BRIDGE COLOMBIA

GROW COLOMBIA

Developing a research cyberinfrastructure in Colombia

August 2018

By C3Biodiversidad, Colombian Cyberinfrastructure Consortium for Biodiversity.

Foreword

In an UKRI-funded international collaborative effort through the 'GROW Colombia' project, UK-based researchers are working with a network of scientists throughout Colombia to study the natural and agricultural diversity within the country's unique biodiversity, documenting its distribution and the threats it faces. 'GROW Colombia' aims for research excellence through promoting innovative technologies, developing resilient research capabilities, building partnerships, and fostering best practice in knowledge exchange, with longer-term goals to stimulate economic and social growth.

The potential for data to transform how we tackle many of the major challenges faced by humanity, from climate change and environmental sustainability to conservation and food security, is huge. However, this can only be achieved through better access to data, enabled through effective and efficient infrastructures. As part of the GROW Colombia project, Earlham Institute organised a workshop in Bogotá in June 2018 which brought together important experts in the country to design a strategy for how to develop a research cyberinfrastructure that can cope with the needs to study the country's staggeringly rich native biodiversity.

GROW Colombia sits alongside the BRIDGE Colombia network (www.bridgecolombia.org) which is a multidisciplinary network of research organisations in the UK and Colombia, founded in March 2017, to develop robust coordinated activities under a shared vision centered on biodiversity as a means to achieve sustainability and peace.






Introduction

Diversity and abundance to foster a bioeconomy

Colombia is one of the 17 countries considered as “megadiverse” by the United Nations Environment Programme (UNEP) and has a national catalogue of biodiversity of up to 62,829 species of all taxonomic groups, 9,153 of them endemic, representing around 10% of all known species on earth. This biodiversity is soon to be revealed in greater depth than ever before. It is now recognised that science and innovation is not a luxury but a prerequisite for social and economic development. A greater understanding of this biodiversity is not only key to conserving and promoting it, but can also drive the economic growth, social equality and a sustainable peace in the country.

This can only be achieved through a greater access to data, enabled through effective and efficient infrastructures - ensuring that we are harnessing the best expertise from all around the world to work on this swathe of grand issues.

To date, Colombia holds a first place in number of species of birds and orchids; second in plants, amphibian, butterflies and freshwater fish; third in palms and reptiles; and fourth in mammals. Bringing biodiversity analysis into the digital world will provide all people and jurisdictions with comprehensive evidence and knowledge to make informed decisions.






Community driven discovery

Exploring and investigating the wealth of information contained within each of the thousands of species in these ecosystems, especially with the advent of modern life sciences methods, starts to stretch the capacity of single research groups.

One way of ensuring the efficient and accurate analysis of such abundant data is to cultivate and foster a research cyberinfrastructure. The main goal of a research cyberinfrastructure is to enable data-driven scientific discovery through not only advanced hardware but a community of people and institutes who can manage and share computational resources in a sustainable, secure, collaborative, and interoperable way. A research cyberinfrastructure aims to meet the needs of the life science community through democratised access to computational resources.

The way researchers interact with computational resources is changing. Cloud computing is a very different model to the 'servers under desks' mentality that has been prevalent up until now. This brings new opportunities and abilities to analyse and share data, but without common infrastructure building blocks it is difficult to make the available resources greater than the sum of their parts. This often makes it hard to combine and reproduce data, especially when teams are using different software, leading to wasted time and effort.

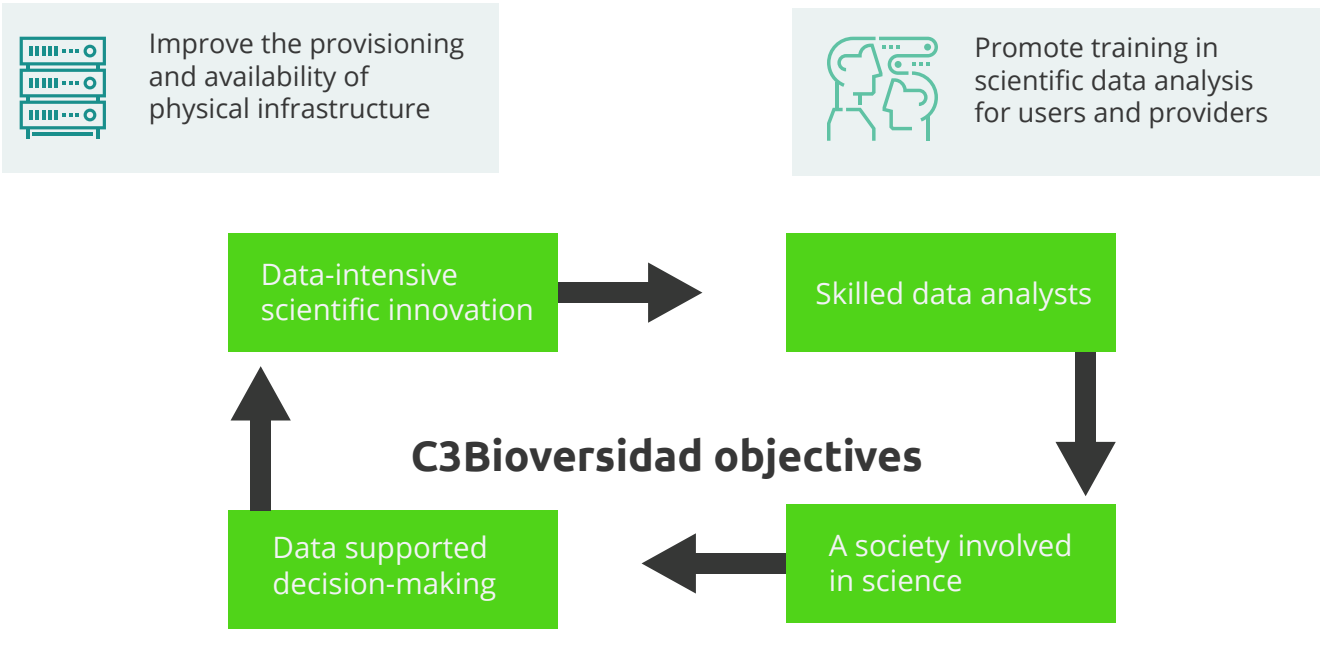
A research cyberinfrastructure:

- Enables data-driven scientific discovery.
- Is a "technological and sociological" ecosystem that facilitates research data services.
- Comprises people with the technical skills necessary to execute and share tools and resources in a sustainable, secure and interoperable way.
- Democratises equitable, fair, and coordinated access to computational resources and relevant scientific data sets.
- Is designed to facilitate remote collaboration and virtual organisations from its basic premise.

A scientific cyberinfrastructure to accelerate the understanding and preservation of biodiversity in Colombia

The Colombian Cyberinfrastructure Consortium for Biodiversity, C3Biodiversidad, aims to develop and promote a scientific cyberinfrastructure in Colombia for analysis of natural and agricultural biodiversity. This Colombian cyberinfrastructure will include the provision of high performance computing (HPC) and large data storage, and allow for the establishment of a virtual organisation around common practices, tools and data. As result, it will necessarily include the community of people and institutions in the country to manage these resources in a secure and interoperable way. C3Biodiversidad has identified the following priorities:

- Improve the provisioning and availability of physical infrastructure.
- Grow training in scientific data analysis for users and providers.
- Design a national policy for research data.
- Develop a scheme to engage diverse stakeholders in research projects and funding planning.



C3Biodiversidad

C3Biodiversidad (Consorcio Colombiano de Ciberinfraestractura para la Biodiversidad) aims to develop a computing intrastructure in Colombia for the analysis of scientific data. Created in Bogota on 28 June 208 by an expert panel from the Science, Technology and Innovation system of Colombia, along with international experts, C3Biodiversidad is open to any stakeholder interested in the development of such a research cyberinfrastructure in Colombia.

Delegates at the C3Bioversidad workshop held in Bogota, June 2018

Enhancing physical infrastructure

The availability in Colombia of capable system administrators and application developers is a key strength to consolidate the physical computational infrastructure of the country. Emigration of these skilled personnel and limited opportunities for technical training of new personnel are consequently the main risks. Specific software tools for High Performance Computing (HPC) administration, planning, and virtual interaction between these personnel, both in Colombia and neighbouring countries, offer further cost-effective opportunities; in some cases these are already available.

A limited infrastructure is the main challenge Colombia presently faces, particularly capacity and physical connectivity between institutions. The existence of a state-sponsored high-speed provider the National Academic Network of Advanced Technology in Colombia (RENATA) can, however, promote collaboration between institutions. This is a pre-requirement for further sharing schemes. Limited funding, vendors, and continuity limit the adoption of new proposals by key institutions which look to prevent or limit uncertainty or external risks.

INTERNAL FACTORS	
STRENGTHS (+)	**WEAKNESSES (-)**
• Capable system administrators. • Experience building with limited resources. • Experience in application development. • Willing to cooperate on human capital.	• High-speed inter-institutional connectivity. • Limited computing and data storage capacity. • Emigration highly trained/skilled. • Limited education and opportunity for training in HPC. • Limited risk management.
EXTERNAL FACTORS	
OPPORTUNITIES (+)	**THREATS (-)**
• RENATA. • Existing collaborations between institutions. • Adoption of software tools for computing management and collaboration. • Experiences from neighbouring countries.	• Institutional hesitation. • Limited funding and vendors. • Long-term continuity. • Government-level involvement.

We recommend building a federated, sustainable and cooperative computational platform to accelerate scientific research and capacity building.

The roadmap would firstly identify the needs and available resources in Colombia via surveys to a diversity of stakeholders, which can be commissioned to a governmental or academic department. It would be followed by assembling a formal advisory committee(s) with experts from key institutions, which would need to agree its own governance and deploy online collaboration tools.

In parallel, RENATA or an equivalent institution would need to coordinate high-speed connectivity between these key institutions, and would involve private Institute Strategic Plans (ISPs) if needed. Eventually, computational resources from the institutions would be progressively shared to the common platform following an ordered procedure (e.g. virtual machines, storage, etc.). This would be catalysed by a social recognition and reward scheme by the Colombian Science Council (Colciencias).

Accelerate data-intensive scientific research

- Explore the biodiversity of Colombia.
- Facilitate skills sharing.

Build a federated sustainable cooperative computational platform

- Commission survey about resources needs and availability.
- Facilitate institutional connections to RENATA.
- Formalise an advisory committee about physical computational resources.
- Implement recognition scheme for resource providers in Colciencias evaluation(s).

Grow training in scientific data analysis for users and providers

In Colombia to date, there are several BSc and MSc programmes in bioinformatics, but a much more limited training offer in advanced computational subjects. There are multidisciplinary research centres and universities with strong diverse departments and a strong bioinformatics network, including a student council, which organises a biannual bioinformatics national conference.

The demand for training is high but opportunities for coordinated training between strong groups have not been fully explored, and internships/visits between groups are uncommon. Further communication between groups, coordination between institutions, funding, facilities and leadership would result in prompt tangible results.

INTERNAL FACTORS	
STRENGTHS (+)	**WEAKNESSES (-)**
• High-level education (BSc/MSc) in bioinformatics. • Multidisciplinary research centres and universities. • International Society for Computational Biology (ISCB) node, student council, National Bioinformatics conference.	• No formal courses in advanced areas for students and researchers. • Lack of a formal training network and communication between groups that can provide training. • Limited opportunities for internal exchanges • Very sporadic internships or visits between national groups. • Lack of cross-disciplinarity in smaller universities and centres.
EXTERNAL FACTORS	
OPPORTUNITIES (+)	**THREATS (-)**
• High demand for training. • Existence of technology-driven communities. • Building training around existing meetings. • Tools to improve communication between researchers. • Subject-focused bioinformatics training in multidisciplinary institutions. • Larger institutions providing training in collaboration with smaller institutions.	• Limited funding, facilities and leadership. • Institutional and leadership opposition. • Limited training programmes for human capital. • Emigration of skilled people who can provide training.

We recommend designing and promoting a coordinated accessible programme of training in scientific data analysis, tailored to different career levels. This programme should enable Colombia to lead the offer of training opportunities in the region.

The roadmap would involve promoting a national network of trainers that would keep an online record of training activities and would aim for some coordination within the country, as well as with equivalent international initiatives, particularly GOBLET (Global Organisation for Bioinformatics Learning, Education & Training).

The network would be promoted and developed within the existing national bioinformatics conference and existing research communities. There is a need for further funding, specifically to facilitate hosting international trainers for courses and staff exchanges between national institutions, particularly from smaller to larger universities and research centres.

Growth people skills in scientific data analysis

- Coordinate advanced tailored training.
- Facilitate institutional cooperation.

Promote coordinated accessible programme of training in scientific data analysis

- Develop a national online network of trainers in bioinformatics within existing communities.
- Coordinate with GOBLET.
- Support hosting of international trainers.
- Support staff exchanges from smaller to larger institutions within the country.

Develop a national policy for research data

The limited enforcement of policy in Colombia that regulates the access, standards, incentives, and retention for research data strangles further developments in the area. Colciencias published its vision for an “open science” in the country at the end of 2018, and prioritised eight areas including open research data, research infrastructures, licences and IPR, metrics, and access to publications. In that context, OECD countries are actually committed to guarantee public access to taxpayer funded research, including both generated data and research publications.

There are Colombian institutions that are already competent in science policy and research data management (Colciencias, BIOS, Observatorio de Ciencia y Tecnología) which can drive the regulatory process and promote incentives among the researchers; as well as specific research communities and projects promoting standards, curation and preservation approaches following international ontologies and best practices. Particular institutions are already actively managing their repositories and databases, while others do not have the capacity or resources.

INTERNAL FACTORS	
STRENGTHS (+)	**WEAKNESSES (-)**
• Existing research community that uses international standards, such as ontology for diversity data. • Existence of the 'science and technology observatory'. • Colciencias' committees and evaluations.	• Compliance with the open research data policy • Lack of incentives for provision and exchange of data. • Lack of standardisation of data. • Variable access policies. • Variable standards for data sensitivity levels. • Limited monitoring and follow up.
EXTERNAL FACTORS	
OPPORTUNITIES (+)	**THREATS (-)**
• Existence of international community-led data standards, ontologies and metadata harvesting. • Data-oriented institutes that can administer repositories and databases. • Federated and centralised models tailored to institutions with different capacities.	• Low recognition for the need of data policies. • Changing governmental agenda. • Need of specialised advice and expertise in key institutions. • Lack of incentives for collaborations, data exchanges and quality control among researchers.

We recommend the promotion and enforcement of the open science policy and a national policy for research data that regulates the access, processing, and sharing of data, particularly biodiversity data, in a standardised way. This would facilitate data-supported decision-making as well as scientific excellence. We recommend a requirement for open access to taxpayer-funded research, including both generated data and research publications.

Emulating previous e-government data policy building exercises in Colombia, the roadmap involves an advisory committee of experts and relevant stakeholders, which would agree the needs for access, curation, retention, traceability, quality, interoperability and availability of the research data, particularly biodiversity research data, as well as the monitoring and enforcement mechanisms of this policy.

As a few self-sufficient institutions in Colombia have already developed their own policies, a federated model between these institutions would coexist with a centralised national data management repository, e.g. RedClara and the Inter-American Development Bank (IADB) are coordinating "Latin America Referencia" (http://lareferencia.info), which would also extent its function to serve institutions without the capacity to do it themselves.

Finally, the evaluation of researchers needs to be extended to incentivise the multidisciplinary collaboration between data creators, analysers, and curators, instead of the scoring scheme that penalises inter-institutional large partnerships.

Facilitate data-supported decision-making

- Incentivise excellence in research.
- Facilitate access to biodiversity data.

Develop a national policy for research data

- Involve stakeholders in new policy design.
- Require open access to taxpayer funded research.
- Assign institution a role in coordinating repositories and databases in the country.
- Reward researchers involved in data partnerships for evaluations.

Secure engagement of diverse stakeholders in research projects and funding planning

There is a limited number of initiatives to engage stakeholders, and a variable interest in research from different stakeholders and sectors. The partnerships between private sector, third sector, government, and society, are more established in the agricultural and environmental sectors. However, the limited resources, funding, and partnerships restrict planning in the big data sector.

There are three positive recent initiatives: specific research funding calls with private institutions, a new funding system based on royalties from the regions, and the international investment promoted by the OECD membership and the peace process.

INTERNAL FACTORS	
STRENGTHS (+)	**WEAKNESSES (-)**
• Existing Colombian Ministry of Information Technologies and Communication. • Well developed communities in the agricultural and environmental sectors. • Increasing foreign investment. • Existing incentives to promote research in the private sector.	• Lack of collaborations between institutions and with different stakeholders. • Data and resources are not usually shared. • Lack of long-term planning and investment policy. • Limiting communication and funding information channels. • Low priority for Science and Technology in the government agenda.
EXTERNAL FACTORS	
OPPORTUNITIES (+)	**THREATS (-)**
• New royalties system. • OCDE membership and new international investment. • Existing international partnerships. • Existing interests in promoting access and sharing of data and resources. • Establish schemes for technology transfer and interest of the private sector in research.	• Limiting opportunities. • Lack of monitoring and evaluation. • Instability.

We recommend promoting private-public networks and partnerships, the involvement of the third-sector (non-profit associations, charities, community groups and cooperatives) in research projects, and data intensive research projects via specific scoped research funding calls, in order to secure the engagement of a diverse range of stakeholders in research planning and execution.

The new royalties scheme would allow implementation of funding opportunities for larger partnerships and data intensive research, and use new funding to promote private-public partnerships, third sector involvement, and multidisciplinary projects. Researchers applying for national funding would need to include a plan for stakeholder engagement in their applications.

In parallel, the Research Council and Universities would extend their support offices to help researchers apply for international funding and engage with different stakeholders; building their network and opportunities.

Promote society involvement and interest in science and technology

- Support private-public partnerships.
- Support third-sector involvement.
- Support multidisciplinary data-intensive projects.

Develop a research support scheme that promotes stakeholder engagement

- Require engagement plans in research projects.
- Implement funding for public-private projects.
- Implement funding calls for data-driven projects.
- Extend researchers' support offices.
- Catalogue networking opportunities.

Acknowledgements:

This publication builds on the analysis from the panel of experts in Bogota on 16-18 June 2018:
Alejandro Caro, **AGROSAVIA**
Alice Minotto, **Earlham Institute**
Andrés Pinzón Velasco, **National University of Colombia**
Anyela Valentina Camargo Rodríguez, **National Institute of Agricultural Botany (NIAB)**
Camilo Corchuelo Rodríguez, **Santo Tomás University**
Carlos Ramírez, **National Academic Network of Advanced Technology of Colombia (RENATA)**
Cesar Orlando Díaz, **Jorge Tadeo Lozano University**
Dairo Escobar, **SiB Colombia – The Alexander von Humboldt Biological Resources Research Institute**
Daniel Fernando López, **The Alexander von Humboldt Biological Resources Research Institute**
Dany Molina, **Colombia's Center for Bioinformatics and Computational Biology (BIOS)**
Diego Rincón, **Catholic University of Colombia**
Emiliano Barreto, **National University of Colombia**
Federica Di Palma, **Earlham Institute**
Gastón Lyons, **University of Los Andes**
Graham Etherington, **Earlham Institute**
Jaime Erazo, **Earlham Institute**
Javier Correa Álvarez, **EAFIT University**
John Jaime Riascos, **Cenicaña**
Jorge Duitama, **University of Los Andes**
Jorge William, **Colombia's Center for Bioinformatics and Computational Biology (BIOS)**
Jose De Vega, **Earlham Institute**
Juan David Pineda Cárdenas, **EAFIT University**
Juan Manuel Anzola, **Corpogen**
Juan Pablo Mallarino, **University of Los Andes**
Julio Marín Duarte, **AGROSAVIA**
Laura Natalia González García, **University of Los Andes**
Leroy Mwanzia, **International Center for Tropical Agriculture (CIAT)**
Luz Miriam Díaz, **National Academic Network of Advanced Technology of Colombia (RENATA)**
Marco Cristancho Ardila, **University de Los Andes**
María Camila Martínez, **Cenicaña**
Monica Munoz Torres, **Ohio State University**
Narcis Fernandez, **Aberystwyth University**
Nelson Enrique Arenas Suárez, **University of Cundinamarca**
Patricia Jaramillo, **National Academic Network of Advanced Technology of Colombia (RENATA)**
Paula Reyes, **AGROSAVIA**
Raúl Ramos Pollán, **University of Antioquia**
Robert Davey, **Earlham Institute**
Romain Guyot, **Autonomous University of Manizales**
Tomás Viloria Lagares, **University of Los Llanos**
Yesid Cuesta Astroz, **University of Antioquia**

The authors would like to acknowledge support from the UK Research and Innovation (UKRI) Global Challenges Research Fund (GCRF) GROW Colombia grant via the UK's Biotechnology and Biological Sciences Research Council (BB/P028098/1).

Contact:
E: info@bridgecolombia.org
W: www.bridgecolombia.org