Flock Together with

CReATIVE-B

A roadmap of global research data infrastructures supporting biodiversity and ecosystem science

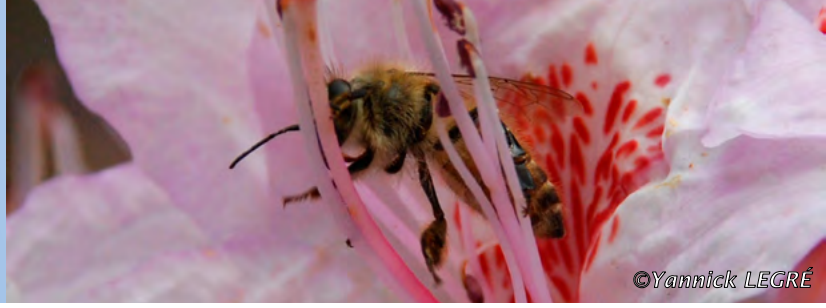
Produced by the project CReATIVE-B

Coordination of Research e-Infrastructures Activities Toward an

International Virtual Environment for Biodiversity

Funded by the European Commission with Grant agreement n°: 284441

1. Introduction



The CReATIVE-B project started in 2011 to consider the "Coordination of Research e-Infrastructures Activities Toward an International Virtual Environment for Biodiversity". CReaTIVE-B supported collaboration between the European LifeWatch ESFRI Research Infrastructure with other large-scale Research Infrastructures (RIs) on biodiversity and ecosystems research across the globe. The immediate objective was to define a roadmap for interoperability on 3 levels:

1. Community Engagement, related to inclusion and serving the demands of the scientific community;

2. Technology, related to data, ICT, e-science services;

3. Legal and Governance, related to property and access rights to data, global policy coordination.

The project aimed to be a catalyst for worldwide collaboration by supporting and initiating coordination activities among these research infrastructures. By doing so, the project explored how the collaboration could best support the ambitions of the Group on Earth Observations Biodiversity Observation Network (GEO BON), one of the societal benefit areas of Global Earth Observation System of Systems (GEOSS). CReATIVE-B also contributed to the priorities as set by the G8 in the 'Carta di Siracusa' in supporting cooperation to further global monitoring of biodiversity, achieving reliable, comparable and interoperable data, developing global approaches to exchange scientific knowledge, best practice, technologies and innovation, fostering comprehensive and focused research and capacity building at all levels on biodiversity and ecosystem services and global environmental assessment[1].

CReATIVE-B organised a number of international workshops to discuss the three levels of interoperability. Several analyses served as input for the conclusions and recommendations in this roadmap document. In addition, the CReATIVE-B project supported 'Global Biodiversity Informatics Conference (GBIC)' as organized by GBIF in 2012. The GBIC conference produced the 'Global Biodiversity Information Outlook' that provided key input for discussion in the CReATIVE-B project.

The organisations composing the partnership of the CReATIVE-B project were the University of Amsterdam; Cardiff University; Gnùbila France; Consiglio Nazionale delle Ricerche, Italy; Universidad de Alcalà de Henares, Franklin Institute; Comunità Ambiente; and the Centre National de la Recherche Scientifique, Institut des Grilles, France.

Besides this European partnership originating from the LifeWatch preparatory project, also Research Infrastructures in other parts of the world and/or with a global orientation were invited to attend the project workshops as "Liaison partners". These are the Atlas of Living Australia, DATA-One (USA), NEON (USA), CRIA (Brazil), SANBI (South Africa), Chinese Academy of Sciences, GBIF (global), World Federation of Culture Collections (WFCC) and GEOBON. This document refers to these infrastructures, together with LifeWatch, as 'cooperating research infrastructures'.

The European Commission supported the project in the Seventh Framework Programme for Research and Technological Development under project number 284441.

[1] https://www.cbd.int/doc/g8/g8-2009-04-23-chair-summary-en.pdf

Table of contents

2. Summary

The Earth's living environment is crucial for buffering extreme hazards of solar radiation, changes in the atmosphere gases, temperature fluxes or fresh water quality. Our planet is the laboratory for biodiversity and ecosystem sciences. The grand challenge for biodiversity and ecosystem scientists is unravelling the complex patterns and processes of life by analysing the large and diverse data sets representing scales of biological organisation (genes, species, populations, ecosystems) at different temporal and spatial scales. Biodiversity research infrastructures are providing the integrated data sets and support for studying scenarios of biodiversity and ecosystem dynamics.

The CReATIVE-B project - Coordination of Research e-Infrastructures Activities Toward an International Virtual Environment for Biodiversity – explored how cooperation and interoperability of large-scale Research Infrastructures across the globe could support the challenges of biodiversity and ecosystem research. A key outcome of the project is that the research infrastructures agreed to continue cooperation after the end of the project to advance scientific progress in understanding and predicting the complexity of natural systems. By working together in implementing the recommendations in this Roadmap, the data and capabilities of the cooperating research infrastructures are better served to address the grand challenges for biodiversity and ecosystem scientists.

Recommendations are directed at promoting users involvement and value delivery by focussing on supporting common and global research goals, joint development of cutting-edge technologies, and involving citizen scientists in research activities with environmental observation and monitoring. While the research infrastructures have a satisfactory level of potential interoperability, there are barriers to global interoperability. Recommended actions are to promote the understanding of the value of interoperable research infrastructures, to develop coordination mechanisms for achieving interoperability with increasing the importance of standards. The challenge is to create a scientific market place allowing users to benefit from workflows of services as served by the cooperating research infrastructures. Sharing data and tools in such workflows with varying provenance of authorship and ownership requires careful and efficient arrangements so that their users can benefit from the combined resources without tedious legal constraints. This even more important with the increasing automatic processing of data supported by "machine-machine" interactions.

The cooperating research infrastructures agreed that each one will explore new funding opportunities to bring the recommendations into effect.

3. Understanding and managing our living environment: Data infrastructures for biodiversity and ecosystem research

3.1. Understand our living environment

The biosphere, the living part of our planet, has shaped to a large extent the stable environment in which we live. The Earth temperature, atmospheric gas composition or freshwater-quality are buffered by the biosphere. The interaction of biological species, with their genetic adaptability, is crucial for the capacity to buffer extreme pressures on Earth. Understanding these processes requires designing a scientific framework for research in all interactions of the biosphere in the Earth System. The grand challenge for biodiversity and ecosystem scientists is to study these system interactions. Increasingly, these complex patterns and processes are studied by analysing big and diverse data sets.

Not a single scientist, project or institute can afford to build and maintain the infrastructure facilities required to support such large-scale research on the biosphere. Large-scale research infrastructures have to provide the facilities and a number of these infrastructures is already serving data and software to scientists across the globe. Ass such, they are also promoting scientific work in support of environmental policies and evidence-based management strategies.

3.2. The role of Research Infrastructures

Research infrastructures are accelerating scientific discovery and understanding. The data infrastructures cooperating in the Creative-B project are supporting frontier research to understand the biosphere and assist in decision support in managing our environment. They provide access to data on baseline observations and provide the models and software tools to run computed 'experiments' to run forecasts into the future.

Such indicators are computed on the basis of a variety of data sets and parameters that together compose a model of reality. Producing indicator maps for different spatial (variation) and temporal (trends) scales requires considerable computational power. Single scientists, research groups or institutes are hampered to enter research on meaningful indicators since it is too difficult to produce or discover the required data, to build and test the significance of alternative models, and to have access to sufficient computational capacity. Research infrastructures are providing such supporting services so that scientists can focus on frontier research with benefits to society. Global cooperation is important to benefit from economies of scale.

Example of biodiversity/ecosystem indicators

A better understanding of the biosphere may lead to developing explanatory indicators of environmental change that for example may assist in predicting the effects of environmental management strategies that are being considered for implementation. Below are a few indicators related to crucial ecosystem services.

Biodiversity/Ecosystem indicator	Example related ecosystem service
Genetic variability	*Genetic pool for food resources or new medicines*
Species richness	*Ecosystem stability; materials for use (timber, biofuels, food)*
Ecosystem functions	*Carbon sequestration; fresh water quality; reducing desertification*

3.3. Defining the research infrastructures

Several categories of research infrastructures are in place or in development to support scientific development:

- Physical sites and transects all over Europe (and beyond) for the systematic production of data.

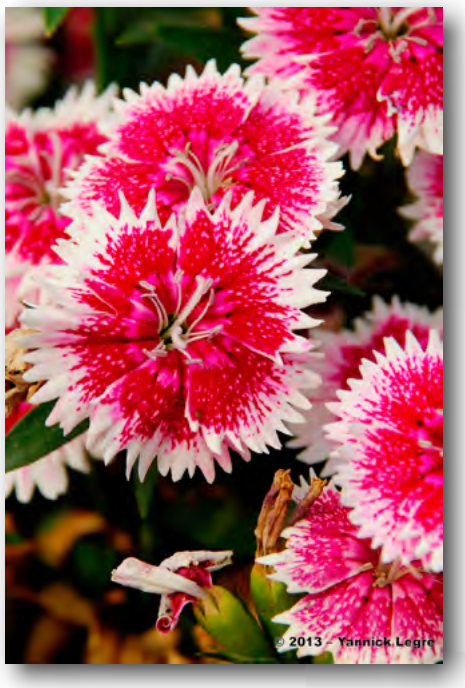
- Instrumentation and other equipment for producing data at site with sensors, images, or DNA sequences and remote data acquisition through airborne sensors and earth observation satellites. Human made observations are required when machine interpretation is not yet possible.

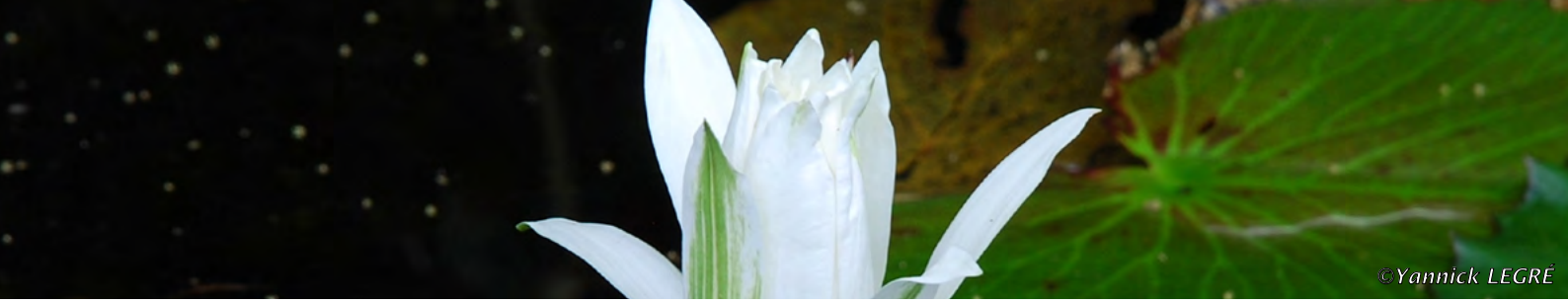
- Digital environments (e-infrastructure) support data storage and preservation, data filtering, data management, and provide services for data analysis and modelling. 'Virtual' laboratories are supporting integrated access to these services using appropriate computational power.

The last category of digital environments is the focus of this Roadmap. These are e-infrastructures or cyber-infrastructures operating in the world-wide-web allowing remote access to their facilities, data and services. Such infrastructures are offering the integrated facilities to enter frontier systems research.

3.4. The landscape of virtual research infrastructures for biodiversity and ecosystem research

A number of research infrastructures with (data) facilities across the world worked together to consider improved services to their scientific user community, as well as interested environmental managers and related policy domains. The cooperation focussed on infrastructure interoperability so that users can benefit from the combined infrastructure facilities through the web portal of each research infrastructure. Interestingly, the cooperating research infrastructures provide already complementary services, which allows each of them to focus on their own strengths, whilst benefiting from the capabilities of the other infrastructures.





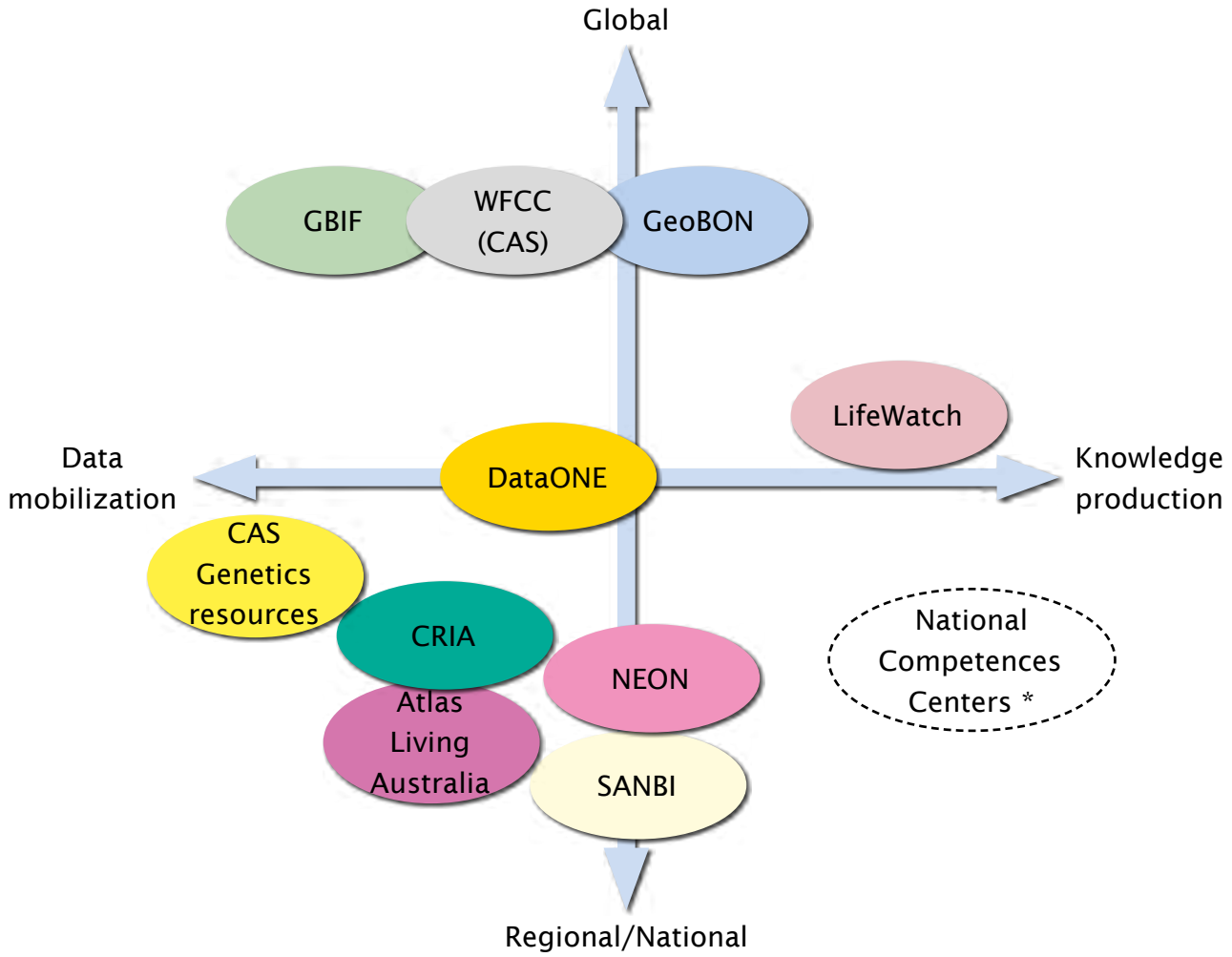


Figure 1: Characterisation of the cooperating biodiversity research infrastructures

The axes in this figure are representing different dimensions to characterize the research infrastructures in general terms. The horizontal axis refers to their operational missions. At the left side are several infrastructures with a strong mission to mobilize data by offering data storage, data sharing and data access services. At the right side are a few infrastructures supporting the use of data for analysis and modelling. The data mobilizing infrastructures are increasingly also offering such data processing services. The vertical axis shows at the bottom side the infrastructures that are mainly operational at the regional or national scale, and at the top other ones with exclusive global services.

* Note that there are currently not much facilities focussing on data processing for knowledge production of regional and national interest. However, some developments are indicating the establishment such facilities as national competence centres.

4. Priorities for the next decade

4.1. Research Infrastructures as new opportunity for understanding our environment

The sustainability of research infrastructures is based on current and/or anticipated demand, which in turn hinges on the active involvement of the scientific community in building the infrastructures. Although the cooperating infrastructures are among the few facilities with the capacity to provide the variety, quantity and quality of research data, this engagement requires an intensive awareness raising effort to convince biodiversity scientists of the benefits of a global virtual facility for data and service space. A common gateway of the cooperating research infrastructures will increase the coordination of worldwide scientific communities in defining and reaching research goals, increasing knowledge and acquiring cutting-edge technology. This is not necessarily a single gateway, but rather an approach whereby through any of the infrastructures also the capabilities of other ones are accessible.

Especially biodiversity & ecosystem research infrastructures have to play a role as a broker between citizens, scientists and other users for the production and use of data, tools and services. In this respect, citizen science is increasingly important for contributing to environmental observation and monitoring, and for contributing to research activities. It is recommended to empower citizen scientists so they can better engage with the supporting facilities of research infrastructures. In this respect the cooperating research infrastructures have to give special attention to remove barriers that prevent user communities from easy access to and use of the infrastructure capabilities.

4.2. Common requirements and selected priorities

The Creative-B project cooperated with GBIF in organising and supporting the Global Biodiversity Informatics Conference (Copenhagen, July 2012). The main outcome of the conference was the publication of the "Global Biodiversity Informatics Outlook" (GBIO) report, a blueprint on biodiversity informatics with short and long-term priorities[2].

The cooperating research infrastructures in the Creative-B project elaborated on this vision through a survey of their key areas of interest, requirements and barriers with their scientific communities. Some of the main requirements are:

- Priority for data discovery and data access technologies across research infrastructures (further details in section 5);
- Cooperation of research infrastructures to provide services with ecological data and with analytical tools to support research, management and conservation;
- Promotion and facilitation of involvement of scientists and decision makers in framing the appropriate research questions, and the provision of decision-support tools;
- Effective governance arrangements facilitating collaboration among research infrastructures with attention to support and feedback from their scientific communities (further details in section 6);
- Research infrastructures should act as a broker between citizens, scientists and

[2] http://www.biodiversityinformatics.org

other users of the data. Citizen science is important for both environmental observations, monitoring and general research support;

- Collaboration between research infrastructures and the private sector is important for developing new tools for the infrastructures and private initiatives.

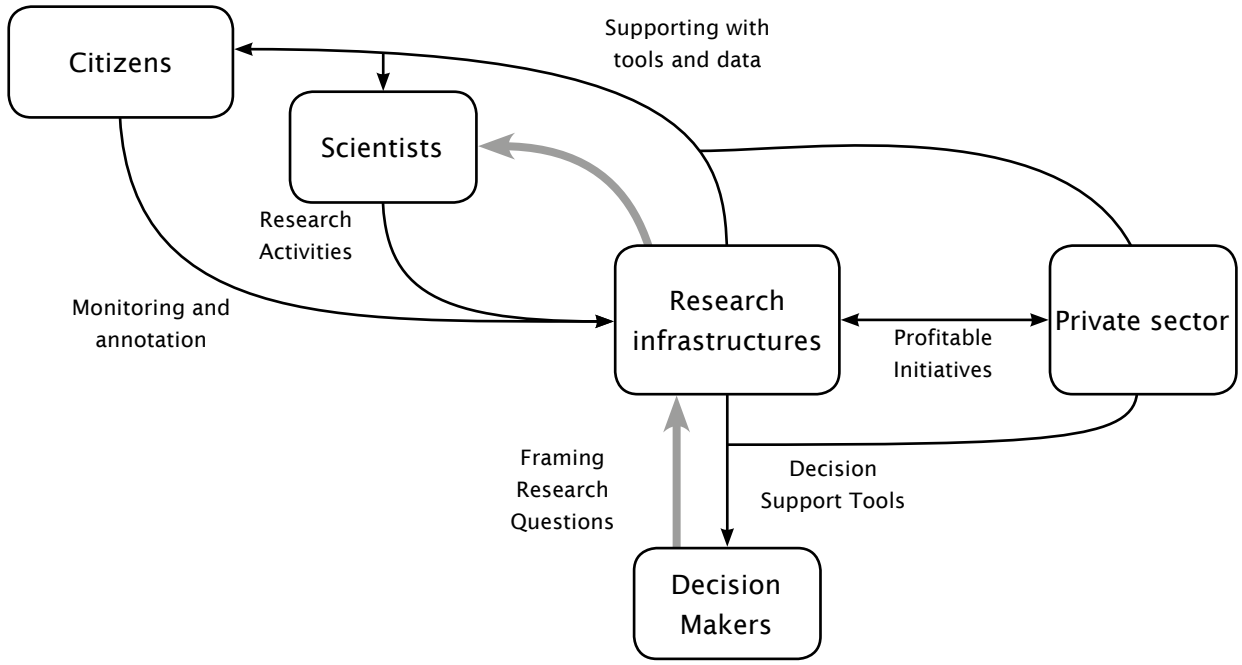
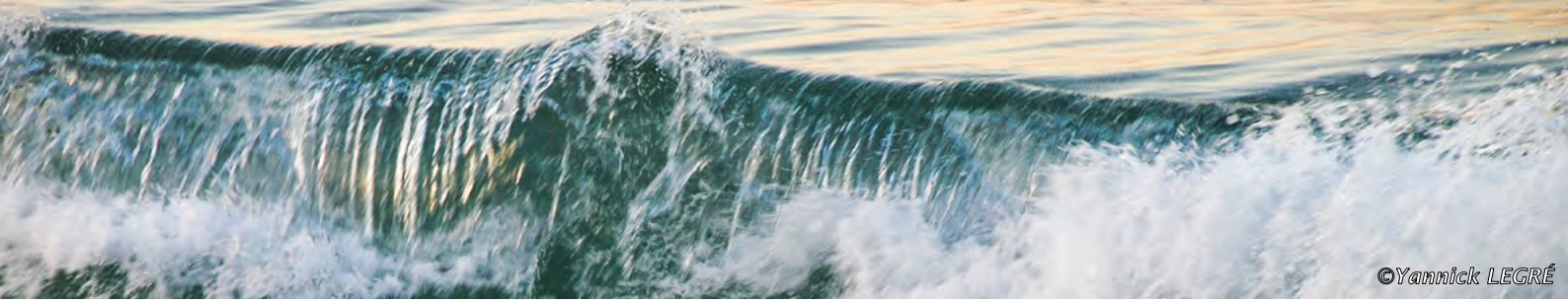


Figure 2: The role of research infrastructures in facilitating collaboration

4.3. Tackling the priorities; opportunities for collaboration

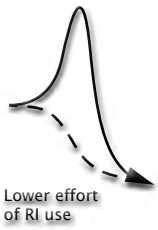
The cooperating research infrastructures are operating with different funding levels, visions, goals, user communities and development strategies. While appreciating the differences, this Roadmap is aiming at transforming some of these into opportunities to reduce duplication and to enhance collaboration for the development and sharing of new data and tools. This includes attention for best practices and common priorities on data quality, integration of data sets, and the involvement of user communities.

A number of actions was identified. For enhancing user involvement, effective strategies have to improve the communication on how biodiversity data and models are relevant for policies on grand challenges such as biodiversity loss, climate change, but also job creation. Some research infrastructures already developed best practices that, if shared, could help raising other awareness activities. The Atlas of Living Australia established a successful strategy with a portal allowing their data providers to see how their data are used; SANBI has experience with linking data to policy related issues (strategy plans, yearly plans, biodiversity serving other policies), showing evidence that biodiversity data are relevant for policy; DataOne User Groups are fostering a worldwide community of Earth observation data authors, and users, assisting in the identification of technical challenges and opportunities in education, research, and policy.

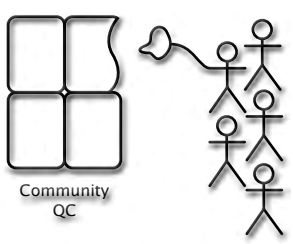


4.4. Gaps and risks concerning the present status and the future of RIs

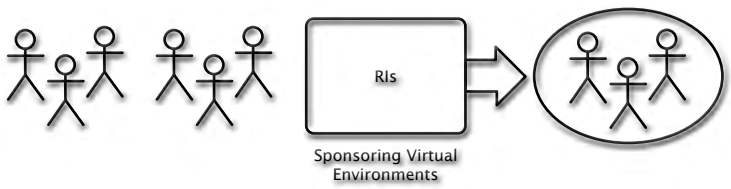
The cooperating biodiversity research infrastructures are facing a number of challenges.

a. 

Reduce barriers for scientific users to benefit from the advantages of 'virtual' infrastructures. User-friendly virtual environments have to simplify access processes, and so reducing the investment of time by researchers to recognize. Training activities and materials are required, but not sufficient as only solution.

b. 

Engage the scientific community in data validation to enhance data quality. Such feedback is an essential addition to automated validation mechanisms in all research infrastructures. The biodiversity research community needs to be motivated and empowered to do its work in an online collaborative way. No such environment currently exists. "(Belbin, et al. 2013)[3]. Such a validation environment is suggested as part of the fundamental backbone of biodiversity infrastructure in 'provision 20' of the Decadal view of biodiversity informatics.

c. 

Create stronger networks of biodiversity and ecosystem researchers by constructing virtual laboratories allowing large research groups to cooperate remotely on grand challenges. This should end up in common tools as part of large-scale services for structured communities.

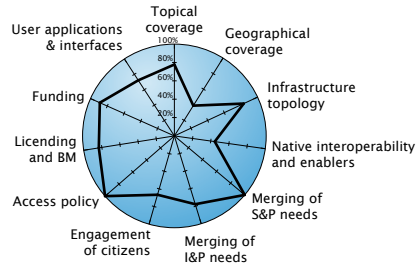
[3] Belbin L, Daly J, Hirsch T, Hobern D, Salle J La. A specialist's audit of aggregated occurrence records: An «aggregator»s' perspective. Zookeys. 2013;305(305):67-76. Available at: http://www.pensoft.net/journals/zookeys/article/5438/abstract/a-specialist's-audit-of-aggregated-occurrence-records-an-»aggregator»s'-perspective

5. Requirements for infrastructure interoperability

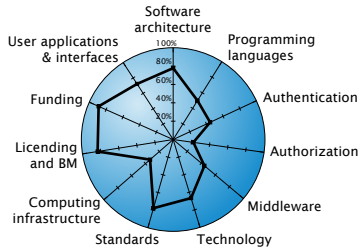
5.1. Potential for achieving interoperability

The diagram with the general overview shows that all the cooperating research infrastructures (RI) exhibit a satisfactory level of potential interoperability, in particular in the way they offer access to biodiversity data, available applications and related resources. Each RI pursues similar objectives in terms of business models, industry and policy involvement and overall sustainability plans. These objectives facilitate achievement of a future international virtual environment (IVE) for biodiversity and ecosystems research and the accompanying governance.

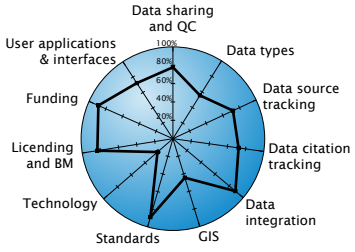
General overview



Service logic



Data



Participating research infrastructures have complementary geographical and topical coverage, while differing in their implementations. The foundations of the infrastructures are the physical topologies of their networks and resources.

As is to be expected, differences in implementation become more obvious in the second diagram “service logic”. Despite similar approaches to software architectures and standards adopted, the service logic in the research infrastructures is the place where most differences can be found. Proprietary middleware’s have been deployed with different security infrastructures, programming languages and technologies - the area where most work is needed to make systems syntactically and semantically compatible.

However, a long-term goal of service orientation is not fundamentally compromised. The third “Data” diagram suggests that some domain-specific standards (e.g., Darwin Core, TAPIR, Ecological Metadata Language (EML) are emerging and that begin addressing the needs of data integration and organisation. Some similar sharing and quality control processes are in place for initiatives dealing with data collection, and traceability is a shared concern for scientific citations and raw data tracking.

5.2. Overcoming barriers to interoperability

Overcoming the barriers to global interoperability in RIs means: (i) promoting understanding of the value of interoperable research infrastructures; (ii) using coordination to achieve interoperability; (iii) emphasising and increasing the importance of standards; and (iv) solving specific technical challenges. Key recommendations in this respect are the following ones.

(i) The value of interoperable RIs: Concrete benefits of interoperability must be visible and promoted to stakeholders in order to encourage and achieve interoperability. Use-cases are important to illustrate these benefits. One significant use-case, presently the focus of wide discussion, is that of "Essential Biodiversity Variables" (EBVs) as introduced by GEOSS-GEO BON. Potentially, EBVs or similar indicators are a core future business for RIs and the converged IVE.

Recommendation:

Illustrate the benefits of interoperability with Essential Biodiversity Variables (EBVs).

Conceived by GEO BON collaborators[4], EBVs are endorsed by the Convention on Biological Diversity (CBD) and in line with 2020 Aichi Targets. They provide a focus for GEO BON and related monitoring activities and have a role to play in biodiversity assessments (e.g. IPBES – the Intergovernmental Platform on Biodiversity & Ecosystem Services) and prediction of the future state of the biosphere.

As a general principle, it should be possible to calculate the value of any chosen EBV:

- For any geographic area, small or large, fine-grained or coarse;
- At a temporal scale determined by need and/or the frequency of available observations;
- At a point in time in the past, present day or in the future;
- As appropriate, for any species, assemblage, ecosystem, biome, etc.
- Using data for that area / topic that may be held by any and across multiple RIs;
- Using a standardised, widely accepted protocols (workflow) capable of executing in any RI;
- By any (appropriate) person anywhere.

(ii) Coordination to achieve interoperability: As the Global Biodiversity Informatics Outlook GBIO2 makes clear, there is considerable complexity to construct an interoperable RI with its interconnected components. There are multiple activities essential to the successful delivery of integrated e-infrastructures for managing and using biodiversity data in support of science and policy. Many of these are already underway but continuous support and increased technical capacity over time are essential. It is necessary to have global coordination and mutual understanding to ensure that the benefits are realised at the lowest possible cost and within a reasonable timeframe. Alongside coordination, investment in training for skills development is also critical.

Recommendation(s):

1. Coordinating interoperability around a consensually agreed technical roadmap of joint and individual actions to be carried out by concerned RIs.
2. Capacity building. Structuring and supporting education and training that encourage intero-

[4] Pereira et al. (2013), Science. **339** (6117): 277-278

perability. Such activities should be organized around/in a specialized biodiversity "market place", e.g. http://www.biodiversityinformatics.org/culture/, together with access to RIs' resources, thus facilitating adoption and harmonizing best practices.

(iii) Emphasising the importance of standards: Regimes of unstable, constantly changing standards are fundamental barriers to interoperability. Stable standards for data formats, data exchange protocols and data discovery protocols with widespread adoption are the basis for good interoperability. Standards however need time to mature and stability accrues when players are actively involved in their simultaneous specification and implementation. Greater clarity is thus needed on standards that should apply in this domain. New standards may not be necessary so the adoption of existing, well-used industry standards should therefore be promoted. Coordinating this process (e.g., through specification in procurement) is essential. Technical enforcement of security, intellectual property protection and data licensing becomes easier when standards are widespread and industrially based.

Recommendation(s):

1. Learn lessons from other domains such as the healthcare sector where the modus operandi has been to solve issues case-by-case. Interoperability "profiles" were introduced in that sector specifying the standards needed at every level (e.g., of an architecture) to be adopted by each provider within the sector. This approach could work for biodiversity science.
2. Publish and promote standards best practices on a central and well-known Website, such as e.g. the GEO BON site at https://www.earthobservations.org/geobon.shtml, and ensure these are considered when roadmapping technical developments across research infrastructures.

(iv) Solving specific technical challenges: Tackling interoperability implies addressing a set of technical challenges internally and externally to cooperating research infrastructures. The biodiversity infrastructure community needs to align and connect their services, and their workflows. There are five key technical recommendations on the roadmap to achieving interoperability.

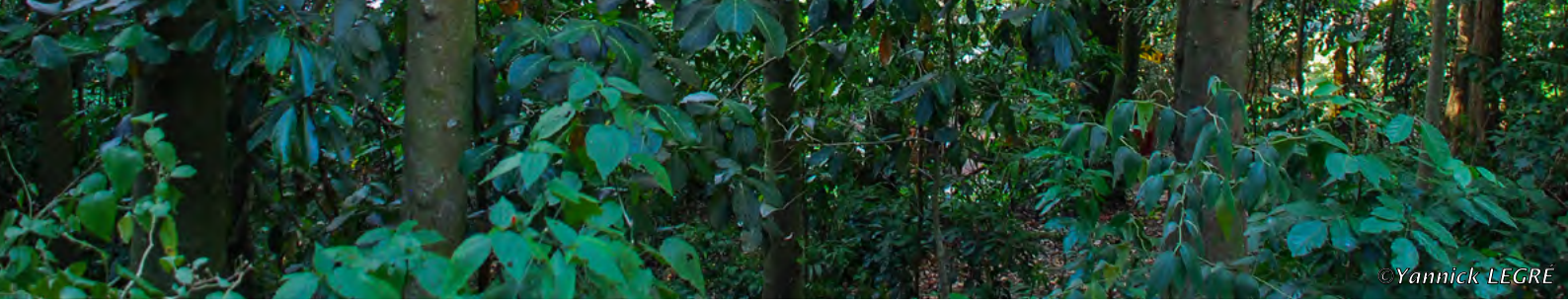
Recommendation(s):

1. Develop enabling, global and federated AAA (Authentication, Authorization and Accounting) infrastructures - 3 years

AAA. Overcoming barriers to AAA when composing complex applications across multiple research infrastructures requires alignment and interworking of security infrastructures. User applications in one research infrastructure should be able to enact services and access data within another infrastructure seamlessly. In practice, AAA interoperability at the global level could be based on the Shibboleth model[5] and on identity federations established more broadly than only biodiversity research infrastructures (e.g. GEANT[6]).

Trust. It will be necessary to establish mutual trust relationships between the cooperating research infrastructures as an essential prerequisite to supporting delegation of users' credentials throughout the flow of enacted services. This is a non-trivial and un-

[5] http://www.internet2.edu/products-services/trust-identity-middleware/shibboleth/
[6] http://www.geant.net/service/eduGAIN/Pages/home.aspx



solved problem arising from multiple levels of trust relationships that exist between: a) users and the applications they use; b) the applications and the domain-specific specialised service providers offering services upon which the applications depend (such as GBIF); and c) service providers and the providers of foundational computing, storage and networking infrastructures (e.g., general-purpose cloud computing and cloud storage).

2. Encourage the use of consistent quality control, semantics - 5 years

Quality control. Data and metadata need to be better qualified in terms of quality, i.e. whether they were quality assured the protocol that was used. Moreover, the granularity between data and metadata also requires a subtle and well-balanced thinking to be turned into meaningful information for users.

Semantics. The lack of applying consistent controlled vocabularies and the absence of a comprehensive and agreed ontology for biodiversity and ecosystem science impedes the semantic integration of data. Alignment of concepts and agreed (meta) structures (copying, for example, the approach of UMLS - Unified Medical Language System) would contribute to better understanding, integration and interoperability.

EBVs. Work in the area of Essential Biodiversity Variables (EBVs) may help in overcoming ontological alignment and complex new developments by introducing an intermediary semantic but simplified layer closer to end-users expectations.

3. Promoting the development, sharing and use of workflows of services - 5 years

Services. Web services play a significant role in separating technological dependencies arising from specific software decisions of research infrastructures. The use of Web services should be encouraged to expose the cooperating research infrastructures functions and to allow their interoperation, without implying intrusive integration nor complex reengineering.

Workflows. As progress is made in exposing data and analytical tools as standard web services, it becomes more important to adopt robust workflow management systems, (e.g. Taverna and Kepler). Workflows make it possible to combine coarse-grained functions into complex applications (such as calculating EBVs) requiring access to resources located in various research infrastructures. Peer-reviewed workflows offer a standard way of doing something or being an approved procedure in a regulatory environment. Workflows have to be repeatable, allowing the same or similar task to be done repeatedly with different data and/or control parameters. Workflows should fit the "ISA" management model of "Investigations, Studies, Assays" that is finding favour in the wider life-sciences[7]. Workflows allow reproducibility and act as a provenance mechanism for capturing the way work was done – provenance of the data, the tools used and the precise steps followed. Workflows offer a faster, cheaper, and integrative way of linking and utilising resources across multiple research infrastructures.

[7] http://www.isa-tools.org

4. Creating a scientific market place for biodiversity services - 5 years

Market. Allowing workflows of services to be composed and executed cross-enterprise and cross-infrastructure require globally accessible catalogues of data, services and associated semantics. Catalogues will be used to publish, discover, share and manage global portfolios of data and services. DataONE and GBIF, for instance have already made much progress in these areas. Multiple catalogues for data lead to the need for federated search and discovery that can be addressed with openSearch technology[8]. Service services, such as the Biodiversity Catalogue[9] should be promoted as well-known and well-founded directories of Web services for biodiversity and ecosystems analysis applications. In both cases, enhanced capabilities permitting semantic searching in and across catalogues are needed for the future.

Enterprise Service Bus (ESB). Comparable to current ESBs, a de-centralised and voluntary "Service Network" approach that accounts for independence and autonomy of individual Service Providers is most likely to find favour among cooperating research infrastructures. Data and service brokering components, such as those investigated by EuroGEOSS[10] take away from Service Providers much of the responsibility for interworking heterogeneous resources – even if they are encouraged to comply with relevant sector standards in order to maximise usage of their service(s).

5. Managing the provenance of resources in RIs - 10 years

Digital Objects Identifiers. All resources of the involved research infrastructures have to be assigned with a unique and global identifier, in the same way that scientific publications (DOIs) and data are. Thus, it would be possible to identify, manage these resources and ultimately to store provenance information when creating, modifying and utilizing them individually or collectively in workflows. A common mechanism across RIs is needed but DOIs appear to be well accepted by the community.

Provenance. Details of all actions carried out in the cooperating research infrastructures, the users involved, as well as all modifications of the state of resources should be monitored, tracked and preserved in order to make it possible to define precisely the provenance of every single digital object, to assure IP ownership, define responsibilities, identify the root causes of problems, improve quality processes and support repeatability of processes. Open models for structuring provenance information, such as the Open Provenance Model (OPM) should be considered.

[8] http://www.opensearch.org/
[9] https://www.biodiversitycatalogue.org/
[10] http://www.eurogeoss.eu/broker/Pages/AbouttheEuroGEOSSBroker.aspx

6. Legal and governmental implications

Apart from technical interoperability, also legal interoperability is a serious issue for biodiversity and ecosystem research infrastructures. Sharing data and tools with varying provenance of authorship and ownership requires careful and efficient arrangements when the cooperating research infrastructures want that their users can benefit from each one's resources. This even more important with the increasing automatic processing of data supported by "machine-machine" interactions. Fortunately, the study of potential issues revealed that there are not serious obstacles, especially not when the cooperating research infrastructures and other appropriate stakeholders will adopt a number of recommendations as explained below.

6.1. Leqal interoperability of the cooperating research infrastructures

Although there are different views on what is meant by "legal interoperability", the cooperating research infrastructures have a common understanding that "legal interoperability" means "ensuring that the data from two or more databases may be combined or otherwise reused by any user without compromising the legal rights of any of the data sources used." (Ref: Legal Interoperability Subgroup of the Group on Earth Observations' Data Sharing Working Group).

Since the issues concerning legal interoperability of biodiversity research infrastructures are similar to those faced by other research infrastructures, the "Research Data Alliance – CODATA Working Group on Legal Interoperability of Research Data" (RDA-CODATA WG) adopted the study for this Roadmap as a case for its work.

6.2. Legal interoperability in the application of technical standards & protocols

There appears not to be no legal barriers to the choice of and use of technical standards & protocols for infrastructure interoperability. Research infrastructures may select their preferred standards & protocols on scientific and technical grounds without interference by States or public funding agencies. (An exception requiring closer evaluation may be GIS related software, i.e. the INSPIRE Directive 2007/2/EC p & s and the Brazilian INDE system Decreto Nº 6.666 de 27/11/2008). The cooperating research infrastructures agreed to a policy of consultation and sharing of experiences before adopting new standards & protocols. This should be part of the training on applying standards & protocols in each research infrastructure. The cooperating research infrastructures also agreed to share awareness about typical contract clauses with the restriction that the use their software or databases cannot be transferred to other users, in particular to scientific communities and other stakeholder users.

6.3. Legal protection of research infrastructures, their data-bases, products and services under IPR law: interoperability and IPR & technical governance for research infrastructures

Currently there are no restrictions in the use of tools for mining texts and data by research infrastructures. Tools developed for one research infrastructure can be adapted by other infrastructures without licensing, since the cooperating research infrastructures are working mainly in open source environments. The cooperating re-

search infrastructures agreed not to impose any obstacles in their negotiations about semi or automatic interoperability mechanisms. This implies that they allow for unlimited (re) licensed rights for use of private software amongst the cooperating research infrastructures. (The European LifeWatch research infrastructure might have to face an exception in the medium term when it would register under database EU IPR law IP, and would decide to commercialize some products directly or by spin-offs companies).

6.4. Access to "Public" Data

All the home countries of the cooperating research infrastructures endorsed the Open Access policies ("Public access" in the US) concerning scientific data obtained through publicly funded research. The research infrastructures have no restrictions in cooperating with data re-use. (The exception is access to genetic resources under article 15 of the Convention on Biological Diversity[11], in particular for Brazil and South Africa). Although the cooperating research infrastructures adopt open/public access policies, they apply a pragmatic approach by not-contesting any claims of data ownership made by those individual scientists or specific scientific communities that still practise the "it´s my data syndrome". Applying other approaches to such scientific communities will be considered when necessary, independently of the standing open/public access policies. Restrictive requirements of external data providers may result in limited access to some data within and between research infrastructures.

Another issue is that the considered new European text and data mining (TDM) mandatory re-licensing policy (the so-called «Licences for Europe, A Stakeholder Dialogue») might become a serious obstacle to data re-use[12]. Such a policy could result in extra costs for the EU-based infrastructures, and make data generated in Europe not accessible for non-European partner infrastructures. The developments will be closely followed by the cooperating research infrastructures, as well by the RDA – CODATA Working Group on Legal Interoperability of Research Data.

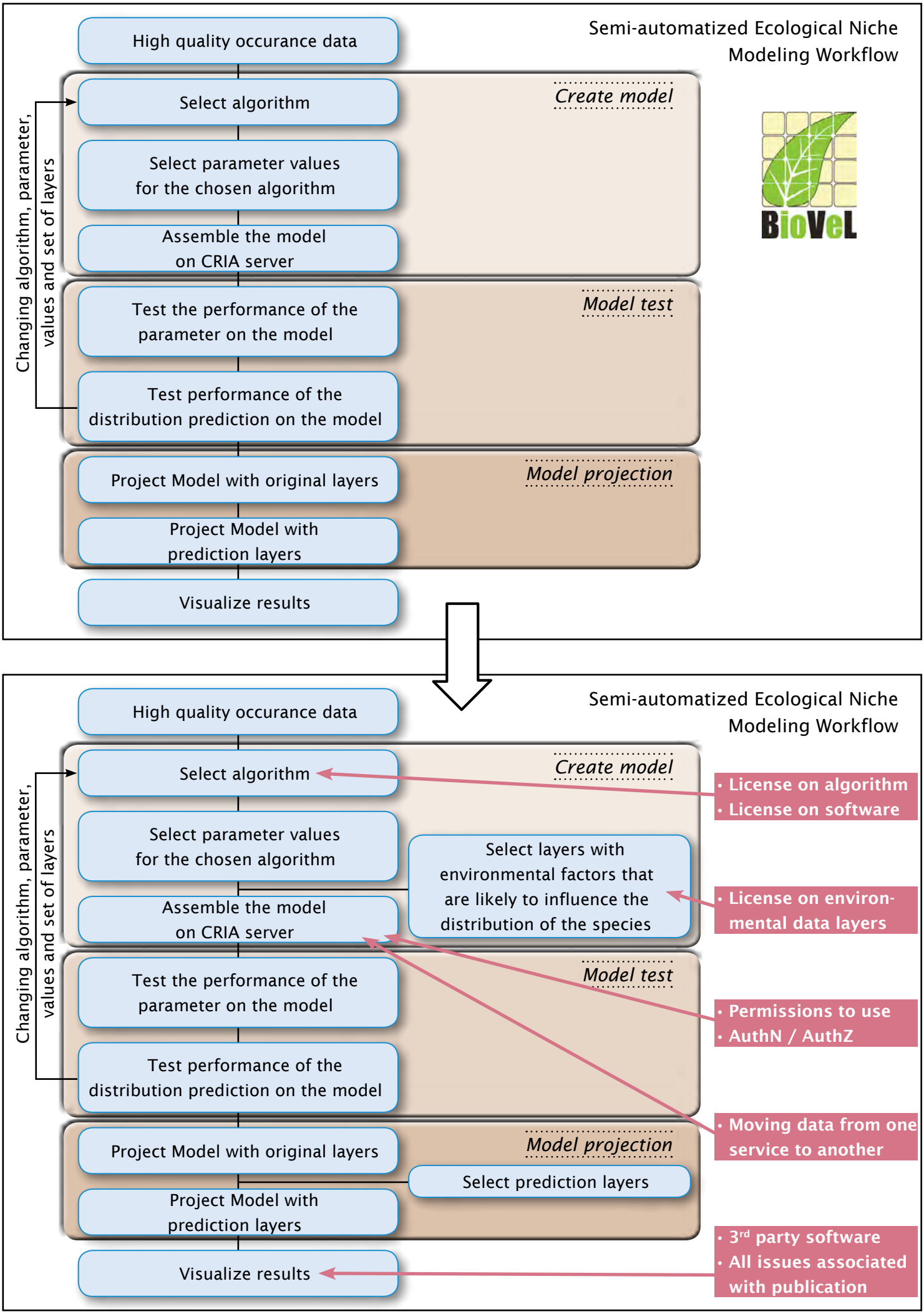
6.5. "Private" data & software protected under IPR law.

Limitations to semi- or automatic interoperability can originate from IPR law governing "data" (e.g. attribution) and from IPR law protecting "software" or other IT developments. Other limitations are known from "data ownership" claims of some communities despite publicly funded research. There is a serious problem that different interpretations and applications of IPR laws, and sometimes with deviations for domestic reasons, is hampering running workflows with multiple data sets and software tools from different sources. When implied licenses would be strictly followed, it is impossible to run workflows automatically. Each step in the workflow will be confronted with different license schemes, sometimes even implying asking prior permission. The findings of the BioVel project on such bottlenecks are illustrative. (An example is in following figure 3). The cooperating research infrastructures agreed these bottlenecks and their potential solution should be evaluated as a common exercise.

[11] http://www.cbd.int/convention/articles/default.shtml?a=cbd-15

[12] http://ec.europa.eu/licences-for-europe-dialogue/en/content/about-site

Figure 3: Example of a simple workflow, but with complicated legal implications



There are still other actions necessary to avoid potential legal obstacles to interoperability.

- Refrain from practices and domestic policies hampering data reuse since excess attribution requirements for aggregated data are imposed;
- Promote similar data quality management techniques and share good solutions (i.e. treatment of aggregated occurrence records);
- Consider the new creative commons 4.0 and CC0 licenses. These licences are intended to provide a normative (versus legal) approach to data attribution;
- Study the implications of the Earth Science Information Partners (ESIP)/ COOPEUS citation protocol[13] (and the GEOSS Data Citation Standard);
- Implement smart solutions for applying waivers of any rights on data served by each research infrastructure so that automatic machine processing of data is supported (as it happens in the medical world).

6.6. Terms of Reference for continued collaboration of the research infrastructures

The cooperating research infrastructures agreed to sign Terms of Reference to continue their collaboration. This allows for establishing a High Level Stakeholders Group (HLSG) serving as a platform for consultation and high-level coordination of activities. More specifically, the HLSG serves as a policy liaison between the cooperating research infrastructures for exchange of opinions and the preparation and dissemination of joint recommendations. Fundraising for cooperative activities will be based on the strategic arguments for realizing interoperable infrastructures, and to achieve economic sustainability.

[13] http://wiki.esipfed.org/index.php/Main_Page

7. A Roadmap for the research infrastructures

The previous paragraphs provide an analysis of required actions to move towards an international collaborative virtual infrastructure environment supporting biodiversity and ecosystem research. These actions result in the following recommendations for the first global Roadmap on cooperating biodiversity & ecosystem research infrastructures.

7.1. Sustain the role of biodiversity & ecosystem research infrastructures

The grand challenge for biodiversity and ecosystem scientists is to understand biodiversity change, and more seriously biodiversity loss. Tackling the grand challenge requires interlinked and interoperable research infrastructures providing the required powerful support services to advance knowledge on larger scales. The production and free accessibility of long-term and broad-spatial data and analysis tools requires sufficiently sustained biodiversity and ecosystem research infrastructures. Currently the funding arrangements are different for the cooperating research infrastructures but most have only guaranteed sustainable funding in the short term. Since the research infrastructures are increasingly mutually dependent - in order to provide the envisaged global infrastructure laboratory - it is recommended to analyse and compare funding principles and mechanisms. Such a study should consider the value of supporting agreements on complementary capabilities and services, budgeting of these services and policies on user fees. This exercise should result in a common view on funding principles, preferably adopted by both funding agencies and research infrastructures.

Data and services of the biodiversity and ecosystem research infrastructures are contributing to societal and economic benefits. The mission of the research infrastructures themselves is in the public domain; exploiting commercial opportunities should be organized 'outside' the infrastructure. Exploiting such opportunities is possible when existing public organisations, private companies or spin-off companies take these up. The cooperating research infrastructures are expected to foster an active policy in this regard.

The research infrastructures cooperating in designing this Roadmap agreed to establish a High Level Stakeholders Group (HLSG), bringing together their leaders for consultation, advice and collaboration. It is recommended that they actively seek funding to enter new collaborative opportunities.

7.2. User interaction and value delivery

Sustaining research infrastructures requires demonstrated demand and use of their services, and therefore the active involvement of their scientific communities should be fostered. Promotion and facilitating of the interactive involvement of scientists and decision-makers in framing the appropriate research questions is a key priority for the cooperating research infrastructures. This involvement will assist in developing targeted capabilities and in the provision of relevant decision-support tools. When each of the cooperating research infrastructures provides a gateway to the colleague infrastructures, worldwide scientific communities can better engage in common research goals, access cutting-edge technologies and increase knowledge.

Biodiversity & ecosystem research infrastructures have to play a role as a broker between citizens, scientists and other users for the use

and production of data, tools and services. It is recommended to empower citizen scientists so they can better engage with the research infrastructures.

Supporting the development and testing of biodiversity indicators is a recommended joint action plan to deliver new services and to demonstrate user involvement and the benefits of interoperability. The concept of Essential Biodiversity Variables (EBVs) as propagated by GEOSS-GEOBON may serve as demonstrator. Cooperating research infrastructures should focus on capabilities to:

- discover and access relevant biotic and abiotic data;
- build the models (and algorithms) to compute EBVs;
- test the sensitivity (and reliability) of EBVs to data and parameter change;
- scale up for use by different areas and times;
- construct virtual laboratories to deploy the services with low-threshold use;
- and finally offering accepted protocols allowing for comparing EBVs.

7.3. Cooperation for infrastructure interoperability

The cooperating research infrastructures exhibit a satisfactory level of potential interoperability, in particular in the way they offer access to biodiversity data, available applications and related resources. This facilitates the achievement of an international virtual environment (IVE) for biodiversity and ecosystems research and the accompanying governance. The recommendations are:

Emphasize and increase the importance of standards

Learn lessons from other domains and proceed on a case-by-case basis; publish and promote standards best practices on a central and well known website.

Solve technical challenges for biodiversity and ecosystem infrastructures

Develop enabling, global and federated Authentication, Authorization and Accounting (AAA) facilities so that users in one research infrastructure can enact services and access data within another infrastructure seamlessly. It will be necessary to establish mutual trust relationships between research infrastructures as an essential prerequisite to supporting delegation of users' credentials throughout the flow of enacted services.

Encourage the use of consistent quality control, semantics

The lack of consistent vocabularies let alone an agreed comprehensive ontology for biodiversity and ecosystem science impedes the semantic integration of data. The recommended work on Essential Biodiversity Variables may assist in dealing with ontologies alignment in complex new developments.

Promote the development, sharing and use of workflows of services

The use of Web services should be encouraged to expose functions of the cooperating research infrastructures and to allow their interoperation. The same holds for exposing analytical tools, data and other resources as standard Web services. It is recommended to adopt robust workflow management systems, since these make it possible to combine coarse-grained and distri-

buted functions into complex applications (such as calculating EBVs).

Create a scientific market place for biodiversity services

A scientific market place allows users to benefit from workflows of services to be composed and executed cross-enterprise and cross-infrastructure but requires globally accessible catalogues of data, services and associated semantics. It is recommended to promote catalogues of services.

Another "market" service is introducing an Enterprise Service Bus (ESB); a de-centralised and voluntary "Service Network" approach that accounts for independence and autonomy of individual Service Providers. It is recommended for adoption by cooperating Research infrastructures.

Managing the provenance of resources in RIs

All resources of the involved research infrastructures have to be assigned with a unique and global identifier in order to identify, manage resources and ultimately to store provenance information when creating, modifying and utilizing resources. Identifiers would allow for the traceability and preservation of every single digital object. Identifiers will also assure IP ownership, define responsibilities, identify the root causes of problems, improve quality processes and support repeatability of processes.

7.4. Legal interoperability

Legal interoperability is a serious issue for biodiversity and ecosystem research infrastructures. Sharing data and tools with varying provenance of authorship and ownership requires careful and efficient arrangements among cooperating research infrastructures. Legal interoperability issues are becoming more significant with the increase of automatic processing of data supported by "machine-machine" interactions. Not all issues on legal interoperability face significant obstacles, but the use of licensed software or middleware and attribution is a source of potential serious problems.

The cooperating research infrastructures and other appropriate stakeholders will adopt the following recommendations.

- Follow and support the work on such legal issues in the Research Data Alliance-CODATA legal interoperability Working Group. The analysis of the Creative-B project is a case study to identify generic solutions to legal aspects of data and licenses interoperability.

- Consider a common policy for the adoption of new technical standards, protocols and knowledge sharing.

- Evaluate the GIS-based data protocols for geo-referencing of biodiversity data in INSPIRE (EU) and INDE (Brazil) regulations on their implications for complicated procedures and costs striking all global research infrastructures deploying these biodiversity data.

- While committing to the Open/Public Access as endorsed by most States, it is recommended to be aware and sensitive to the existing varying cultures of scientific communities on data protection.

- Follow the global developments dealing with data attribution mechanisms and policies, in particular to imposed excessive attribution resulting in restricted or blocked data re-use.

- Follow the new creative commons 4.0 and CC0 licenses and the implications of providing a normative (versus legal) approach to attribution. Propose alternatives for universal waivers.

- Analyse in depth the implications of diverging data and software licenses when running workflows. Currently even simple workflows cannot run "on a click" or automatically since a suite of agreements with different licenses must be processed manually. Propose standard machine-readable solutions.

- Give special attention to the "Re-licensing Europe" dialogue[14] and the potential limitations for text and data mining activities, harming the operations of the cooperating research infrastructures. The European research infrastructures are recommended to keep their global colleagues informed about impeding developments in this regard.

7.5. Education and training

Research infrastructures are operating at the edge of current knowledge and technology, and they seek for and support excellence in science. As such it is conditional to invest in training and capacity building. In their social environment of both collaboration and competition, training and capacity building should specifically be directed at young and new researchers to enable better use of the research infrastructures. It is recommended that the cooperating research infrastructures and their funding bodies:

- develop and support training and capacity building, including arrangements for exchange of staff in order to learn about best current practices.

- structure and support education and training that encourages interoperability. Such activities should be organized around specialized biodiversity "market places" while providing access to infrastructure resources, so that such training contributes to the adoption and harmonizing of interoperability practices.

- communicate how biodiversity data and models are relevant for advanced research in support of environmental policies. In turn, this will provide feedback to data resources on data use and on required new data delivery.

[14] http://ec.europa.eu/licences-for-europe-dialogue/en/content/about-site

Annexes

I. The cooperating biodiversity research infrastructures in Creative-B

The Europe based LifeWatch infrastructure for biodiversity and ecosystem research is in development to provide virtual environments, enabling integrated access to data, analytical and modelling workflows and computational capacity. It is a new approach for large-scale cooperation in simulation and scenario development experiments.

The Group on Earth Observations Biodiversity Observation Network – GEO BON – coordinates activities relating to the Societal Benefit Area (SBA) on Biodiversity of the Global Earth Observation System of Systems (GEOSS). Some 100 governmental, inter-governmental and non-governmental organizations are collaborating through GEO BON to organize and improve terrestrial, freshwater and marine biodiversity observations globally and make their biodiversity data, information and forecasts more readily accessible.

In Australia, the Atlas of Living Australia (ALA) contains information on all known living species in Australia, aggregated from a wide range of data providers: museums, herbaria, community groups, government departments, individuals and universities.

In Brazil, the Reference Centre on Environmental Information (CRIA) aggregates and disseminates biological information of environmental and industrial interest, as a means of organising the scientific and technological community of the country towards conservation and sustainable use of Brazil's biological resources.

CAS is China's government organisation, founded in Beijing on 1 November 1949, as the nation's highest academic institution in natural sciences and its supreme scientific and technological advisory body, and national comprehensive research and development centre in natural sciences and high technologies. The CAS Biodiversity Committee oversees the operations of biodiversity infrastructures in China. The CAS Germplasm Bank of Wild Species (GBoWS) is one of the 11 large research infrastructures managed by the Chinese Academy of Sciences (CAS). CAS is hosting the World Data Centre for Microorganisms.

In the USA, the Data Observation Network for Earth (DataONE) is developing the future foundations for environmental sciences with a distributed framework and sustainable e-infrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered earth observational data.

Through a global network of countries and organizations, the Global Biodiversity Information Facility (GBIF) encourages free and open access to biodiversity data, and promotes and facilitates the mobilization, access, discovery and use of information about the occurrence of organisms over time and across the planet.

The South African National Biodiversity Institute (SANBI) leads and coordinates research, monitors and reports on the state of biodiversity in South Africa. Providing biodiversity information is central to SANBI's mandate and it does this by providing several databases and other resources developed by SANBI and its partners.

II. Authors - Contributors list

Authors	Organization	Country
ALONSO Enrique	Universidad de Alcalá	Spain
BELLISARI Livia	Comunità Ambiente	Italy
DE LEO Francesca	Consiglio Nazionale Delle Ricerche	Italy
HARDISTY Alex	Cardiff University	United Kingdom
KEUCHKERIAN Samuel	Centre National de la Recherche Scientifique	France
KONIJN Jacco	University of Amsterdam	The Netherlands
LOS Wouter	University of Amsterdam	The Netherlands
MANSET David	Gnubila	France
SPINELLI Oliviero	Comunità Ambiente	Italy
VICARIO Saverio	Consiglio Nazionale Delle Ricerche	Italy

Contributors	Organization	Country
BELBIN Lee	Atlas of Living Australia	Australia
CANHOS Vanderlei	CRIA Reference Center on Environmental Information	Brazil
CANHOS Dora	CRIA Reference Center on Environmental Information	Brazil
COOK Bob	DataOne	USA
GELLER Gary	GEO BON	USA
HOBERN Donald	Global Biodiversity Information Facility	Denmark
JI Li-Qiang	Chinese Academy of Sciences	China
KOSKELA Rebecca	DataOne	USA
LaSALLE John	Atlas of Living Australia	Australia
LEGRÉ Yannick	Centre National de la Recherche Scientifique	France
LLOSENT Maria	Universidad de Alcalá	Spain
MA Keping	Chinese Academy of Sciences	China
MA Juncai	Chinese Academy of Sciences	China

Contributors	Organization	Country
MICHENER William	*DataOne*	*USA*
ROBERTSON Tim	*Global Biodiversity Information Facility*	*Denmark*
SAARENMAA Hannu	*GEO BON*	*Finland*
SANCHEZ Francisco	*LifeWatch Spain/CWI*	*Spain*
UHLIR Paul	*National Academy of Sciences*	*USA*
VIEGLAS Dave	*DataOne*	*USA*
WILLOUGHBY Selwyn	*South African Biodiversity Information System*	*South Africa*
WISSEL Silvia	*University of Amsterdam*	*The Netherlands*
XIANG-YUN Yang	*Chinese Academy of Sciences*	*China*

NOTES

CREATIVE-B

Related initiatives

ATLAS OF LIVING AUSTRALIA
sharing biodiversity knowledge

CHINESE ACADEMY OF SCIENCES

GBIF
www.gbif.org

cria

DataONE

GEO GROUP ON EARTH OBSERVATIONS

LifeWatch

SANBI
Biodiversity for Life

Photo Credit: with courtesy of ©Yannick LEGRÉ