Interoperability Standards in LifeWatch

Nicola Fiore[1], Daniel Fuentes[2], Alberto Basset[3], Juan Miguel González-Aranda[4]

LifeWatch in a nutshell

LifeWatch[5] is a collaborative project, now entering its construction phase, to involve researchers and engineers from across the European Union. It includes a range of new services and tools to help the researchers communicate, share data, analyze results, create models, manage projects and organize training.

Biodiversity and ecosystem research did already benefit at an early stage from the new opportunities provided by information and communication technology. Especially the possibility to build relational databases proved to be an important tool to capture biodiversity data and to organize information differently. The LifeWatch ICT Infrastructure (or infrastructure for short) shall be a distributed system of nodes that provide access to and processing of biodiversity data from a variety of sources through common open interfaces for several decades.

Data requirements

LifeWatch has to manage data from almost all areas of biodiversity. Many data sources will be integrated in a non-distributed way. These sources have different structure, design, data and the management tools which work them are different. Clearly, various systems are nested in
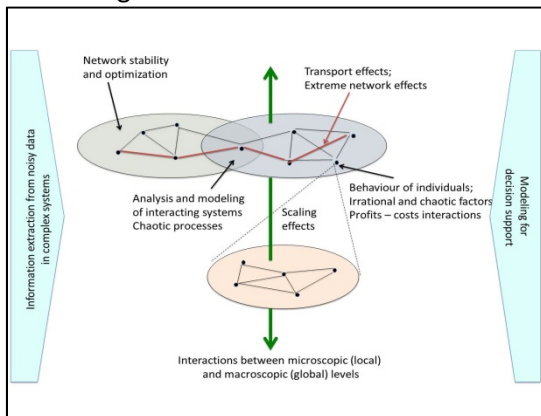

Figure 1: Scheme to cover multidimensional biodiversity challenges. (Source: Wouter Los.)

higher-level systems, and often with not interpretable linear relations, partly since relations should be described over extreme spatial (nanometers to hundreds of kilometers) and temporal (milliseconds to hundreds of million years) scales. In addition abiotic systems are acting on the biodiversity systems with equally chaotic processes. Figure 1 holds a scheme with an abstract representation of the strategies in biodiversity research to deal with these multidimensional challenges.

The data relevant for LifeWatch can be classified into three main groups. First, the **primary data**, which is the information taken in an automatic way (i.e. sensors) or manually (i.e. scientist). It can be divided into the biotic (species, genetics, ecological information) and abiotic data (soils, ocean circulation, weather, climate, humans). The second is the **processed data**, the information extracted from files that remains in a not digitalized and structured format (articles, books, websites and so on). And the third is the software, all the **code** from applications to manage the data from the two previous points.

State of the art

The LifeWatch interoperability is based on standards. Standards are necessary to manage and share data coming from different providers. In the state of art there are a set of proposed

[1] Università del Salento (Italy), email: nicola.fiore@unisalento.it
[2] Estación Biológica de Doñana (CSIC, Spain), email: dfuentes@ebd.csic.es
[3] Università del Salento (Italy), email: alberto.basset@unisalento.it
[4] Ministerio de Economía y Competitividad (Spain). Email: juanmiguel.gonzalez@mineco.es
[5] http://www.lifewatch.eu

methodologies and solutions to describe data and to make them interoperable, each one of these comes from a different biodiversity area. **Darwin Core**[6] includes a glossary of terms (properties, elements, fields, columns, attributes, or concepts). It describes data properties of biodiversity resources, such as information about the location and identification of the occurrence of an organism. The number of concepts that Darwin Core can describe could be a strength for the usability of the standard but a weakness for the description of the observation. **ABCD**[7] (Access to Biological Collections Data) schema is another highly-structured standard for data exchange and access model for taxon occurrence. The main advantage of ABCD is the availability of a big set of concepts related to a specie observation but, sometimes, it is very difficult to use due to the ambiguity of the concepts. **INSPIRE**[8], the Infrastructure for Spatial Information in Europe (INSPIRE) directive, aims to create a European Union (EU) spatial data infrastructure. It is based on the spatial ISO standards from the ISO191XX series (so based on GML). However, the data specifications are not yet finalized, the biodiversity occurrence data does not directly fall into a specific INSPIRE theme and there are not developed tools yet to map the data into INSPIRE.

Standard	Lastest version	Biodiversity Areas	Number of Concepts	Extensions	Related organizations	Related tools
ABCD	2.06 (2005)	Preserved, living and data collections, DNA samples	1200 aprox.	Geosciences (EFG) DNA(ABCDDNA), Herbarium (HISPID)	TDWG, Universität Berlin	Biocase
Darwin Core	1.4 (2009)	Spatiotemporal occurrences of biological specimens	90 aprox.	Curatorial, Geospatial, Paleontology, Interaction	TDWG, GBIF	GBIF tools (IPT,)Tapir, DiGIR*

*Digir and Tapir are not yet supported.

LifeWatch approach for interoperability

For the LifeWatch strategy, two parallel ways to go have been considered. For short-term purposes, it may be convenient to use the standard ABCD to describe the data and design some extension/s for the uncovered concepts. Other standards, like Darwin Core, have a limited number of fields and can be used for a discovered service to ensure the interoperability with other initiatives (i.e. GBIF data portal). Tools designed by BioCASE such as the Provider Software tool are appropriated to map the data to the standard ABCD.

However, for the longer term, it is necessary to adopt a very different approach based on search and mapping with the terms of the user. This can be from controlled vocabularies, but also with some intelligence that takes into account synonyms of terms. This should finally end in (semi) automatic data discovery and data integration. For biodiversity data this is the only way to proceed, as we will see increasingly machine-machine interactions.

Developing a new standard is a not an immediate thing. It is needed a scientific community to approve each element that has to be incorporated to the standard and each concept has to be well defined avoiding any ambiguity. If the standard changes, the applications that already use it has to be interoperable with the new ones. Adding and removing elements of a new standard can give flexibility to the research community but that implies many technical problems to ensure the interoperability between applications.

Acknowledgements

---

[6] http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome
[7] www.tdwg.org/activities/abcd
[8] http://inspire.jrc.ec.europa.eu