

THE IMPACT OF BRAZIL'S VIRTUAL HERBARIUM IN E-SCIENCE

Date: February 16, 2017

Reporting Period: February 2015 - February 2017

Report Due: February 20, 2017

Suggested Citation: Canhos, Dora Ann Lange, Sidnei de Souza, Vanderlei Perez Canhos, and Alexandre Marino. "The Impact of Brazil's Virtual Herbarium in E-Science." Final Report for Open and Collaborative Science in Development Network (OCSDNet), February 20, 2017.

Report Prepared by:

Dora Ann Lange Canhos

Sidnei de Souza

Vanderlei Perez Canhos

Alexandre Marino

Summary

The impact of Brazil's Virtual Herbarium in e-Science	1
1. Executive Summary	1
2. Research Problem	2
3. Research Objectives and Findings	4
4. Project Implementation and Management	6
5. Project Outputs and Dissemination	7
6. Community Building	8
7. Impact	10
8. Reflective Learning on Internal Dynamics:	10
9. Recommendations (for OCSDNet):	11
10. Additional Comments	11
Annex 1. Identifying outcomes from on-line data sharing	12
1. Methodology	12
2. Results and Discussions	13
3. Final Comments	15
Annex 2. Analysis of users participation through the annotation system and BioGeo	17
1. Annotation System	17
2. BioGeo (Biogeography of the Flora and Fungi of Brazil)	19
Annex 3: blocked data	25
1. Introduction	25
2. Herbaria with blocked data fields	26
3. Zoological records	28
4. Final comments	33
Annex 4. Analysis of Users and Usage	36
1. Usage	36
2. Users	38
Annex 5. Contextualizing Openness	40
1. Is science open?	40
2. The Project	40
3. What strategies contributed to <i>openness</i> ?	41
4. Results	42
5. Impact on Data Providers	43
6. The strength of the network	44
7. Final Comments	44
8. References	44

1. Executive Summary

This project's main research theme is to identify motivations and outcomes from open and collaborative science through e-infrastructures. The e-infrastructure under analysis is Brazil's Virtual Herbarium that has herbaria as data providers and the scientific community in Brazil as target users.

A questionnaire was sent out to 99 Brazilian herbaria to identify drivers for sharing data and outcomes from this activity. Questions also included a SWOT Analysis (strengths, weaknesses, opportunities, and threats) concerning Brazil's Virtual Herbarium. This was followed by a face-to-face meeting where the preliminary report was presented and discussed. Fifty seven herbaria were involved in this study. According to the vision of these professionals, the Virtual Herbarium brought great benefits. Outcomes derived from sharing data through a public e-infrastructure included greater institutional recognition; greater involvement with graduate courses, increased number of visits to the herbaria; increase of the holdings; and, increase of grants.

The human network established is considered the project's most important asset. The capillarity of the network, with the participation of at least one herbaria from every state of the country, including small and regional herbaria, and the involvement with graduate courses are an important contribution of the project. Open sharing of textual data and images was viewed as a strength due to greater visibility and acknowledgement of the role and importance of herbaria.

The greatest threat mentioned was the discontinuity of the project due to the economic situation of the country and the lack of long-term public policies to support such an e-infrastructure.

The project also analyzed the contribution of users through an annotation system and a workflow to produce and publish ecological niche models on-line. For those that contributed their knowledge through the annotation system, the most important driver for participation is to contribute to improve data quality and to enable the use of the data in research. As to the distribution model, the motivation to produce and publish these models was for their own usage in new collecting efforts, for research, to publish articles, and for public policies.

An analysis was carried out to identify the reasons for not sharing (blocking) specific data fields. We believe that the best way to stimulate openness is that each data provider determines its own data policy, blocking what is considered sensitive data and only sending to the network open data that can be publicly viewed by all interested. Through this analysis, we found that, in some cases, the data provider is not documenting its data policy. We found a number of collections where, not only the reasons for blocking data were unknown, but also the fact that there was data being blocked was also unknown. This study showed that we have to improve the procedures and monitor blocked data.

An important product of this project was the analysis of usage and users that resulted in an on-line system with data usage statistics and a dynamic report with users' profile and what data is being used for. A greater involvement with policy makers and the private sector, identified in the user survey, may lead to better informed decisions and the development of open monitoring and evaluation mechanisms.

A report was also submitted to OCSDNet contextualizing openness using Brazil's Virtual Herbarium as a study case. A detailed description of the work and results of these studies are included in this report.

This project represented a unique opportunity to analyze outcomes from open data sharing. This work showed that not only is it important to develop and maintain e-infrastructures to increase access and usability of data for scientific development, but also to improve the quality, reliability, and completeness of data and information. It also showed the importance of the social network behind this process.

As to the future, there are still great challenges to overcome such as governance of such a network of people and institutions and its sustainability.

2. Research Problem

This project's research theme was motivations and outcomes in data sharing and open collaboration through e-infrastructures. Research questions were addressed to all data providers, users, and contributors to understand the motivations and outcomes from open data and expertise sharing. Questions envisaged when the project was proposed included:

- Has data-sharing through Brazil's Virtual Herbarium (BVH) lead to more recognition and support for data providers?
- Are official evaluating mechanisms considering data sharing as an important element and do they result in incentives to collect, organize, qualify and share data?
- Is data-sharing being affected by the way scientific production is evaluated and has this lead to inter-institutional, multi-discipline projects?
- What motivates crowd sourcing through the *Annotation system* and *BioGeo*?
- Are there mechanisms that could be used to increase participation?
- What are the reasons for blocking data?
- Who are the data users?
- For what purpose do they use the data and tools?

Research approach to answer these questions

In order to understand drivers that motivate herbaria to openly share their data and outcomes from this activity, an extensive analysis was carried out, first through the elaboration and application of a semi-structured questionnaire together with a SWOT analysis, indicating strengths, weaknesses, opportunities, and threats concerning Brazil's Virtual Herbarium. The answers to the questionnaire (39 answers out of 99) were analyzed and presented and discussed at a face-to-face meeting (35 herbaria represented out of 99) and a report was submitted to OCSDNet with the views of 57 herbaria (Annex 1).

As to crowdsourcing, the Virtual Herbarium developed two important mechanisms to allow users to collaborate: an annotation system and a workflow (BioGeo¹) that enables users to produce and publish species ecological niche models.

When the survey was prepared for the analysis of contributors of the annotation system, 622 comments had been received from 141 people. This tool is used for communicating and correcting errors or in identifying the material on-line. An example can be seen in Figure 1.

¹ BioGeo (Biogeografia da Flora e Fungos do Brasil) - <http://biogeo.inct.florabrasil.net>

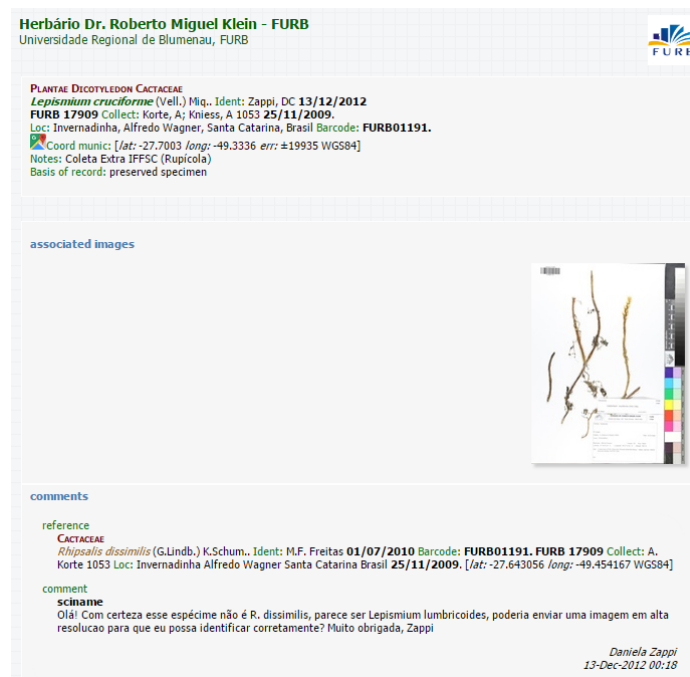


Figure 1. Example of a record with an associated comment

An analysis on the nature of the comments was first carried out and then an email was posted to all 141 users asking what was their motivation in using this tool. Only 20% answered. The answers were analyzed and a report was submitted to OCSDNet (Annex 2).

As to BioGeo, a specific survey was sent to all 177 specialists registered, even though only 43 had actually developed and published distribution models on-line. We received 17 answers, less than 10%, but with important contributions that will help guide future activities. A report with the analysis of the answers was submitted to OCSDNet (Annex 2).

The analysis about blocked data was expanded to include all biological collections that share their data through the *speciesLink* network and have identifiable blocked data fields on-line. An analysis of all blocked data was carried out for each data provider as to what data fields were blocked, if the species was included in a red list and the collection date. An email was sent to all 24 responsible curators, asking the reasons why the data was blocked. This study was important to realize that blocked data must be monitored as some curators didn't even know that the data was being blocked. A full report was submitted to OCSDNet and can be found in Annex 3.

As to the analysis of Data Users and Usage, the first action was to analyze usage statistics using: *AWStats*, a logfile analyzer (parameters include unique visitors, visitors, pages, hits, and bytes); and *Google Analytics* (geographic distribution of users). Analyzing log data for the year 2015, *speciesLink* network showed an average of 29 thousand users per month (*AWStats*), with about 95% of the users from Brazil (*Google Analytics*). These statistics do not show the amount of data that is being used, nor what percentage refers to the Virtual Herbarium, nor does it show who is using the data and for what purpose.

An interface to measure data usage based on logs of the search interface was developed with the support of the Virtual Herbarium project, motivated by the OCSDNet project (www.splink.org.br/showUsage). The report is presented in Annex 4. These statistics are also being used by individual herbaria to show their importance to their institutions.

The next step was to qualify usage. For this purpose, we studied the user survey carried out by Atlas of Living Australia, one of the most important biodiversity information e-infrastructures of the world. A simpler survey

was developed and published on-line. The results were dynamically shown on-line as users filled out the form. The survey was closed on January 09, 2017 and the results as online². This was publicized through the search interface, posted on CRIA's blog, and emails were sent to data providers. An analysis of the result of the survey was carried out and the report is presented in Annex 4.

Besides developing the project proposed to OCSDNet, we also attended demands from OCSDNet producing a position paper on how openness is unfolding in the political, legal, technical, social and cultural context in our community. This position paper was submitted to OCSDNet and is available in Annex 5.

3. Research Objectives and Findings

CRIA's project proposal indicated that outputs of this project should contribute to the following OCSDNet thematic research areas:

- Theme 1 (T1) - Motivations: the project should identify drivers that motivate herbaria as data providers to share their data through the e-infrastructure;
- Theme 2 (T2) - Infrastructure and technologies: The e-infrastructure and online tools available, local connectivity and IT support would be evaluated to indicate possible barriers to full participation;
- Theme 3 (T3) - Communities of Practice in Open and Collaborative Science: herbaria will be evaluated as to institutional policy and legal impediments concerning open data sharing; and
- Theme 4 (T4) - Potential impacts (positive and negative) of open and collaborative science: outcomes from participation to data providers and the diversity of uses and users of the e-infrastructure shall indicate potential or real impacts of open and collaborative science.

As to *motivations to openly share data*, the project identified drivers that motivate the participation of Brazilian herbaria in the network (Annex 1) and the motivation of data users to send their comments through the annotation system and to produce and publish species distribution models (Annex 2).

Herbaria at first participated in order to obtain resources to organize and digitize their holdings. With time they realized that by openly sharing data on-line they also obtained more recognition from their own institutions and from researchers from Brazil and abroad. This brought more collaboration and more involvement with students and, in some cases, with the private sector.

The motivation to send contributions through the annotation system was to improve the quality of the data. Whereas participation in developing and publishing species distribution models was seen as an important tool to plan new collecting efforts.

As to *infrastructure and technology*, the document *Contextualizing Openness - Brazil's Virtual Herbarium as an Open Collaborative Science Infrastructure* (Annex 5) presents a number of strategies or lessons learned that contributed to eliminate or reduce barriers to open data sharing. The design of the network is fundamental. Lessons learned includes data policy being determined by each data provider, but all data that is sent to the e-infrastructure will be openly shared. For this, the tools to be able to retain sensitive data were made available. In other words, each data provider must have full control over his/her data, determining what will be openly shared. Another lesson learned is that the complexity of the network in informatics must lie at the e-infrastructure's end and when possible, internationally agreed standards and protocols must be used.

As to Communities of Practice in Open and Collaborative Science, Brazil's Virtual Herbarium is structured as a collaborative network, where each "node" is important. As a result, there are more than 100 national herbaria openly sharing their data on-line, with at least one participating herbarium in every state of the

² See <http://inct.splink.org.br/dataUse?lang=en>

country, plus 21 herbaria from abroad. Ninety five percent of the Brazilian herbaria are associated to postgraduate courses, increasing the network to include teachers, researchers and students.

specieslink

191 datasets of Plants & Fungi in 2017

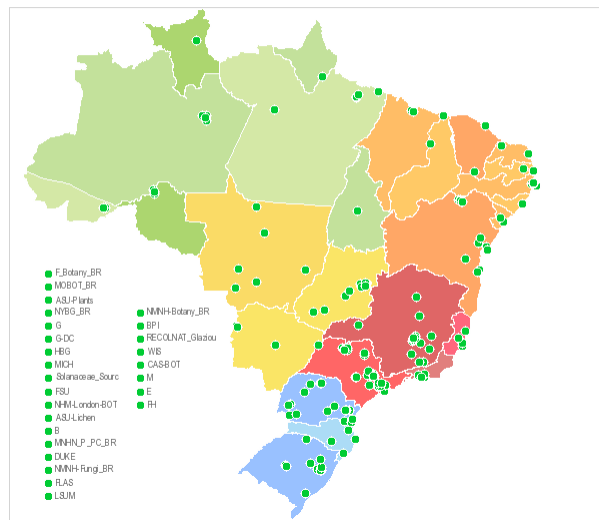


Figure 2. Geographic distribution of the participating herbaria of Brazil

As to *potential impacts (positive and negative) of open and collaborative science*, outcomes from participation were discussed by the participating herbaria and the results are presented in Annex 1. Figure 3 reflects the movement of data (entry and removal) in the network. Monthly averages are presented for both total online records and total georeferenced records. The red line shows the number of data providers per month.

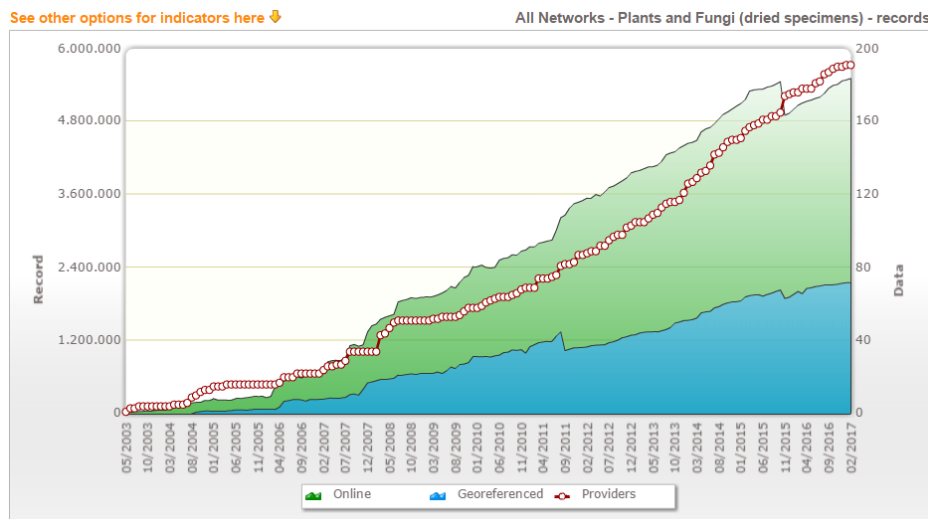


Figure 3. Evolution of the Virtual Herbarium

One can see the constant increase of data and data providers, with the exception of October 2015 when one of Brazil's largest herbaria decided to remove its data because they felt that they were losing visibility. This shows that there are still some cultural barriers to overcome.

Another important impact can be seen in Figure 4.

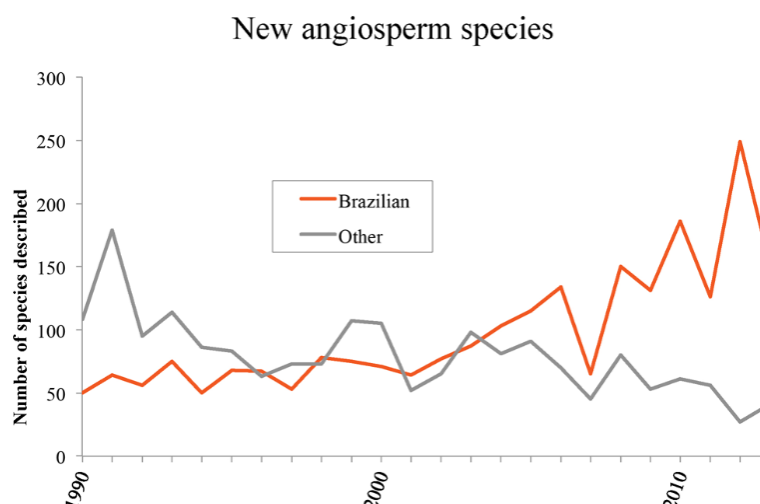


Figure 4. Number of angiosperms species described by Brazilian (orange line) and foreign (grey line) scientists from 1990 to 2013. (Canhos, et al. 2015))

The growth of angiosperm species in Brazil described by Brazilian scientists is clear. Public policies, the availability of data and the network of people surely contributed to this change.

When developing the e-infrastructure our main aim was to make data available on-line to all interested. A path between the data providers and the users of this data. We now realize that this is not an easy task as it implies in a cultural change and, in the beginning, it is not clear what the data provider, in this case, the team responsible for the herbaria, gains from this. To most, it seemed as much more work with very little to gain. Now it is clear that a lot is gained from organizing and sharing one's own data. Most data providers also became intensive users of data and were able to also share knowledge and benefit from other user's knowledge to improve their own data. Providing data is not a one-way road. Besides using and benefiting from feedback mechanisms, most herbaria found that visits to the facilities increased as did joint projects and research.

4. Project Implementation and Management

Type of Activity	How does this activity relate to your research and OCSDNet's objectives
1 Preparing a questionnaire to identify data sharing outcomes, sending it out, analyzing, and preparing a report	Identifying outcomes from on-line data sharing (Annex 1) is directly linked to the Virtual Herbarium and to OCSDNet's objectives
2 Analyzing user's participation through feedback mechanisms and crowdsourcing workflows (Annex 2)	Identifying motivations for voluntary participation helps in the design of new tools and is also linked to mechanisms of open collaboration
3. Analyzing blocked data (Annex 3)	Understanding why data is blocked helps in the development of mechanisms to unlock data
4 Developing tools to help analyze users and usage (Annex 4)	By creating tools to measure usage, and making this available online not only fulfilled the needs of the project but allows other types of analysis and uses by all interested

5 Contextualizing openness (Annex 5)	Helped analyzing practices that either benefits or may become barriers to open data sharing
--------------------------------------	---

Any Future Activities Planned?

The continuation of Brazil's Virtual Herbarium project in Brazil was approved. From CRIA's perspective, this includes a very heavy routine linked to maintenance and further development of the e-infrastructure. OCSDNet's project offered the necessary resources to study this initiative under a policy perspective and this was fundamental for a more profound understanding of what this project really achieved. We hope to find funds to continue this type of analysis.

5. Project Outputs and Dissemination

Participation in Workshops and Conferences where BVH was presented and specific products derived from the OCSDNet were addressed

Name of workshop	Objectives of workshop	Outcome(s) of workshop	Number of participants present	Any relevant links to event information
1 IUBS 2015 – Frontiers in Unified Biology	Symposium: Building a Biodiversity Informatics Agenda that will deliver a Unified Biology	As part of the round table “Challenges in building an infrastructure for all of biology”. CRIA presented <i>The importance of local infrastructures</i> , based on the Virtual Herbarium. A slide on the outcomes from sharing data with the work carried out for OCSDNet was presented	~80	
2 IDigBio Summit 2015	The Summit focuses on discussions of shared goals, challenges, opportunities, and collaboration.	CRIA's presentation was about Brazil's Virtual Herbarium: Outputs, Outcomes, and Challenges. The OCSDNet study was presented		https://www.idigbio.org/wiki/images/9/96/IDigBio-Summit-V_Brazil-Virtual-Herbarium_Canhos.pdf
Virtual Herbarium Planning Meeting	Discussion of the achievements, difficulties and future of the Virtual Herbarium	Presentation and discussion of the Outputs and Outcomes and SWOT Analysis	50	Report available on Annex 1.

Force16	Building bridges and connecting communities	Presentation: The impact of Brazil's Virtual Herbarium in e-Science		https://www.force11.org/media/video/impact-brazils-virtual-herbarium-e-science
---------	---	---	--	---

List of relevant publications

Name of Publication	Type (book, journal article, newspaper, blog, etc.)	Authors	Link
The Importance of Biodiversity E-infrastructures for Megadiverse Countries	Journal article (Plos Biology)	Dora A. L. Canhos , Mariane S. Sousa-Baena, Sidnei de Souza, Leonor C. Maia, João R. Stehmann, Vanderlei P. Canhos, Renato De Giovanni, Maria B. M. Bonacelli, Wouter Los, A. Townsend Peterson	http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002204
Destaques da rede <i>speciesLink</i> em 2016	Blog CRIA	Dora A. L. Canhos	http://blog.cria.org.br/2017/01/normal-0-21-false-false-false-pt-br-x.html
Uso dos dados da rede <i>speciesLink</i>	Blog CRIA	Dora A. L. Canhos	http://blog.cria.org.br/2016/08/uso-dos-dados-da-rede-species-link.html
Assessing the benefits of data sharing: the experience of Brazil's Virtual Herbarium	Blog (OCSDNet)	Dora A. L. Canhos	http://ocsdnet.org/assessing-the-benefits-of-data-sharing-the-experience-of-brazils-virtual-herbarium/

Other important links that capture project findings/impact. (Ex: Social Media activity, photo galleries, educational resources, websites used for project dissemination, blogs, etc.)

URL	Content of URL
http://inct.splink.org.br/showUsage	Statistics of Data Usage for the <i>speciesLink</i> network, informational base of Brazil's Virtual Herbarium
http://inct.splink.org.br/dataUse	Survey about the use of data from the <i>speciesLink</i> network

6. Community Building

Brazil's Virtual Herbarium was built upon three existing initiatives that were decisive to the success of the project:

- Brazil's National Research and Education Network – Rede Nacional de Ensino e Pesquisa (RNP);
- The *speciesLink* Network; and,
- Brazil's Botanical Society – Sociedade Botânica do Brasil and its network of herbaria.

RNP provides internet connectivity to practically all university and research centers of the country. It also hosts CRIA's information system of public interest in its Internet Data Center. Due to this partnership, all hardware are installed in one of the best places possible in terms of connectivity (stability, speed, and uninterrupted operation) and infrastructure (electricity, refrigeration and security). CRIA continues to develop and maintain all systems.

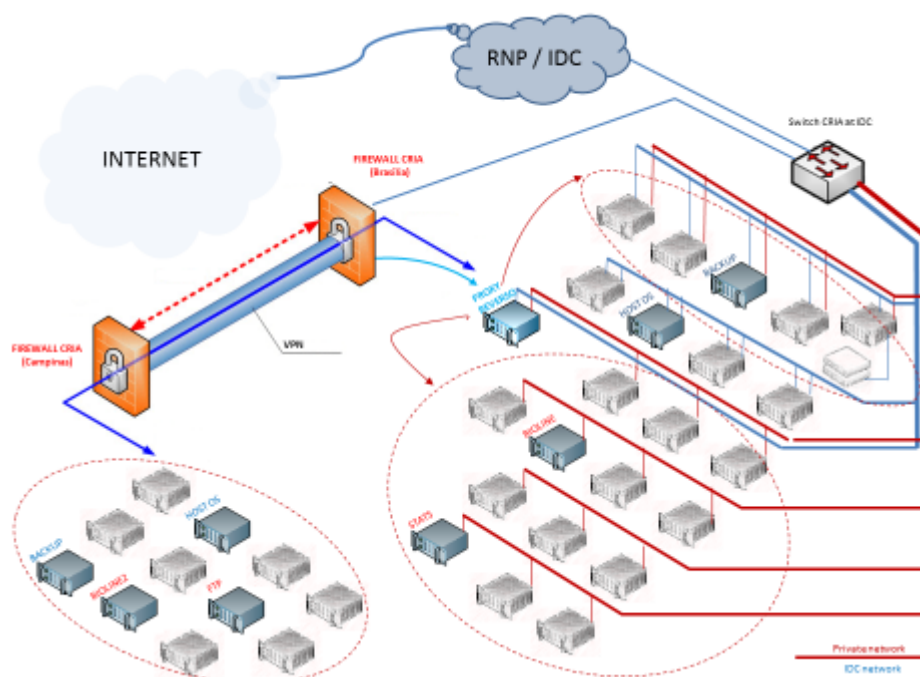


Figure 5. Network with CRIA's information systems at IDC/RNP

speciesLink is the information platform adopted by the Virtual Herbarium. Its development began in 2001 to integrate and openly serve data from biological collections first from São Paulo State and latter expanded to the whole of the country. In 2006, thanks to collaboration with the Botanical Gardens of New York and Missouri, *speciesLink* began integrating data of material collected in Brazil that is held in collections in other countries (data repatriation). Therefore, the Virtual Herbarium project began with a network of 25 herbaria and an information platform with 1.9 million data records on-line. Establishing new partnerships was far easier with a working e-infrastructure in place, as many cultural barriers had already been overcome.

Brazil's Botanical Society also plays an essential role in Brazil's Virtual Herbarium in organizing the botanical community. One of its commissions is the Brazilian Network of Herbaria³ that maintains a catalogue with contact information of all herbaria in the country. It also holds its national congress every year, which is an enormous opportunity for the Virtual Herbarium community to congregate and present new developments. At the annual congress, Brazil's Virtual Herbarium steering committee holds a face-to-face meeting, open to all interested and a round table session, where new developments and achievements are presented.

The e-infrastructure also brings the community together, both as data providers and users. An important activity carried out within the context of this project that certainly helped the community was the

³ See <http://www.botanica.org.br/rbh>

questionnaire and SWOT analysis carried out to identify possible outcomes from data sharing and to plan future developments.

The user survey carried out within this project's scope also brought the user community and its needs closer to CRIA's development team and the project's steering committee.

7. Impact

This project's greatest impact was on the herbaria (data providers), that had the opportunity to reflect upon outcomes derived from their participation in the network. Strengths and weaknesses of the Virtual Herbarium as well as possible or real opportunities and threats (SWOT Analysis) were also discussed. This reflection was at first individual, answering a questionnaire and at a second moment at a face-to-face meeting carried out in three phases: (1) presentation of the answers to the questionnaire; (2) small group meetings to discuss the main points; (3) group presentations and plenary discussion. All contributions were received and a report submitted the Virtual Herbarium Coordinating Group and to OCSDNet. These activities helped program the Virtual Herbarium's next phase.

The Virtual Herbarium already had a number of indicators that helped evaluate its developments and outputs. Quantifying and Qualifying usage was an important indicator added thanks to this project. An important indicator of this project's relevance and impact was the fact that all analysis produced for OCSDNet were translated to Portuguese and added to the Virtual Herbarium's project reports.

8. Reflective Learning on Internal Dynamics:

- a. What has been the successes and main challenges of your project (in terms of the way that the project was conducted)?

Our main success or achievement was the analysis carried out to identify outcomes from data sharing involving almost 60% of the participating herbaria. Perhaps even more important than the results, was the process. The network is stronger with most herbaria completely involved in achieving common goals.

Our main challenge was to follow and meet the demands of OCSDNet, as these activities were not included in our proposal, which, in itself, was very demanding. We are a very small team, with only one person working on the interface between information technology and policy. Therefore, our focus was to carry out the activities indicated in the proposal, while OCSDNet demanded a more dynamic interaction.

- b. Does your team have mechanisms in place to capture these lessons and share them internally? If so, which mechanisms and how have they benefited your project?

All developments involved the team directly as they required the development of tools. Besides CRIA's staff, this project benefitted the steering committee and involved herbaria curators and users in general that answered the survey and with this were able to make their needs and ideas known.

- c. Please reflect on internal project power dynamics and its influence in project development and outcomes. How did you observe power dynamics to play out? How might north-south relations within the project have also played a role? *(Note, this is not for individual naming and blaming but rather a self-reflective way to try to understand how power affects research collaboration.)*

Practically, all that was developed and discussed was openly shared. The process empowered the participating herbaria to make their ideas better known and to discuss the future, and most probably, impact the future of Brazil's Virtual Herbarium.

9. Recommendations (for OCSDNet):

- a. How did you find the (experimental) network model that was used by IDRC to administer the OCSDNet subprojects? What were notable strengths and weaknesses you experienced?

There may have been an incomprehension from our side. When the project began, we thought our project would offer insights on open data sharing, just by carrying out the activities planned and reporting the results. We did not realize it would be expected of us to have a more participative role in OCSDNet. For this reason, with the exception of face-to-face meetings, our direct participation was kind of marginal. This is not due to lack of interest. We had to focus on our project's activities to be able to deliver what had been proposed. We unfortunately did not have the extra time to interact with the OCSDNet team and other subprojects.

- b. In your experience, how might a culture of shared learning be fostered/improved for future iterations of a network such as OCSDNet?

It is interesting to involve very different projects from different countries and cultures to learn from different experiences, but this is not an easy task. In our case, perhaps we could have contributed more if the expected interactions and direct contributions to OCSDNet during the project had been made clearer upfront as part of the project's activities.

- c. Has feedback from members of the network had an impact on your research? (consider insight from the coordination team, advisors and peers in the network).

Even though we did not have enough time to become involved with the team as we should have, our deliverables were affected by the insights of the coordination team, advisors, and peers. Questions asked and documents requested certainly made us think about our practice and how this affects or affected open data sharing.

- d. Do you have any other advice/feedback that you would like to provide to OCSDNet or IDRC? (consider modes of communication, evaluation, etc.)

I think I have already stated what I would like to share.

10. Additional Comments

We would like to thank IDRC and the OCSDNet team for this fantastic opportunity. This project gave us the resources and therefore the time to reflect about the outcomes of data sharing within the framework of the Virtual Herbarium. During these 15 years of development of the *speciesLink* network, our focus has been on outputs such as the amount of data shared, the number of data providers, the number of hits, visits and so on. The possibility to look into data usage and on the effects of open data sharing for data providers has made the Virtual Herbarium much stronger and integrated.

ANNEX 1. IDENTIFYING OUTCOMES FROM ON-LINE DATA SHARING

One of the project's objectives is to identify possible drivers that motivate herbaria to openly share their data through an e-infrastructure and possible outcomes of this participation. One of the central research questions of this project is "Has data sharing through the Brazilian Virtual Herbarium (BVH) led to more recognition and support for data providers?" The Brazilian Virtual Herbaria is one of the country's National Institutes of Science and Technology, a program of the Ministry of Science, Technology, and Innovation.

1. Methodology

Together with the BVH's steering committee, we elaborated a semi-structured questionnaire with both open and multiple-choice questions concerning possible outcomes from sharing data on-line.

We also carried out a SWOT analysis, requesting of each curator the Strengths, Weaknesses, Opportunities, and Threats concerning the Brazilian Virtual Herbarium.

The questionnaire was sent by email by the project's coordinator to curators of all 99 herbaria associated to the network. We received 39 answers. Each herbaria was classified in 5 separate groups, according to the size of its holdings:

1. Up to 10 thousand vouchers;
2. Between 10 and 50 thousand vouchers;
3. Between 50 and 100 thousand vouchers;
4. Between 100 and 200 vouchers; and,
5. With more than 200 thousand vouchers.

The purpose of this "classification" was to evaluate if there were issues specifically related to the size of the herbarium.

All answers were tabulated and a report with the analysis of the answers was prepared and discussed with BVH's steering committee and presented at the evaluation and strategic planning meeting held in Belo Horizonte on April 14 and 16, 2015⁴.

Thirty five herbaria were present at the meeting, and a new round of discussions was carried out in smaller groups and were presented in plenary. All this material was handed in and used to produce this report.

This report is the result of the opinions of 17 herbaria that answered the questionnaire and participated at the meeting in Belo Horizonte, 22 herbaria that answered the questionnaire but were not present at the meeting, and 18 herbaria that did not answer the questionnaire but took part of the meeting. Therefore, this study includes the opinion of curators from 57 herbaria, which at the time represented 58% of all associated herbaria of the network.

⁴ See program in Portuguese at http://cria.org.br/eventos/inct_2015/program

2. Results and Discussions

Outcomes

Table 1 presents the answers given by curators concerning possible outcomes derived from sharing their data through the e-infrastructure BVH.

Size		< 10.000	10.001 - 50.000	50.001 - 100.000	100.001 - 200.00	>200.000	All
Number of Herbaria	No.	27	39	17	9	7	99
Number of Answers	No.	12	12	7	5	3	39
	%	44%	31%	41%	56%	43%	39%
Greater institutional recognition	No.	11	10	3	5	3	32
	%	92%	83%	43%	100%	100%	82%
Greater involvement with graduate courses	No.	9	8	5	2	2	26
	%	75%	67%	71%	40%	67%	67%
Increase in the Number of Visits	No.	10	12	6	2	3	33
	%	83%	100%	86%	40%	100%	85%
Increase of the holdings	No.	11	8	6	2	3	30
	%	92%	67%	86%	40%	100%	77%
Increase amount of grants	No.	6	6	3	2	3	20
	%	50%	50%	43%	40%	100%	51%

39% of all herbaria associated to the BVH answered the questionnaire. Outcomes derived from sharing data through a public e-infrastructure included (1) greater institutional recognition; (2) greater involvement with graduate courses, (3) increased number of visits to the herbaria; (4) increase of the holdings; and, (5) increase of grants.

As to being recognized or acknowledge by their own institution, the answers show that larger institutions are fully recognized. This makes total sense as the costs are much greater and an annual budget must be secured. This is not necessarily true for small herbaria, mostly in universities, that do not even have a position for curators. The lack of recognition of the work or even of the existence of these herbaria by the host institution was always presented as a mayor problem of smaller collections. Therefore, the result of the survey indicating that 92% of herbaria with holdings of up to 10 thousand vouchers stated that sharing their data through the e-infrastructure gave them more visibility and institutional recognition is an important outcome of the project.

An important aspect of the network is that 95% of the participating herbaria are associated to graduate courses. The use of data and tools available in the Virtual Herbaria have become a routine in graduate courses such as botany, taxonomy, and ecology. By organizing and publicizing data of its holdings, herbaria have become more involved with graduate programs. Once again, looking at the smaller herbaria with up to 10 thousand vouchers, one can see that their involvement with the graduate courses increased. Many also indicated that by exposing the data of small, but geographically specific holdings, they attracted the interest of students and specialists. With this, the number of visitors increased as did the number of new samples

deposited in their herbaria. These are important outcomes directly influenced by sharing data through the e-infrastructure.

Another major problem for smaller herbaria is external funding. With greater visibility and, in many cases, by submitting proposals as a network, 50% of the smaller herbaria with holdings under 50 thousand vouchers were successful in receiving external grants. However, not only did the small herbaria benefit from sharing the data of their holdings in an open platform, larger herbaria also acknowledged a great impact in the number of visits, holdings, and grants. Larger herbaria also manifested that their internal organization was improved and overall planning and setting goals to be achieved was also enhanced as data was made available on-line. By sharing their data on-line and by using all tools available for analysis, herbaria could work on data quality and plan future collecting efforts.

The increase of the holdings (77% of the herbaria) can be attributed to its greater visibility, its increased involvement with graduate students, and the increased interest of specialists in visiting the area where the herbaria is based. Some herbaria answered that besides the increase of the number of visitors, these are more diverse – both from different fields of knowledge and from different geographic areas.

SWOT Analysis

Curators were requested to indicate what they considered were strengths, weaknesses, opportunities and threats concerning the Brazilian Virtual Herbarium. Strengths and weaknesses referred to actions within the control of the network and opportunities and threats referred to external factors that are not within the control of the network but are important elements that must be monitored.

STRENGTHS

All herbaria emphasized as strengths the social network, the value of data sharing, and the financial, technical and scientific support offered through the project.

Social Network

The social network established and strengthened throughout the project promoted increased interaction between curators and technicians from different institutions. Answers indicated that there was a change in the mindset of the professionals involved that now feel more valued and part of the achievements of the project. Increased self-esteem and a constant search for improvement was also noted. The increased geographic coverage of the network, with the participation of small herbaria, was emphasized, as many of these are regional collections, whose copies are underrepresented in other collections. Curators also indicated increased collaboration with students and researchers from other courses and institutions, and the visit of foreign researchers.

Data Sharing

Open sharing of textual data and images was viewed as a strength due to the greater visibility and acknowledgement of the role and importance of herbaria. Outcomes such as greater institutional recognition and deposits of new material (graduate students and researchers) were once again mentioned. On-line organization of data and the availability of tools to help find errors and inconsistencies were also mentioned and contributed to the improvement of the quality of the data that is being shared. An important observation mentioned was that data organization and on-line sharing also increases the knowledge curators have of their own holdings and enables better planning and the elaboration of strategies to increase and improve these holdings.

Project Support

The existence of the project with the support of the Brazilian government (CNPq⁵) with funds for grants, materials, equipment, and for courses and visits of specialists was pointed out as being fundamental for the organization, digitization, and improvement of the holdings.

The fact that the project was developed from existing initiatives was considered a strength. These initiatives are the Brazilian Network of Herbaria of the Brazilian Botanical Society; the speciesLink network developed by CRIA; and, the Brazilian National Research and Educational Network (RNP).

WEAKNESSES

The most important weakness cited by all curators refers to human resources. Not only are they insufficient, but specialists that are retiring are not being replaced. Even though the grants to hire students and technicians to work on the organization and digitization of the collections were mentioned as a strength of the project, here they state that these grants are transitory and for limited periods.

The same applies to infrastructure and the necessity of more training programs. The project promoted yearly meeting at the Congress of Botany and also held 2 general meetings (the first with representatives of 70 herbaria and the second with 35) to present and evaluate what was done and to help plan the future. Many herbaria indicated that it would be important to hold more such meetings and this way guarantee a more participatory process.

OPPORTUNITIES

The possible continuation of the federal government's program of National Institutes of Science and Technology is seen as an opportunity for continuity.

Making data freely and openly available on-line is seen as an opportunity for new research insights and for the advancement of e-taxonomy, valuing the role of herbaria.

The possibility of hiring professionals that were trained throughout the project is also seen as an opportunity to ensure the transfer and multiplication of acquired knowledge.

The advancement of information and communication technology is also seen as an opportunity to enhance the quality of the content shared on-line and to increase the interaction between herbaria and data users.

THREATS

Perhaps the greatest threat mentioned was the discontinuity of the project. Within this line, another point was the duplicity of similar projects, as opposed to collaborating and networking with existing initiatives.

The economic situation of the country and the lack of long-term public policies to support such e-infrastructures were considered threats.

3. Final Comments

This document synthesizes the opinion of curators from 57 herbaria associated to the Brazilian Virtual Herbarium. According to the vision of these professionals, this initiative brought great benefits and should continue.

The human network that was established is considered the project's most important asset. The "visiting specialists program" that used on-line data to identify the herbaria to be visited and specialists required, the

⁵ National Council for Scientific and Technological Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*)

courses and technical visits and remote support given were actions that strengthened this human network with a sense of being part of the group.

The capillarity of the network, with the participation of at least one herbaria from every state of the country, including small and regional herbaria, and the involvement with graduate courses are an important contribution of the project.

Therefore, returning to our central research question “Has data sharing through the Brazilian Virtual Herbarium (BVH) led to more recognition and support for data providers?” the answer is yes.

ANNEX 2. ANALYSIS OF USERS PARTICIPATION THROUGH THE ANNOTATION SYSTEM AND BIOGEO

The Brazilian Virtual Herbarium (BVH) is one of the country's National Institutes of Science and Technology, a program of the Ministry of Science, Technology, and Innovation. It integrates 160 datasets from more than 100 herbaria of Brazil and 11 from abroad. More than 4.8 million records associated to more than 685 thousand images of vouchers, live material and pollen are freely and openly available to all interested.

The project developed two important mechanisms to allow users to collaborate:

- An annotation system; and,
- A workflow (BioGeo) to enable users to produce and publish species ecological niche models.

Questions include:

- What motivates users to send their comments (*Annotation System*)?
- What motivates researchers to produce and publish their models through *BioGeo*?
- Are there mechanisms that could be used to increase participation?

1. Annotation System

Within the Brazilian Virtual Herbarium project, feedback mechanisms were developed to allow users to send their comments about specific data records. When clicking on the “new comment” icon, a window pops up for users to provide their input (figure 1).

comment to the curator

reference **PLANTAE FABALES FABACEAE**
Inga edwallii (Harms) T.D.Penn.. Det: Mello, AS 26/10/2013 FLOR 55741 Collect: Guedes, C; Mello, AS 01 26/10/2013.
Loc: RYY, borda da floresta, trilha Apekatu, Itapoá, Santa Catarina, Brasil
[lat: -26.1169 long: -48.6161 err: ±18121 WGS84] Altitude: 5m.
Notes: hábito arbóreo, 7m, flor branca (estames), sem fruto, poucos folíolos; Floresta Ombrófila Densa de Terras Baixas

name

email

subject scientific name

comment scientific name
identification
geography
other

send

Figure 1. Popup window to enable users to send comments to curators

As can be seen on figure 1, the system presents the full data record and users must fill out the form with their name and email, select the subject – scientific name, identification, geography, and other – and write their comment. When clicking on **send** the comment is sent to the email indicated by the user for

confirmation. Once confirmed, the curator receives the email that is also archived in a database. Figure 2 shows a record with an associated comment.

Herbário Rondoniense - RON
RON Universidade Federal de Rondônia

PLANTAE SAMAMBAIA BLECHNACEAE
Blechnum
RON 8523 Collect: Bigio, N.C.; Daly, D.; Lima, A.S.; Oliveira, E.C.; Silveira, A.L.P. & Coelho, H.A. 1173 **05/12/2013**.
 Loc: Beira da Estrada, mata de Igapó., Vilhena, Rondônia, Brasil Barcode: **RON00008523**.
 Coord orig: [lat: -12.998514 long: -60.391578 WGS84]
 Basis of record: preserved specimen

associated images

comments

reference
PLANTAE SAMAMBAIA BLECHNACEAE
Blechnum **RON 8523** Collect: Bigio, N.C.; Daly, D.; Lima, A.S.; Oliveira, E.C.; Silveira, A.L.P. & Coelho, H.A. 1173 **05/12/2013**.
 Loc: Beira da Estrada, mata de Igapó.
 Vilhena Rondônia Brasil Barcode: **RON00008523**.
 [lat: -12.998514 long: -60.391578 WGS84]

comment
identification
 Camarada Narciso, Isto não é um *Blechnum*, mas sim *Thelypteris interrupta* (Willd.) K. Iwats. (Thelypteridaceae). Abraços!

Vinicius A.O. Dittrich
 10-Oct-2015 01:24

new comment

Figure 2. Record RON 8523 with an associated comment.

Figure 2 shows that the herbarium RON has voucher 8523 identified as *Blechnum* and the comment indicates that it is *Thelypteris interrupta* (Willd.) K. Iwats. (Thelypteridaceae). This comment was sent on October 10, 2015, but the last time this database was updated was in June 2015. Even though the record was not altered by the curator, it appears as an on-line annotation associated to the specific record. Users can therefore benefit from a specialist comment even before the data has been revised and altered. Users can also check this information as there is an image associated to the record.

Methodology

When this survey was prepared, the system had received 622 comments from 141 people. 473 comments referred to the scientific name, 68 to the identification of the material, 59 to the geographic data, and 22 classified as "other". 85.5% of the comments referred to data records of plants, 13.5% to animals and 1% to microorganisms.

An email was posted to all 141 users who sent their comments through the annotation system asking what was the motivation for using this tool. In order to facilitate the analysis, six options were offered:

- (1) Contribute to the improvement of the quality of the data;
- (2) Correct errors in order to enable the use of the data in their research;
- (3) Correct errors in order to use the data in the BioGeo workflow;

- (4) Check the determination and/or geographic information to use this information in the List of Species of the Brazilian Flora;
- (5) Check the determination and/or geographic information to use this information in the red list assessment (CNCFlora);
- (6) Others. In this option, users were asked to specify what other reasons they had.

People could choose more than one option. We also asked whether the herbarium accepted their comments and corrected possible errors, asking them to choose one of the four options below:

- All records were corrected
- Most records were corrected
- Some records were corrected
- No record was corrected

Results

Of the 141 emails sent, we received 20 answers, around 14% of the total.

- 85% indicated that their motivation was to contribute to the improvement of the quality of the data
- 50% to correct errors in order to enable the use of the data in their research
- 5% to correct errors in order to use the data in the BioGeo workflow
- 5% to check the determination and/or geographic information to evaluate the species' distribution and include this information in the List of Species of the Brazilian Flora

No one indicated the use of the tool to use the data in the red list assessment and no other motivation was included.

As to whether, to their knowledge, the collections are benefiting from their comments to correct possible errors, only 16 of the 20 specialists answered this question.

- 15% indicated that all records were corrected
- 15% indicated that most records were corrected
- 15% indicated that some records were corrected
- 15% indicated that the records were not corrected
- 20% indicated that they do not know whether the data was corrected

Comments

The most important driver for participation is to contribute to improve data quality and to enable the use of the data in research. It is probable that the 20% that did not answer the second block probably did not check to see whether the records were modified. However, we can conclude that 60% did not only contribute with their comments but also checked to see if the data was changed.

2. BioGeo (Biogeography of the Flora and Fungi of Brazil)⁶

BioGeo is a system developed to expand the knowledge about biogeography of plants and fungi of Brazil, using modeling techniques of potential distribution and counting with an active participation of specialists. A diagram of the workflow is presented in figure 3.

⁶ See <http://biogeo.inct.florabrasil.net> (in Portuguese only)

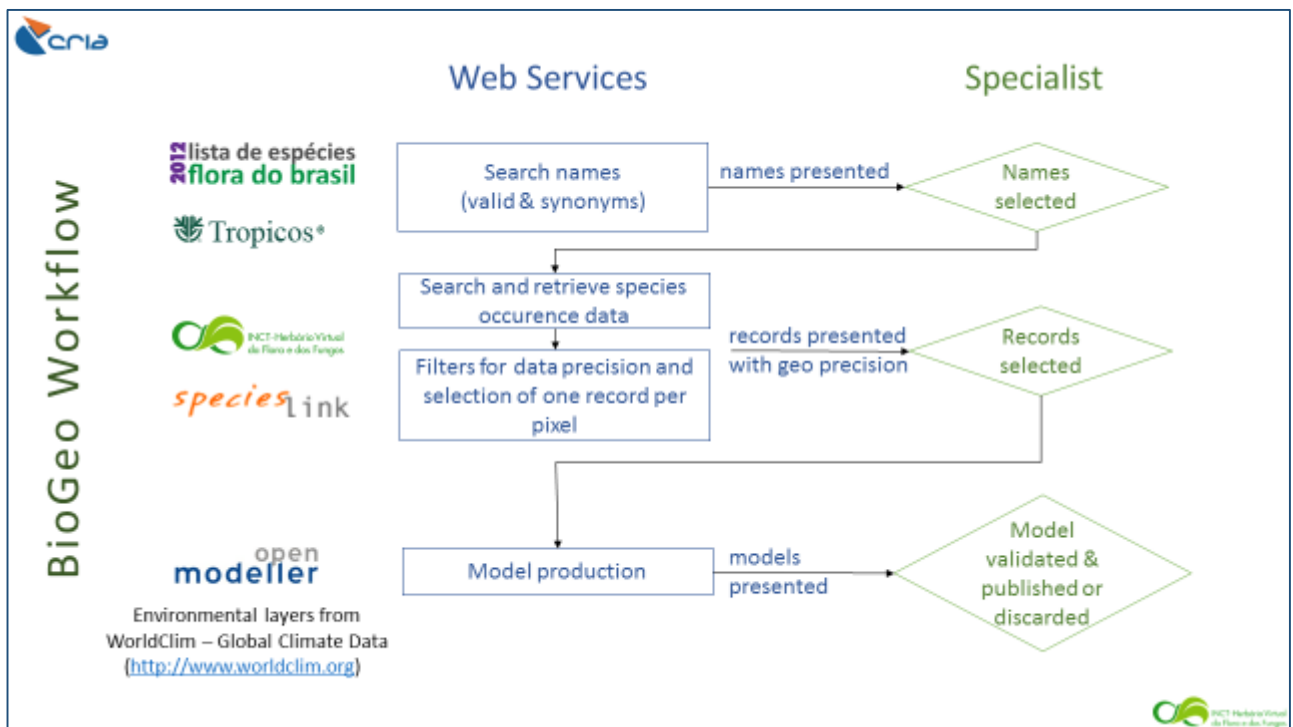


Figure 3. BioGeo Workflow

When a specialist registers in the system, he/she indicates the species or genera that he/she wants to model. The process begins when a specific species is selected. Through the workflow, the system using web services for the List of Species of the Brazilian Flora (provided by CRIA) and Tropicos (provided by Missouri Botanical Garden) presents a list of names (valid names and synonyms) to the specialists who selects those to be included for searching. The name as searched through the speciesLink web services and the results go through a filter for data precision (lat/long) that selects one record per pixel. Records selected by the system and all other records found are presented to the specialist who defines which records will be used in modeling. Depending on the number of point data, 1 to 5 different algorithms are used and models are produced using the openModeller web services and WorldClim data. The resulting models, together with a consensus model, are presented to the specialist who then decides whether it should be published or discarded (Figure 4).

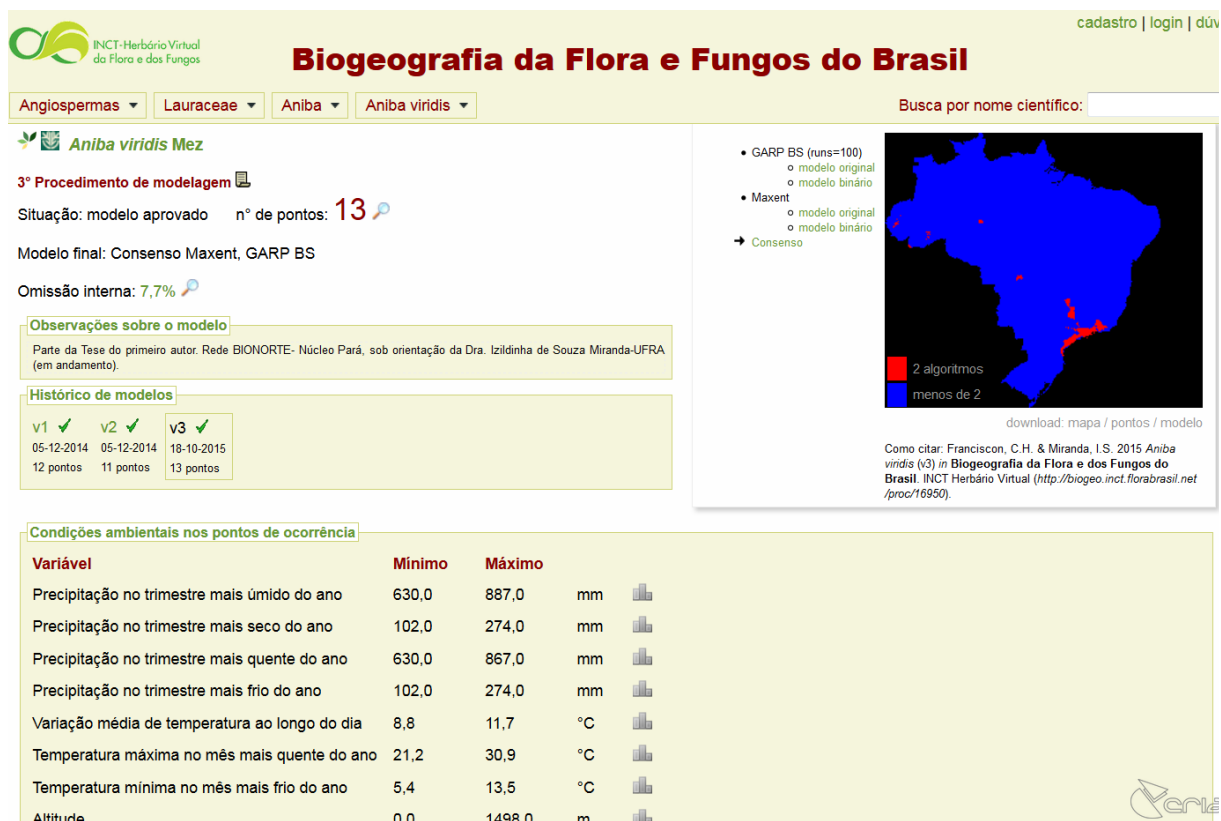


Figure 4. Example of a distribution model published on-line

Methodology

When the survey was sent out, there were 177 specialists registered in the system, meaning all were authorized to produce distribution models. Nevertheless, only 43 had actually published models on-line.

An email was sent to all 177 specialists who were asked to fill out the following information:

- (1) Institution
- (2) Academic level
- (3) Field of interest
- (4) If registered and did not publish any model, what was the impediment
- (5) Use of the model published through Biogeo
 - a. Planning new collecting efforts
 - b. Public policies
 - c. Articles
 - d. Others (please specify)
- (6) Weaknesses of BioGeo
- (7) Strengths of BioGeo
- (8) What would you like to see in BioGeo in the future

Results

We received 17 answers from 10 PhDs, 4 students doing their PhD, 2 Masters and 1 Bachelor (biology).

Nine (53%) did not publish their models for diverse reasons:

- They are still compiling the data
- They just carried out some tests

- They do not like the idea of having another specialist authorizing their participation – NOTE: this is not correct and an email was sent to this specialist to clarify this point.
- Problems with understanding the system
- The resulting model did not answer the hypothesis raised about the species
- Lack of time
- I am not a specialist

As to the use of the models:

- Two (12%) have not used any model available
- Four (24%) have used the models to plan new collecting efforts
- Three (18%) for public policies
- Seven (41%) to publish articles
- Two (12%) indicated using the models for their research (this was under “others”)

Weaknesses

- The system does not include data from other South American countries (restricted to Brazil)
- The fact that non specialists can generate models
- Many specialists in biogeography are not participating. Suggests a better communication strategy to make BioGeo known.
- Data of the models can only be exported in xml. It would be interesting to have other formats such as ascgrid and float.
- The models use a fixed set of environmental layers. It is not possible to select layers, algorithms, and other parameters.
- Not intuitive
- Insufficient data, this is not a problem of the system but it affects the quality of the models generated
- Not being able to project future scenarios with climate change

Strengths

- Easy to use (5 – 29%)
- Available distribution models
- The system that generates the maps is very good
- The system is fast and the graphic output of the models is good
- Extremely easy to use, principally for decision making, policy or research
- Many species have models
- Data sharing without restrictions
- Standardization, enabling the comparison of models
- Treatment of the occurrence points to generate the models
- Reduction of processing capacity of desktops to generate models
- Utility in planning new collecting efforts
- Visualizing the models, data used, liberty to select the data (validate or not and justify)
- Facility to manipulate and interpret
- Potential for diverse uses, both academic and for public policies

Future Requirements

- Inclusion of data from other South American countries
- Historical data about areas where specimens were collected
- Specialists that participate in the Flora of Brazil initiative should be invited to participate

- A link to data use, restrictions, models and results to give proper credits and stimulate new contributions
- A feedback mechanism for public policies
- Distance training in GEOstatistics
- Possibility of selecting geographic areas (such as states, regions, phytogeographic limits, among others) and bioclimatic layers.
- Where specialists are listed, include the species and families they are modeling
- Enable projection for scenarios of climate change

Comments

Although the number of answers was small (less than 10%) there are important contributions that can help guide future activities.

The number of people that registered compared to the number of specialists that are actually producing models indicate that there may have been a misinterpretation as to the usage of the system. It seems that people may have thought it necessary to register in order to access the models. This seems plausible when some of the answers received by those that have not published any model includes “I am not a specialist”.

One of the motivation in participating is obviously to use the model that was produced. Our main focus as to data users is the scientific community. Therefore, it is interesting to recognize the usage of the model to plan new collecting efforts, for research, to publish articles, and for public policies. These were all aims of this tool.

As to weaknesses, as the focus of the system was to help plan new surveys in Brazil, the geographic scope was Brazil and with current environmental conditions. It would be interesting to further develop the system increasing the geographic scope to South America – more data and possibly a better resulting model – and to build scenarios considering climate change.

A weakness mentioned referred to the necessity of a better communication strategy, as important specialists are not involved. This is true. The system was not publicized as it was under development and now, as the project ended, it is just being maintained. If we are able to obtain new grants, a communication strategy must be in place.

As to strengths, although one user said the system is not intuitive, five said it is easy to use. Given the fact that the system shares distribution models for 3.562 species without having provided any training courses, one can assume that it is intuitive for a knowledgeable person. Table 1 shows the number of species with models per taxonomic group.

Table 1. Species distribution models produced through BioGeo

Taxonomic Group	No. of Species in the List of Brazil	Species with Distribution Models	%	No. of Approved Models	No. of discarded models	Models awaiting approval
Algae	4.747		0			
Angiosperms	32.831	3,471	10.6%	4,046	126	147
Bryophytes	1.524	5	0.3%	5	3	12
Fungi	5.712	10	0.2%	10	6	1
Gymnosperms	30	4	13.3%	4		
Ferns and Lycophytes	1.253	59	4.7%	68	5	7
Total	46.097	3,549	7.7%	4,133	140	167

Despite the lack of a more substantial dissemination of BioGeo, almost 8% of all fungi and plant species that occur in Brazil have a distribution model published and openly available on BioGeo.

It is interesting to see that one user expressed as strength of the system the fact that data is shared without restrictions.

The answer as to future requirements certainly will help us when writing a new proposal for funds to enable the continuity of this initiative.


1. Introduction

When participants of the *speciesLink* network began a more systematic discussion on sensitive data, collections were basically concerned in not sharing geographic coordinates of endangered species data. To preserve this information, the general procedure was to not send the data record to the network. On the other hand, when a species is included in a so-called Red List, the government must plan specific actions viewing its conservation. Therefore, the occurrence data of threatened species is fundamental in planning their conservation.


However, some consider informing precise geographic coordinates of endangered species something that might contribute to their extinction, whereas others consider this a vital information for social protection of that species. Therefore, there is no right or wrong. Both visions have good arguments to block or share the data. This is when *speciesLink* changed from a centralized data policy to a distributed one, only making it clear that all data sent to the network would be freely open and accessible to all.

When changing its strategy from “all data must be open” to allow curators to hold back data they considered sensitive, CRIA wanted a system that would allow a partial retention of data, as opposed to retaining full datasets or data records. There were no mechanisms to block specific data fields. If the data provider did not want to share the geographic coordinates of a record, the whole data record would be retained, and users simply would not know that that data existed.

When the *speciesLink* network began to deal with sensitive data, there were two reasons for not sharing geographic coordinates: data of endangered species or data that had not been published. As of 2008, if, for example, a curator wishes only to hold back precise geographic coordinates, this can be done. An extra column is added to the spreadsheet or database used by the collection, where curators can mark the records or just the geographic coordinates that will not be sent to *speciesLink*. When data fields are marked, information that they have been blocked is sent, so users can distinguish “blocked data” from “no data” (figure 1).



Herbário Vale do São Francisco - HVASF
Universidade Federal do Vale do São Francisco



UNASF
UNIVERSIDADE FEDERAL DO
VALE DO SÃO FRANCISCO

PLANTAE PTERIDACEAE
Acrostichum danaeifolium Langsd. & Fisch.. Det: S.R.S. Xavier **23/08/2010**
HVASF 8818 Collect: A.P. Fontana 6652 **13/05/2010**.
 Loc: Fazenda Bom Sucesso, propriedade do Neto, Sobradinho, Bahia, Brasil
 Altitude: 500m.
 Sex: U
 Notes: Ereta, ca. 1,8m. Soros acrosticoides marrons.
 Basis of record: preserved specimen Coord.: Bloqueadas

associated images










Figure 1. Data record with blocked coordinates – “Coord.: Bloqueadas” (speciesLink, 2016)

As this tool was developed about nine years ago, as part of our OCSDNet project, we evaluated all collections that have blocked data and asked curators why the data was retained. The intention is to see whether there is a policy issue that could help open the data for some types of data retention.

2. Herbaria with blocked data fields

ESA – Herbário da Escola Superior de Agricultura Luiz de Queiroz

- Estimated holdings: 145,000
- On-line records: 123,405
- Blocked locality data fields: 1,035 records (0.8% of on-line records)
 - 704 records also have blocked coordinates

Characteristics of these 1,035 records:

- 735 with images
- 177 distinct species
- 3 records included in Brazil’s list of endangered species
- 690 data records of Orchidaceae
- 344 data records of Bromeliaceae
- 1 unidentified material

The endangered species are *Epidendrum zappii* (2 records) and *Pseudolaelia cipoensis* (1 record), and have their locality data blocked.

Curator’s answer: Material that has not been published does not enter the herbaria. Data that is blocked (locality and geographic coordinates) is of plants of commercial interest, so these data fields are blocked to avoid exploitation.

HUESB - Herbário da Universidade Estadual do Sudoeste da Bahia

- Estimated holdings: 11,649
- On-line records: 11,649
- Records with blocked data fields: 281 (1.6% of on-line records)

Characteristics of these records:

- 81 distinct species
- 1 included in Brazil's list of endangered species
- 105 data records of Orchidaceae
- 101 data records of Bromeliaceae
- 25 data record of Cactaceae
- 8 records of unidentified material
- 21 data records of another 19 families

The endangered species is *Echinopsis oxygona*, a species endemic to the Pampa biome of Rio Grande do Sul.

Curator's answer: They block the coordinates and/or locality data of some species of the families Orchidaceae, Bromeliaceae and Cactaceae as they receive visitors that are not associated to research institutes or universities and are concerned that these species may be taken from their natural environment.

HVASF – Herbário do Vale e São Francisco

- Estimated holdings: 23,174
- On-line records: 23,174
- Records with blocked data fields: 3,033 (13% of on-line records)

Characteristics of these records: 122 families

- with imagens: 725
- distinct species: 809
- included in Brazil's list of endangered species: 41
- data records of Bromeliaceae: 855
- data record of Cactaceae: 588
- data records of Fabaceae: 243

The endangered species include 41 records of species of the following families:

- Bignoniaceae (1 record): *Handroanthus spongiosus* (Rizzini) S.Grose
- Bromeliaceae (23 records): *Aechmea muricata* (3); *Aechmea werdermannii* (4); *Canistrum alagoanum* (1); *Canistrum aurantiacum* (3); *Canistrum montanum* (1); *Cryptanthus burle-marxii* (2); *Cryptanthus zonatus* (3); *Guzmania monostachia* (1); *Guzmania sanguinea* (1); *Lymania globosa* (1); *Neoregelia pernambucana* (1); *Orthophytum grossiorum* (1); *Vriesea cearensis*.
- Cactaceae (9 records): *Discocactus bahiensis* (7); *Espostoopsis dybowskii* (2)
- Rubiaceae (5 records): *Mitracarpus rigidifolius* (5)
- Rutaceae (3 records): *Pilocarpus trachylopus*

Curator's answer: no answer

JPB – Herbário Lauro Pires Xavier

- Estimated holdings: 61,518
- On-line records: 61,518

- Records with blocked data fields: 751 (1.2% of on-line records)

Characteristics of these records:

- distinct species: 177
- Only 3 data records refer to *Apuleia leiocarpa*, a species included in Brazil's list of endangered species

Apuleia leiocarpa is considered vulnerable for its commercial use as wood.

Curator's answer: her recommendation is not to include geographic coordinates of endangered species, but only three records are of a specie included in Brazil's red list. Her technician informed that the data that is blocked refers to unpublished material of a specific protected area. They protect this data because they are concerned that it may be used without giving them credit.

SPSF

- Estimated holdings: 48,315
- On-line records: 48,315
- Records with blocked data fields: 11 (0.02% of on-line records)

Characteristics of these records:

- distinct species: 9
- included in Brazil's list of endangered species: 0

Curator's answer: The curator did not know that some records were blocked. He thinks it was an error of the person that entered the data and has already released the data. He also informed that data is only blocked if the depositor expressly requests this.

UPCB

- Estimated holdings: 88,841
- On-line records: 88,841
- Records with blocked data fields: 3,669 (4% of on-line records)

Characteristics of these records:

- distinct species: 885
- included in Brazil's list of endangered species: 129 records of 45 species
- 34% of blocked data records of Melastomataceae

Curator's answer: He did not know why the geographic coordinates were blocked (he was not the curator at the time) and asked CRIA to release the data. We informed him that they are responsible for blocking the data by marking a specific column in their spreadsheet indicating that a data field of that specific record should not be sent to the network. He informed us that they lost all their data some time ago and downloaded the data from *speciesLink*. The curator will have to go back to the voucher and manually enter the data in the database, as the original coordinates are not sent to the network.

3. Zoological records

CEMEC (457 – locality field)

- Estimated holdings: 40,000
- On-line records: 18,656

- Records with blocked data fields – State & Municipality: 457 (2%)

Email sent on November 24, 2016 and no answer was received.

CFBH – Coleção “Célio F. B. Haddad” – Universidade Estadual Paulista – UNESP

- On-line records: 30,593
- Records with blocked data fields: 30,336 have the collector’s name blocked

Email sent to curator on November 22, 2016. The curator of the collection says that the collector’s name is blocked to protect the person from members of the animal rights movement.

CIAMT- Collection of Aquatic Invertebrates of the Federal University of Mato Grosso

- Estimated holdings: 1,753 lots with 128,312 individuals
- On-line records: 1,606
- Records only identified to the genus level: 53 accepted names and 5 names not found

This collection not only retains information about the geographic position of the specimen, but also the name of the species, only revealing the identification to the genus level.

Under the on-line description of the collection it is stated that: “Information related to the georeferenced location, date of collection, sampling effort, abundance and density, shape file and other observations related to each record can be obtained upon request and authorization”.

Curator’s answer: email sent 11/10/2016 no answer

CMPHRM (1735 – notes)

- Estimated holdings: 7,000
- On-line records: 1,746
- Records with blocked data fields: 1,735 (98% of on-line records)

Practically all records have the field notes blocked.

Curator’s answer: email sent on 14/12/2016

CMUFMT – Mammals Collection of the Federal University of Mato Grosso

- Estimated holdings: 3,818
- On-line records: 3,042
- Records with blocked data fields: 2,995 (98% of on-line records)

Characteristics of these records: 47 accepted species names (405 records), 259 names not found of 2,125 records, and 55 names only identified to the species level (2,125 records), and 65 record without any identification.

- 47 distinct species (accepted names)

19 records of the following species included in Brazil’s list of endangered species: *Blastocerus dichotomus*, *Leopardus pardalis*, *Leopardus tigrinus*, *Leopardus wiedii*, *Myrmecophaga tridactyla*, *Panthera onca*, *Priodontes maximus*, *Puma concolor*

Curator’s answer: email sent on 11/10/2016 no answer

DZUP-COLEOPTERA – Entomological Collection Padre Jesus Santiago Moure - Coleoptera

- Estimated holdings: 700,000

- On-line records: 116,360
- Records with blocked data fields: 19,406 (17% of on-line records)

Characteristics of these records:

- 183 distinct species
- 210 names not found in the available dictionaries
- 13 records of species included in Brazil's list of endangered species
- 14,208 records of unidentified material
- 299 data records identified to the genus level

The endangered species are *Agacephala margaridae*, *Doryphora reticulate*, and *Ensiforma caerulea*.

Curator's answer: email sent on 11/10/2016 no answer

DZUP-DIPTERA - Entomological Collection Padre Jesus Santiago Moure - Diptera

- Estimated holdings: 1,500,000
- On-line records: 36,088
- Records with blocked data fields: 5.671 (16% of on-line records)

Curator's answer: The curator said that the blocked data were related to an ongoing projects, but that some projects had ended and the data could be opened. CRIA's staff helped the collection in unblocking the data of three families. The result is that now 2,119 records have blocked locality data and/or coordinates, which represents 6% of the records on-line. All are diptera from the Anthomyiidae family.

Characteristics of these records:

- 11 distinct species
- There are no records of species included in Brazil's list of endangered species
- 1,017 records of material identified to the genus record
- 42 records of unidentified material

This is a very clear case of holding back data of an ongoing project. It also shows the need to monitor this, as it is easy to forget to open the data when the project has ended.

FIOCRUZ-CAVAISC - Collection of Apterous Arthropod Vectors of Community Health Importance of Oswaldo Cruz Institute

- Estimated holdings: 28 thousand ticks, mites, fleas, and lice specimens
- On-line records: 6,459
- Blocked data field: Collector (5,421 - 84%), Identifier (6,085 - 94%)

Email sent on November 22, 2016 and answered on November 23 requesting that the collector field be opened. However, this is an action that depends on the collection and the team responsible for the database did not know how to do it. CRIA's team had to provide on-line training to capacitate the team in blocking and unblocking data.

FIOCRUZ-CMIOC – Mollusca Collection of the Oswaldo Cruz Institute

- Estimated holdings: 150 thousand records
- On-line records: 10,448
- Records with blocked data fields: 10,448 (100% of on-line records)

Characteristics of these records:

- 80 distinct species
- 4 species (487 records) included in Brazil's list of endangered species
- 923 records of unidentified material

Curator's answer: They retain the data while the material is still being studied, but they do not specify a maximum period of time for publication. In reality, all coordinates are blocked, so I believe they do not feel confident enough for a full disclosure of the data.

FIOCRUZ-CMM - Collection of Medical Malacology

- Estimated holdings: 15,112
- On-line records: 15,112
- Records with blocked data fields: 15,111 (100% of on-line records)

Characteristics of these records:

- 21 distinct species
- 1 record of species included in Brazil's list of endangered species
- 217 records of unidentified material

Curator's answer: They retain the data until the results of the projects are published. But many samples were collected more than 20 years ago, so it seems to be that there has not been a cultural change to liberate the data.

FIOCRUZ-CSIOC - Coleção de Simulídeos do Instituto Oswaldo Cruz

- Estimated holdings: 35,000
- On-line records: 16,067
- Records with blocked data fields: 10,109 (63% of on-line records)

Characteristics of these records:

- 60 distinct species
- 3 records of unidentified material
- 37 types

Curator's answer: email sent on 11/10/2016 and not answered

MBML-ANFIBIOS

- Estimated holdings: 9,800
- On-line records: 9,800
- Blocked locality, coordinates, and species name: 116 (1.2 % of on-line records)

An email was sent to the curator on October 28, 2016 and no answer was given, but the data is marked as blocked due to a request from the collector.

MBML-PEIXES

- Estimated holdings: 12,165
- On-line records: 12,165
- Blocked coordinates and species name: 2,262 (19% of on-line records)

Curator's answer: no answer was given to the email sent on October 28, 2016, but the data is marked as blocked due to a request from the collector.

MBML-REPTEIS

- Estimated holdings: 3,879
- On-line records: 3,879
- Blocked coordinates and species name: 247 (6% of on-line records)

Curator's answer: no answer was given to the email sent on October 28, 2016, but the data record states that it is blocked due to a request from the collector.

MCP – Bee Collection of the Catholic University of Rio Grande do Sul

- Estimated holdings: 60,000
- On-line records: 30,635
- Records with blocked data fields: 17,238 (56% of on-line records)
 - Blocked Coordinates: 4,686
 - Blocked Municipality: 17,238

Characteristics of these records:

- 271 distinct species
- 37 type records
- 202 records are of *Arhysosage cactorum* a species included in Brazil's list of endangered species

Most of these specimens (57%) were collected between 1998 - 2000.

Curator's answer: There are a number of papers in preparation and the authors ask to retain the data. We also lack personnel for data entry and for the organization of the collection.

UFMG-AMP – Collection of Amphibians from the Center of Taxonomic Collections of the Federal University of Minas Gerais

- Estimated holdings: 18,680
- On-line records: 18,680
- Blocked collector: practically all on-line records (18,655) have the collector's name blocked
- Blocked Coordinates: the collection has 3,906 records with coordinates and all are blocked (21% of on-line records)
- The on-line data includes 36 records that were canceled.

Curator's answer: email to the curator's returns due to exceeded disk quota.

UFMG-GIR - Collection of specimens of amphibians larvae of the Federal University of Minas Gerais

- Estimated holdings: 1,976
- On-line records: 1,976
- Blocked Data:
 - Collector: 1,740 (88% of on-line records, all other records are blank for this field)
 - Coordinates: 977 (49% of on-line records, represents all georeferenced records by the collection)
- Records canceled: 1

Characteristics of these records:

- 152 distinct species
- 9 records of 7 species included in Brazil's list of endangered species

Curator's answer: email to the curator's returns due to exceeded disk quota.

UFMG-REP

- Estimated holdings: 2,946
- On-line records: 2,946
- Blocked Data:
 - Collector: 2,716 (92% of on-line records, all other records are blank for this field)
 - Coordinates: 420 (14% of on-line records, represents all georeferenced records by the collection)

Characteristics of these records:

- 220 distinct species
- 9 records representing 5 species included in Brazil's list of endangered species

Curator's answer: email to the curator's returns due to exceeded disk quota.

4. Final comments

When the *speciesLink* network began considering sensitive data, there were basically two reasons for not sharing data: full records of endangered species or of material that had not been published. The only concern at the time was in publicizing geographic coordinates.

With this experience in mind, *speciesLink*'s search interface easily separates all records with blocked geographic coordinates (fig.2).

Figure 2. Searching for records with blocked geographic coordinates (*speciesLink*, 2016)

For this reason, when beginning this study, we only analyzed this group of records, thinking that we had all blocked data mapped. Repeating this analysis (only records with blocked geographic coordinates), we have 82,755 records, 90% (74,448) Animalia and 10% (8,271) Plantae. For this reason, we decided to include blocked data for animal records in our analysis.

Analyzing animal records with blocked coordinates, less than 1% are of species included in Brazil's list of endangered species. For plants, a little more than 2% of these records are of endangered species. This was

an obvious indication that what before was said to be the main reason for blocking geographic coordinates, was not so. Therefore, we began a more detailed analysis of these blocked records.

When studying each collection we found that other data fields were being blocked, such as collector and identifier (specialist who identified the sample). This analysis became somewhat complex because geographic coordinates have a different treatment in the database and are easily separated from unblocked geographic coordinates. What we found were records with other fields blocked and even with more than one data field blocked.

Therefore, what seemed to be something easy to identify and analyze, had to be carried out individually, through specific searches on each data set and analysis of the inventories to find which data fields had been blocked.

Searching for blocked data fields (any field = bloqueado | bloqueada), 191 thousand records were found, 185 thousand of animal records and 5.6 thousand records of plants. This represents 7.8% of on-line animal data records and 0.1% of on-line plant data records. As the botanical community is working in very close collaboration with CRIA to build the Virtual Herbarium e-infrastructure, we believe that for this group, there has been a faster cultural change, which led to this result of less plant data records with blocked data.

As stated before, reasons for blocking data fields were more diverse than expected. Besides endangered species and not publicizing material that has not been published or studied, such as material of ongoing projects and of not indicating the geographic coordinates or locality data of plants of commercial interest to avoid exploitation some new reasons for not sharing the data were found such as:

- Fear for not receiving due credit: by blocking some data fields, users have to contact the collection if they need the data;
- collector's name is blocked to protect the person from members of the animal rights movement;
- data is blocked due to a request from the person who collected the sample in the field; and,
- lack of personnel for data entry and for the organization of the collection

We also found two herbaria with blocked data fields where the curators did not know the reason why the previous curators had blocked them. One released the data fields immediately, while the other had had a problem in the past and lost all their digitized data and used the online data to recuperate their database. Through this survey, they realized that the network did not have all the collection's data. In the case of blocked data fields, what is sent to the system is the information that that data was blocked, and not the actual data. This collection will have to re-digitize all blocked data fields.

Many lessons were learned from this analysis. *speciesLink*'s policy is that each data provider must determine its own data policy, blocking what is considered sensitive data and only sending to the network open data that can be publicly viewed by anyone interested. This, in our view, is the best way to stimulate openness.

What we now realized is that, in some cases, the data provider is not documenting this data policy. We found a number of collections where, not only the reasons for blocking data are unknown, but also the fact that there is data being blocked is also unknown. In many cases, the new curator saw no reason why the data could not be shared and made it available. There is also a case where the curator thought that only geographic coordinates of endangered species were being protected, but the technician responsible for the data input "protected" other datasets for other reasons.

We also found that data collected more than 20 years ago is still being protected because it has not been published. It would be important for the collection to stipulate a maximum period for the publication of the data, after which the data will be shared.

This study has also shown that we have to improve the procedures and begin monitoring this data. Just by carrying out this survey and contacting the curators, many data that before were blocked is now being openly shared.

ANNEX 4. ANALYSIS OF USERS AND USAGE

A completely different approach was given to this theme.

1. Usage

The first action was to analyze usage statistics using: *AWStats*, a logfile analyzer (unique visitors, visitors, pages, hits, and bytes); and *Google Analytics* (geographic distribution of users). Analyzing log data for the *speciesLink* network for the year 2015 showed an average of 28.65 thousand users per month (*AWStats*), with about 95% of the users are from Brazil (*Google Analytics*). These statistics do not show the amount of data that is being used, nor what percentage refers to the Virtual Herbarium, nor does it show who is using the data and for what purpose.

In October 2012 we launched a new search interface that enabled the use of a number of tools to produce maps, charts, visualize images as catalogues, compare images and carry out full downloads of the requested data. We found that that the log of this usage could be used as an interesting parameter to measure actual usage, but there was a technical constraint to overcome. All logs of usage since October 2012 meant handling a very large data set so the challenge was in producing an efficient, dynamic, and understandable data usage report online.

We contacted the team from iDigBio (Integrated Digitized Biocollections), an NSF funded project that holds a database with over 75 million specimen records and 18 million media records and presents an impressive performance in its search interface. With their guidance, we studied an open source search engine developed in Java called Elasticsearch and developed an interface to measure data usage⁷ based on logs of the *speciesLink*'s search interface. This development also had the support of the Virtual Herbarium project. Data usage can now be analyzed for all of *speciesLink*, just for the Virtual Herbarium or for each individual data provider.

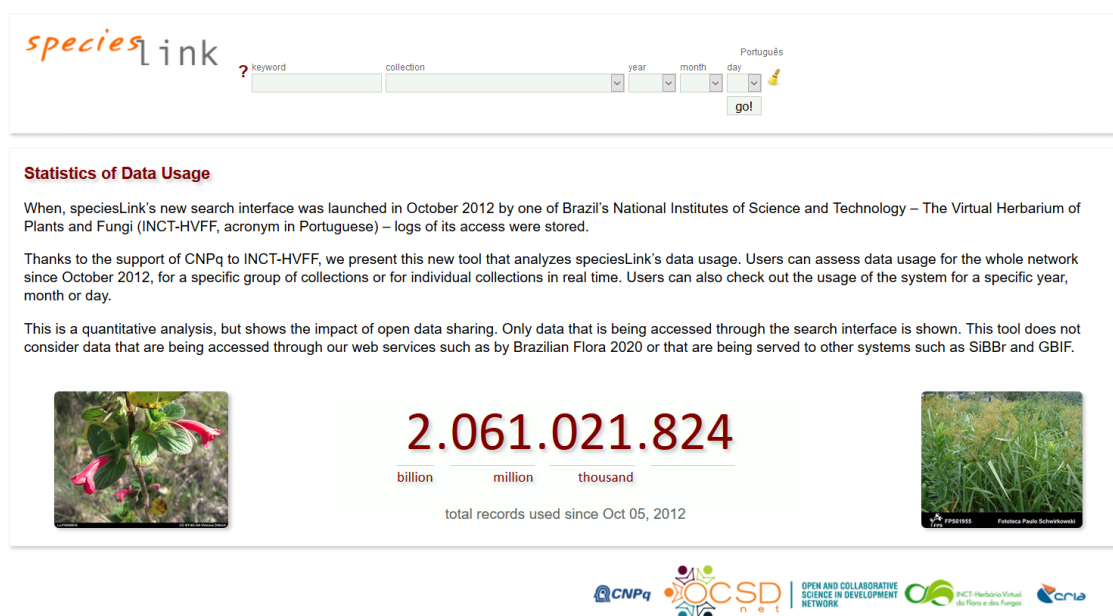


Figure 1. Front page of the Statistics of Data Usage interface

When using “plants” as keyword, the system presents the usage statistics for the Virtual Herbarium. Figure 2 presents the number of plant data records and images used since the interface was launched in October

⁷ See www.splink.org.br/showUsage

2012. For the last 3 years, more than 400 million data records and 3 million images were used. This represents an average of 1.2 million data records used per day.

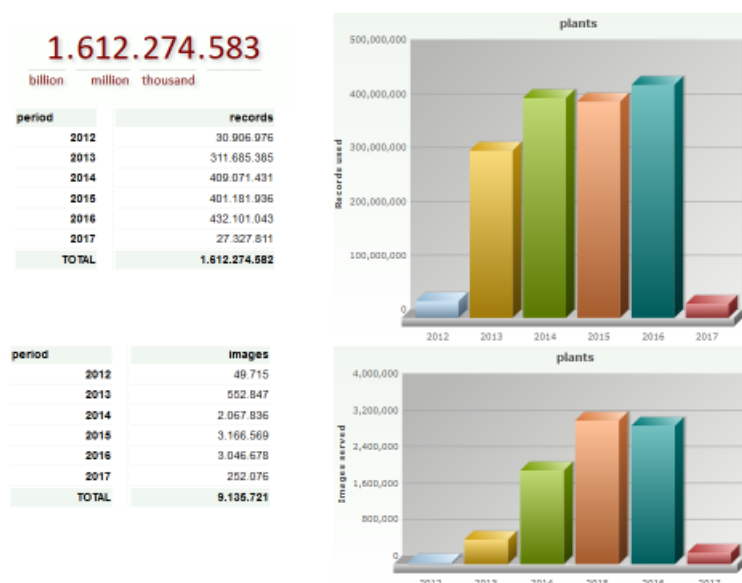


Figure 2. Usage statistics for plants – textual records and images used (11/02/2017)

The statistics also compares the number of data records used with the average number of records shared during the period analyzed. Figure 3 shows that the data used represents more than 600 times the data shared. This shows the strength of data sharing and the reuse of data. One can also evaluate which visualization tools are most used. Figure 3 shows as most used, maps, graphs and download.

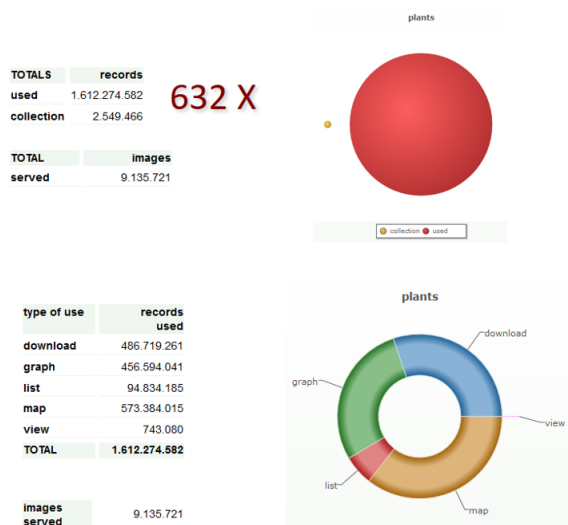


Figure 3. Records used compared to offered and tools used

It is important to bear in mind that this usage only represents usage through the search interface. Data shared through web services and shared with other e-infrastructures such as GBIF, iDigBio, and SiBB through GBIF's Internet Publishing Toolkit (IPT) are also not considered in this statistics.

2. Users

The next step was to study how to analyze who are the users and for what purpose is the data being used. We based our study on the online survey carried out by Atlas of Living Australia (ALA) in May 2015⁸. They received 833 answers and asked where ALA users come from, how the ALA contributes to their work and life, and which features of the website they used, liked most and would improve. ALA is one of the most complete biodiversity information systems of the world, with more than 67 million occurrence records, with great visibility and usage. Analyzing the report, we found that the survey was very extensive and this could be a problem for our small staff to analyze the answers and may well be the reason why the number of users who responded was relatively low.

So, a very much simpler online survey was developed, where users could easily select their answer from the options shown and write their comments in free data fields. The results were dynamically shown on-line in both English and Portuguese as the survey was answered. (<http://inct.splink.org.br/dataUse?lang=en>). The survey was publicized through the search interface, posted on CRIA's blog, and emails were sent to all data providers.

The survey was launched in April 2016 and closed on January 09, 2017. There were basically two inquiries, one on the users profile and another on the purpose of using the data. Although a controlled vocabulary was used with “clickable” answers, all themes included “other” as an option and presented a field where users were free to write what they wanted. With this we had a very quick visual analysis of both the user profile and what the data was used for and, at the same time, we had very rich material of the comments, criticisms, and suggestions made.

Figure 4 presents the *Users Profile* as shown on-line.

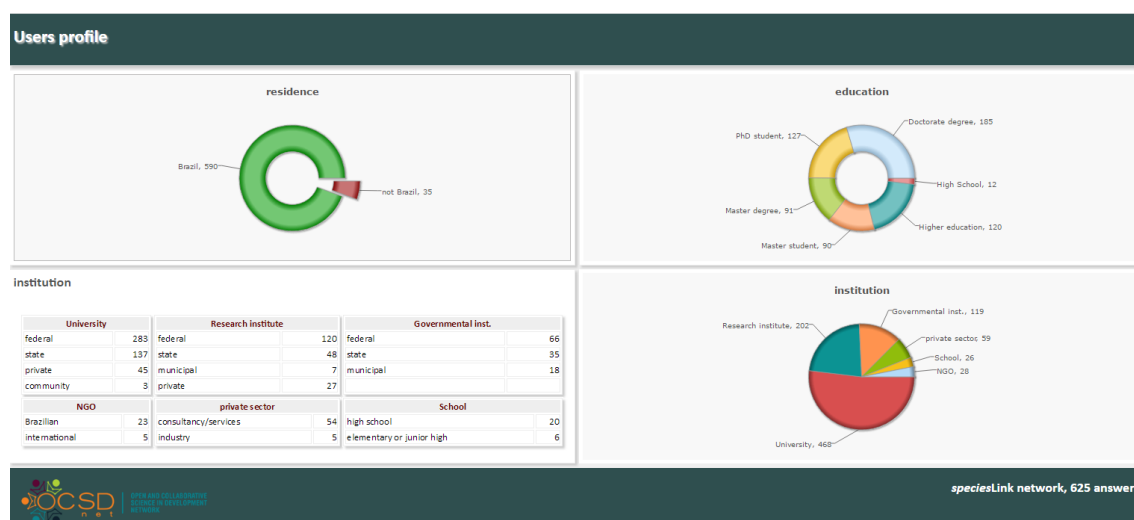


Figure 4. Users profile – speciesLink network (Jan 09, 2017)

speciesLink's main goal is to promote the development of science and informed decision making. Target users are the scientific community and policy makers in Brazil. This survey showed that 94% of the 625 users who answered the survey are residents in Brazil. As to their education, 99% have at least a university degree, and about 50% of the users have a doctorate degree or are PhD students. So we can tell that are users as residents of Brazil with a high education level. As to their institution, over 50% are from universities, 22% from research institutes and 13% from governmental institutes. However, there is an important segment (11%) from the private sector, NGOs and schools that we had not acknowledged before.

⁸ See ALS User Survey 2015 - <http://www.ala.org.au/blogs-news/ala-user-survey-2015/>

The survey also analyzed what the data was being used for. As with the user profile, we presented a questionnaire with controlled vocabulary with options of usage in research, education, and “other” uses. For all groups, a free field was added where other usages could be expressed. When a type of usage became more frequently mentioned in the free field, it was added as an option. The result is shown in Figure 5.

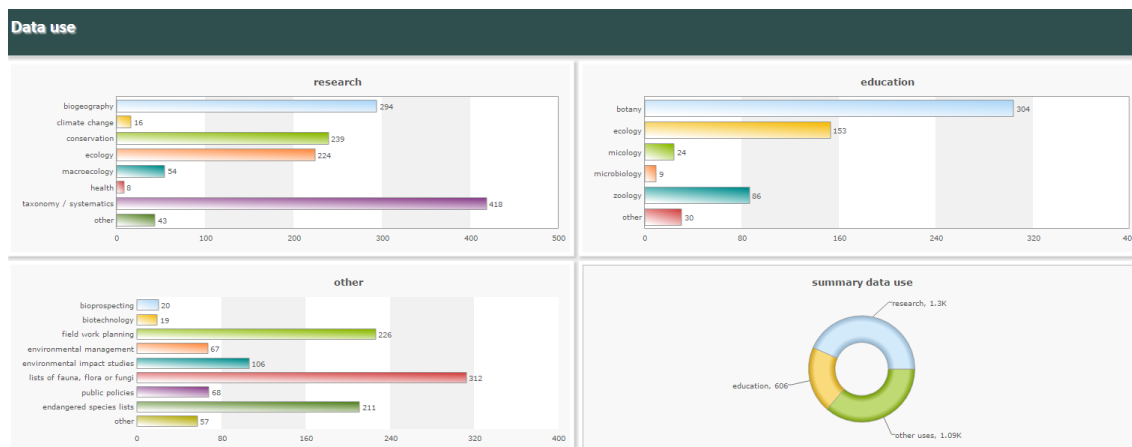


Figure 5. Data use - speciesLink network (Jan 09, 2017)

43.3% of usage is on research, 20.3% on education, and 36.4% on other uses. When analyzing *Research* one can easily see the importance of the data for taxonomy and systematics, which was expected. But it is interesting to verify the use in biogeography, conservation, ecology, and macroecology. In education, we expected to find a large use in botany and ecology, as those are the communities we most work with, followed by zoological and microbial collections. The results for “other uses” were very interesting, the most relevant being: lists (flora, fungi, fauna, and endangered species) and field work planning. However, we were surprised with the usage in environmental impact studies, public policies, and environmental management.

We were also positively surprised with the comments, suggestions, and new demands received. 25% of the answers received included comments and/or suggestions. All were analyzed and now serve as the basis of our work plan for new developments for the next 6 years. We intend to repeat this survey every two years.

Brazil's Virtual Herbarium as an Open Collaborative Science Infrastructure

1. Is science open?

The basis of modern science began during the Renaissance with the concept of scientific experimentation by Francis Bacon (1561-1626), and with the idea that humanity would benefit from a collective, organized public knowledge system. With this need, various initiatives began to promote scientific collaboration such as the organization of scientific societies (e.g. Royal Society, 1660), journals (e.g. *Philosophical Transactions of the Royal Society*, 1665), and the great exhibits (e.g. *The Great Exhibition of the Works of Industry of all Nations*, 1851).

Before mass education, scientists were the holders of knowledge, and scientific communication was practically restricted to the scientific community. Many scientific developments today aim at solving specific problems involving specialists from different fields of knowledge, working in different countries and with different cultures. The evolution of information technology and communication is changing not only the way knowledge is produced but also as to how it is being communicated (Gibbons, Limoges, et al., 1994). The dissemination of results is not sufficient. Science is an object of public interest, subject to public discussions, so its language becomes vernacular with a greater dissemination of scientific data and information to society (Nowotny, Scott and Gibbons, 2001). There are growing demands for on-line, dynamic, real-time, and two-way information and communication systems, carried out throughout the process, and not restricted to scientists. Communicating science and knowledge must reach out to all the community of specialists that necessarily must be part of the process (Hobsbawm, 2008).

This evolution of scientific communication is especially true for botany and its importance to sustainable development. Challenges range from local to global and “openness” is vital at all levels. However, there are many hurdles to overcome. Evaluation systems in universities and research centers are mostly based on individual metrics, when working as a team is essential. Publishing in journals of great international impact is what counts, even for developing countries, and this reduces the importance of local journals in local languages, with a focus on local problems. Networking and providing significant scientific services such as publishing and curating data are normally not valued, when the availability of quality data is the basis for the advancement of science and for policy and decision-making processes.

2. The Project

The increase of knowledge on Brazilian biodiversity, associated with scientific advances to understand the evolutionary processes that generate and maintain this diversity, are fundamental to the sustainable use of this natural capital. Samples and associated information on plants and fungi collected in Brazil in the last three centuries are stored in herbaria in Brazil and abroad. Brazil's Virtual Herbarium of Flora and Fungi (BVH) was established in 2009 to document, store, disseminate, and increase the knowledge base on the diversity of plants and fungi of Brazil.

Large investments are continuously made in developing cyber infrastructures to support research (Barjak et al., 2013). Examples from Brazil include Brazil's National Education and Research Network (RNP) and the National Centers for high performance processing (Cenapad). However, engineering breakthroughs alone are not enough to achieve the outcomes envisaged for the undertaking of e-Science and other global collaborative activities supported by the cyber infrastructure. If it is to be achieved, it will more likely be the result of a nexus of interrelated social, legal and technical transformations (David, 2005; Tenopir et al., 2011).

BVH has undoubtedly benefited from the advancements of RNP and of *speciesLink*⁹, the e-infrastructure used as its information base, which is under development since 2001. However, its major achievement was to integrate institutions and people as a network, with different roles but with common aims.

BVH's continuous success depends on consolidating the social network established and its e-infrastructure as a platform for e-science to boost frontier developments in taxonomy, ecology, biogeography, and biodiversity informatics.

3. What strategies contributed to *openness*?

During *speciesLink*'s early stages of development, participating biological collections had to openly share all data available. There were no mechanisms in place to hold back data considered sensitive or confidential. In the name of openness, to participate, all had to be shared. Sharing data with its own community was a normal practice among biological collections, but making data available to anyone interested without knowing who was accessing it and for what purpose meant an enormous cultural change. When mechanisms were built to ensure that data providers could easily send only data that they selected as open data, more collections were willing to share their data through the network.

Lesson learned: data policy, including decisions as to what data can be shared openly must be carried out at the data provider's end. The e-infrastructure adopts a general policy (CC BY-NC-SA 3.0¹⁰), and all data that is shared must follow the specific license.

Another important feature refers to expertise in informatics. Since the beginning it was clear that most biological collections had very little expertise and inadequate infrastructures concerning informatics. Therefore, the strategy was to adopt a simple architecture at the data provider's end and reduce demands, trying not to alter the collections routine.

Lesson learned: the complexity of the network in informatics must lie at the e-infrastructure's end.

The use of internationally accepted data standards and communication protocols was fundamental. *speciesLink* began its development in collaboration with SpeciesAnalyst, a network in the US developed at Kansas University. GBIF, the Global Biodiversity Information Facility was also just beginning. All these initiatives got together and defined a common data model (DarwinCore) and a protocol (DiGIR – Distributed Generic Information Retrieval). The use of common standards and protocols is what enabled integration of data from other networks, facilitating the work on data repatriation.

Lesson learned: the use of internationally agreed standards and protocols is essential.

Brazil's Virtual Herbarium project began with existing infrastructures, developed by CRIA, responsible for the development and maintenance of the *speciesLink* network, and RNP, the Brazilian National Research and Educational Network (*Rede Nacional de Ensino e Pesquisa*), responsible for the backbone of the national academic network. Members of BVH's steering committee are members of the Brazilian Botanical Society (SBB – *Sociedade Botânica do Brasil*) and its network of Brazilian Herbaria (*Rede Brasileira de Herbários*). These three initiatives, CRIA, RNP, SBB, and, evidently, the botanical community are the pillars of this project that would not have progressed as it has were it to start disregarding existing initiatives.

Lesson learned: when developing an e-infrastructure, focus on establishing strategic alliances with successful initiatives.

⁹ <http://splink.cria.org.br>

¹⁰ Creative Commons license: Attribution (BY), Non-commercial (NC), Share-alike (SA)

4. Results

BVH's project began in December 2008, with 25 national herbaria associated to the project (of which only 18 were sharing their data on-line), two herbaria from abroad repatriating their data of samples collected in Brazil, and 16 non-associate herbaria that were sharing their data through the speciesLink network. The total amount of data records shared through the *speciesLink* was about 1.8 million.

All project targets were surpassed. Today, BVH integrates data from more than 100 associate national herbaria and 21 from abroad. Besides herbaria data, BVH also integrates data from a pollen collection and two taxonomic databases. Images of vouchers (845 thousand), live plants (28 thousand), and pollen (3 thousand), all associated to data records, are also shared openly through the network, that today shares 5.2 million data records on-line.

Many tools were developed in close partnership with the herbaria and the user community. The search interface¹¹ was largely enhanced, and allows users to produce maps, charts, and inventories with the result of their search. It also enables users to compare images and to produce catalogues on-the-fly. An annotation system was developed to provide users the means to help curators in improving the quality of the data.

The result is an impressive statistics of usage. Over 400 million records were *used* in 2015¹². What is meant by *used*, is data that is retrieved to meet various user demands: production of maps, charts, viewed in lists or as an individual record (specimen card), or downloaded. This represents 76 times the amount of data records available on-line. It is important to note that these statistics do not include data served through web services, just through the search interface.

Since the beginning of *speciesLink*'s development, various tools were developed to help curators improve the quality of their data. Records are not modified; the system just presents "suspect" records, recommending that they be checked by the curator. The data provider must correct and update the records. Reports with all suspect or inconsistent records are available on-line¹³ for both curators and users to attest the quality of the data. These tools were greatly enhanced due to the close proximity to the herbaria.

Another important strategy was to use of data to determine priorities. Herbaria were asked to include all data on-line, even of material that was not identified. This data was used to help structure the program of visiting specialists and, when images were also available, promoted on-line determinations by taxonomists of the world (*cybertaxonomy*).

A system called *Lacunas*¹⁴ (Canhos et al. 2014) was developed to help identify taxonomic and geographic information gaps of plants and fungi of Brazil. This tool helps the project's steering committee in prioritizing taxonomic groups for digitization and to identify understudied groups, indicating the need for training of taxonomists. Curators also use this information to develop strategies to guide new fieldwork.

BioGeo, *Biogeography of Flora and Fungi of Brazil*¹⁵, was developed to help guide fieldwork and improve data quality. The system presents two interfaces, one open to all interested where all published species geographic distribution models are available and another for registered specialists that want to produce such distribution models. Table 1 shows the number of models publically available per taxonomic group.

¹¹<http://inct.splink.org.br>

¹²See usage statistics at <http://inct.splink.org.br/showUsage>

¹³Select a dataset to see a report at <http://splink.cria.org.br/dc>

¹⁴See <http://lacunas.inct.florabrasil.net>

¹⁵See <http://biogeo.inct.florabrasil.net> (only in Portuguese)

Table 2. Number of species with distribution models per taxonomic group (BioGeo, June 2016)

TaxonomicGroup	No. of species in the Brazilian List*	Specieswithdistributionmodels	%
Algae	4.746	0	0,0
Angiosperms	32.824	3.627	11,0
Bryophytes	1.524	5	0,3
Fungi	5.710	10	0,2
Gymnosperms	30	4	13,3
FernsandLycophytes	1.253	67	5,3
Total	46.087	3.713	8,1%

*IPT of March 18, 2015

The system was launched experimentally in September 2012 and is the work of one developer and over one hundred voluntary specialists. For more details read CRIA's blog (in Portuguese)¹⁶. The system has helped improve data quality and target fieldwork.

5. Impact on Data Providers

All evaluation parameters follow strict metrics such as the number of data records and images on-line, the number of georeferenced records, data quality, and usage. *The impact of Brazil's Virtual Herbarium in e-Science* is a project within OCSDnet – Open and Collaborative Science in Development Network. In this project, one of the objectives was to identify possible drivers that motivate herbaria to openly share their data through an e-infrastructure and possible outcomes of this participation. One of the central research questions was “Has data sharing through the Brazilian Virtual Herbarium (BVH) led to more recognition and support for data providers?” This study included the opinion of curators from 57 herbaria, which at the time represented 58% of all associated Brazilian herbaria of the network.

Outcomes informed by the herbaria, derived from sharing data through a public e-infrastructure included (1) greater institutional recognition; (2) greater involvement with graduate courses, (3) increased number of visits to the herbaria; (4) increase of the holdings; and, (5) increase of grants.

Smaller herbaria consider the lack of recognition of the work or even of the existence of herbaria by the host institution a major problem. This survey indicated that 92% of herbaria with holdings of up to 10 thousand vouchers stated that sharing their data through the e-infrastructure gave them more visibility and institutional recognition. This undoubtedly is an important outcome of data sharing through BVH.

Another important aspect of the network is that 95% of the participating herbaria are associated to graduate courses. The use of data and tools available in BVH have become a routine in graduate courses such as botany, taxonomy, and ecology. By sharing their data on-line, herbaria have increased their involvement with graduate programs. Many also indicated that by exposing the data of small, but geographically specific holdings, they attracted the interest of students and specialists. With this, the number of visitors increased as did the number of new samples deposited in their herbaria. Some herbaria answered that besides the increase of the number of visitors, these are more diverse – both from different fields of knowledge and from different geographic areas. These are important outcomes directly influenced by sharing data through the e-infrastructure.

Another major problem for smaller herbaria is external funding. With greater visibility and, in many cases, by submitting proposals as a network, 50% of the smaller herbaria with holdings under 50 thousand vouchers were successful in receiving external grants. However, not only did the small herbaria benefit

¹⁶ Biogeografia da Flora e dos Fungos do Brasil at <http://blog.cria.org.br/2013/11/bioge.html>

from sharing their data in an open platform, larger herbaria also acknowledged a great impact in the number of visits, holdings, and grants. Larger herbaria also manifested that their internal organization was improved and overall planning and setting goals to be achieved was also enhanced as data was made available on-line. By sharing their data on-line and by using all tools available for analysis, herbaria could work on data quality and plan future collecting efforts.

6. The strength of the network

The social network established and strengthened throughout the BVH's project promoted increased interaction between curators and technicians from different institutions. There was a change in the mindset of the professionals involved that now feel valued and part BVH's achievements. The increased geographic coverage of the network, with the participation of small herbaria, is a very important asset, as many of these are regional collections, whose copies are underrepresented in other collections. Participation in BVH promoted increased collaboration with students and researchers from other courses and institutions, and the visit of foreign researchers.

The impact of this collaborative network - involving data providers, data users, and the information technology team - can also be measured by the number and significance of tools and applications available. When considering e-infrastructures one tends to focus on hardware and in developing systems to facilitate the access to data. BVH found that the key of innovation is in working collaboratively, as a true interactive network.

7. Final Comments

It is important to develop local e-infrastructures as the organization and dissemination of data increases its usability and usefulness locally. Many funding agencies worldwide request that project proposals include strategies for managing data and sharing it on-line. This is an important step, but not sufficient. For users to be able to rely on information systems, it is crucial for them to operate with uninterrupted, long-term funding, and these agencies operates through project-based strategies. For data that is permanent and must be kept and offered over time, an e-infrastructure must be in place and must provide services to projects that produce such data. E-infrastructures require long-term maintenance and constant development, continuous and dynamic evaluation and planning, and efficient governance models to assure continuity of the network and its services (Canhos et al., 2015).

8. References

- Barjak F, Eccles K, Meyer ET, Robinson S, Schroeder R. The emerging governance of e-infrastructure. *J Comput-Mediat Comm*. 2013; 18: 1–24.
- Canhos DAL, et al. Lacunas: a web interface to identify plant knowledge gaps to support informed decision-making. *Biodiv Conserv*. 2014; 23: 109–131.
- CANHOS, DAL; et al. The Importance of Biodiversity E-infrastructures for Megadiverse Countries. *PLoS Biology* (Online), v. 13, p. e1002204, July 23, 2015. <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002204>
- David PA. Towards a cyberinfrastructure for enhanced scientific collaboration: providing its “soft” foundations may be the hardest part. In: Kahin B, Foray D, editors. *Advancing knowledge and the knowledge economy*. Cambridge: MIT Press; 2006. pp. 431–453.
- DAVID, P. A. The Historical Origins of 'Open Science': An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution. *Capitalism and Society*, 3, n. 2, 2008. <http://www.bepress.com/cas/vol3/iss2/art5>

- GIBBONS, M. et al. **The New Production of Knowledge**: Dynamics of Science and Research in Contemporary Societies. London: SAGE Publications Ltd, 1994. 179 p. ISBN 978-0-8039-7794-5.
- HOBBSAWM, E. J. Era dos extremos: o breve século XX; 1914-1991. Tradução de Marcos Santarrita. 2. ed. [S.l.]: Companhia das Letras, 2008. 598 p. ISBN 9788571644687.
- NOWOTNY, H.; SCOTT, P.; GIBBONS, M. Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty. Cambridge: Polity Press, 2001. 278 p. ISBN 978-0-7456-2608-6.
- Tenopir, C. et al. Data Sharing by Scientists: Practices and Perceptions. PLOS ONE, 2011. <http://dx.doi.org/10.1371/journal.pone.0021101>