1.3 Scientists/PIs

This group is a particular type of data provider or originator because PIs define the objective of a particular glider mission and evaluate the results based on detailed knowledge of the environment to produced delayed mode data. This implies a finely tuned quality control procedure is overseen by the PI. As mentioned, a PI may also operate the glider in practice. Often gliders are only part of a larger planned experiment that addresses a particular process or feature of scientific interest. Recently, the OceanGliders task team identified three such topics in which gliders can provide valuable insight: boundary currents, deep convection, and ocean-atmosphere interaction (data management was identified as a critical cross-cutting issue). Examples, including the three above, of how gliders fit into the global ocean observing system (GOOS), which includes several observing platforms in addition to gliders, are described in [10]. Research scientists predominantly use netcdf formats: either EGO or another CF compliant standard. If no local toolbox is available for conversion or quality control, PIs work directly with the manufacturer netcdf file using their own codes or fine-tuning manufacturer codes if available (e.g. basestation python routines). Some have converted manufacturer formats to ship-based formats such as MedAtlas ascii or Ocean Data View (ODV) for easy integration into hydrographic databases. For example, SeaDataNet II project has been undertaken within the EU FP7 framework from 2011 to 2015 as successor to the earlier SeaDataNet project in EU FP6 (http://www.seadatanet.org/). The project aimed at providing a research infrastructure and tools for archiving and recovering and ultimately providing marine data, including the development and adoption of standards.

For data exchange formats, it developed and endorsed ODV4 ASCII format and also accepts its own flavour of Netcdf CF format and MedAtlas ASCII as an optional format. In some cases, unique formats are used (e.g. Matlab binary in the case of UEA).

The mix of formats above can be a disadvantage when PIs share data with each other during an experiment because often the observations made by the various platforms must be viewed in relation to each other to further plan the experiment. If the different platforms provide widely ranging formats, it becomes difficult to overlay one data set with another. This can happen even when different types of gliders are used in the same experiment. Another obvious disadvantage is when PIs share their data with the research or wider community, either public or private. There is no guarantee that the file provided, even if it is netcdf CF compliant, contains the necessary metadata. At least in the case of EGO, IMOS, and IOOS netcdf, a particular minimum set of metadata is required.

2.1.4 Data Centers and Developers

This group consists of experts in handling data sets from a variety of marine platforms and formats. This includes developing and maintaining codes to convert to widely-accepted formats and to perform quality control. These experts also develop interfaces which allow users to discover and download data, and they continuously develop the standards and formats themselves. Data Acquisition Centers (DACS) often convert files sent by operators to a standard netcdf format (EGO in the case of Coriolis Data Center and BODC [2, 3], and other versions of CF 1.x for IMOS/ANFOG [8] and U.S. IOOS National Glider Data Assembly Center (NGDAC) [9]). A generic quality control is typically run and data transmitted to real time channels (like GTS or their own web site). The files can be further examined by PIs and re-submitted as delayed mode after detailed quality control. These data centers can be found at: http://www.coriolis.eu.org, http://www.bodc.ac.uk/, https://portal.aodn.org.au/, https://gliders.ioos.us/index.html.

In BRIDGES, T3.3 identified the current state of glider (meta-) data management, including the assessment of data and metadata storage and provision as well as requirements analysis and system design to achieve interoperability. It was found that storage and discovery are not streamlined and the usage of ontologies (e.g. for instrument types or measured phenomena) are not established. However, as described in D3.3, the netcdf files can be used to populate more generic, data and metadata model structures. These are Open Geospatial Consortium (OGC) standards, and ISO approved, and form part of the larger Sensor Observation Service for both observations data and metadata and sensor metadata. This will be discussed as a recommendation in Sec 3.5. The point here is to realize that data scientists deal with markup languages (SensorML) and generalized formats (Marine Sensor Web Enablement (SWE) profile) in order to make a wide variety of data sets interoperable. Ideally this can happen without changing the fundamental data file format for a particular platform like gliders, if there is one, but be mapped or linked to it. In fact there is such a candidate for the glider community: it has developed a netcdf format in the framework of the EGO (Everyone's Gliding Observatory) and FP7 Gliders for Research Observation and Ocean Management (GROOM) project (2). A best practices document for data management is already in development (3).

2.1.5 End users

Blue industry, researchers, authorities, public organizations and individuals are all stakeholders when it comes to marine data, including data from gliders. A range of applications can be imagined: from monitoring a mining or offshore platform operation, to measuring baseline conditions for environmental regulatory agencies (See Deliverable D2.1). With well-documented quality-controlled data in the public domain, in particular, the data may be used in the future for some presently unknown purpose such as to develop a new product or service, or extract information about the environment previously ignored.

Currently, end users must accept whatever format is provided to them. From the previous descriptions, this could be a manufacturer's format, a scientist's or research group's favourite format, a community-accepted format, depending on the originator and the pathway by which the data arrived. It is acceptable that there are at least two pathways: a real time and a delayed mode. However, it is not acceptable that each of these pathways have several variants in format. Of utmost importance is the fact that many of the currently-used formats do not contain proper metadata, or if they do, it is not done in a standard way in order to be joined together in a common interface. It is also undesirable that end users may have to visit several data center portals to find glider data, since no one portal currently provides a full picture, even of open data provided by the originators to their favourite DAC.

2.2 Summary of existing practices and pitfalls

This review of standards reveals weaknesses at all stages of the lifetime of a given data set. Firstly, it is possible that data are not properly matched with metadata by the manufacturer or operator. Diligent effort by both actors is required to avoid this. The scientists and PIs, assuming the data and metadata are present, will process and control them, including adding and/or controlling the metadata before providing the delayed mode data, possibly in a different and/or "community" format. The pitfalls seen here are that these formats may not be interoperable with each other or with other marine platform data sets, and that the metadata are not adequately included. It also appears there has been a duplication of effort in generating toolboxes to convert or process data sets. At the data center level, workload around glider data is high as a result of the lack of standardization. Much work goes into ensuring observations are stored and provided, but it appears metadata are not yet addressed sufficiently based on the varied status provided by different data centers. While many actors use netcdf formats, the range of flavours, in particular regarding metadata, is an obstacle to interoperability. A number of recommendations to address the observed pitfalls follows.

3 Recommendations for formatting

3.1 Sensor Level

After discussions with ALSEAMAR, and based on first hand experience, the experts in T3.2 recommend that sensors that can store and provide their metadata should be strongly preferred, in particular if sensors are queried and respond using a standard protocol. This should be the PUCK protocol [11]. Very few commercially-available sensors achieve this, but partner CSCS is developing a wet payload pluggable interface that can turn almost any sensor into a 'smart' sensor with a small penalty in power and mass to the platform. Further work with the NexOS project will be required to elaborate the next steps. The characteristics such as calibration information, hardware identification, instrument status, etc., should be stored as a SensorML file, the the sensor should respond to PUCK commands by extracting the information requested from the SensorML file.

3.2 Platform level

Again after discussions with ALSEAMAR, and based on first hand experience with other glider platforms, the experts in T3.2 recommend that gliders should contain a standard set of information about the glider instance on board. Much like the sensors attached, this should be a SensorML file describing basic information about the glider and the default mission behaviour. It should contain information about the sensor payloads, but if the payloads meet the recommendation in Sec. 3.1, this amounts to a link to the relevant SensorML file residing on the sensor. The glider should also be able to be queried from a local or remote computer using a standard protocol, which could also be the PUCK protocol.

3.3 Basestation or Operator level

At the operator level, complexity arises because it cannot be assumed that recommendations in Sec 3.1 and 3.2 have been followed. Naturally, this implies effort on the part of the operator to make up for a lack of standardization at the sensor or platform level. In any case, it is recommended that the operator NOT develop a new toolbox, but acquire one of the freely available ones in order to properly match up and/or generate the metadata to pair with each observation. A number of toolboxes have been developed to convert manufacturer format to other formats, but are not widely used outside the groups that developed them. They often add metadata fields and quality-control processes [4, 5, 6]. Modification of the toolbox will no doubt be required, but this should be provided back to the community using an open source paradigm. It is also recommended that the operator provide Netcdf CF compliant files, preferably a well-documented fit for purpose version, such as the EGO format [2, 3] or IMOS/ANFOG [7], or . A minimum set of metadata is required, but the operator should not be limited in the type and amount of metadata to include. In the case where recommendations of Sec 3.1 and 3.2 are met, a mapping or link to the SensorML files must take place (see implementation plan in Sec 4). This does not yet exist, but when it does, it should require much less effort to match hardware metadata to observations than the case-by-case toolbox development described above.

3.4 Investigator level

Investigators who are not operators (see above) should not be required to make format changes, but should use the provided files to produce delayed mode versions of them. This means the delayed mode file may differ only in the addition or update of data streams (derived products) or metadata streams (QC flags and comments). If these are not foreseen in the original format, the operator should add them following the Netcdf CF or EGO standard.

3.5 Data Center or provider level

The recommendations for this actor were described in detail in BRIDGES task T3.3. In that task, 52N and NOC identified basic concepts to be used in specifying an interface to glider data. In particular these protocols are part of the Open Geospatial Consortium (OGC) Sensor Web Standards:

- Sensor Observation Service (SOS)
    - Storage and retrieval of observation time series data (in O&M encoding)
    - Provision of sensor metadata (could be in SensorML)
- Observation & Measurements (O&M)
    - Model and encoding for data measured by sensors

- Data model also adopted by ISO
- Stores measurement itself as well as related metadata
- SensorML
  - Model and encoding for sensor metadata only
  - Describes (internal) processes that affect the measurements
  - Describes nature of inputs and outputs of a sensor or a sensor system

Based on the current state, 52N and NOC performed a requirements analysis for interoperable system design. Those conclusions foresee the adherence to the recommendations for sensor and platforms with easy integration into the proposed design, but would not require it, as long as the operator recommendations are followed (regarding netcdf files with complete metadata and data). Of course the operator workload would be drastically reduced if sensor and platform recommendations for SensorML are followed.

The main highlights are that it should naturally make use of OGC Sensor Web standards (above), it should be compatible with existing solutions (CF or EGO NetCDF), the sensor metadata shall use a common structure and description, and ontologies and catalogues for descriptions of instruments, phenomena, areas of interest should be used (NERC Vocabulary Server suitable). Please see Fig. 2. The Marine SWE profile is a harmonised implementation of SWE ensuring interoperability between participating data groups by including semantic mark up within metadata. It has been adopted here, and developed in various ways by a number of cooperating projects.

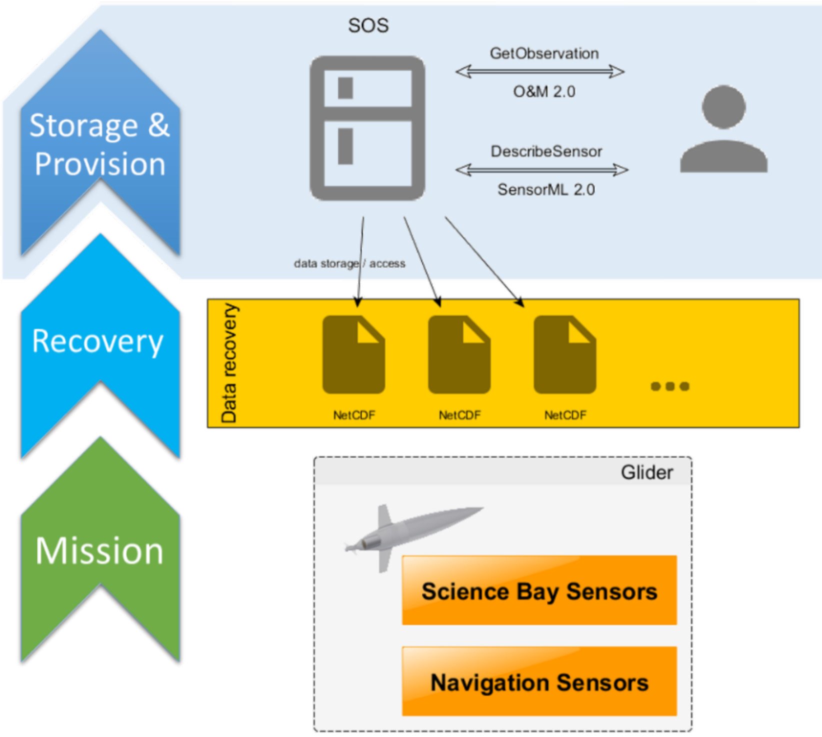

Figure 2. Proposed interface for Storage & Provision of glider data (top layer). Mission and Recovery layers already exist.

4 Implementation plan overview

While a number of protocols and standards exist and are already used to some extent, implementation on a large scale has not occurred, with the exception of NetCDF CF compliant files, which can vary widely. Widespread adoption requires some development to stitch those protocols together for the glider data pathways as follows:

1. Operator-friendly **ways to generate the recommended SensorML files** for glider platforms, and for sensors that do not support it.
2. **After-market solutions for making a 'dumb' sensor 'smart'** by allowing it to store a SensorML file and communicate using the PUCK protocol.
3. User-friendly **ways to map the SensorML data for the glider+sensors to a NetCDF CF file at the operator** level.
4. A **tool for mapping from the NetCDF glider file to the observation (data and metadata) model of O&M and the sensor metadata model of SensorML** (both also proposed in D3.3). The proposed metadata model allows the use of a common structure and description, with accepted ontologies and catalogues for descriptions of instruments, phenomena, areas of interest (e.g. NERC Vocabulary Server). It remains to be seen exactly what specifications the initial NetCDF file will have, most likely an EGO NetCDF is the best place to start, with more general CF compliant files to come later.
5. **Software that will provide a Sensor Observation Service (SOS) for storage and retrieval of observations and metadata.** The proposed solution is the Marine SWE profile using OGC Sensor Web Standards, which applies the above data and metadata models (see D3.3).

For the remaining period of BRIDGES WP3, it is planned to implement a demonstration of steps 4 and 5. Work on the interface for step 2 is already well underway at CSCS, and steps 1 and 3 will be investigated further by requesting information from the NexOS project and ALSEAMAR, and the community at large. Already, 52N has established a shared working space (Twiki) to document and discuss common approach: metadata models, ontologies and vocabularies.

After the demonstration, feedback will be requested from the actors above. In particular:

1. To what extent are OGC standards used in your work, in particular for glider data?
2. Describe specific advantages of the proposed structure relevant to your work.
3. Describe any drawbacks or problems with OGC standards for glider data.
4. Please suggest alternatives or solutions to the problems/drawbacks above.

5 Summary and Conclusion

The goal of the larger community of organizations handling glider data is to promote and establish the widespread use of existing tools and formats for harmonizing data sets collected by autonomous underwater gliders. This will result in more quality-assured observations of the seas, with increased visibility, availability, and usability. In the long term, adoption of higher level standards for glider metadata should be harmonized with European and global efforts. A number of H2020 European projects (AtlantOS, NEXUS) are working on this aspect of integrating many types of data (BRIDGES for glider data) into a Sensor Web Enabled system to allow widespread discovery and download services through web interfaces in a standard way.

The effort here, therefore, is to promote standardization at every level of handling, from data collection at sea to provision to the end user. Protocols exist for each level, but are not widely used. It will take a cooperative interaction among all actors to implement the available protocols, but little new development is expected. Simply adopting the present recommendations as a community and making small collaborative efforts to implement in the frame of current future projects will ultimately result in the desired effect. It is important that data management be a central issue in future meetings, workshops, and conferences, not only for progress in implementation, but to explain usage of the protocols and tools that are proposed so they can be more widely accepted and used.

The BRIDGES consortium will work closely with the EuroGOOS Glider Task Team (http://eurogoos.eu/gliders-task-team/) and expected JCOMM-OPA OceanGlider Teams. In fact, the leader of WP3, Daniel Hayes, is a member of both of these teams, and is designated as responsible for data management on the JCOMM-OPA OceanGliders team. These teams have also explicitly mentioned the intention to work towards the promotion of open source tools and best practices (7). Synergy with other projects ensures longevity: development of SensorML profiles for marine applications (COMMON SENSE, NeXOS, SCHeMA projects) for the marine community (ships, gliders, floats, buoys).

6 References

1. AtlantOS Project Deliverable 7.1, “Data Harmonization Report, accessed 27 October, 2016. https://www.atlantos-h2020.eu/download/deliverables/7.1%20Data%20Harmonization%20Report.pdf

2. Everyone's Gliding Observatory gliders NetCDF format reference manual, accessed 1 August, 2016. http://www.ego-network.org/dokuwiki/lib/exe/fetch.php?media=public:datamanagement:ego_gliders_netcdf_format_manual_v1.2.pdf

3. Simple procedures for glider data management in Everyone's Gliding Observatory, accessed 1 August, 2016. http://www.ego-network.org/dokuwiki/lib/exe/fetch.php?media=public:datamanagement:data_management_good_practices_manual_v1.0.pdf

4. Loic Houpert, oceano_data_toolbox to read glider data from the seaglider basestation and EGO netcdf format, accessed 1 August, 2016. https://bitbucket.org/Lhoupert/oceano_data_toolbox

5. C. Troupin, J.P. Beltran, E. Heslop, M. Torner, B. Garau, J. Allen, S. Ruiz, J. Tintoré, A toolbox for glider data processing and management, Methods in Oceanography, Volumes 13–14, September–December 2015, Pages 13-23, ISSN 2211-1220,

http://dx.doi.org/10.1016/j.mio.2016.01.001http://www.sciencedirect.com/science/article/pii/S2211122015300207

6. Bastien Queste, The UEA Seaglider Toolbox, accessed 1 August, 2016. http://www.byqueste.com/toolbox.html.

7. The IMOS Matlab Toolbox, accessed 27 October, 2016. https://github.com/aodn/imos-toolbox/wiki

8. Australian National Facility for Ocean Gliders (ANFOG), Data Management User's Manual, Version 3.0, April 12th 2010, accessed 26 October, 2016, http://imos.org.au/fileadmin/user_upload/shared/IMOS%20General/documents/Facility_manuals/ANFOG_data_management3_0.pdf.

9. The US IOOS NGDAC NetCDF File Format Version 2, accessed 27 October, 2016, https://github.com/ioos/ioosngdac/wiki/NGDAC-NetCDF-File-Format-Version-2

10. T. Liblik, J. Karstensen, P. Testor, P. Alenius, D. Hayes, S. Ruiz, K.J. Heywood, S. Pouliquen, L. Mortier, E. Mauri, Potential for an underwater glider component as part of the Global Ocean Observing System, Methods in Oceanography, Volume 17, December 2016, Pages 50-82, ISSN 2211-1220, http://dx.doi.org/10.1016/j.mio.2016.05.001.

11. OGC® PUCK Protocol Standard, accessed 27 October, 2016, http://www.opengeospatial.org/standards/puck.

Appendix: Example Configuration Files

sea.cfg (SeaExplorer):

```
#-----------------------------------------------------------------------------
#  SeaExplorer Configuration File (compatible with SeaExplorer soft V1.1.1)
#-----------------------------------------------------------------------------

# Global config.
id=SEA013
# Freq. (sec)
period1=20
period2=0.1
#mode 0 : Nominal, 1 : Simu with asserv, 2 : Simu
mode=0

# Communication timeout (restart modem)
com.inactivity.timeout1=180
# Communication timeout (resume mission)
com.inactivity.timeout2=600

# Security
security.depthLimit=725
security.globalTimeout=5
security.lowBatteries=25
security.nocom=3
security.vacuumLimit=90000
security.flyTimeout=5
security.depthStabilityLimit=10
security.depthStabilityDuration=180
security.depthStabilityError=2
security.abortMask=0

# Heading regulator
heading.deadzone=0
heading.pid.kp=0.2
heading.pid.ki=0.002
heading.pid.kd=15
heading.pid.il=15
heading.pid.el=15
heading.pid.hi=45
heading.pid.lo=-45
heading.pid.ka=0

# Iridium
iridium.call.number=00881600005346
iridium.call.number2=00881600005346
iridium.call.tries=5
iridium.call.timeout=50
iridium.cmd.tx=sendByIridium.sh

# To load data files = 1
pld.loadfile=1
pld.asp=0

# Ballast
```

ballast.lo=168
ballast.hi=693
ballast.hy=5
ballast.coef=-0.505
ballast.zero=431

 # Moving Mass
mass.lin.lo=9
mass.lin.hi=718
mass.lin.hy=5
mass.lin.coef=-6.71
mass.lin.zero=712

mass.ang.lo=333
mass.ang.hi=698
mass.ang.hy=5
mass.ang.coef=2.216
mass.ang.zero=511

 # Depth
pressure.zero=0.492
pressure.maxvoltage=2.512

 # Flasher
flr.install=1

 # Altimeter
alt.install=1
#alt.simu.seabed=700
#alt.range=50
#alt.pulse=100
#alt.gain=6

 # Batteries
batt.coef=78.473
batt.zero=-1327.8

sea.msn (SeaExplorer):

```
#---------------------------------------------------------------------------------
#          SeaExplorer Mission File
#---------------------------------------------------------------------------------

# Mission
msn.id=101

# Default Heading
#msn.heading=-9999

#zero setup
msn.lin.base=43
msn.ang.desc.base=0
msn.ang.asc.base=0
msn.bal.base=0

# Surfacing parameters
msn.srf.lin=100
msn.srf.ang=0
msn.srf.bal=+500

# Inflection parameters
msn.inflection.altitude=15
msn.inflection.surfaceDepth=10

# Cycles parameters
msn.cycle.total=500
msn.cycle.surfacingRate=5

[profiles]
p1=saw(PU-10; PD+10; ZT10; ZB30; BU+200; BD-200)

[path]
g1=wp(x552.277, y4349.2197);p1;
```

sg_calib_constants.m (Seaglider):

```
%sg_calib_constants.m
% values for basestation calculations, diveplot.m, etc.
% last edited 23-Feb-15, F.Stahr

% basic glider and mission params
id_str   =        '150'    ;

%mass   =        53.652   ; % in kg, for PortSusan
mass    =        53.912   ; % for open ocean waters

mission_title =  '20160831_010';
%mission_title   = 'Univ of Cyprus';

%rho0   =        1022.8   ; % in kg/m3 for Puget Sound
rho0    =        1029.2   ; % for open ocean waters

%volmax          =        52859    ; % projected for Puget Sound
%volmax =   52904.3;   ; % from Puget Sound regression, 23-Feb-15
volmax  =        52935    ; % projection for ocean ballast

% regressed from PS 23-Feb-15, hydrodynamic model params
hd_a = 1.69279e-03;
hd_b = 1.29775e-02;
hd_c = 1.04316e-05;

pitchbias = 0;      % pitch reference in regressions

% CT sensors cal constants
calibcomm =   'SN 0071 cal 14-Jan-15';        % SN and cal date
t_g       =        4.36930384E-03  ;
t_h       =        6.36992015E-04  ;
t_i       =        2.57456023E-05  ;
t_j       =        2.83939021E-06  ;
c_g       =        -1.00037548E+01 ;
c_h       =        1.13004296E+00 ;
c_i       =        -1.12554014E-03 ;
c_j       =        1.76934797E-04  ;
cpcor     =        -9.5700000E-08   ;
ctcor     =        3.2500000E-06    ;

%below splits apply at 500m level by default
QC_cond_spike_shallow=0.005; % 0.15/ARGO_sample_interval_m, # [S/ml/m] Carnes 0.02
QC_cond_spike_deep=0.0009;   % 0.025/ARGO_sample_interval_m, # [S/ml/m] Carnes 0.01
QC_temp_spike_shallow = 0.10;
QC_temp_spike_deep = 0.050; % below 500m; %added 20150309 DRH (J. Bennett suggestion .01 for both)
QC_cond_spike_depth = 600;
QC_temp_spike_depth = 600;
QC_spike_comm = 'Spike defined as (abs(v2-0.5*(v3+v1)) - 0.5*abs(v3-v1)) / 0.5*abs(d3-d1)';
QC_median_comm = 'Median Filter on C and T as follows: marked probably bad if a point falls outside 2 std of
the local 9 point window';

QC_temp_min=10; %':-2.5, # [degC] Carnes, compare global Schmid -2.5 (labsea?) MDP -4.0
QC_temp_max=40; %':43.0, # [degC] Carnes, compare global Schmid 40.0
QC_salin_min=35.0; %':19.0, # [PSU] was 2.0 per Carnes; ditto Schmid but we can't fly in waters that fresh
QC_salin_max=41.0; %':45.0, # [PSU] Carnes, compare global Schmid 41.0
```

```
% SBE oxygen cal constants
comm_oxy_type= 'SBE 43f';
calibcomm_oxygen = 'SN 129, 04-May-11'; % SN and cal date
Soc     = 2.4977E-04     ;
Foffset = -8.1950E+02    ;
o_a = -9.9371E-04;
o_b = 1.3160E-04;
o_c = -1.4794E-06;
o_e = 3.60E-02;
PCor = 0;  % used as flag to force usage of new algorithm

% this glider also carries WET Labs BB2F-VMG SN 450, last cal 15-Jun-11
calibcomm_scatterometer =       'WET Labs BB2F-VMG SN 450, 07-Jun-11'; % SN and cal date
scale_470 = 1.28e-5; %(m^-1 sr^-1)/counts
dark_counts_470 = 51;
resolution_counts_470 = 1.0;
%calibration to be implemented in processing:
%Vol_scatter_470 = scale_470*(OUTPUT - dark_counts_470); % (m^-1 sr^-1)

scale_700 = 3.250e-6; %(m^-1 sr^-1)/counts
dark_counts_700 = 49;
resolution_counts_700 = 2.2;
%calibration to be implemented in processing:
%Vol_scatter_700 = scale_700*(OUTPUT - dark_counts_700); % (m^-1 sr^-1)

scale_fluor = 0.0161; %(micrograms / liter)/count
dark_counts_fluor = 56;
resolution_counts_fluor = 2.2;
max_counts_fluor = 4121;
%calibration to be implemented in processing:
%Chl_conc_fluor = scale_fluor*(OUTPUT - dark_counts_fluor); % (micrograms/liter)
% end of fil
```

Table 1. Groups that handle glider data and the primary formats and tools they currently use. Note that the EGO NetCDF is a NetCDF CF compliant format, and has not been explicitly listed here.

Groups	Formats		Tools				
	Data	Metadata	Conversion	Quality Control	Storage	Provision	Discovery
Platform Developer	ASCII, NetCDF	ASCII, SensorML	N/A	Own	N/A	N/A	N/A
Sensor Developer	ASCII, SensorML	ASCII, SensorML	N/A	N/A	N/A	N/A	N/A
Glider Operator	ASCII, NetCDF	ASCII, NetCDF	Community or own	Community or own	Community or own	Community or own	Community or own
Scientist-PI	NetCDF	ASCII, NetCDF	Community or own	Community or own	Community or own	Community or own	Community or own
Data Center	NetCDF	NetCDF	Own	Own	Community	Community	Community
Interface Developer	NetCDF	SensorML	N/A	N/A	N/A	N/A	N/A
User: Industry	All of above	All of above	Own	N/A	Own	N/A	N/A
User: Academia	NetCDF	NetCDF	Community or own	N/A	Own	N/A	N/A
User: Government	All of above	All of above	Own	N/A	Own	N/A	N/A
User: Public	All of above	All of above	Own	N/A	Own	N/A	N/A