Project	AtlantOS – 633211
Deliverable number	D7.3
Deliverable title	Full life cycle Report: Report on AtlantOS Networks full life cycle data flow including Data Policy and Intellectual property rights
Description	Assessment of the full data lifecycle of observing data and information from data acquisition to long-term archiving to detect fields for improvement that may be common for a data-providers in AtlantOS
Work Package number	7
Work Package title	Data flow and data integration
Lead beneficiary	UniHB
Lead authors	Koop-Jakobsen (UniHB). Waldmann (UniHB). Huber (UniHB)
Contributors	All WP7 data networks
Submission data	N/A
Due date	30.09.2016
Comments	N/A

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 633211.

Last updated: 04 October 2016

Contents

Executive summary:

This report assesses the full data lifecycle of data and information generated during the AtlantOS project with the goal of detecting fields for improvement. The assessment takes its starting point in the Data Management Plan (DMP), which was produced as part of the Open Research Data Pilot that AtlantOS comply with. The DMP sets the framework for the handling of data produced in AtlantOS from acquisition over curation to dissemination, and shall thereby assure full lifecycle management of AtlantOS data also beyond the lifetime of the project. The first version of the DMP was installed in September 2015. However, the DMP is a living document that will evolve as the work of AtlantOS, and in particular WP7 progresses. The goal is to continuously implement agreements on data management made by WP7 into the DMP. This report evaluates the DMP in the light of full life cycle data management principles.

In general, the current version of the DMP was found to be very generic and lack concrete details. The cause for this generic approach comes down to the diversity and the maturity of the AtlantOS data providers and the international outlook. AtlantOS is a large project involving many data providers and data integrators, ranging from large research observatories to mature data integrators, which already has advanced data-policies and workflows installed. Consequently, a prescriptive DMP with a one-model-fits-all approach cannot be applied in AtlantOS. Through the work of WP7, data networks and integrators in AtlantOS have come to an agreement on a set of essential minimum applicable standards, which shall ensure cross platform coherence. The efficiency of the WP7 harmonization process shall be demonstrated by making AtlantOS data available through integrators like Copernicus or SeaDataNet and the EMODnet portals. To improve the data lifecycle aspects of the DMP, these initiatives shall be implemented in the next version of the DMP planned for December 2016.

The current version of the AtlantOS DMP was based on the first version of the EC DMP guidelines. The most recent EC-guidelines updated in July 2016 have more explicit recommendations for full life cycle management through the implementation of the FAIR principles, which states that the data produced shall be **Findable, Accessible, Interoperable and Reusable** (**FAIR**). The implementation of the FAIR principles is intended as a conceptual integration rather than a technical integration. It is concluded that the work of WP7 on the data management in the AtlantOS project comply with the FAIR-principles as a concept, in the sense that all aspect of the FAIR

principles are being considered. The next version of the DMP shall specifically seek to implement the FAIR principles.

Data management in AtlantOS – the work of WP7

Within the AtlantOS project, WP7 is dedicated especially to improve the data management and interoperability among the observation networks involved in AtlantOS. Hereby, WP7 shall ensure availability of the diverse and interdisciplinary data pool collected from the Atlantic Ocean by different ocean observing platforms. The overall goal is to make these data freely available, in a readily usable format, with sufficient information attached in the form of metadata for a scientific use as well as monitoring purposes.

The strategy in WP7 is to work towards an integrated data system within AtlantOS that [Task 7.1] harmonises work flows, data processing and distribution across the in-situ observing network systems (GEOMAR for GOSHIP and Sea floor Mapping, University of Exeter for SOOP, SAHFOS for CPR, ICES for Fish+plankton, Ifremer for Argo, NERC for OceanSites, CNRS for Glider, EUMETNET for Surface drifter, HZG for Ferry box, BODC for Tide gauges) and [Task 7.2] infrastructures that integrates in-situ observations in existing European and international data infrastructures (Copernicus Marine Service, SeaDataNet NODCs, EMODnet, EurOBIS, GEOSS) so called Integrators.

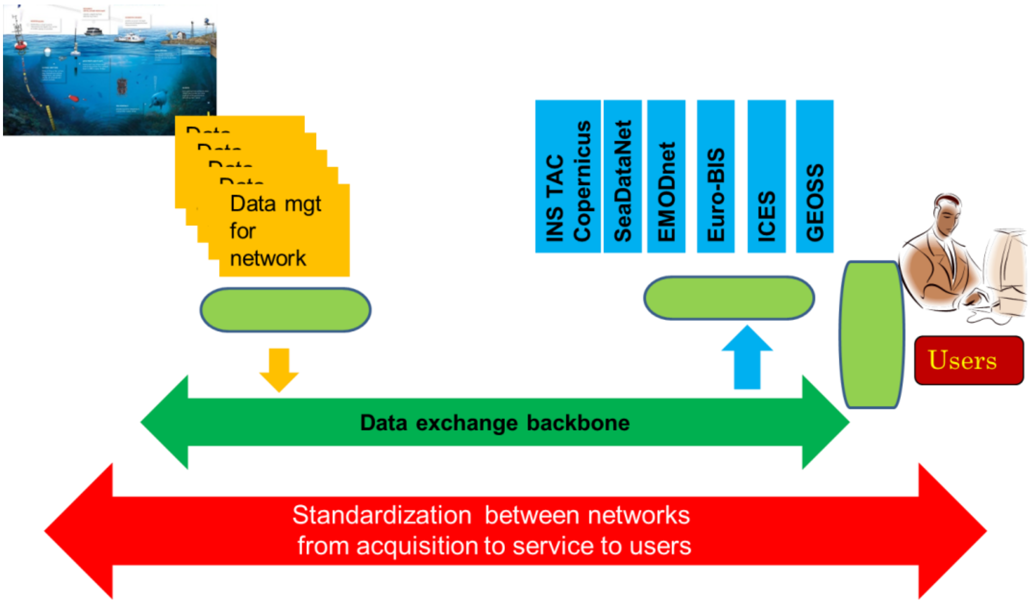

Figure 1: The Data system for AtlantOS. The AtlantOS networks shall feed their information to the Data integrators through a Data exchange backbone. AtlantOS WP7 is in task 1+2 setting up the framework for the data exchange backbone.

WP7 shall improve the interoperability among data networks and data integrators considering all aspects of the full data life-cycle; from data-acquisition over data-curation to dissemination of data to stakeholders around the Atlantic Oceans as well as the general public. In this report, we evaluate how well AtlantOS comply with the aspects essential to full data-lifecycle management.

Introduction to Full life-cycle data management

With the ever increasing amount of data collected in environmental science as a result of the installation of automated observation platforms, there is a growing need for identifying the most efficient strategy of processing the data all the way from the initial planning of data collection to the availability of the data products and their dissemination. Management of large pools of data resources benefits significantly from a well-established and standardized approach starting even before the data is collected. However, at present time best practice for life data management is often largely undefined and is generally left as a decision for the data curator and/or data publisher. In the AtlantOS project, WP7 is working on identifying a common data management framework involving both data providers as well as data integrators (Fig. 1)

Only with a standardized, traceable workflow throughout the lifetime of the data resources can data quality and interoperability and good discoverability be assured. It is in the light thereof that recent conceptual data lifecycle models have evolved. Data lifecycle models have the purpose of providing a structure for the many operations that is needed for a data record throughout its life in order to assure the best possible use of the data.

This need for data management strategies has resulted in the description of multiple data lifecycle models over the past decades, and the concept of data management in a scientific context have evolved from a general perception of information lifecycle management to specific lifecycle models for scientific data. Ball (2012) and Crowston and Qin (2011) recently reviewed the available literature on Data management lifecycle models, and found that only a few of these models were published in the scientific literature, and many exist as thematic working documents from individual data management projects, which only available via project- and datacenter-webpages and university archives. Hence, this is an evolving field and the concept of data lifecycle models is changing.

The lifetime data-management models available for scientific data has various origins ranging from individual data management projects to long-term scientific data curators and publishers like the UK Data Archive and DataONE in the USA (Ball 2012). These models are developed to accommodate different needs; a small temporary research project needs a different data management lifecycle plan than a large data observation network. Hence, a generic model that covers all aspect for Data lifecycle management for all purposes is not available and potentially not probable. However, among the Data lifetime models available, there are certain commonalities in the sense that they describe a series of steps standardizing the handling of data from discovery to publication. These steps include a) **Data acquisition, processing and quality assurance b) Data description and representation** c) **Data dissemination** d) **Repository services/ preservation** (Crowston & Qin 2011).

Major global scientific data management initiatives, like RDA and Belmont Forum continues to discuss and provide guidelines for full lifecycle data management and most scientific funding agencies have installed obligatory data management planning as part of their funding scheme. However, there has been no broader consensus on the explicit content of lifecycle data management plans. In this regard, in a recent communique from the European commission it is highlighted that Europe is not yet fully tapping into the potential of its data resources, primarily due to the lack of clear incentives promoting data sharing, fragmentation of resources and lack of interoperability of data-resources (European-Commision Brussels, 2016 19.4.2016 COM(2016) 178 final)

Full life-cycle data management in H2020

In Horizon 2020, the biggest EU Research and Innovation programme, Data Management Plans (DMPs) were introduced in the Work Programme for 2014-15, and installed as a requirement for projects participating in the Open Research Data Pilot. In H2020, the DMPs describe the data generated and plans for their exploitations in terms of their curation, preservation and accessibility. In other words, the intention is a full lifecycle data management plan. The DMP is meant as a living document, which shall be improved continuously as the project progresses.

A recent initiative from the EC to unlock the potential of the data resources is the implementation of the FAIR-principles in H2020. The FAIR-principles were initiated at a workshop held in Leiden, where various stakeholders in the scientific community including academia, industry, funding agencies, and publishers drafted a new set of data management guidelines. This resulted in a concise and measureable set of principles referred to as the FAIR Data Principles abbreviating that data should be **Findable, Accessible, Interoperable and Reusable** (**FAIR**) (Wilkinson 2016.). The fair principles were generated to improve the practices for data management and data-curation, and FAIR intends to describe the principles in a way that is domain-independent and hence can be applied to a wide range of data management purposes, whether it is data collection of individual researcher or data management of larger research projects regardless of scientific disciplines.

The FAIR principles have recently been endorsed by H2020 and are now implemented in the updated guidelines for the data management plan in H2020 (Guidelines on FAIR Data Management in Horizon 2020 Version 3.0 26 July 2016). This is intended as an implementation of the FAIR concept rather than a strict technical implementation of the FAIR principles. However, with the FAIR principles now being directly described as part of the biggest EU Research and Innovation programme, this could be a landmark for the streamlining of full data lifecycle management for the scientific community in Europe.

With the endorsement of the FAIR-principles by H2020 and their implementation in the guidelines for H2020 DMPs, the framework for full-lifecycle data management in AtlantOS has now been prescribed and must be included in the DMP. Consequently, in the following, we shall evaluate the AtlantOS DMP as well as the work of WP7 on harmonization and standardization and evaluate the compliance with the FAIR principles as well as with full lifecycle data management in general.

Full life-cycle data management in the AtlantOS through The Data Management Plan

AtlantOS DMP – the current version

The AtlantOS project complies with the Open Research Data Pilot. Hence, open access to AtlantOS data for free access, reuse, and redistribution shall be the code of practice in AtlantOS, and only when plausibly justified may restrictions apply; like risks of IPR infringements.

The first edition of the AtlantOS Data management plan was produced and installed within the first 6 months of the project based on the guidelines provided at that time (Guidelines on Data Management in Horizon 2020 Version 16 December 2013). This was the first step towards improved harmonization of data produced within the framework of AtlantOS. The DMP contains essential elements from the work of WP7 in general and particularly in regards to harmonization as it describes the data that will be authored and how the data will be managed and made accessible throughout the lifetime of AtlantOS and beyond. The current version includes information on the following topics (for details see www.atlantos-h2020.eu/download/deliverables/11.2%20Data%20Management%20Plan.pdf) (AtlantOS_DMP 2015)

- **Sources and types of data resources produced AtlantOS – Overview of the data AtlantOS landscape**
- **Prioritization of data resources – AtlantOS EOVs**

- **Standardization – Recommendation on formats**
- **Strategy for Data reuse and exploitation – Giving guidelines on Discoverability, accessibility, reusability, quality assurance and time compliance.**
- **Principles of access and sharing – AtlantOS data sharing strategy**

However, as the DMP is meant as a continuously evolving document WP7 shall implement the agreements on harmonization and standardization made under the progress of WP7. Following we shall analyse the DMP in light of full life cycle data management principles and give recommendations for improvement.

Analysis of the "AtlantOS DMP – the current version", and recommendation for improvements.

General impression

The general impression of the first version of the AtlantOS DMP, which has also been put forward by comments from internal and external reviewer, is that the DMP comes across as rather generic setting the frame for data handling in AtlantOS by stating general principles on data management in regards to discoverability, accessibility, reusability, and quality control rather than providing concrete details on standardization and harmonization.

However, in order for a Data management plan to have effect, it must be formulated in a way that all data providers and data integrators can comply with. In this regard, it is worth noticing that most of the data-providers as well as data integrators in AtlantOS are mature infrastructures with long-term experience, which administer a considerable part of the data from the Atlantic Ocean produced by leading ocean sciences institutions in Europe. These data handling entities have a high maturity level and have long-term existing data-policies and integrated workflows assuring high standards of data handling and dissemination. Due to the high maturity of the data centers and great diversity of the data resources that they administer imposing a prescriptive one-model-fits-all Data Management Plan is not a desirable solution if even probable. Hence, the challenge for AtlantOS, in regard to the DMP, is to find a pragmatic way, which can accommodate all data providers and data integrators in AtlantOS, while at the same time setting the frame for full lifecycle data management; this is essentially the work of WP7 as illustrated in figure 1 built as a system of systems focussing on interfaces on which higher level of interoperability recommendations are defined.

The first 18mo of the WP7 has been dedicated to finding common grounds for data handling in AtlantOS, and through a pragmatic approach identifying commonalities among the involved data networks and data integrators and identifying area for improvement. This has resulted in concrete initiatives, which are shortly for referred here, for full overview see AtlantOS D7.1

an agreement on a minimum set of applicable standards in regards to unique identifiers for Platforms **(**WMO or ICES code) and Institutions (EDMO codes) use of standard vocabularies in AtlantOS including a standardized vocabulary matrix for AtlantOS EOVs, recommendation Quality control for Real-time and Near Real-time data and minimum requirements for distribution means. Furthermore guidelines were generated for DOI assignment and recommendation on catalogue technique

These initiatives shall be implemented and concretize the upcoming version of the AtlantOS DMP and hereby set the framework for improved interoperability in AtlantOS

Standardization:

The current DMP has sought to implement guidelines recommending to follow the requirements of the INSPIRE directive for data and metadata formatting. This recommendation may be improbable at present times due to the already installed data policies by the individual AtlantOS data providers organized in International Network under JCOMM umbrella and data integrators. Furthermore, recent development in data brokering technology can help overcome many of the standardization barriers rendering prescriptive standardization somewhat superfluous. For example, AtlantOS - data integrator; SeaDataNet, has made use of the GEOSS brokering services to upgrade to the new SeaDataNet ISO 19139 Schema in order to comply with the EU INSPIRE Directive Implementing Rules. In the near future, AtlantOS shall revise the standardization-section of the current version of the DMP implementing the all the harmonization agreements accomplished among data networks and data integrators in WP7 within the first 18mo of AtlantOS.

Compliance of the Data Management Plan with the FAIR principles

With the indorsement of the FAIR principles and it incorporation into the guidelines for DMPs in H2020, the FAIR principles hereby serve as a template for a full-lifecycle data management. Although the FAIR principle does not serve as an independent lifecycle data model, it assures that the most important components of a full life cycle model is covered. Hence, these FAIR principles shall be implemented in the next version of the AtlantOS DMP provided that the implementation of the FAIR principle is intended as a conceptual integration rather than a technical integration.

In the following, we shall show that AtlantOS has made progress on all aspect of the FAIR principles.

Findability

- The AtlantOS DMP gives recommendation on Formats for data and metadata based on standards for data and metadata formats as well as exchange protocols that have been established within the marine community such as SEADATANET or Copernicus

- AtlantOS is working on applying permanent identifiers on all levels. A document describing the general principles of Digital Object Identifiers (DOI) was formulated. It provides examples of DOI implementation useful for AtlantOS networks.
- In AtlantOS WP7, there is agreement that Platforms should have a unique identifier that will be either WMO or ICES code for ships. Furthermore, Institutions used in a Data File shall have an EDMO code (European Directory of Marine Organisations). An EDMO catalogue is in making.

Accessibility

- AtlantOS comply with the open data pilot. Hence all data produced under the AtlantOS shall be aimed at open access. The currently installed Data management plan clearly states the open access foundation saying that *Free and open access without any restrictions shall be granted to the metadata of the data* and *Observing nodes under the umbrella of AtlantOS will follow the principle of free and open access to data produced by their facilities*. Furthermore, it *aims to identify unnecessary or obsolete barriers towards open access*.

- AtlantOS WP7 is exploring opportunities in GEOSS to make AtlantOS data and services openly available in an interdisciplinary global context through the GEOSS GCI. A workshop on the use of the GEOSS GCI is in planning for spring 2017 and the outcome shall be incorporated into the data management plan.

Interoperability

- AtlantOS WP7 gives recommendations data and metadata standardization. These shall be implemented in the AtlantOS DMP.

- AtlantOS WP7 has agreed that Metadata used by the networks should be “mappable” on standard vocabularies existing and EU (SeaDataNet) or international (CF or WoRMS for Taxa). A vocabulary matrix for AtlantOS EOVs has been generated and is available for AtlantOS WPs https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/A05/

- AtlantOS WP7 is exploring opportunities in GEOSS to make use of the GEOSS common infrastructure in particular the GEOSS data broker services to improve the interoperability in a global perspective and to assure the best possible interdisciplinary interoperability.

Re-use

- AtlantOS Open Research Data Pilot and all data produced under the framework of AtlantOS shall be available to third parties free of charge for scientific purposes (restrictions may apply for commercial use) and in compliance with open access regulations.
- In regards to time restrictions on data re-use. It is stated in the DMP that Observing networks contributing to AtlantOS will make data and metadata publicly available without undue delay
- WP7 has given recommendations for minimum distribution means saying that providing an FTP service at the level of network data management as the minimum delivery service. Additional services such as WEB services can also be provided, but are not mandatory.
- AtlantOS are giving recommendations for quality control measures for near-real time data acquisition.

Conclusions on Full lifecycle management in AtlantOS

Whereas the FAIR principles are mostly concerned with assuring best possible dissemination of data resources for future use, *full lifecycle data management* also includes guidelines on Data acquisition and processing. Through the identification of the AtlantOS data landscape and the prioritization of essential variables, as well as the formulation of recommendation for quality control measures for near-real time data acquisition, AtlantOS is involved with all aspects of the FAIR principles as well as full life-cycle data-management in general. However, as the AtlantOS project consists of many data providers and data integrators with different maturity levels and data resources to be administers, a prescriptive one-model-fits-all Data Management Plan is not probable. WP7 shall continue to work on selected aspects of all FAIR principles improving; Findability, Accessibility, Interoperability and Reusability, one step at a time. The outcome shall be implemented the coming versions of the DMP.

References

Guidelines on Data Management in Horizon 2020 Version 1.0, 11 December 2013. European Commision - The EU framework for Reseaerch and Innovation - Horizon 2020

Guidelines on FAIR Data Management in Horizon 2020 Version 3.0 26 July 2016. European Commision - The EU framework for Reseaerch and Innovation - Horizon 2020.

AtlantOS_DMP (2015) AtlantOS_Data_Management Plan D11.1 https://www.atlantos-h2020.eu/download/deliverables/11.2%20Data%20Management%20Plan.pdf. Accessed 01.09.2015.

Ball A (2012) Review of Data Management Lifecycle Models. University of Bath
Unpublished http://opus.bath.ac.uk/28587/

Crowston K, Qin J (2011) A capability maturity model for scientific data management: Evidence from the literature. American Society for Information Science and Technology Annual Meeting, New Orleans, LA

European-Commision (Brussels, 2016 19.4.2016 COM(2016) 178 final) COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS.

Wilkinson MDea (2016.) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3