EGI-Engage

Service Delivery Model

from datacentre to community

Date	19 October 2017
Activity	WP2
Lead Partner	EGI-Engage
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2699

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	E. van Maanen & A. Ellenbroek	FAO	31-AUG-2017

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.0	13/07/2017	E. van Maanen & A. Ellenbroek	E. van Maanen & A. Ellenbroek
v.0.1	15/07/2017	Review Pasquale Pagano, processed comments	E. van Maanen & A. Ellenbroek
v.0.5	21/07/2017	Review Sy Holsinger, processed comments	E. van Maanen & A. Ellenbroek
v.0.9	16/08/2017	Review Yannick Legré processed comments	E. van Maanen & A. Ellenbroek
v.1	28/08/2017	E. van Maanen & A. Ellenbroek	E. van Maanen & A. Ellenbroek
v.1	29/08/2017	Review Vivi Katifori (Agroknow), processed comments	E. van Maanen & A. Ellenbroek
v.2	31/08/2017	Final version	E. van Maanen & A. Ellenbroek

TERMINOLOGY

A complete project glossary and acronyms are provided at the following pages:

- https://wiki.egi.eu/wiki/Glossary
- https://wiki.egi.eu/wiki/Acronyms

Contents

Executive summary

Data managed in the fishery and marine sciences sector is growing exponentially in both size and frequency, and it is expected to keep this pace in future years. This exponential growth brings forth a variety of data challenges including interoperability when sharing, using and reusing data.

An important finding related to interoperability is the need to clearly segment the roles and responsibilities of different actors in a service delivery across infrastructures. With the aim to support this need, we propose to develop a Service Delivery Model (SDM) that facilitates the community process to describe a multi-stakeholder and cross-infrastructure service delivery through a layered infrastructure.

This SDM consists of four layers: (inter) national datacentre, infrastructure mediator, service mediator and the research community.

The first layer is at the (Inter)National Datacenter where data and metadata are stored and processed. The Datacenter exchanges data with an Infrastructure Mediator to enable collaboration. An Operational Level Agreement, which defines the responsibility of parties, provision and support of the defined services, supports this exchange of data.

The Infrastructure Mediator exchanges data with the Service Mediator where data are processed, organized, structured or presented in a given context. Within this step data becomes information ready-to-use for analyses. A Service Level Agreement covers the IT Services, IPRs, service level targets and responsibilities of the Service Provider and the customer.

The Service Mediator is responsible for data transfer to the Research Community where information is transformed and modelled with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. The transfer of data is covered by Terms of Use and Privacy Policy that gives definitions, user rights and responsibilities and the (ownership of) IPRs of (various parts of) the content. If necessary, a Memorandum of Understanding can be used to express a convergence of will between relevant parties and intended common line of action and direction.

The application of the proposed SDM on a use case with FAO, EGI and CNR revealed several flaws that need improvement: (i) there are no concrete legal terms on relevant fields such as ownership of IPRs, (ii) there is a lack of concrete consequences when the Provider does not meet the service level targets, (iii) lack of clear legal definition of terms such as ‘Intellectual Property Rights’, and (iv) no mention of options and consequences to distribute data to third parties.

After consulting various stakeholders, other recommendations and guidelines for further development of this SDM were proposed: (i) define data, (ii) define data ownership, and (iii) define citation, attribution and credit, (iii) clarify citation of compound datasets, and (iv) prevent long tail research data. The development of a more refined SDM will facilitate the formulation of exploitation plans and data policies across data-platforms, and will contribute to build trust and confidence in research communities that need to exploit these platforms.

EGI is in the position to support organizations with the development of data policies by co-developing a clear and ready-to-implement SDM. The SDM will help the communities to understand and formulate technical support in data management. This study proposed an SDM, and tested this in a FAO use case. EGI can use this study by elaborating these recommendations into a next version of the proposed SDM.

1 Introduction

The data managed in the fishery and marine sciences sector is growing exponentially in both size and frequency, and it is expected to keep this pace in future years. Technological innovations, such as mobile phones and satellites (to e.g. monitor and detect illegal fishing activities[1]), require Big Data collection, management and processing, and interoperability across current institutional and infrastructure boundaries.

To optimize the effective and efficient use of Big Data, it is important that these are available in the correct format, accessible, well organized and properly annotated. This practical challenge requires the availability of a technical infrastructure and skills to process large amounts of data in analytical services and feed these into decision-making tools, among others. Interoperability of data is required to reliably combine data from multiple data services. Interoperability can be categorized in organizational, semantic, technical and legal interoperability as explained by the European Interoperability Framework[2].

All types of interoperability are important, but this report relates to EGI-Engage Deliverable 2.6[3] on legal interoperability, and thus considers only legal interoperability. Based on feedback received on that Deliverable, the community objective was to further investigate options to establish a legal interoperability framework. This document can be considered as an Annex to the original Deliverable; it includes feedback from EGI Engage members, reflects changes in interoperability technical support for specific use cases in an established collaboration with CNR, and collects progress in and feedback from the FAO community in establishing a comprehensive data policy.

An important community finding was the need to clearly segment roles and responsibilities of different actors in a service delivery across infrastructures. We used that recommendation to develop and propose a Service Delivery Model (SDM) to the community to examine if that presented a viable method to describe a multi-stakeholder and cross-infrastructure service delivery. This document contains both the SDM and the comments from the community on an abstract use-cases defined around it.

Chapter 2 presents an outline of the proposed Service Delivery Model and all associated (legal) documents. Chapter 3 tests the proposed SDM with a use case between FAO as end customer (research community), through D4Science (BlueBRIDGE and iMarine data services) as service mediator and EGI as the infrastructure mediator. However, this SDM can be generalized to ensure its generic application in different sectors. Chapter 4 tests the overall applicability of the SDM model to gauge if the proposed SDM is understandable for representatives of a research community. The focus is on developing a framework, rather than to test specific services delivery mechanisms.

[1] http://www.ship-technology.com/features/featureusing-big-data-to-combat-illegal-fishing-5688984/

[2] This framework gives specific guidance on how to set up interoperable digital public services. Retrieved on July 17th 2017 at: https://ec.europa.eu/isa2/sites/isa/files/eif2.png

[3] https://documents.egi.eu/public/ShowDocument?docid=2699

2 The Service Delivery Model

The figure below presents the Service Delivery Model (SDM) with for each step in the research lifecycle the minimum required agreement along with a brief description.

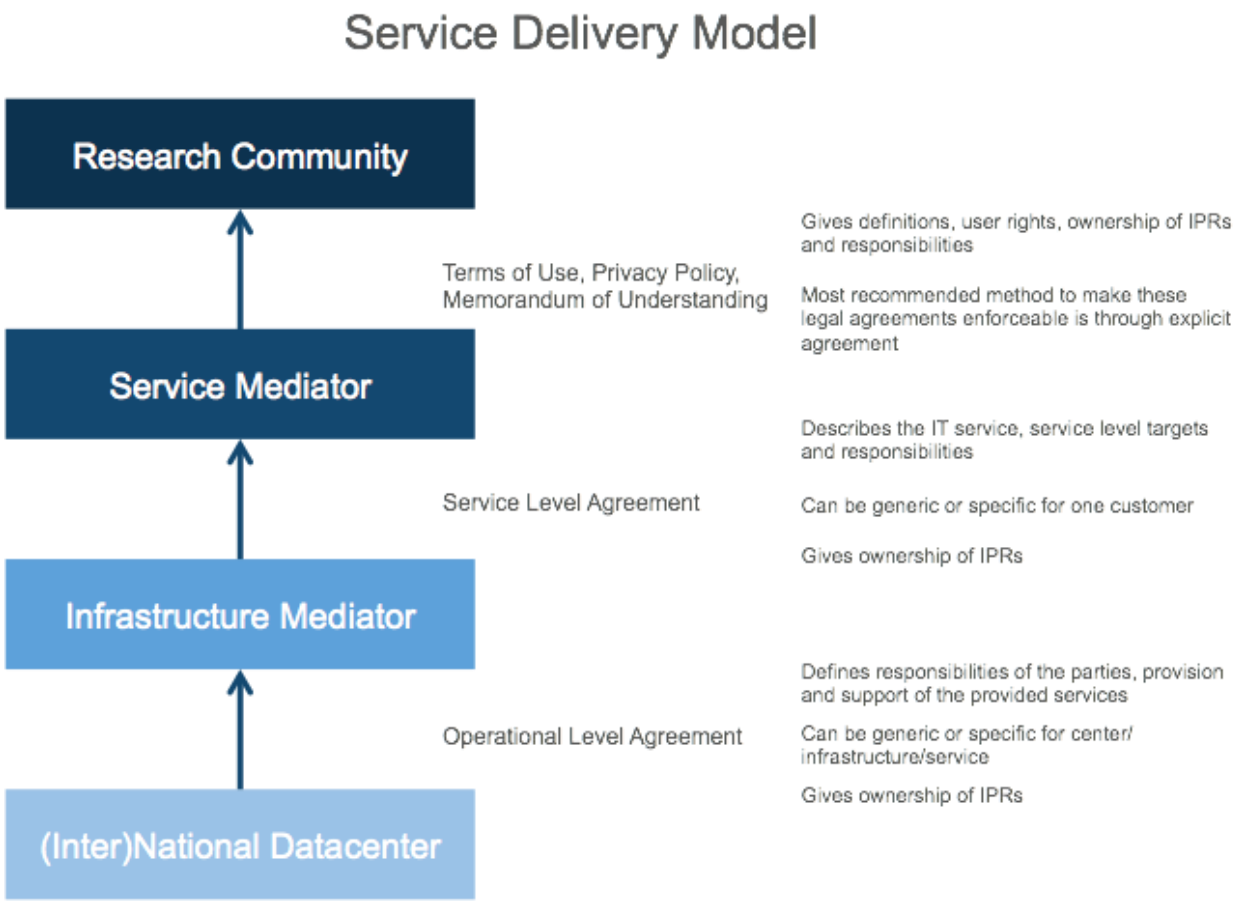

Figure 1 Overview of the Service Delivery Model

The first layer is at the (Inter)National Datacenter level where data and metadata are stored and processed. Metadata is structured data about data, of any sort in any media, which imposes order on a disordered data or information. In database management systems, metadata are index files and data dictionaries that store administrative information, while for textual, image and geospatial information metadata records can be stored as metadata headers or as independent data.

The Datacenter exchanges data with an Infrastructure Mediator to enable collaboration on shared resources to encompass the sharing needs of research activities. Examples of such Infrastructure Mediators are EGI and EUDAT. The data policy is captured by means of an Operational Service Agreement (OLA) which defines the goods or (data) services to be provided, (ownership of)

Intellectual Property Rights (IPRs) of (various parts of) the content and the responsibilities[4] of the parties involved in this agreement.

The Infrastructure Mediator exchanges data with the Service Mediator where data are processed, organized, structured or presented in a given context that makes the data useful. Within this step data becomes information[5] ready-to-use for analyses. Examples of such Service Mediators are D4Science, Global Biodiversity Information Facility (GBIF) and Copernicus Marine Environment Service. The exchange of data is carried out by means of a Service Level Agreement (SLA) which describes the IT Service, (ownership of) IPRs of (various parts of) the content, service level targets and responsibilities of the Service Provider and the customer[6].

The Service Mediator is responsible for data transfer to the Research Community where information is transformed and modelled with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. The transfer of data is carried out by means of Terms of Use and Privacy Policy which gives definitions, user rights and responsibilities and the (ownership of) IPRs of (various parts of) the content. If necessary, a Memorandum of Understanding can be used to express a convergence of will between relevant parties and intended common line of action and direction. This document is not binding unless it meets the four criteria of contracts: offer and acceptance, consideration and intention to be legally bound. Due to the complexity of this system it appeared to be beneficial to bundle in each layer all conditions and terms set in the lower layers. In this way the complexity of lower layers are 'hidden' in upper layers. For example, the Terms of Use reflect all underlying conditions and terms set in both the SLA and OLA.

[4] https://wiki.egi.eu/wiki/EGI_OLA_SLA_framework#Operational_Level_Agreements

[5] Data in themselves are fairly useless. But when these data are interpreted and processed to determine their true meaning, they become useful and can be called information. Data is the computer's language. Information is our translation of this language.

[6] https://wiki.egi.eu/wiki/EGI_OLA_SLA_framework#Service_Level_Agreemens

3 Application of Service Delivery Model in FAO use case

The focus of chapter 2 was to develop a generic SDM to create an overview of the research lifecycle and the required legal documents. The Food and Agriculture Organization of the United Nations (FAO) participates to the EGI-Engage project to support the creation and validation of a potential SDM. The SDM in the FAO use case brings together:

- FAO as the end customer with the need for legal interoperability;
- The D4Science infrastructure as a mediator for data services (developed in the BlueBRIDGE H2020 project); and
- EGI as the infrastructure and computational resource provider.

With the legal interoperability objective in mind FAO reviewed two agreements to identify potential gaps and to see whether current agreements can be modified or whether completely new agreements are required to fit with the SDM. To test the proposed SDM and to find out where the gaps are in FAO's current service delivery, five FAO officials were interviewed. These FAO officials were selected from different departments that are concerned with data policies and sharing but all with a different perspective. Officials from the following departments were interviewed: fisheries and aquaculture (FIAS), communications (OCCI), legal (LEGN) and statistics (ESS) (please see the Annex for a summary of the interviews). It is important to note that while organizations like FAO highlight the need for comprehensive data policies, few have implemented binding and formal policies.

At first, it is important to consider what type of data is included in any SDM. There are three broad types of data in research, which are all included:

1. Primary/raw data: data coming directly from the source;
2. Aggregated data/statistics: data processed from primary/raw data; and
3. Metadata: reference data describing either the primary/raw data or aggregated data/statistics.

In most data projects, there are likely to be two components. The first is the data collected, assembled, or generated, i.e. the raw content in the system (e.g. fish catch from a specific vessel or reported by a country, the readings of an instrument). The second component is the data system in which the data is stored and managed. The SDM focuses on the policy and legal aspects of both components.

We usually do not think of data content separate from the system in which it is stored, but the distinction is important in terms of IPRs. The question is what is protected by copyright. Factual

data has no copyright protection since it does not meet the criterion of creativity; it is after all not possible to copyright facts (Carroll, 2015).

Because of the different copyright status of databases and data content, different mechanisms are required to manage each. Copyright and Terms of Use can manage the use of databases and some data content (that which is itself original), but contract law, trademarks, and other type of management are required to regulate factual data. All FAO interviewees responded that to facilitate the use, reuse and sharing of data a uniform FAO data policy is required. FAO's data policy currently depends on case-by-case assessments, which is highly time-intensive. The interviewees also agreed that the awareness of the need for a FAO data policy is increasing, but nothing concrete has been established yet. The proposed SDM can be a first step towards a uniform data policy, facilitating the compatibility of aforementioned types of mechanisms.

As set forth in EGI-Engage D2.6, legal interoperability intends to provide clarity on the rules and rights for the use, reuse and sharing of data. After interviewing multiple experts varying from an IP lawyer to policymakers the following definition of legal interoperability was agreed upon: *'the compatibility of legal rights, terms, and conditions of databases from two or more sources so that the data may be combined and integrated by any user, without further permission and without compromising the legal rights of any of the data sources used'* (van Maanen & Ellenbroek, 2016).

A SDM covers more than legal interoperability. It includes the creation, organisation, documentation, storage and share of data services delivery. A SDM takes into account issues such as data protection and confidentiality, data preservation and curation, and provides a framework that supports researchers and their data throughout the course of their research and beyond. There are various benefits of a complete and applied SDM such as: traceability of data, continuity if project staff leave or new researchers join, avoid unnecessary duplication (e.g. re-collecting or re-working data), maintain underlying publications of data, more collaboration in the research and data sector, easier to cite data for all data users and clarity on legal aspects of data sharing, access and reuse. It is worth mentioning that these benefits might require different types of SDMs. For example, the difference between publication[7] and sharing[8] of data require different licensing schemes and therefore types of SDMs (DMP Edinburgh, 2015)[9]. The overall structure of the SDM is developed in such a way that these different types do not require a different overall structure.

The SDM was used to scrutinize agreements through interviews of several relevant parties:

1. the SLA[10] between D4Science and the EGI Foundation on e-Infrastructure computing resource provisioning, and

[7] The act of making data accessible under a certain set of conditions without restraining policies for entities.
[8] The act of making data accessible under a certain set of conditions to selected people either directly or indirectly (such as all members of a group, all users having a specific role, all members of a VRE) identifiable.
[9] Based on University of Edinburgh Data Management Plan, Retrieved on July 19th 2017 at: http://www.ed.ac.uk/information-services/research-support/research-data-service/planning-your-data/writing-a-data-management-plan
[10] https://documents.egi.eu/public/ShowDocument?docid=2875

2. the MoU[11] between FAO and CNR to enable communities to exploit e-infrastructure services. These agreements in particular were selected due to their legal character. The SDM model provides a template against which these agreements can be assessed to identify gaps in the legal framework.

Several gaps were identified.

Service Level Agreement (D4Science and EGI Foundation)

- There are no concrete legal terms on relevant fields such as ownership of IPRs in both the document itself and the reference to EGI's Policies and Procedures website, these should be secured from back-end to front-end, and
- There is a lack of concrete consequences when the Provider does not meet the service level targets. A Service Level Agreement is not binding per se. It has to be appended to a contract to gain this capacity. The contract will then regulate the consequences of breaching the agreed SLA, which is usually a fraction of the cost paid per month, limited to this cost.

Memorandum of Understanding (FAO and CNR)

- The term 'Intellectual Property Rights' (Art 6) is not defined in this MoU. As such, it is unclear what the scope of the term 'Intellectual Property Rights' is (this is not a clear-cut term, for example with regards to the protection of know-how). A more common, though fairly broad, definition of Intellectual Property Rights is:

"Intellectual Property Rights means all copyrights, personality rights, patents, utility models, design rights, trademarks, domain names, database rights, rights concerning confidential information (including know-how) and other intellectual property rights and similar rights, both registered and unregistered and including all applications and rights to be applied for and/or to be granted, the renewal and/or expansion of, and rights to claim the priority of, such rights, and all similar or equivalent rights or other forms of protection which exists now, or in the future in any part of the world.";

- The term "originating Party" (Art 6) is unclear. It is therefore recommended to use the following definition of originating Party:

"Parties remain the sole owners of the Intellectual Property Rights that Parties use to carry out activities under this MoU. Intellectual Property Rights created in the course of this MoU, or arising out of this MoU, by both Parties, jointly or separate, will vest in FAO. If necessary, Parties will promptly, upon request, assist in the transfer of such rights to FAO";

- There is no mention of IPRs when content is developed, either jointly or separate, with third parties, either created in the course of this MoU or arising out of this MoU. FAO officials from the fisheries and aquaculture and the communication departments

[11] https://goo.gl/UgXmUq

confirmed that this lack of clarity of the origin of this work complicates attribution and citation when reusing data; and

- There is no mention of options and consequences to distribute data to third parties. The interviewed representative of the communications department adds to this that the purpose of the reuse of data should also be included in the options to distribute data to third parties. This representative presented an example where an external party wanted to conduct a market analysis for investment in agro-alimentary industry. This party wanted to include FAO's commodity prices in its algorithm to assess where and when to buy food products. The aim of this party appeared not to be aligned with FAO's strategic objectives, as this external party would merely financially exploit periods of food shortage.

Other recommendations to be included in all agreements within this SDM are:

- Include a fine-grained identity policy within all agreements. Research in the fishery and marine sciences sector is mostly subject to international organizations that can have different statuses with regard to IPRs. The interviewee working within the legal department confirmed this recommendation and mentioned that the status of FAO in this respect is more extraterritorial than the status of other organization such as the International Maritime Organization (IMO). Such choices are reflected by policy-decisions of these Organizations;
- Develop and document a standardized process to analyze whether standards for ownership of IPR are included in the metadata;
- The Infrastructure Mediator together with the Service mediator should approach the Research Community to collaborate on user-based management[12]. D4Science is the ideal partner since the decisions with respect to software are being made in this layer.

These current agreements can be modified to develop a complete and concrete SDM. This is stated in the SLA in Section 9[13], and in the MoU in Article 12[14]

[12] Such as security and quota management, CRUD, provenance/business/ownership metadata, and the responsibility of dynamic metadata.

[13] Section 9: 'Amendments, comments and suggestions must be addressed to the Provider and the Customer (...)'

[14] Art 12: 'This MoU may be modified by the written mutual consent of the Parties, in accordance with their respective rules. Such amendments will enter into force one month following notifications of consent by both Parties.'

4 General recommendations for data management

To develop a successful SDM all elements included in policies and procedures have to be clear and concise. Consultation of EGI, the fisheries and marine science and research community, and some institutional stakeholders showed that definitions of some basic terms have been the source of misunderstandings within service deliveries. These terms are 'data', 'data ownership' and 'data citation, attribution and credit'. This chapter aims to define these terms, and gives other recommendations and guidelines based on the interviews taken at FAO, when developing a SDM.

Define data

There are many legal definitions of data; unfortunately there is not one uniform legal definition of data that is applicable for all types of data managed in an infrastructure. For this study, two practicing IPR lawyers from leading international law firms were interviewed about the definition of 'data'. Both answered that 'there is no uniform definition of data', that 'satisfies everybody and that covers all syntactic and pragmatic aspects of the use of data'. Both IPR lawyers state that 'data' is 'information' or a 'value' in the broadest sense.

To provide more insight in the definitions used, this study presents two definitions and briefly discusses the benefits and drawbacks.

The Information Commissioner's Office (ICO), UK's independent body set up to uphold information rights, defined data as follows[15]:

> **'Data** means information which:
>
> (a) is being processed by means of equipment operating automatically in response to instructions given for that purpose,
>
> (b) is recorded with the intention that it should be processed by means of such equipment,
>
> (c) is recorded as part of a relevant filing system or with the intention that it should form part of a relevant filing system,
>
> (d) does not fall within paragraph (a), (b) or (c) but forms part of an accessible record as defined by section 68, or
>
> (e) is recorded information held by a public authority and does not fall within any of paragraphs (a) to (d).'

This definition is comprehensive but its applicability in practice is low due to its complexity. The objective is to develop a definition of data that is understandable for individuals varying from data managers to legal professionals. Therefore a definition such as the above is not preferable.

[15] https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/

Furthermore, this definition raises questions such as: (i) is information that has not been collected or mined yet, defined as data, and (ii) in which stage does information become 'data'?

Whereas ICO's definition is rather generic, this study will also present a definition tailored to the context of this study: research data. Leiden University defined 'research data' as follows:

> 'Research data can be defined as recorded information that supports or validates the observations or conclusions from a research project. It takes many forms, ranging from numbers and measurements to documents, publications and images. Research data covers both collected, unprocessed data as well as analysed, generated data. This can be in digital and non-digital formats (e.g. samples, completed questionnaires, sound recordings, etc.).[16]'

The positive aspect of this definition is that it is generic and can be applied to many cases and situations of research data. The drawback of this generic character is that it lacks clarity that legal situations require.

Researchers assume that they control the data and have the intellectual property rights and that they can decide what terms to impose on their data. Often, however, researchers do not, in fact, have these rights. It is therefore important to define the frequently used term 'data ownership'.

Define data ownership

"Data ownership" is a relatively new term that was not commonly used before 2000. There is much discussion amongst IPR lawyers on this subject for two main reasons.

At first, data is an intangible good. The notion of "ownership" carries with it a sense of ownership akin to that applied to tangible goods. This may lead to the presumption that property rights apply. Such rights may be used to assess a monetary value, limit access, and prescribe how goods may or may not be reused by others. Every physical thing can by default be an object of property. But property rights apply only to tangible goods (Egloff, 2014). Which means that from a strictly legal perspective 'ownership of data' does not exist.

The protection of non-tangible goods is always limited to specific areas and specific objects; there is no "general protection". If national laws do not specify that particular non-tangible goods are objects of intellectual property rights, then no rights apply. Individuals cannot claim intellectual property rights over items that are not covered by the relevant national laws (Egloff, 2014).

Secondly, the basic concept of ownership means to have legal title and full property rights to something. If we accept this as the correct definition of ownership, then data ownership must be to have legal title to one or more specific items of data. However, this cannot be what is generally meant. If it were, then anyone assigned as a data owner could take the data they were told they owned and sell it. In many cases however this is impossible because 'data owners' in its popular

[16] http://law.leiden.edu/organisation/meijers/research-institute/data-management-2016.html

use are not authorized to sell the data. Data ownership therefore does not have a literal meaning, and is rather an analogy instead of a defined term.

Yet there is one area where data ownership is 'real'. This 'real' data ownership concerns data that is purchased as a service from data vendors. Licensing of such data has many issues around intellectual property. Basically, the data vendors regard the data as their property and limit redistribution and derivation in the licensing agreements.[17]

Such restrictions present data management challenges. It is not easy to prevent redistribution – after all, how do you know what downstream applications are doing with the data? The same is true of using vendor-supplied data to derive or compute other data. Yet the vendors do enforce their contracts and are always on the lookout for violations, some of which can result in hefty settlements.

The problem here is determining to what extent data is the property of a supplier, particularly if the data can be found in the public realm (which it often can be). It is true that the vendor is collecting, standardizing and attaching metadata, but do these actions really make the data the property of the vendor? There is much debate on this topic.

Another term frequently used is 'rights holder'. Rights holders control the use of their exclusive rights, including reproduction and distribution.

When a third party uses this data, the rights holder should be cited, attributed or given credit.

Define data citation, attribution and credit

The terms attribution, credit, and citation all have distinct meanings. Attribution refers to the legally imposed requirement to attribute the rights holder when data are copied or reused in a specified manner. The remedy against someone who fails to attribute is to file a lawsuit. Either based on breach of contract or infringement of an IPR, depending on the legal mechanism used to impose the attribution requirements. Credit, on the other hand, is explicit recognition for the contribution to someone else's work. Finally, there is citation, which is rooted in norms of scholarly communication. The purpose of citation is to support an argument with evidence. However, citation has also become a proxy for credit, albeit an imperfect one (Pearson, 2012).

In the case of copyrighted work, citation is a legal obligation. As is stipulated in Art. 6bis of the Berne Convention (World Intellectual Property Organization 1979a)[18], every author shall have the right to claim authorship, "independently of the author's economic rights, and even after the transfer of the said rights" (Egloff, 2014).

[17] http://www.b-eye-network.com/view/15697
[18] http://www.wipo.int/wipolex/en/details.jsp?id=12214

Clarify citation of compound dataset

Title 17 of the US Code (U.S.C. 17) refers to compound data products as works formed by the collection and assembling of pre-existing materials. Surprisingly, this topic has received very little attention. There appear to be no generally useful standards for citations of compound products.

Whenever an individual within the research community uses parts of a dataset, it is not always clear how to cite this part of the dataset. This happens frequently, especially in international organization where data is sourced from many countries and different institutes.

Compound datasets are common in statistical systems, where for instance data from different countries are combined. These datasets often cannot be simply considered as ownership of the contributing parties as they are processed, filtered, or sometimes even estimated. In many cases, the producer of the compound dataset becomes the rights owner, but in many cases this is not agreed with the data contributors, and thus an unclear situation ensues. The increased availability and use of on-line datasets make it more important to understand the rights associated with compound datasets, if only to prevent unintended use. This brings a responsibility to the organizations that compile these datasets to inform their contributors, and agree with them on the terms of use of their data in a data policy. This is a complex and labour-intensive task.

There currently is no comprehensive technology that can describe compound datasets as a whole, or track the content and modifications to it.

Prevent long tail research data

While sophisticated research infrastructures assist scientists in managing massive volumes of data, the so-called long tail of research data frequently suffers from a lack of such services. This is mostly due to the complexity caused by the variety of data to be managed and a lack of easily standardiseable procedures in highly diverse research settings (Pröll et al., 2016). Long tail datasets are heterogeneous, small of size, unique standards, not-integrated and have an individual curation (Heidorn, 2008).

The term ‘data provenance’ refers to the process of tracing and recording the origins of data and its movement between databases. Scientific research is generally held to be of good provenance when it is documented in detail sufficient to allow reproducibility (Altintas et al., 2004). Various initiatives such as DataONE and the Research Data Alliance are working to tackle issues of provenance.

The long tail of data is evident in global statistical datasets where data pass sometimes 5-6 stages before they end up as an aggregate data-point in a global dataset. It is impossible to properly trace all contributions, and at a certain stage the tail has to be cut. The interviews learnt that there are several options to manage the long tail, but no single approach is currently applied. An infrastructure can provide essential services to organizations that are in need of managing (part of) the tail. This will require international standards to ensure the tail supports interoperability across organizations.

5 Conclusion

Organizations that wish to exploit cloud-based services, often find it difficult to obtain an overview of the options and opportunities. With a SDM, they can obtain a quick overview of the implications of a specific architecture, but also avoid overlap or even conflicts between different legally relevant documents. A quick scan of an emerging exploitation of EGI resources by a community through an intermediate showed the benefits of the approach.

This study mostly serves as fact finder, a brief study to develop a framework as a first step towards a SDM that has a more generic character with the aim to establish collaborations between the different layers. A next step for this proposed SDM would be to implement the recommendations and guidelines learned during the development of this study. The following major points were discovered.

The application of the proposed SDM on the FAO (chapter 3) use case made clear that there is a lack of concrete and clear regulations (such as consequences when one party does not meet the agreement), mention of options to distribute data to third parties, and standardized process to analyze whether standards for ownership of IPR are included in the metadata. These aspects can be included in a next step of the proposed SDM.

It appeared that parties involved in data management amongst others encounter more fundamental questions such as defining ‘data’, ‘data ownership’, and ‘data citation, attribution and credit’. This study has tried to find answers in chapter 4 on a selection of these questions, which can be implemented in a next version of the proposed SDM.

The FAO use case and the interviews revealed that the development of comprehensive data policies could be challenging for organizations working in data management. EGI is in the position to support these organizations with the development of a data policy by providing a clear and ready-to-implement SDM. The development of such a SDM will benefit the communities in formulating technical support in data management.

6 Bibliography

Altintas, I, C Berkley, E Jaeger, M Jones, B Ludascher, and S Mock (2004) Kepler: an extensible system for design and execution of scientific workflows. Proceedings of 16th International Conference on Scientific and Statistical Database Management, pages 423–424

Carroll, M.W. (2015) Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol 13(8): e1002235.

Heidorn, B. (2008) Shedding Light on the Dark Data in the Long Tail of Science

Jones, S. (2011) How to Develop a Data Management and Sharing Plan. DCC How-to Guides. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

Lammerant, H., Galetta, A., De Hert, P. (2014) Legal issues in big data in Anna Donovan, Rachel Finn & Kush Wadhwa (eds.), Report on legal, economic, social, ethical and political issues regarding big data. BYTE Deliverable D2.1, 35-58. BYTE Deliverable D2.1: Report on legal, economic, social, ethic ed, 2014.

Van Maanen, E. & Ellenbroek, A. (2016) Report on data sharing policies and legal framework in fishery and marine sciences data sector. EGI-Engage D2.6. https://documents.egi.eu/public/ShowDocument?docid=2699

Magron, F. (2015) Towards a data sharing policy for survey and monitoring data collected and/or stored by the SPC Coastal Fisheries Programme on behalf of member countries, SPC Coastal Fisheries Science and Management Section, Working Paper 14.

Pearson, M., Uhlir, P. (2012) "Developing Data Attribution and Citation Practices and Standards." Board on Research Data and Information; Policy and Global Affairs; National Research Council

Proell, Stefan, Kristof Meixner, and Andreas Rauber (2016) "Precise Data Identification Services for Long Tail Research Data." 13th International Conference on Digital Preservation (iPRES 2016). 2016.

W. Tan P. Buneman, S. Khanna (2000). Data provenance: Somebasic issues. In Proc. of FSTTCS, 2000.

7 Annex

During a visit to FAO - HQ in Rome from July 12th until July 14th E. van Maanen updated the interviews he took with four FAO Staff members in March/April 2016 on the legal aspects of data sharing. The fifth interviewee at FAO was not interviewed before by E. van Maanen.

FAO interviewees are represented by their department names, as information is not shared as a person.

Function	Abbreviation used
FAO Office of Corporate Communications	OCCI
FAO Legal Department	LEGN
FAO Statistics department (2 persons)	ESS
FAO Fisheries and Aquaculture information branch	FIAS
CNR-ISTI, Pisa	CNR

The main findings of the interviews are mentioned below.

- FAO recognizes the need for a comprehensive data policy;
- The development effort for a FAO data policy is substantial, and a multi-year effort;
- The interviews (March 2016) on the development of a FAO data policy were helpful in fostering a cross-departmental dialogue;
- At this stage, a representative of ESS is developing a data policy for FAOSTAT data which is considered as 'guinea pig' covering a substantial amount of FAO data-flows;
- The appointment of a Chief FAO Statistician is seen as an opportunity to place relevant data policies under his/her auspices, and thus achieve a corporate data policy;
- For FAO, this process is connected to, but not only driven by technology, and the technology options are considered along with legal and organizational aspects. This shift is better described as a process than as a project.

OCCI

OCCI starts with mentioning that since the pervious interview FAO's data policy has not further developed officially. OCCI sees a growing awareness about the need for a (electronic) data policy.

OCCI observes there is 'a recognition of the need for a data policy, but this is a long process and not yet brought into practice'. OCCI spends a substantial amount of time assessing whether and how data can be published. This is a tedious process and should be streamlined according to OCCI, and here a data policy is needed.

The two main challenges when formulating a data policy are (i) the origin of the data and (ii) the purpose of the data. In some cases it is not clear what the exact origin of the data is. Before deciding what to do with a copyright, you first have to know who owns the Intellectual Property Rights. Therefore clarity on the origin of data is required. Secondly, due to the public and non-profit character of FAO of the UN it is important to assess the purpose of the data requested. OCCI mentions an example where a party wanted to conduct a market analysis for investment in agro-alimentary industry. This party wanted to include FAO's commodity prices in its algorithm to assess where and when to buy food products. The aim of this party appeared not to be aligned with FAO's strategic objectives, as would merely financially exploit periods of food shortage.

Furthermore, OCCI mentions that OCCI has decided to publish all internal administrative data in compliance with the International Aid Transparency Initiative (IATI). Although this is merely one dataset, it can be seen as a good step towards publication of FAO data. OCCI also states that FAO is currently working on the publication of all SDG indicators. These experience could also be used when developing a broader FAO data policy.

OCCI states that more background is needed to assess whether the proposed Service Delivery Model (SDM) is clear enough for implementation.

LEGN

The interview with LEGN confirmed that little has changed in data policies, and that the effort to prepare this policy is ongoing. Since the previous interview of March 2016 concerning legal interoperability and data related policies FAO has taken steps to organize itself, e.g. through the appointment of Chief Statistician Officer (also mentioned by ESS) as a significant change towards a uniform data policy in FAO.

LEGN explains there is concern with ensuring the legal status of FAO as an international organization (not subject to national laws) in data sharing issues. The International Maritime Organization (IMO) was given as an example of an organization that is not fully ensuring its legal status as international organization (it also has a different mandate) and FAO is studying other organizations to prepare its data policy.

ESS-I

ESS confirms that there have not been any many changes in FAO's data policy over the year 2016/2017. FAO has become more active in data policy related initiatives, such as GODAN. ESS-I

does notice a growing interest in data sharing and open data since the interviewed of March 2016. ESS-states that the presentation given last year about legal interoperability evoked the interest of more people in this topic, but this has not resulted in any concrete changes yet.

ESS mentions 'a major internal change within ESS', i.e. the appointment of a Chief Statistician Officer. The aim of this appointment is to decentralize data policy within ESS, which could be beneficial towards developing a data policy. According to ESS, this data policy should include accountability and should be implemented for every step in the 'data value chain'.

The current developments around the data policy of FAOSTAT data can be used as 'guinea pig'. When ESS-I pioneers a data policy with FAOSTAT, experience can be used to develop a broader data policy. ESS-I mentions that there should be a balance between pioneering and collaboration. All relevant stakeholders should be included in the process of developing a data policy.

ESS states that he is not the right person to assess whether the proposed SDM is clear enough for implementation.

FIAS (Fisheries Statistics and Information Branch)

FIAS states that there is a lack of clarity of the origin of data, and thus the ownership of Intellectual Property Rights. An example of this is FIAS's work on VRMS where he sometimes experiences a lack of clarity of the provider of data. 'Sometimes a data provider delivers an observation, and FAO adds an image to this observation: who has the ownership of the Intellectual Property Rights of this combination?'

FIAS also states that there is a lack of clarity on (i) the applicability of FAO's terms of use, (ii) the selection of licensing options FAO is offering, and (iii) the legal approach of definitions of content/data/information.

FIAS is convinced that legal aspects (including a clear disclaimer) of data sharing should be included in the business metadata.

FIAS furthermore states that FAO's terms of use should be reconsidered to assess whether it is complete and clear enough.

ESS-II

ESS-II mainly works on topics such as metadata for text, linked data and the semantic web. ESS-II states there is a lack of clear data policy within FAO. ESS-II mentions that a colleague from ESS is currently developing a data policy for FAOSTAT statistical data. OCCI is aware of this development and might want to apply this data policy to other types of data according to ESS-II.

ESS-II needs more background to assess whether the proposed SDM is clear enough for implementation.