HORIZON 2020

Coordination and Support Action

Grant Agreement No: 652641

EU-PolarNet

CONNECTING SCIENCE WITH SOCIETY

Deliverable No. 3.1

Survey of the existing Polar Research data systems and infrastructures, including their architectures, standard/good practice baselines, policies and scopes

Submission of Deliverable

Work Package	WP 3
Deliverable no. & title	D3.1: Survey of the existing Polar Research data systems and infrastructures, including their architectures, standard/good practice baselines, policies and scopes
Version	Final
Creation Date	25.03.2015
Last change	11.06.2016
Status	☐ Draft ☒ WP lead accepted ☒ Executive Board accepted
Dissemination level	☒ PU-Public ☐ PP- Restricted to programme partners ☐ RE- Restricted to a group specified by the consortium ☐ CO- Confidential, only for members of the consortium
Lead Beneficiary	CNR-DTA
Contributors	☐ 1 – AWI, ☐ 2 – CNRS, ☒ 3 - NERC-BAS, ☐ 4 - CNR-DTA, ☐ 5 – SPRS, ☐ 6 – IPEV, ☐ 7 - IGOT-UL, ☐ 8 – RUG, ☐ 9 - RCN, ☐ 10 – MINECO, ☐ 11 – CSIC, ☐ 12 - UW-APRI, ☐ 13 – BAI, ☐ 14 – GEUS, ☐ 15 – VUB, ☐ 16 – UOULU, ☒ 17 – RBINS, ☐ 18 - IGF PAS, ☐ 19 - IG-TUT, ☒ 20 – AMAP, ☐ 21 – WOC, ☐ 22 - GINR
Due date	29.02.2016
Delivery date	04.07.2016

Table of Content

Introduction

Abstract

Final report of the, questionnaire-based, survey of the existing polar research data systems and infrastructures, including their architectures, standards/good practice baselines, policies and scopes.

Objectives of the survey

The main objectives of the Deliverable are to:

- Provide a survey of existing European and international polar data management systems and infrastructures.
- Analyze the technical solutions presently in place that facilitate (open) access to data –e.g. network protocols, (meta)data standard models etc. This also includes the current Data Management Plans set up to ensure data protection, accuracy and long-term use.
- Analyze the legal, regulatory and, possibly, contractual rules that currently frame access to the data and explore the possibilities of harmonization, including the promotion of open access and use.

Adopted Methodology

The survey is based on the analysis of the responses provided by the international Polar Community to an online questionnaire targeted to system/network managers. The questionnaire is enclosed in Annex A.

The questionnaire response statistics are reported in Annex B.

The Questionnaire scope and content

The Questionnaire is still online and will remain at least until the end of 2016. The Web address is:

https://eupolarnetd31.typeform.com/to/hcxYeH

What was asked for

To survey the existing Data Management Systems by collecting information on the adopted data management practices (e.g. data policy and interoperability standards) in order to facilitate Polar data access and Interoperability (i.e. discovery, access, and use) across the International Community.

The information collected will be analyzed and presented to the European Commission and the International Polar Community in order to promote and improve Open Access and Interoperability for the Polar research domain, building on the existing Data Systems. A set of recommendations will be provided to the European Commission, paving the way towards a coordinated European Polar Data Infrastructure.

Who was asked to respond to the Questionnaire

The managers of existing Data Management Systems (from simple research data systems to the more complex operational data sharing infrastructures, aka e-infrastructures) dealing with Polar research data. The questionnaire addresses all branches of science that generate polar-related data, which are made available through data systems or e-infrastructures: social sciences, ecology, chemistry, etc.. There are no constraints on the "nationality" of the data system that can be described.

Questionnaire Structure

The questionnaire consists of four sections:

1. Who are you? (Respondent information, Data System identification, and the description of the organization managing the Data System) – Parts I and II
2. Which data are you sharing or managing? (Data content managed by the Data System) – Part III
3. Which technology do you apply to allow data accessibility and interoperability? (Description of the technology implemented by the organization for the Data System, including metadata, data formats, and interoperability protocols for data discovery, access, and use) – Parts IV to VII
4. Which Data Policy do you apply? (Description of the data policy applied by the organization for the Data System, including long-term preservation and accuracy) – Part VIII

References and useful links

- H2020 EU-PolarNet project Website[1]
- European Commission -DG-R&I. (2015, October 01). Guidelines on Data Management in H2020[2]

[1] http://www.eu-polarnet.eu/

[2] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

- GEO. (2015, October 01). GEOSS Data Management Principles[3]
- GEO. (2015, October 01). The GEOSS Data Sharing Principles[4]

Contact point for the Questionnaire

For any questions about the survey and the questionnaire, please, contact:

- Department of Earth System Science and Environmental Technologies (DTA) of the National Research Council of Italy (CNR)

Questionnaire response metrics

From the 01st of January 2016 to the 19th of February 2016, more than 200 unique visitors viewed the online questionnaire website; 29% of them (i.e. 58) responded to the survey. Figure 1 reports these metrics.

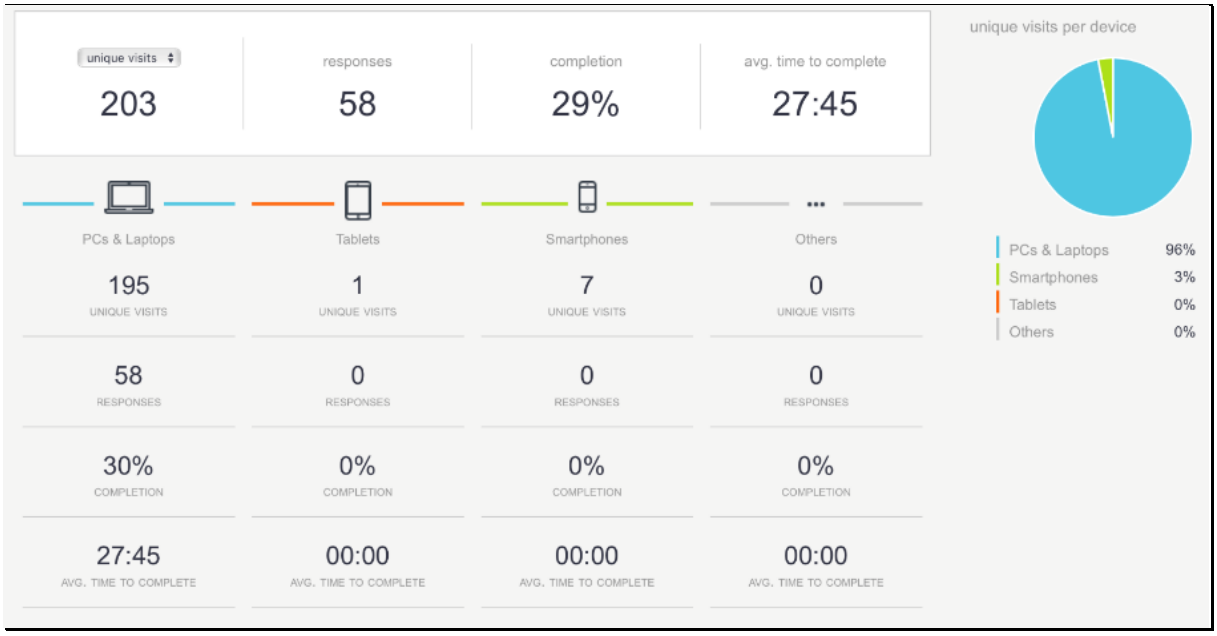

Figure 1. Metrics characterizing the responses received.

[3] http://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf

[4] https://www.earthobservations.org/geoss_dsp.shtml

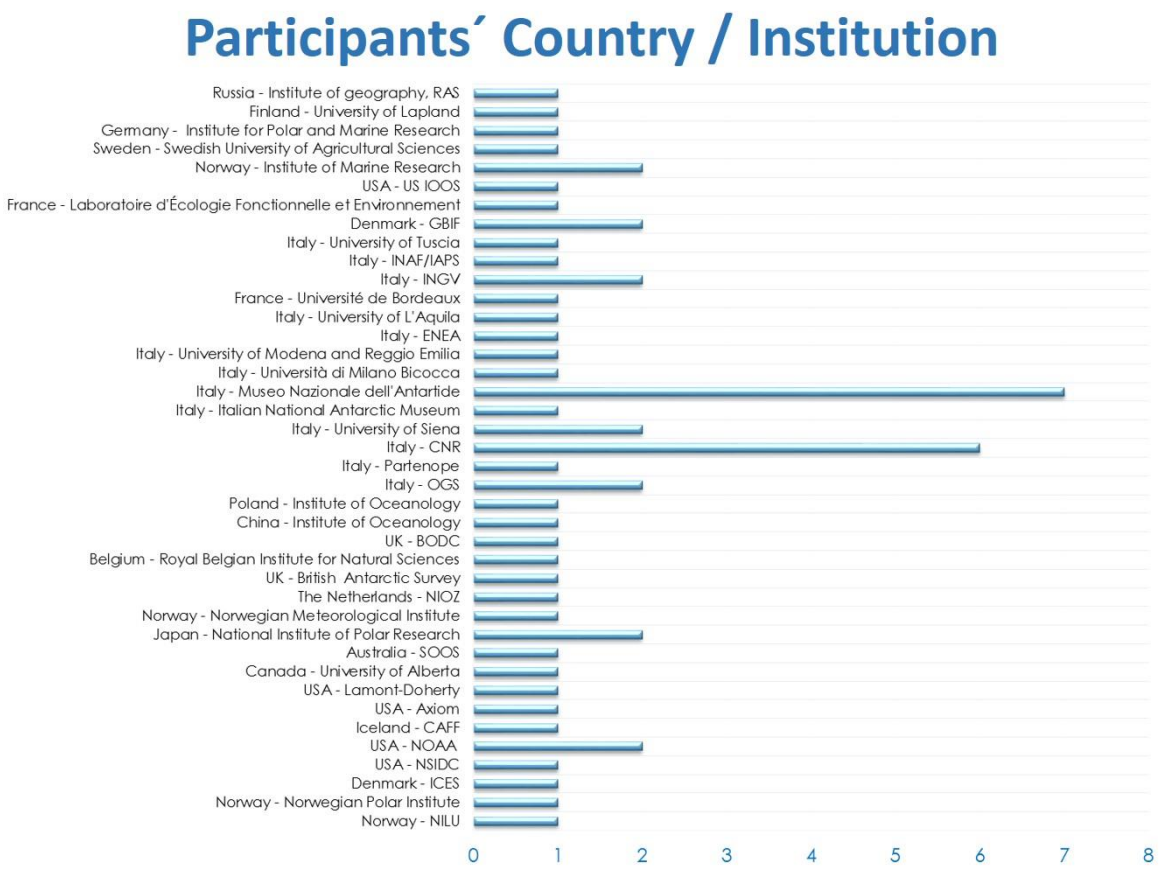

Figure 2. Organizations that responded to the questionnaire and the number of questionnaires filled in by each of them

Responses per Country and Organization

Figure 2 reports the organizations that responded to the questionnaire, while Figure 3 shows the geographical distribution of the responses –Italy, USA and Norway are the three countries that provided the most responses.

Responses came from 4 continents (only Africa is missing) and 21% of them were submitted by non-European Organizations.

Most of the responses were provided by research agencies and universities, as shown in Figure 4.

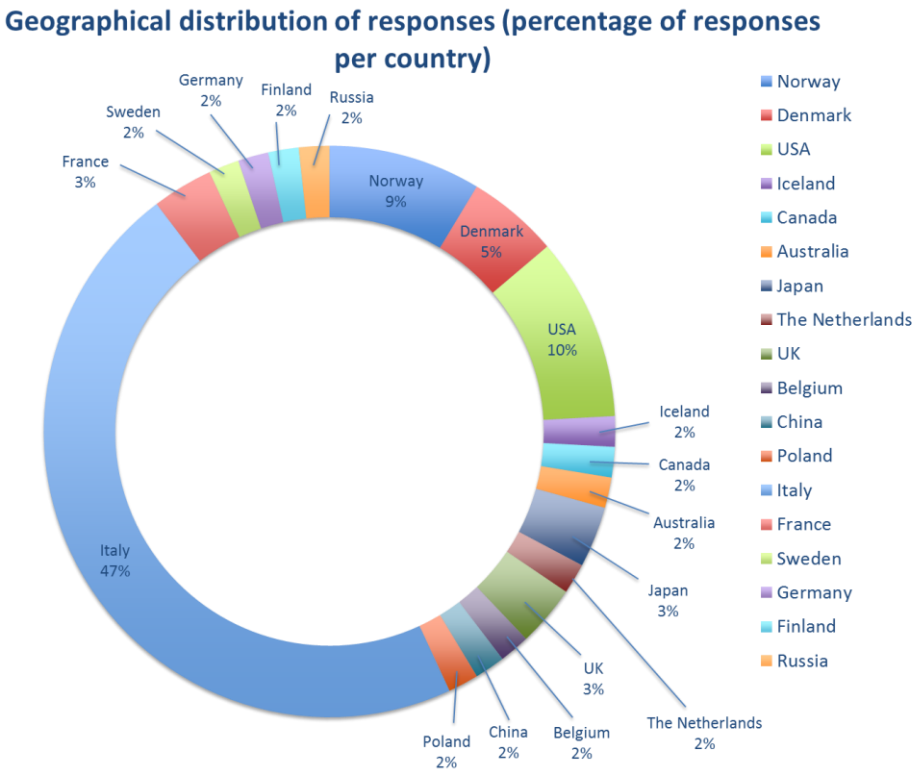

Figure 3. Geographical distribution of responses

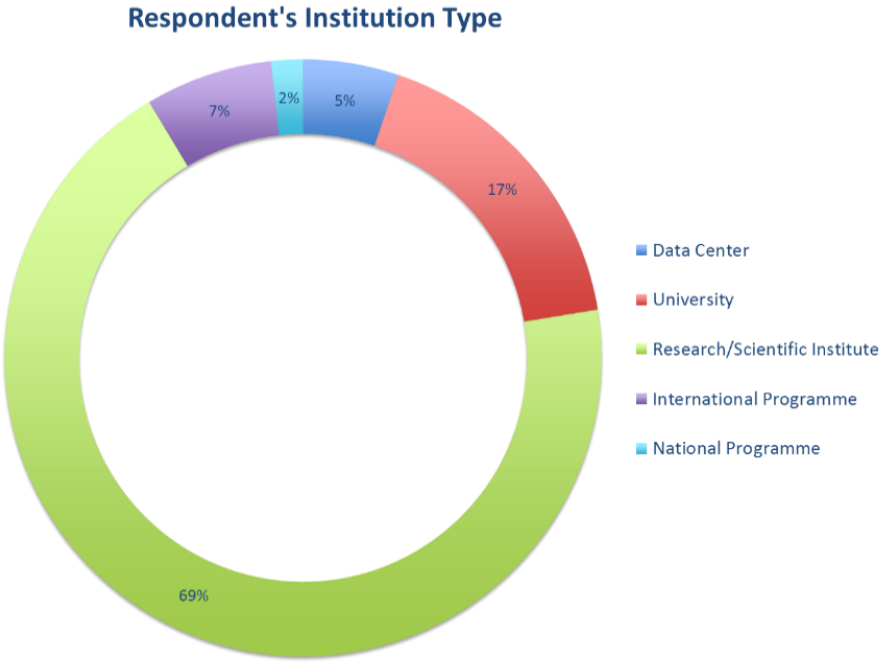

Figure 4. Organization types that responded to the questionnaire

General findings

Data Systems types

Out of the total responses, 12% of the data systems described are self-assessed as being "not relevant for the European polar research" –see Figure 5. It is worthy to note that this percentage is less than the 21% of answers provided by organizations outside Europe.

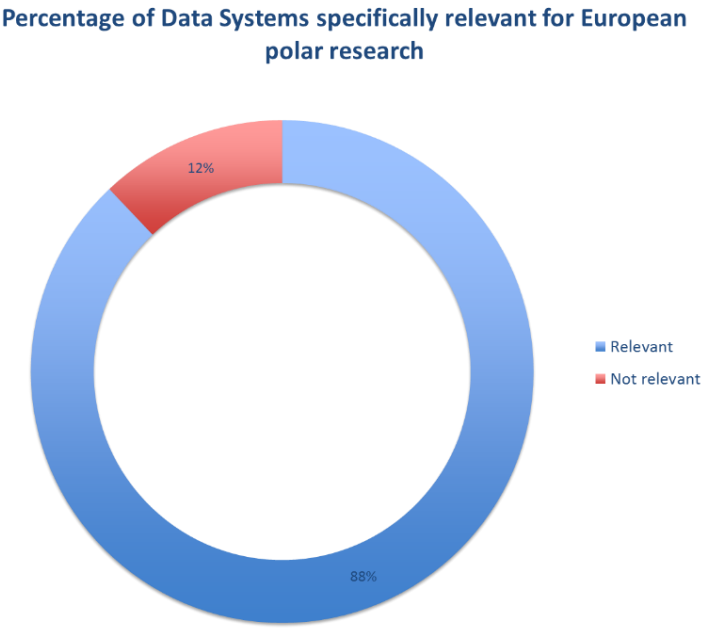

Figure 5. Data System relevance for European Polar research

Almost all of the systems described deal with research data. In addition, a large majority manages monitoring data, while about half of the systems analyzed deal with processed data –see Figure 6.

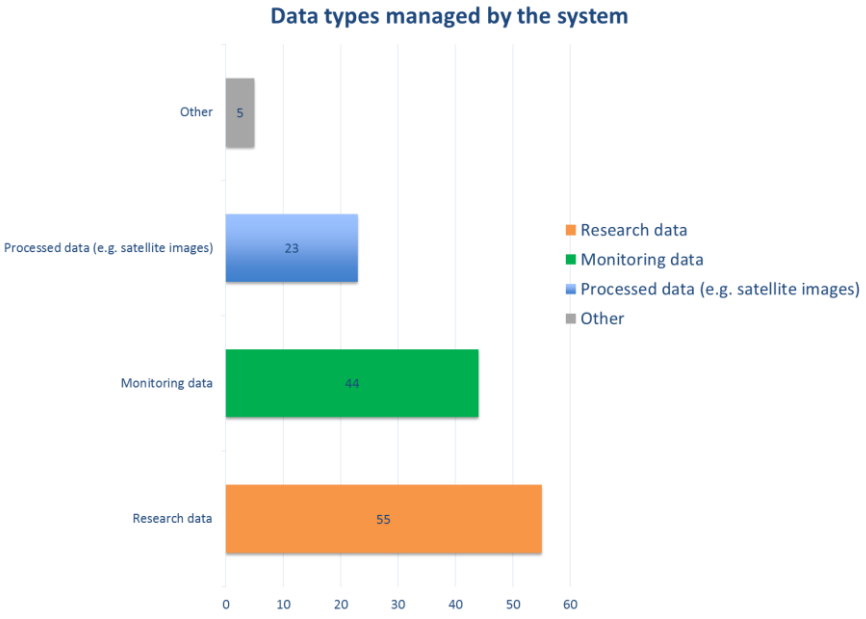

Figure 6. Data types managed by the systems

Disciplines covered by the data systems

Atmospheric, biodiversity, ecosystems and biological sciences are the most popular areas covered by the data systems –comprising almost 50% of the responses, as depicted in Figure 7. The variety of disciplines considered by the data systems is wide including also social, legal and health data, as shown by Figures 7 and 8.

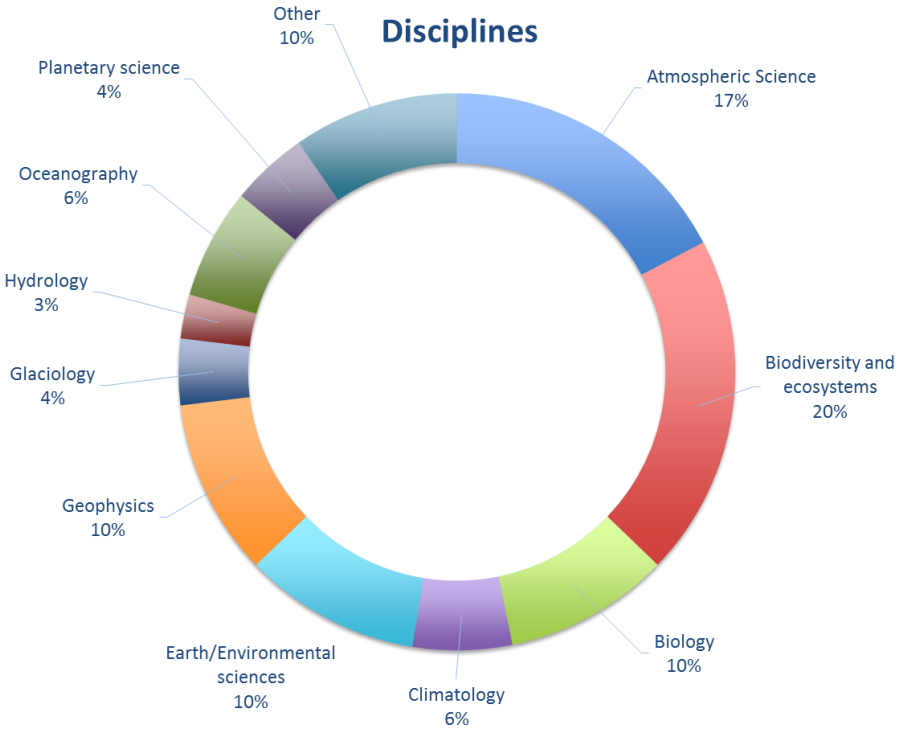

Figure 7. Disciplinary Data managed by the systems (as percentage of the total)

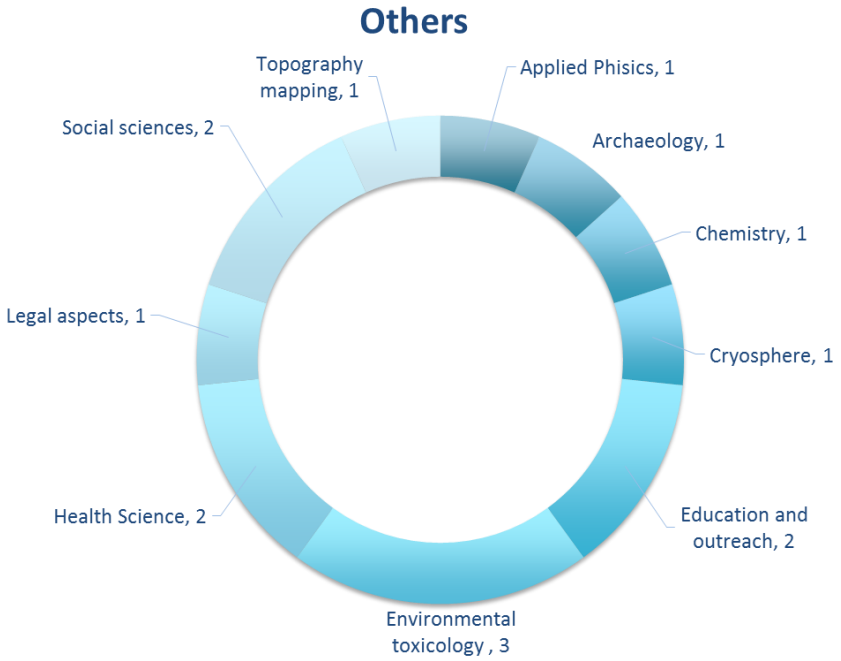

Figure 8. "Other" (Fig. 7) disciplinary data managed by the systems

Data Discoverability, Accessibility and Interoperability

This section deals with: Metadata, Data encoding, Discovery, Access and Processing services implemented by the data systems.

As depicted in Figure 9, a large majority of the data systems analyzed implement a formalized data encoding specification, but only half of them make use of a standard specification; they do not implement any processing interface.

More than half of the systems implement a metadata scheme, but only half applies a standard specification. Fewer of them (i.e. half of the responders) make use of a metadata specification to underpin a discovery service/interface. The systems that implement discovery services also publish an access interface. Figure 10 reports percentages.

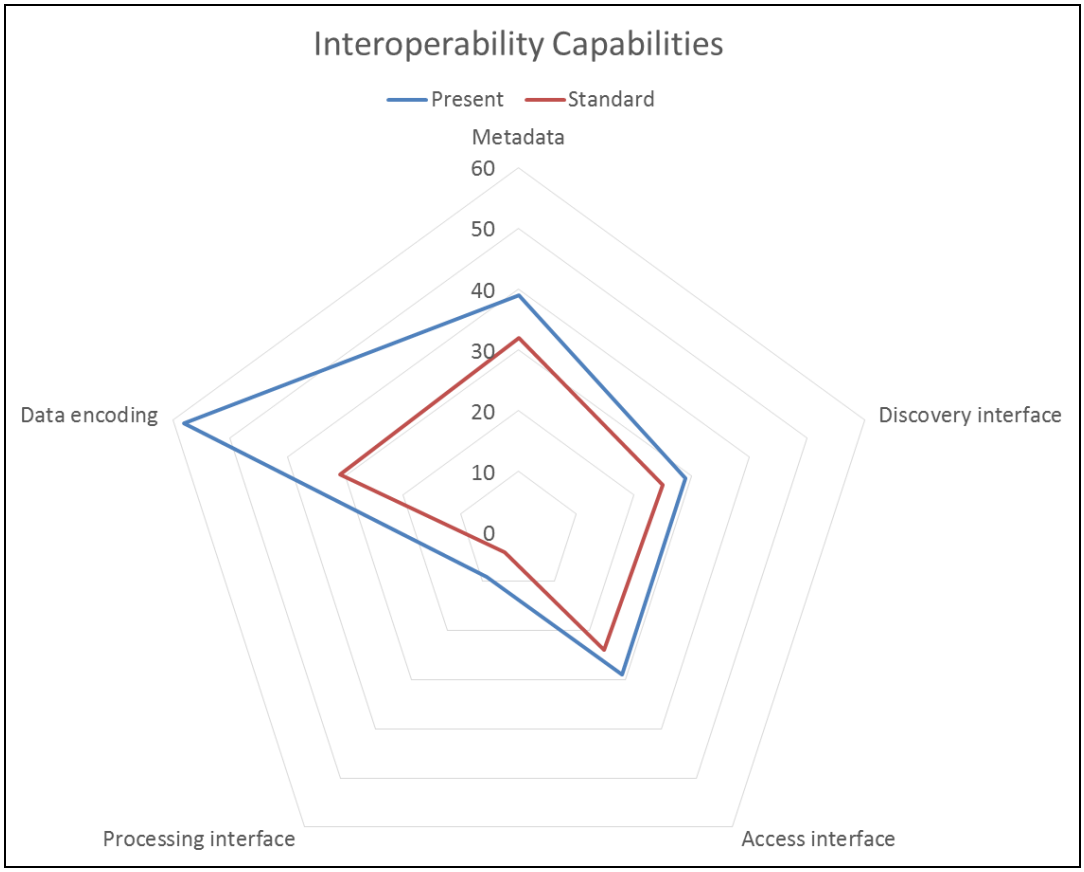

Figure 9. Interoperability artefacts implemented by the Data Systems

Metadata specification

As for the Data Systems implementing a metadata standard specification, the majority applies ISO TC211 metadata specification (either abstract or encoding). More than a quarter (25%) are compliant with the INSPIRE Metadata specification. More than 20% implement a Community metadata specification. Figure 11 shows the complete results.

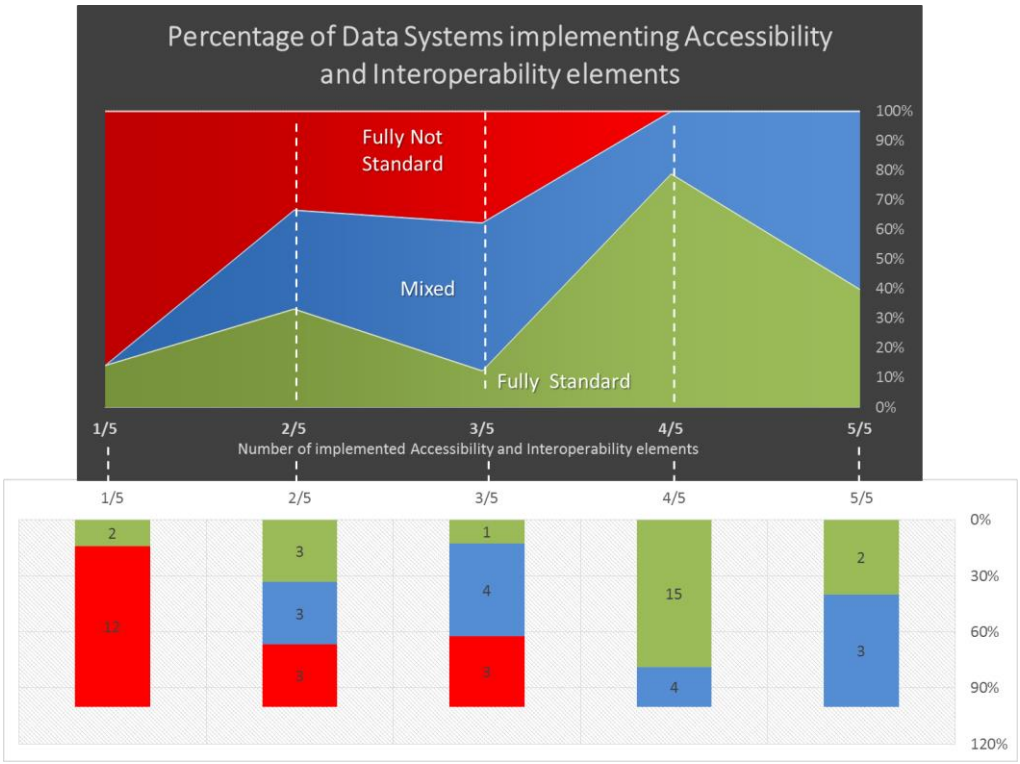

Figure 10. Percentage of data systems implementing Data Discoverability, Accessibility and Interoperability

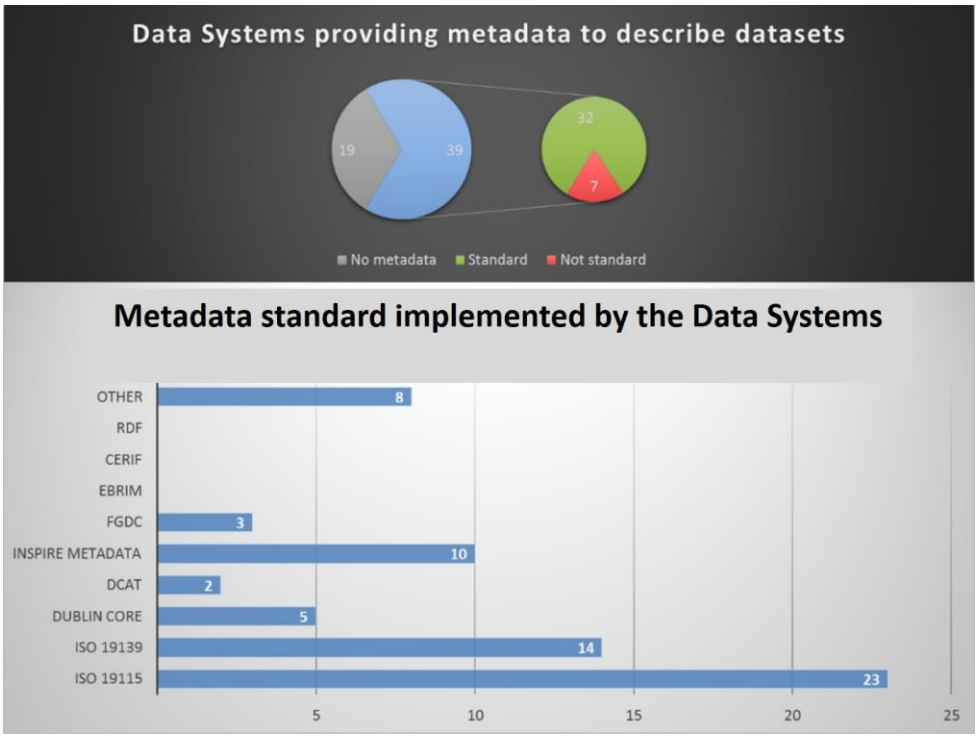

Figure 11. Metadata standard specifications implemented by the Data Systems.

Slightly more than half of the systems implement a permanent ID (PID) characterizing each dataset (Fig 12).

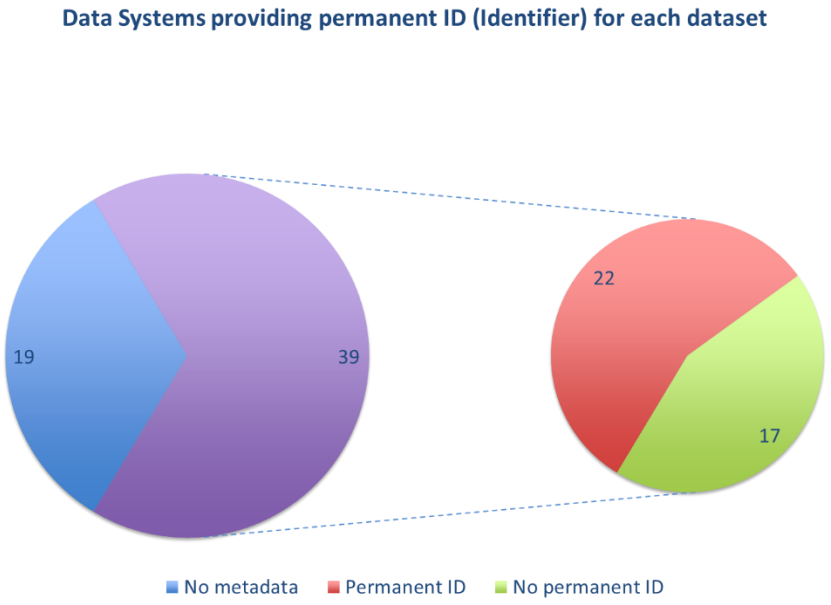

Figure 12. Data Systems providing a Permanent ID

Figure 13 shows the advanced metadata elements supported by the Data Systems: a large majority provides information on the data origin (67%); a majority provides a description of data nature and scale (60%); only a third (33%) provides a link to publications; only 10% of the systems inform about possible users.

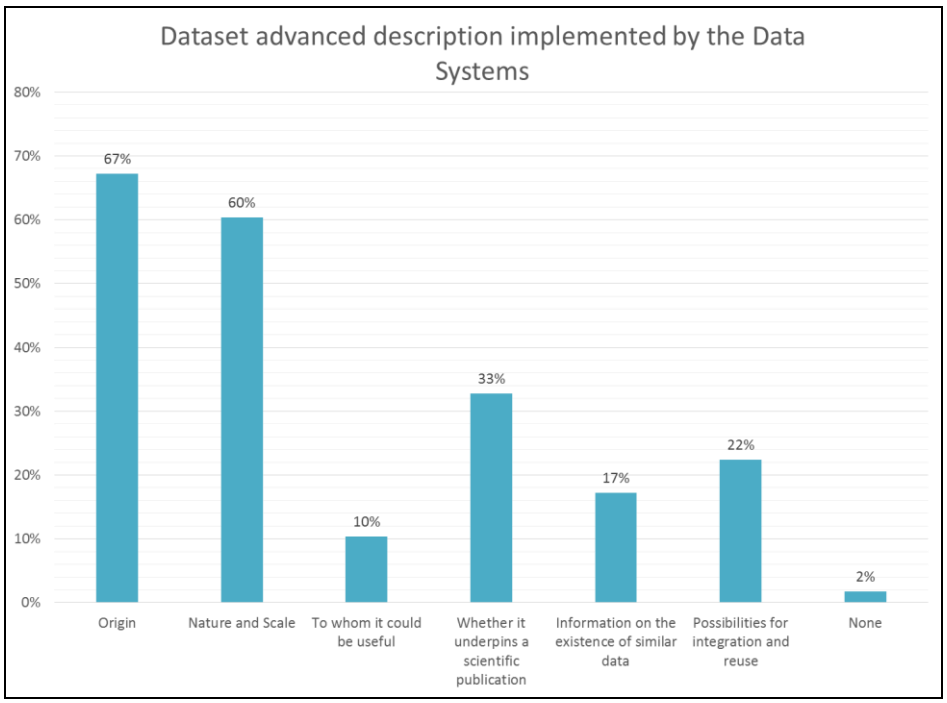

Figure 13. Advanced metadata implemented by the Data Systems

Data Encoding

As for the Data Systems implementing a standard specification for data encoding, there is no significant preference: the majority makes use of CSV encoding followed closely by DIF, CF-netCDF, netCDF, and SHP. A good number of systems makes use of encodings based on Community specifications. Figure 14 reports the full statistics.

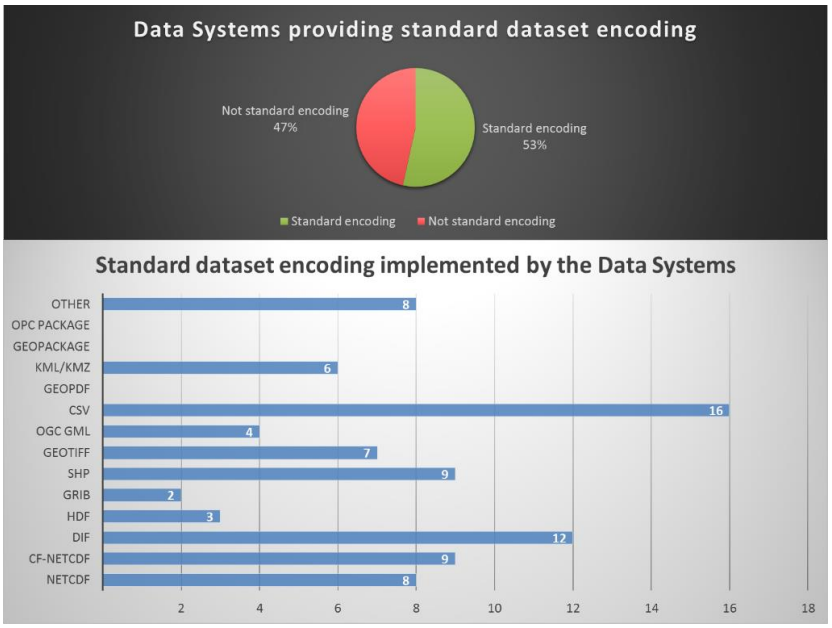

Figure 14. Standard specifications used to encode datasets

Dataset Discovery service

As depicted in Figure 15, the Data Systems analyzed implement several different standard interfaces for data discovery and harvest: the most used are OAI-PMH (for harvesting) and OGC CSW (for online query); FTP and THREDDS are well-used as well as other Community catalogue specifications.

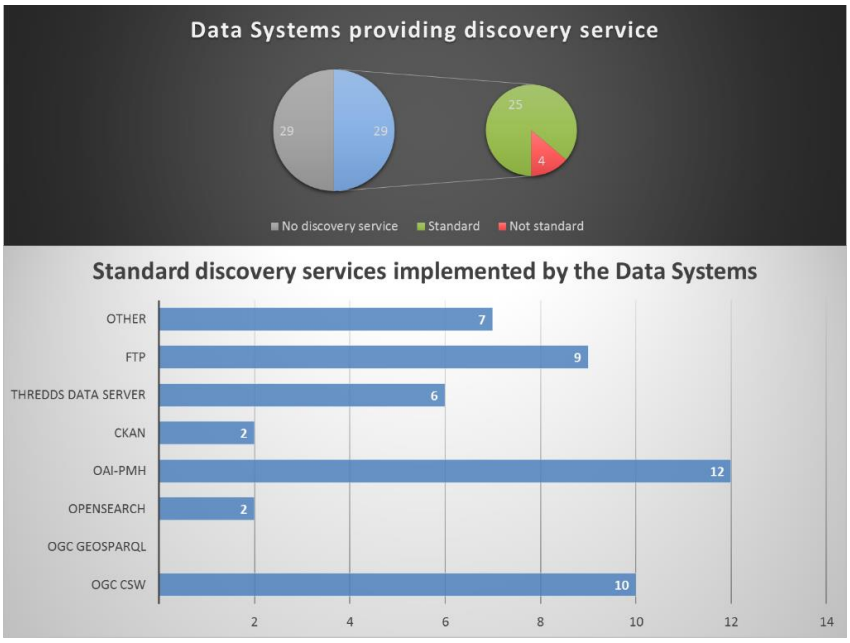

Figure 15. Standard discovery interfaces used by the Data Systems

Dataset Access service

As depicted in Figure 16, the Data Systems analyzed implement several different standard interfaces for data access and download: OGC WMS (Web Map Services) is the most used followed by FTP, OPenDAP, THREDDS and OGC WFS (Web Feature Service).

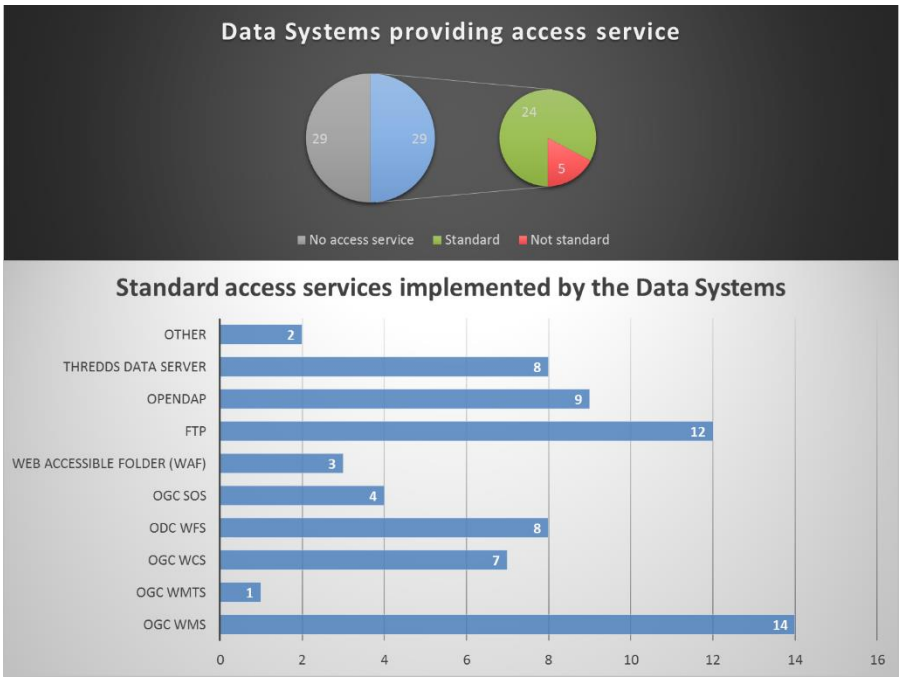

Figure 16. Standard access interfaces used by the Data Systems

Dataset Processing service

Figure 17 shows that only a few of the data systems analyzed (15%) support a data processing interface. They use the capabilities offered by tools like R language and Matlab.

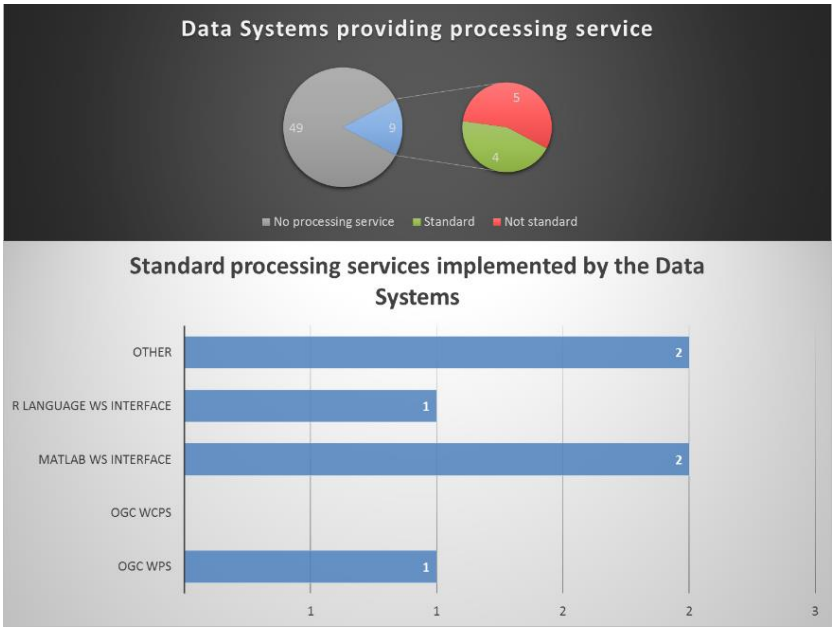

Figure 17. Standard processing interfaces used by the Data Systems

Data Policy and Archiving

The majority (71%) of the data systems analyzed are characterized by a formal data policy –see Figure 18.

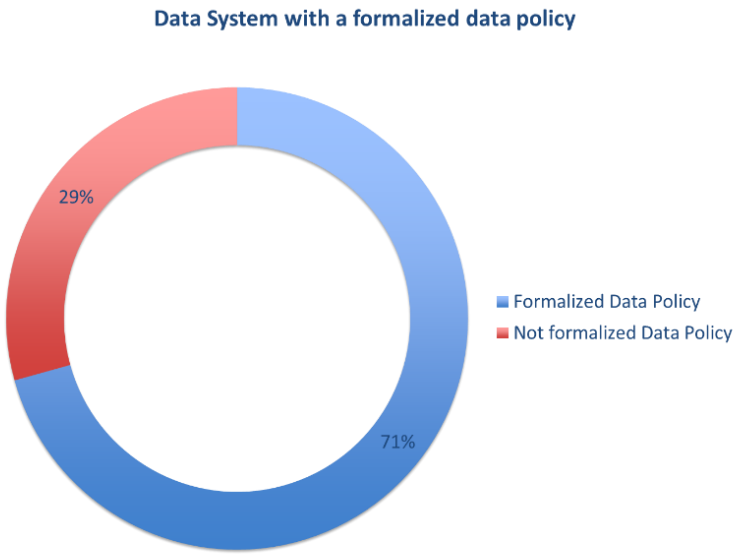

Figure 18. Percentage of systems analyzed with a formal data policy

Almost all the data systems provide information about the access procedures, but they do not inform users about the instruments to be used for data access. Most of them inform users about the access constraints, but they do not provide explanations about the reasons for not sharing data and more than 50% do not clearly specify an embargo time. Figure 19 depicts the data sharing information considered and the resulting percentages.

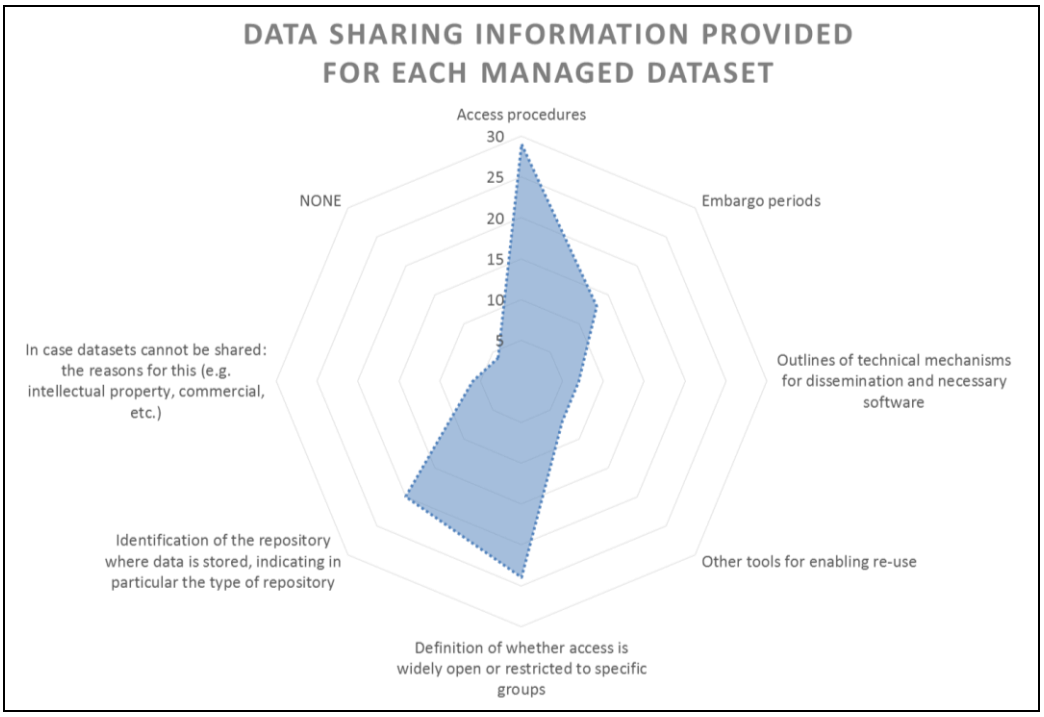

Figure 19. Sharing Information provided by the data systems

As to archiving and preservation information, only 8% of all the analyzed data (i.e. 11.3% of the systems including some data policy and archiving information) inform users on how long the data will be preserved and its approximate end volume. About a quarter (25.8%) of the all the analyzed data (i.e. 34% of the systems including some data policy and archiving information) provides information about the procedures set up for long-term preservation –see Figure 20.

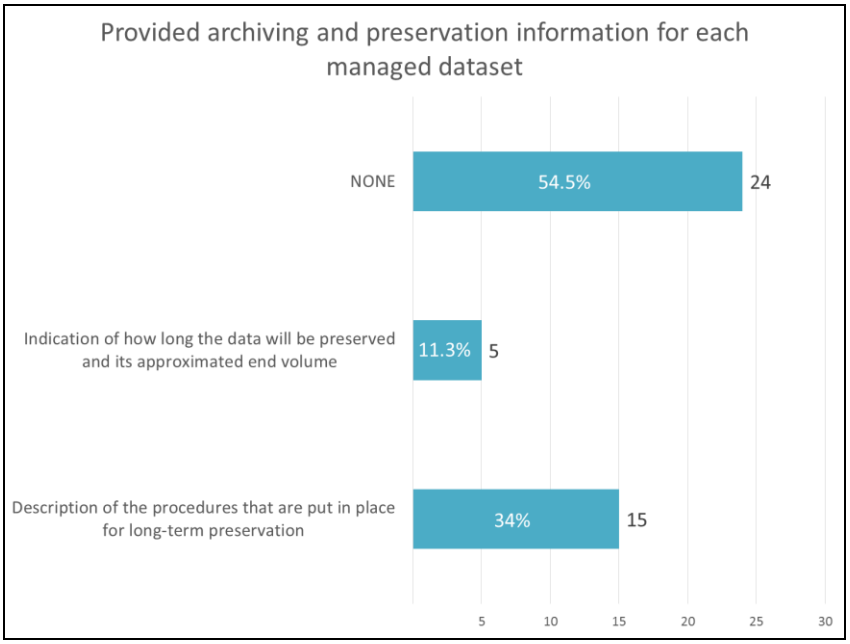

Figure 20. Archiving and preservation information provided by the data systems

Main areas of discussion

Looking at the questionnaire results, it is possible to recognize a set of areas that may deserve a follow up and/or a discussion.

- It is worthy to note that 21% of the responders belong to organizations outside Europe. However, only 12% of the total number of responders described their data systems as “not relevant for the European polar research”.
- The variety of disciplines considered by the data systems analyzed is wide, including also social, legal and health data –percentages are shown by Figures 7 and 8. Polar research is a clear example of a multidisciplinary research area.
- A large majority of the data systems analyzed implements a formalized data encoding specification and fewer a metadata specification. Still metadata specification is an issue.
- The utilization of international (or well-used and documented community) standards for data and metadata specification is not well supported –about 50% of the analyzed systems.
- Metadata and data encodings are not commonly used to expose online interfaces for data discovery and access –fewer than 50% of the systems analyzed.
- Many international (or well-used and documented community) specifications are utilized for the data discovery and access interfaces –this is likely bound to the multidisciplinary nature of Polar research.
- Most of systems inform users about possible access constraints, but they do not provide explanations about the reasons for not sharing data and more than 50% of the systems analyzed do not clearly specify an embargo time.

- Few systems (15%) support data processing and expose an interface for that; half of them (7%) make use of a standard interface.
- Only 8% of all the analyzed data inform users on how long the data will be preserved and its approximated end volume. About a quarter (25.8%) provides information about the procedures set up for long-term preservation.

Final Recommendations

This section summarizes some possible recommendations emerging from the presentation of the survey results. It is recommended:

1. To engage the Polar Research Community in defining and adopting a set of Data Management principles to enhance the current data descriptions and ensure an effective sharing and (re-)usability.
    a. Provide these principles in a clear format;
    b. Provide a set of implementation guidelines for the principles.
2. To preserve and leverage the European landscape of systems/infrastructure “diversity” in order to ensure sustainability and evolution for Polar Research.
3. To consider interoperability solutions specifically developed for large multi-organizational system-of-systems and addressing highly multi-disciplinary research.
4. To continue the online survey with the aim to prepare a second report at the end of 2016.
5. To recognize the “missing questions” and include them in an enhanced version of the questionnaire.

ANNEX A: The Questionnaire

EU-PolarNet

Co-funded by the European Union

H2020 EU-PolarNet: Survey of the existing Polar Research Data Systems

Including their standards, good practice baselines, policies, and interoperability solutions

Part I: Who Are You?

The Respondent information.

a. Name *

b. Company *

c. Address

d. Address 2

e. City/Town

f. State/Provence

g. Country

h. Email Address

i. Phone Number

Part II: Data System Description

Data System identification, and the description of the Organization managing the Data System.

a. Name of the Described Data System *

b. Is the Data System of specific relevance for European polar research? * (Yes/No)

1. If you answered “Yes”, please explain why

c. Name of Organization managing the Data System *

d. URL of the Organization managing the Data System

e. Name and contact information of a contact person for the Organization managing the Data System * (You/Other)

If you answered “Other”, please fill contact information below

1. Name
2. Company
3. Address
4. Address 2
5. Country
6. Email Address
7. Phone Number

Part III: Which data are you managing?

Data content managed by the Data System.

a. Which branch(es) of Sciences are you dealing with? *

b. Please provide an example of the key parameters you are handling *

c. Data types * (Choose as many as you like)

- Research Data
- Monitoring Data
- Processed Data
- Other

Part IV: Which technology do you apply to allow Data Accessibility and Interoperability?

Description of the technology implemented by the Organization for the Data System, including metadata, data formats, and interoperability protocols for data discovery, access, and use.

a. Does the Data System implement any metadata to describe datasets? * (Yes/No)

If you answered “No”, please go to question b. of this part.

1. Is it an international/community standard? * (Yes/No)

   If you answered “Yes”, please choose as many as you like from the following list *

- ISO 19115
- ISO 19139
- Dublin Core
- DCAT
- INSPIRE Metadata
- FGDC
- ebRIM
- CERIF
- RDF
- Other

If you answered “No”, please describe the metadata encoding utilized by your Data System:

2. Does the Data System implement any permanent ID (Identifier) for each dataset? * (Yes/No)

   If you answered “Yes”, please describe which is the strategy used to provide a permanent ID to each Dataset

3. Dataset advanced description * (Choose as many as you like)

- Origin
- Nature and Scale
- To whom it could be useful+
- Information on the existence of similar data
- Possibilities for integration and reuse
- None

b. Does the Data System implement any international/community standard to encode managed datasets? * (Yes/No)

If you answered “Yes”, please choose as many as you like from the following list *

- netCDF
- CF-netcdf
- DIF
- HDF
- GRIB
- SHP
- GeoTIFF
- OGC GML
- CSV
- GeoPDF
- KML/KMZ
- GeoPackage
- OPC Package
- Other

If you answered “No”, please describe the dataset encoding utilized by your Data System *

Part V: Discovery Service

a. Does the Data System provide any discovery service to managed datasets? * (Yes/No)

If you answered “No”, please go to part VI.

1. Is it an international/community standard? * (Yes/No)

If you answered “Yes”, please choose as many as you like from the following list *

- OGC CSW
- OGC GeoSPARQL
- OpenSearch
- OAI-PMH
- CKAN
- THREDDS Data Server
- FTP
- Other

If you answered “No”, please list the non-standard (internet) protocols implemented and published by the Data System for dataset discovery *

Part VI: Access Service

a. Does the Data System provide any access service to managed datasets? * (Yes/No)

If you answered “No”, please go to part VII.

1. Is it an international/community standard? * (Yes/No)

If you answered “Yes”, please choose as many as you like from the following list *

- OGC WMS
- OGC WMTS
- OGC WCS
- OGC WFS
- OGC SOS
- Web Accessible Folder (WAF)
- FTP
- OpenDAP
- THREDDS Data Server
- Other

If you answered “No”, please list the non-standard (internet) protocols implemented and published by the Data System for dataset access *

Part VII: Processing Service

a. Does the Data System provide any processing service to managed datasets? * (Yes/No)

If you answered “No”, please go to part VIII.

1. Is it an international/community standard? * (Yes/No)

If you answered “Yes”, please choose as many as you like from the following list *

- OGC WPS
- OGC WCPS
- Matlab WS Interface
- R language WS Interface
- Other

If you answered “No”, please list the non-standard (internet) protocols implemented and published by the Data System for dataset processing *

Part VIII: Data and Sustainability Policy

Description of the Data and Sustainability Policy applied by the Organization for the Data System, including long-term preservation and curacy.

a. Do you have a formalized Data Policy for the managed datasets? * (Yes/No)

If you answered “No”, please go to part IX.

1. Data sharing information, please indicate whether the following information are provided by the Data System for each managed dataset* (Choose as many as you like)
    - Access Procedures
    - Embargo Periods
    - Outlines of technical mechanisms for dissemination and necessary software
    - Other tools for enabling re-use
    - Definition of whether access is widely open or restricted to specific groups
    - Identification of the repository where data is stored, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.)
    - In case datasets cannot be shared: the reasons for this (e.g. Ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related)
    - NONE
2. Archiving and preservation, please indicate whether the following information are provided by the Data System for each managed dataset* (Choose as many as you like)
    - Description of the procedures that are put in place for long-term preservation
    - Indication of how long the data will be preserved and its approximated end volume
    - NONE

Part IX: End

EU-PolarNet Co-funded by the European Union

Thank you

Annex B: Responses Statistics

Is the Data System of specific relevance for European polar research

58 out of 58 people answered this question

1	Yes	51 / **88%**
2	No	7 / **12%**

Name and contact information of a contact person for the Organization managing the Data System

58 out of 58 people answered this question

1	You	55 / **95%**
2	Add Contact	3 / **5%**

Is it an international/community standard?

39 out of 58 people answered this question

1	Yes	32 / **82%**
2	No	7 / **18%**

Data types

58 out of 58 people answered this question

1	Research data	55 / **95%**
2	Monitoring data	44 / **76%**
3	Processed data (e.g. satellite images)	23 / **40%**
4	Other	5 / **9%**

Does the Data System implement any metadata to describe datasets?

58 out of 58 people answered this question

1	Yes	39 / **67%**
2	No	19 / **33%**

Which ones?

32 out of 58 people answered this question

1	ISO 19115	23 / **72%**
2	ISO 19139	14 / **44%**
3	INSPIRE Metadata	10 / **31%**
4	Other	8 / **25%**
5	Dublin Core	5 / **16%**
6	FGDC	3 / **9%**
7	DCAT	2 / **6%**
8	CERIF	0 / **0%**
9	RDF	0 / **0%**
10	ebRIM	0 / **0%**

7	None	1 / **3%**

Does the Data System implement any permanent ID (Identifier) for each dataset?

39 out of 58 people answered this question

1	Yes	22 / **56%**
2	No	17 / **44%**

Dataset advanced description

39 out of 58 people answered this question

1	Origin	39 / **100%**
2	Nature and Scale	35 / **90%**
3	Whether it underpins a scientific publication	19 / **49%**
4	Possibilities for integration and reuse	13 / **33%**
5	Information on the existence of similar data	10 / **26%**
6	To whom it could be useful	6 / **15%**

Does the Data System implement any international/community standard to encode managed datasets?

58 out of 58 people answered this question

1	Yes	31 / **53%**
2	No	27 / **47%**

Which ones?

31 out of 58 people answered this question

1	CSV	16 / **52%**
2	DIF	12 / **39%**
3	CF-netCDF	9 / **29%**
4	SHP	9 / **29%**
5	Other	8 / **26%**
6	netCDF	8 / **26%**
7	GeoTIFF	7 / **23%**
8	KML/KMZ	6 / **19%**
9	OGC GML	4 / **13%**
10	HDF	3 / **10%**
11	GRIB	2 / **6%**
12	GeoPDF	0 / **0%**
13	GeoPackage	0 / **0%**
14	OPC Package	0 / **0%**

Does the Data System provide any discovery service to managed datasets?

58 out of 58 people answered this question

1	Yes	29 / **50%**

Is it an international/community standard interface/protocol?

29 out of 58 people answered this question

1	Yes	25 / **86%**
2	No	4 / **14%**

Which ones?

25 out of 58 people answered this question

1	OAI-PMH	12 / **48%**
2	OGC CSW	10 / **40%**
3	FTP	9 / **36%**
4	Other	7 / **28%**
5	THREDDS Data Server	6 / **24%**
6	CKAN	2 / **8%**
7	OpenSearch	2 / **8%**
8	OGC GeoSPARQL	0 / **0%**

Does the Data System provide any access service to managed datasets?

58 out of 58 people answered this question

1	Yes	29 / **50%**
2	No	29 / **50%**

Is it an international/community standard interface/protocol?

29 out of 58 people answered this question

1	Yes	24 / **83%**
2	No	5 / **17%**

Which ones?

24 out of 58 people answered this question

1	OGC WMS	14 / **58%**
2	FTP	12 / **50%**
3	OpenDAP	9 / **38%**
4	ODC WFS	8 / **33%**
5	THREDDS Data Server	8 / **33%**
6	OGC WCS	7 / **29%**
7	OGC SOS	4 / **17%**
8	Web Accessible Folder (WAF)	3 / **13%**
9	Other	2 / **8%**
10	OGC WMTS	1 / **4%**

Does the Data System provide any processing service to managed datasets?

58 out of 58 people answered this question

1	No	49 / **84%**
2	Yes	9 / **16%**

Is it an international/community standard interface/protocol?

9 out of 58 people answered this question

1	No	5 / **56%**
2	Yes	4 / **44%**

Which ones?

4 out of 58 people answered this question

1	Matlab WS interface	2 / **50%**
2	Other	2 / **50%**
3	OGC WPS	1 / **25%**
4	R language WS interface	1 / **25%**
5	OGC WCPS	0 / **0%**

Do you have a formalized Data Policy for the managed datasets

58 out of 58 people answered this question

1	Yes	41 / **71%**
2	No	17 / **29%**

Data sharing information

41 out of 58 people answered this question

1	Access procedures	29 / **71%**
2	Definition of whether access is widely open or restricted to specific groups	24 / **59%**
3	Identification of the repository where data is stored, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.)	20 / **49%**
4	Embargo periods	13 / **32%**
5	Other tools for enabling re-use	7 / **17%**
6	Outlines of technical mechanisms for dissemination and necessary software	7 / **17%**
7	In case datasets cannot be shared: the reasons for this (e.g. Ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related)	6 / **15%**
8	NONE	4 / **10%**

Archiving and preservation

41 out of 58 people answered this question

1	NONE	24 / **59%**
2	Description of the procedures that are put in place for long-term preservation	15 / **37%**
3	Indication of how long the data will be preserved and its approximated end volume	5 / **12%**