

## **RDA-CODATA**

### **INTEREST GROUP ON LEGAL INTEROPERABILITY OF RESEARCH DATA**

#### **Case study: iMarine**

Date: March 2014

#### **Background**

iMarine - [www.i-marine.eu](http://www.i-marine.eu) - is a project Co-funded by the European Commission under Framework Programme (FP7). The project was launched in November 2011 and will end in September 2014.

The project aims to establish and operate an e-infrastructure supporting the principles of the ecosystem approach to fisheries management and conservation of marine living resources. The infrastructure is expected to enhance science-based data processes in support to implementation of the Ecosystem Approach and, where possible, support open-data.

Partners currently part of the the iMarine Consortium are:

- Centro de Referencia em Informacao Ambiental (CRIA), Brasil
- Consiglio Nazionale delle Ricerche (CNR), Italy
- Engineering Ingegneria Informatica SpA (E-IIS), Italy
- European Organisation for Nuclear Research (CERN), Switzerland
- Fishbase Information & Research Group Inc. (FIN), Philippines
- Food and Agriculture Organisation of the United Nations (FAO), Italy
- Foundation for Research and Technology Hellas (FORTH), Greece
- GEIE ERCIM (ERCIM), France
- Institut de Recherche pour le Developpement (IRD), France
- National and Kapodistrian University of Athens (NKUA), Greece
- OBIS, United Nations Educational, Scientific and Cultural Organization – UNESCO (UNESCO), France
- Terradue s.r.l. (Terradue), Italy
- Trust-IT Services Ltd. (Trust-IT), United Kingdom

#### **The Ecosystem Approach Community of Practice (EA-CoP)**

The users of this infrastructure are divided in several groups; “Communities of Practice”. An example is the Ecosystems Approach to Fisheries (EAF) community; the EA-CoP of Regional Fishery Bodies, FAO staff, Universities, Research centres, etc. Rather than serving content, iMarine serves these Communities of Practice; with data and the tools to manage and modify data. It is not intended to build websites, but provides tools to develop quality content. The EA-CoP is thus equipped with tools to view and review, modify, merge, and validate content.

The tools enable the EA-CoP to disseminate “open data” but also to prepare, manage and control the data flow in a secure and confidential environment.

## The Data

Typically data are contributed by partners and cover statistical data, species occurrence data, species taxonomic data, geospatial products, and oceanographic data. Other content is also gathered from public data sources.

Data mostly come under a license or usage policy, but iMarine is not aiming for a formal technical support structure to either maintain or enforce these. We rely on policy support structures that are provided through the data formats themselves, e.g. for geospatial data, a policy can be attached through its metadata, and the same is true for statistical data. However, these are informative meta-data, and not systematically managed during computational processes (e.g. merging datasets).

All data remain in the infrastructure unless a user (a human) moves them to a publicly accessible area (i.e. share by email, a type of dropbox environment, a shared folder, etc.). Individual users' activities are not tracked nor are any constraints set on their rights. So far the main management feature on data is that the person that entered the data is the same that decides where the data will be stored and who can access there.

## Type of Outputs

- Statistical computations (e.g. statistical reallocations)
- Maps
- Reports, Web pages [e.g. fact sheets]

## The policy challenge

iMarine offers an e-infrastructure for viewing, modifying, generating, merging, and validating content. The content can be stored in the infrastructure or external repositories. Policies are needed to marshal those steps, and attach to the final product a meta-information on provenance, use-rights, ownership, and validity. These policies set the principles and boundaries which govern the flows of data for the EA-CoP, define data security and confidentiality aspects, describe the collaborative work through data sharing, and manage publishing.

Particular policies apply to:

- data access and persistence
  - guarantee the integrity of all data populating iMarine
  - facilitate the access and the utilization of data in cost effective ways
- data sharing
  - manage private and confidential data
  - share data among users according to negotiated sharing principles
  - properly inform users on data provenance and conditions of use, including:
    - propagate copyright and license settings
    - manage the document metadata
    - provide documentation of sources

## Status of the iMarine policy development

The iMarine policy is captured in three documents: Terms of Use, Privacy Policy and Best Practices.

The Terms of Use are in an advanced draft status and the document has been revised by different actors including legal experts. At the end of the project it is expected to produce a further version consolidating several feedback collected in these months including whatever may come out by the compilation of this case study.

The privacy policy depends on information that has to be discussed with potential clients. The project so far has focused on developing the technology framework in which these policies can be applied.

The best practices document is in working progress and will be further formulated around individual workflows to ensure that the intended service adheres to the clients requirements.

### **1) The legal frameworks and specific policies (or lack of them) governing different types of research data important to a specific scientific domain or problem area important to each case study.**

The iMarine project is developing tools and services that promise to manage data workflows. As such, it does not adhere to or is subject to a specific legal framework. The data managed through it are owned by the data providers who can define a policy on data access. One such example is the INSPIRE directive, where iMarine aims to offer tools and services that enable data providers to comply with parts of this directive.

The domain of the project; the Ecosystem Approach to Fisheries and the Conservation of Marine Living Resources, has several policy needs related to data management. Here, a need is felt for more Open Access policies surrounding data held in legacy systems. The need for a legal framework is perceived where users wish to access and use data that they can not freely distribute. Such framework would define the a re-use policy defining the conditions for use of sensitive de-aggregated data in space and time, e.g. through a non-disclosure agreement specifying the acceptable aggregation level of distributed products.

Such policies that users of the infrastructure are expected to adhere to refer to the use of specific data and components made available through this infrastructure. These infrastructure driven policies should also refer to the data access, acceptable formats, descriptive metadata, disclosure and embargo periods, copyrights and licenses.

## **2) The perceived barriers to data sharing or interoperability and perceived needs for increased interoperability that spurred interest and investment in new legal interoperability approaches.**

### **Legal Barriers**

iMarine stakeholders that act as data providers in the infrastructures address the Legal Interoperability problem in their own context; they use iMarine tools to manage access and sharing of their own data to a restricted group. By retaining all control of their data, they avoid interoperability issues. The main reason is that they are faced with limitations on the scope with whom they can share data, e.g. because their institute's policy imposes access limitations. Changing the organization's mandate to adapt a more open data policy is often considered too expensive.

### **Formatting barriers**

iMarine specifically aims to unlock currently inaccessible datasets by offering harmonization features. These features allow organizations to convert their data from current restricted formats to internationally accepted standards. However, there are no accepted formats in all domains that are precise enough in meeting all needs.

### **Financial barriers**

The costs of establishing interoperable data environments are often high for individual organizations that in turn will see little benefit of their contribution. At the same time, the costs of data acquisition and aggregation are often geared towards an existing workflow, and adopting an interoperable data approach often necessitates replacing considerable amounts of these workflows with serious implications on staff and training, tools and appliances, responsibilities and management. With iMarine, several successful examples of cost-effective transformations can be provided, yet pooling of resources requires further co-ordination efforts to lower the financial barriers.

### **Tooling barriers**

The use of data in interoperable environments supposes that they are of comparable quality and resolution. However, also calibration of instruments and methodology has an impact on the re-usability of data and the reproducibility of experiments. With iMarine this issue can not yet be addressed, as it requires that formatting issues are resolved.

### **3) A description of any effective legal interoperability processes, techniques and institutions that have been developed or adopted to overcome the barriers that have been identified.**

**Creative Commons** (CC) is the current adopted license scheme. The iMarine data access and sharing policies are compiled upon that logic and the formulation is tested with case studies to verify that CC licenses are successfully covering all needs. Sharing data and release of derivative products are among the most challenging aspects. A derivative product inherits copyright obligations from its sources and a cascading mechanism is applied to any new product. From the original terms of use, when combining different data sources, the more restricted source of information drives the overall possibility of public dissemination, e.g. if only one of the utilized sources is not for commercial use, the derivative product can not be for commercial as well. CC licenses are in this view adopted as an additional layer on top of the original terms of use trying to simplify the norms with few and clear options.

With this strategy it is crucial that institutions are open to accept and adhere to CC licenses. However, shifting to customized terms of use to more generic and standard protections like Creative Commons may be not a simple process for many institutions. A possible way could be to map specific products from the original terms of use to one of the CC options so as to facilitate the re-utilization without imposing the adoption of the CC licenses for the entire institution or organization.

### **4) The stakeholders involved in developing, testing, and implementing legal interoperability approaches and their roles, level of engagement and investment, and impact.**

The [iMarine Board](#) defines the data e-Infrastructure governance model including legal interoperability aspects.

This section summarizes how the FAO Fisheries and Aquaculture Department and other institutions part of the iMarine consortium are dealing with products produced utilizing the iMarine e-infrastructure.

The stakeholders are grouped in various 'identities'; Developers, Owners, and Users. For each of these, different approaches apply in the establishment of an interoperability framework or services using such framework.

For **developers**, the keyword is security. In iMarine, a lot of responsibility is given to

developers, who have developed a technology framework in support of interoperable systems. This includes advanced authentication and security issues. However, the technology per se does not constitute a legal framework, as this is the role of the owners.

For **owners**, the keyword is trust. This usually implies they rely on the iMarine infrastructure to provide back-end management of their front-end data. This trust is built by reassuring the stakeholder with documentation around which a reliable exploitation scenario is built. This exploitation level framework is not legally binding yet.

For **users**, the keyword is reliability. Once a work-flow is delivered, it is validated by a group of users to verify if their expectations surrounding this workflow have been met. Users validate a service, which is not the same as testing or proofing. Their validation will give the trust needed by the owners.

The above stakeholder driven approach to develop interoperability is thus 'develop as you go', which proved to be effective, even if for users not accustomed to Agile development it may be difficult to grasp at the beginning.

Examples of stakeholders are best given for this use-case at the level of owners, and these include the FAO Fisheries Department for its activity on the Vulnerable Marine Ecosystems database (VME-DB), the CGIAR related FIN for its AquaMaps, etc.

##### **5) Progress to date in implementing legal interoperability approaches, including identification of criteria or metrics used to assess success or impact, use of technology or other mechanisms to promote adoption, and estimates of funding and other resources provided to support implementation.**

iMarine by itself does not aim to implement a legal framework yet aspires to support challenging interoperable data-flows, and these include legal aspects. These aspects are addressed at data level, technology infrastructure and user-community that collaborate to define, implement and validate these aspects. This has resulted in significant progress:

All **content** shared or publicly exposed through iMarine is foreseen to be described by Business Metadata. This meta information can be also considered as a means to facilitate interoperability. Business metadata are supposed to cover the following: ownership and context, authorship, copyright licenses, content description. Any user accessing content will have the possibility to have a glance view on the related business metadata, including the type of license attached to that particular content. This information is foreseen to be machine readable as well (e.g. CC REL metadata). An automatic citation mechanism is also foreseen based on business metadata.

At **technology** level, interoperable technologies have been implemented for geospatial, statistical and biodiversity data, greatly reducing the costs to participants to expose their

data under international de-facto standards. Most technological problems have been solved in iMarine, yet conversion of data to these interoperable formats remains challenging in costs and quality.

iMarine has made good progress in developing and maintaining a **community** around interoperable data frameworks, and without an active community identifying needs and validating results proved to be nigh impossible. Developing a legal framework through an intense collaboration between technology providers and data owners based on proven iMarine technologies is currently ongoing through the definition of an exploitation model.