Sistema de Informação Distribuído para Coleções Biológicas:

A Integração do Species Analyst e SinBiota

FAPESP/Biota process no. 2001/02175-5
March 2003 – March 2004

Arthur D. Chapman

10 March 2004

Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota
Project Report – Introduction

Arthur D. Chapman
Mar 2003-Mar 2004

1. Contents

Report

1. Background

Following approaches by Vanderlei Canhos and others, on 11 October 2002, I accepted a consultancy to carry out twelve months work at CRIA on FAPESP/Biota process no. 2001/02175-5. The purpose of the Grant was

- to look at issues of modelling at CRIA, especially in relation to SinBiota and speciesLink;
- to discuss issues with staff and researchers across the State, and train people in aspects of environmental modelling;
- to examine interactive modelling systems used at CRIA for SinBiota and Species Analyst;
- to examine methods of identifying errors in biodiversity databases and recommend on possible "data cleaning" tools;
- to study aspects of the proposed Virtual Herbarium of the State of São Paulo;
- to examine GIS systems being used at CRIA and train staff and others in various aspects of GIS;
- to carry out visits, talk to researchers and students and train researchers and students in various aspects of environmental data management and analysis.

2. Executive summary

Most of the proposed objectives of the project have been satisfactorily carried out, and a number of significant achievements made in the way environmental data is managed and analyses carried out on biological data at CRIA. During the twelve months at CIRA, I have prepared (some in conjunction with other staff and researchers) 13 Reports, 7 scientific papers, and 2 on-line data management tools. For a detailed list, see page 8 and appendices to this report.

There is still a long way to go, however, and I am confident that CRIA will continue to advance and lead the world in the development and use of simple, user-friendly biodiversity informatics tools and methodologies. CRIA is fortunate to have an extremely talented group of people working for it, and has managed to retain most of those talented people in face of what must be significant challenges to poach them from places such as the United States and elsewhere. Part of that has come about by the close-knit feeling of staff in the organization, and the continued achievements being made that all staff are made feel a significant part of. Some of this is, of course, a result of being a small organization, with most researchers being able to work on their projects uninterrupted, and with little distraction from bureaucratic procedures.

I have made a few recommendations below on additional aspects that may be looked at in a similar manner to the issues examined during this project. They include the introduction of climate-change modelling, reserve and conservation selection methodologies, decision-support tools, and methods for determining environmental risk. In addition, the aspects covered under this project require constant monitoring and assessment as new techniques and methodologies become available, and as user requirements change – possible including over the longer term from a mainly research driven agenda to public policy, environmental management, ecotourism, and education driven agendas with scientific and research linkages. This is a challenge for the future, but one that CRIA, and the FAPESP/Biota program are obviously aware of.

I have recommended to the Directors of CRIA, and in this report, that a follow up study should be conducted in about twelve months time to examine how the recommendations I have made have been implemented. At the same time, some aspects of the other criteria I have mentioned above (and expanded on below) could also be examined and recommended on. This may require a further 3-6 months work.

I strongly endorse the process whereby external people are brought into the program and can spend some time working with staff and researchers on various aspects of the program. The type of systems and tools that CRIA is developing and working with, are universal, and the ideas are developing rapidly (after all, it is only just 10 years ago that Tim Berners-Lee introduced the concept of URLs and the World Wide Web began). No one place or organization will have all the good ideas. It is thus important that collaboration and interaction with people doing similar work around the world be encouraged and supported. This can be done in two ways – in bringing people in to CRIA for short periods as has been done with myself and Dr Townsend Peterson, or by sending CRIA staff to other institutions for short periods to work with experts there as was done with Ricardo Scachetti-Pereira. I wholeheartedly endorse this process, and believe without it CRIA would rapidly become of less value to the

FAPESP-Biota program, and the program itself would become of less value to the community through not being able to effectively disseminate the results of its many valuable research projects.

3. Summary of Work Carried out and Significant Achievements of the Project

Item	Work	Significant Achievements	Reports and Publications
Modelling Algorithms	– Conducted a review of modelling being carried out at CRIA – looked at the data layers being used, algorithms and methods and made a number of recommendations. – Carried out a review of the Lifemapper system (University of Kansas) which has been developed in conjunction with CRIA. – Assisted CRIA staff with modelling problems, especially with discussions on Scale that will lead to further research and additional papers. – Held discussions with Barry Chernoff, Wesleyan University, Connecticut, and others on methods for modelling in aquatic environments.	– Obtained new environmental layers (esp. climate layers) that improve the scale of modelling being carried out firstly by up to 25 times, and then by up to 3600 times. – Modified the types of layers being used to use layers that are more in keeping with the environment. – Brought a change in methods being used for resampling grid data that has led to an improvement in the integration of different data layers. – Assisted in planning of a modelling framework for CRIA which will lead to a broadening of the type of modelling being carried out and the types of modelling algorithms being used. – Suggested some new research directions which are being taken up at CRIA – Developed ideas for a workshop to be conducted in the USA in 2004 on Modelling in Aquatic Environments. – Held discussions with staff on modelling, and especially on the effects of scale on modelling, etc.	– **Report 3** – The Case for a 3-minute Climate Surface for South America (**Appendix E** to this report). – **Report 3b** – A 1 km Climate Surface for South America (**Appendix F** to this report). – **Report 4** – Environmental Modelling in CRIA (**Appendix G** to this report). – **Report 7** – Lifemapper (**Appendix K** to this report). – **Chapman, A.D., Muñoz, M.E.S. and Koch, I**. (in press). Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* **1:** – Preparing paper on Environmental Modelling for possible submission to *Biota Neotropica.*

Interactive Mapping	– Conducted a review of SinBiota's Atlas Biota – looked at the interface and useability from a user's perspective. – Examined a number of other mapping systems on the Internet and made recommendations. – Discussed issues and ideas on data visualisation with CRIA staff and others	– A decision has been made at CRIA to rewrite completely the map interface being used for the Atlas, as well as for other interfaces at CRIA. This will take into account the recommendations made in my report. – Invitation to attend a workshop in New Hampshire in September 2003 to discuss mapping interfaces as well as other issues. Alexandre Marino attended and represented CRIA. – Invitation for CRIA (and myself) to collaborate on an international project looking at data visualisation and internet mapping interfaces, especially for use with biological data. – A number of minor modifications have been made to the current interface.	– **Report 1** – SinBiota: Atlas Biota (**Appendix A** to this report).
Virtual Herbarium	– Held discussions with a number of people on the development of a Virtual Herbarium/Museum São Paulo.	– The work being carried out by CRIA staff on the distributed *species*Link system, is world leading. The developments being carried out with DiGIR routines and add-ons to this have been recognised internationally. – XML scripts being used by the Australian Virtual Herbarium were obtained and passed on to CRIA staff.	– No separate reports or publications were prepared, however, the other reports cited above, have aspects that overlap with, and impact on, speciesLink and the idea of a distributed virtual herbarium or museum. These include: the modelling algorithms, the mapping interface, and data cleaning and validation tools.

Automatic Identification of Errors	– An examination of a range of software, on-line services, guidelines and standards was carried out with the aim of developing a data-cleaning toolkit – A number of software products were obtained and the algorithms tested using data from São Paulo. – New on-line algorithms were written to assist users in entering, checking and validating museum and herbarium data.	– Several software products were found to be available that carry out error testing of species data. Several of these were found to be worthwhile for inclusion on a data-cleaning toolkit. – Permission was obtained to include some software products on a data-cleaning toolkit if this is developed. – Permission was obtained to include the HISPID standard on any data-cleaning toolkit that is developed. – An algorithm for finding the latitude and longitude of a point a distance and direction from a gazetted locality was developed in conjunction with CRIA staff and made available on the internet. The algorithm also reports on the error inherent in determining the geocode. – An algorithm was developed and made available on the internet to identify outliers in latitude, longitude and/or altitude in existing databased records, as well as determining records that may be located wrongly either on-shore or off-shore. – Request from the Global Biodiversity Information Facility to write a Best-Practice document on Data Cleaning and validation in conjunction with scientists at Yale University and the Museum of Vertebrate Zoology in California. – Presented a number of seminars on data quality to University of Campinas, University of São Paulo, and to the Biota Symposium in Água de Lindóia.	– **Report 5 –** Environmental Data Quality – a discussion paper (**Appendix F** to this report). – **Report 6 –** Environmental Data Quality – Data Cleaning Tools (**Appendix G** to this report). – **Chapman, A.D.** Guidelines on Biological Nomenclature – Brazil edition (**Appendix H** to this report). – **Pereira, R. Scachetti,, Soberón, J. and Chapman, A.D.** (in prep). Data Enhancement II. Detecting and Eliminating Errors from Biodiversity Datasets. *Biodiversity Informatics* **1:** – **Beaman, R., Chapman, A.D. and Wieczorek, J.** (in prep.). *Spatial accuracy assessment for biological collections: Best practices for collecting, managing, and using biodiversity data*. 6th International Symposium on Spatial Accuracy Assessment, Portland, Maine, 28 Jun-1 Jul 2004. – **Marino, A., Paverin, F., de Souza, S. and Chapman, A.D.** (in prep). *Simple on-line tools for geocoding and validating biological data*. To be submitted to CODATA Journal.

Geographic Information System	– An examination was made of the GIS systems being used at CRIA. A decision had been made, however, just prior to my arrival to obtain licences for ESRI's ArcView 8, so it was decided that this Item was a lower priority.	– CRIA obtained several ArcView 8 licences in early 2003, just prior to my arrival. This GIS system is the leading GIS in the world at the moment, and in-spite of the cost, I believe is a wise decision. At the moment, no suitable public-domain GIS software can do the job required by CRIA. – Some modification was made to the methods used to resample grid data (moved from Nearest Neighbour to Bilinear Interpretation and Cubic Convolution. – A number of datasets were obtained for use in the GIS – including Gazetteers, climate layers (as mentioned above), and a number of other South American data layers. – Staff were assisted in the use of the ArcView GIS, and a number of methodologies and algorithms and advice were obtained from previous colleagues to improve its usability.	– No separate reports were prepared.
Visits, Talks and Courses	– To give talks on my work at CRIA, on modelling and data validation, etc.	– A number of seminars were given, including to the University of Campinas, University of São Paulo, and to the Biota Symposium in Água de Lindóia. – Many one-on-one discussions were held with both scientists and students throughout the State during the twelve month stay in Campinas.	– **Chapman, A.D.** (in press). Qualidade e validação dos dados ambientais - Metodologias e ferramentas. Powerpoint presentation to Biota Symposium, Água de Lindóia 8-1- December 2003.

Others	– Examined data management as carried out in CRIA and prepared a report. – Assisted CRIA directors in preparing a report on Data Sharing for the Global Biodiversity Information Facility (GBIF). – Assisted in the design and running of the CODATA sponsored Inter-American Workshop on Environmental Data Access. – Examined the BioLink software – museum database management software developed by CSIRO in Australia and prepared a report – Reviewed and commented on a number of papers written by staff of CRIA and other researchers in the State, prior to their submission for publication. – Evaluation of the FAPESP/Biota Program was carried out as a member of the Scientific Advisory Committee – Extensive collaborations were carried out by email with international colleagues	– A Standard 'README' file was prepared for use by CRIA Staff in managing their file systems. However, this has yet to be adopted universally by the organization (Report 2a). – A draft data-structure was designed for consideration by CRIA staff (Report 2a). – An examination of metadata standards was made, and recommendations for a possible metadata standard for CRIA (and possible for brazil) (Report 2b, Guidelines 2) – En examination of on-line metadata clearinghouses, and a suggestion made as to a possible solution for use by CRIA – perhaps with some modification (Report 2b) – A report on data-sharing was prepared by CRIA and submitted to the Global Biodiversity Information Facility (Canhos *et al.* 2003). – A report was prepared in conjunction with other members of the FAPESP/Biota Scientific Advisory Committee and presented to FAPESP. – A successful Inter-American Workshop on Environmental Data Access was held in Campinas in March 2004. – Valuable contacts were developed for CRIA through my previous international contacts, contacts made via email while at CRIA, and contacts made through attendees at the Workshop on Data Access. – Submitted a chapter for the 5th Anniversary Volume of the Biota Program.	– **Report 2a –** Data Management Standards – a. Internal File Systems (**Appendix B** to this report). – **Report 2b** – Data Management – b. Metadata (**Appendix C** to this report). – **Guidelines 2** – Guidelines for Documenting Species and Vegetation Data (**Appendix D** to this report). – **Report 8 –** Biolink (**Appendix L** to this report). – **Canhos, D.A.L., Chapman, A.D. and Canhos, V.P.** (2003). Studies on data Sharing with Countries of Origin. Contract No. GBIFS 2003/04. Report to GBIF. Nov. 2003. – **Chapman, A.D.** (in press). *The human legacy – reversing the trend*. Submitted to Biota 5th Anniversity volume. –

4. Papers, Reports and Publications

Papers

Chapman, A.D. (in press). *The human legacy – reversing the trend*. Submitted to Biota 5th Anniversity volume.

Chapman, A.D., Muñoz, M.E.S. and Koch, I. (in press). Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* **1:**

Pereira, R. Scachetti, Soberón, J. and Chapman, A.D. (in press). Data Enhancement II. Detecting and Eliminating Errors from Biodiversity Datasets. *Biodiversity Informatics* **1:**

Beaman, R., Chapman, A.D. and Wieczorek, J**.** (in prep.). *Spatial accuracy assessment for biological collections: Best practices for collecting, managing, and using biodiversity data*. 6th International Symposium on Spatial Accuracy Assessment, Portland, Maine, 28 Jun-1 Jul 2004.

Marino, A., Paverin, F. de Souza, S. and Chapman, A.D. (in prep). *Simple on-line tools for geocoding and validating biological data*. To be submitted to CODATA Journal.

Chapman, A.D. (in prep). *Environmental Modelling in Brazil – issues and ideas*. For possible submission to *Biota Neotropica.*

Reports

Chapman, A.D. (2003a). *Guidelines on Biological Nomenclature – Brazil Edition.* Campinas: CRIA. Jun 2003.

Chapman, A.D. (2003b). *The Case for a 3-minute Climate Surface for South America*. Internal report no. 3 to CRIA. Draft - May 2003; Final 28 Jul 2003.

Chapman, A.D. (2003c). *Lifemapper – Comments and Ideas*. Internal eport no. 7 to CRIA. 23 Jul 2003.

Chapman, A.D. (2003d). *BioLink 2.0. A preliminary evaluation.* Internal report no. 8 to CRIA. Sep. 2003.

Chapman, A.D., Chernoff, B. and Schalk, P.H. (2003). *Report of the Fourth Evaluation of the BIOTA-FAPESP Program by the Scientific Advisory Committee. Água de Lindóia – 8 to 17 December 2003*. Report to FAPESP. [To be put on-line 2004]

Chapman, A.D. (2004a). *SinBiota – AtlasBiota: Interactive Mapping in CRIA*. Internal report no. 1 to CRIA. Draft Mar 2003; Final 28 Jan. 2004.

Chapman, A.D. (2004b). *Data Management Standards*. Internal Report no. 2 to CRIA. Draft May 2003, Final Jan. 2004.

Chapman, A.D. (2004c). *1 km Climate Surface for South America.* Internal Report no. 3b to CRIA. Jan. 2004.

Chapman, A.D. (2004d). *Environmental Modelling in CRIA. Issues and ideas*. Internal report no. 4 to CRIA. Version 1 – June 2003; Final Version 29 Jan 2004.

Chapman, A.D. (2004e). *Environmental Data Quality. a. Discussion Paper*. Internal report no. 5 to CRIA, 20 Jan. 2004.

Chapman, A.D. (2004f). *Environmental Data Quality. b. Data Cleaning Tools.* Internal report no. 6 to CRIA Draft Jun 2003; Final 26 Jan 2004.

Chapman, A.D. (2004g). *Data Management – b. Metadata*. Internal report no. 2b to CRIA. 12 Jan. 2004.

Canhos, D.A.L., Chapman, A.D. and Canhos, V.P. (2003). *Studies on data Sharing with Countries of Origin. Contract No. GBIFS 2003/04*. Report to GBIF. Nov. 2003.

On-line Publications and Tools

Chapman, A.D. (in press). *Qualidade e validação dos dados ambientais - Metodologias e ferramentas.* Powerpoint presentation to Biota Symposium, Água de Lindóia 8-10 December 2003. [To be put on-line 2004]

CRIA (2004a). *geoLoc*. Campinas, Brazil: CRIA. http://splink.cria.org.br/tools/

CRIA (2004b). *spOutlier* Campinas, Brazil: CRIA. http://splink.cria.org.br/tools/

5. Background to Data and Data Issues – Principles of Data Management

The value of Brazil's (or São Paulo's) biological data would be hard to estimate. The cost of acquisition of biological information is high. In Australia, for example, it is not unusual for a single survey to exceed $1 million (Burbidge, 1991). The process of data acquisition, description and enhancement, including digitisation, represents a significant value-added component to an already valuable resource. Biological collections in museums and herbaria, collected over a period of 2-300 years, represent a valuable resource that cannot be neglected, and its digitisation opens up that vast resource to a myriad of uses.

The FAPESP-Biota Program has taken into account the importance of data and data management and have funded a number of projects, including SinBiota and SpeciesLink to support the data aspects of the program. These programs are managed by CRIA - Centro de Referência em Informação Ambiental, an NGO organization based in Campinas. This report examines the way CRIA manages and uses the biological data it collects and accesses, and makes a number of suggestions and recommendations.

The data being gathered by CRIA includes data of a number of different types. The majority of data being accessed through the *species*Link project is opportunistic data of a 'presence-only' nature. On the other hand, data being acquired through SinBiota often includes stratified survey data, while other data may be of the nature of presence or absence in a particular area or grid. Each of these data require handling in different ways.

a. *Opportunistic data* – opportunistic 'surveys' are collections generally made in a haphazard manner. Most herbarium and museum data falls into this category. The species collected are often ones of interest to the collector, and the places they are recorded from are often those places where the species is expected to occur i.e. the collector goes out looking for the species in areas where he/she expects to find it. This data rarely, if ever supplies in formation on absences. Because this type of data constitutes by far the largest resource of data available, it is the data most commonly used in environmental studies, and especially in environmental models.

b. *Survey data* – stratified survey data is one of the most difficult to store and manage, because each survey is undertaken with specific requirements in mind. It is thus often difficult to combine data from one survey using one technique with data from other surveys conducted using different techniques.

c. *Presence/absence within an area* – data on presence or absence within an area, such as a species list from a National Park or presence or absence within a ten-minute grid etc. is easy information to store but has limitations on its use. The size of the grid, for example, or the shape of the polygon (National Park) can cause significant limitations on how (and if) the data can be used for a particular purpose.

Seldom are records available in datasets of the absence of a species at a particular location. Lack of information concerning absences severely restricts the potential of attributes for statistical modelling. Presence-only data cannot be used to say anything about absences unless assumptions are made regarding the sampling strategy used to obtain the data. Too often the lack of a record is scored by analysts as a genuine absence, where in fact it may be that the location has not been surveyed for the species leading to massive bias in the data. Grid or transect data are rare with species data and are more usually applied to environmental rather than biotic data.

"Because presence-only datasets are usually derived from opportunistic or *ad hoc* surveys they are particularly prone to bias in geographical and/or environmental coverage. The extent of bias is difficult to assess due to a lack of recorded absences. (Ferrier and Watson 1997).

There are a number of principles for good management of data, and CRIA has adopted many of these principles in its data management strategies.

Custodianship and ownership

A key aspect of good data management involves the clear identification of the owner of the data. In most cases this is the organization or group who originally commissioned the data and has managerial and financial control of the data. The data owner generally has legal rights over the data, along with copyright and intellectual property rights. This applies even where the data is collected, collated or disseminated by another party as part of contractual agreements (NLWA 2004).

As such it is important for data owners to establish and document the ownership, intellectual property rights and copyright of their data in order to safeguarded their rights.

The other key aspect to data management is custodianship. Custodianship refers to the organization or group that has the responsibility for maintaining the data. In many cases the custodian of a dataset may be the same as the owner, but this is not always the case. A dataset may include data from a number of owners, and the custodian thus may have responsibility for the combined dataset. For a number of datasets, it may be that CRIA has a custodianship role while not having any ownership in the data.

Documentation

All datasets should be identified and documented to facilitate their subsequent identification, proper management, effective access and use.

To provide access to the dataset owned or managed by an organization, a catalogue of data should be compiled. This is a collection of discovery level metadata for each dataset, in a form suitable for users to reference. These metadata should provide information about the content, geographic extent, currency and accessibility of the data, together with contact details for further information.

Proper access to the data, and to the data's documentation (metadata) is essential for modelling. It is important that there is a thorough understanding of the nature of the data and its properties for use in any analysis. Without adequate supporting information, the data may be not be usable, or more likely, be used wrongly.

The process of establishing metadata about datasets also provides an opportunity to substantially enhance the quality of the information. For example, the process of determining attribute types may highlight inconsistencies in the data. Such inconsistencies and omissions may then be flagged for the custodians to examine and address. The result of this process is higher integrity data for all.

Use of Data for Spatial Modelling

"Any records of a biological entity can be used to produce distribution maps which are a crude predictive model of the spatial distribution of that entity. This is subject, of course, to assumptions about spatial generalisation of the data, adequacy of sampling, changes in distribution over time, accuracy of identification and completeness of data coverage. No survey or collection of records can hope to provide a complete census or inventory, except perhaps for very localised populations of endemics. Surveys and other collections are samples and, as such, knowledge of the sampling methodology is required to determine confidence in the resultant spatial model." (Belbin, *et al.* 1994).

More details on aspect of data management and use are given under the various reports attached.

6. Reports Summary

Report 1: Atlas Biota (see Appendix A)

AtlasBiota is one of the key outlets for disseminating information from the FAPESP Biota program. It was developed in the early stages of the Biota Program using open-source technologies that were available at the time.

The Atlas is based on Mapserver (http://mapserver.gis.umn.edu/) technology from the University of Minnesota. MapServer is not a full-featured GIS system, nor does it aspire to be. It does, however, allow for basic mapping on the internet, has the advantage of being Open Source software, and provides the ability for wrapping with Java, Javascript, Perl and more recently PHP MapScript.

The Atlas is a good start, and carries out what it does quite well. One of its strengths is the extremely quick redraw time. It could be extended, however, to include more flexibility and useability from a users point of view.

The technology for on-line mapping services has recently moved on considerably and quite sophisticated on-line mapping systems can now be developed with high functionality. CRIA staff are currently redeveloping the Atlas, and current prototypes are proving more than promising. It is likely that the mapping interfaces developed by CRIA over the next month or so, will prove to be world-leading in their simplicity and functionality.

Aim: To examine the SinBiota – Atlas Biota; make comments on the interface and useability from a general user's point of view; and make recommendations for possible modifications to the interface.

Method:

a. Examine SinBiota's Atlas Biota,
b. Examine comparable Atlas sites,
c. Make recommendations.

Results: A detailed report was prepared and presented to CRIA (see Appendix A – Report No. 1 SinBiota-AtlasBiota) in March 2003.

Key points included:

- Need for dentification of the purpose of the Atlas
- Comments on existing interface
- Comments on other on-line mapping systems
- Recommendations

Significant Achievements

- Some minor changes have already been made to the SinBiota Atlas as a result of the report, but it has been decided by CRIA staff to examine the Atlas as a whole and to rebuild it, taking into account the recommendations made in the report.

- Considerable progress has been made in redevelopment of the Atlas, and this redevelopment takes into account recommendations made in my earlier draft report.

Report 2a: Data Management Standards (See Appendix B)

CRIA is beginning to accumulate Spatial data from a range of sources, as well as increasingly creating Spatial GIS layers in-house.

The present storage arrangements are ad-hoc with data stored in a range of locations, in a range of formats, and with varying levels of documentation. Often data is stored on the C-Drive of individual staff members, which creates a problem with back-up and access.

CRIA has also recently purchased ESRI's ArcView 8 for use by CRIA staff.

It would appear to be an opportune time for a set of internal data-storage and management standards to be adopted by staff in order for data to be generally available, be fully documented, and to be in a consistent format.

> **Method:** A brief examination of the current situation with data storage and management in CRIA was conducted, and a report prepared.
>
> **Results:** A draft report (Appendix B) and presented to CRIA in May 2003. The report was an adaptation from the Data Management Standards prepared for Environment Australia. Some of the issues in the Australian Standard will not be applicable to the CRIA situation, however there are a number of key issues that need to be considered and adopted.
>
> Key issues include:
>
> - The encouragement of a change in culture such that key data is not stored on the C-drive of individual researchers where it is not subject to backup, and is unavailable to other researchers
> - The need to develop a structured series of data directories so that data is easily able to be found by multiple users and is user-independent (so that when individual users are on leave or leave the organization, data is not lost); and is intuitive in its structure and naming.
> - That data be documented – both as meta-data for long-term completed datasets and as "readme" files for in-progress data sets.
>
> **Significant Achievements**
>
> - A standard 'README' file structure was developed, but as yet has not been universally adopted within the organization. A number of datasets that I have developed or obtained have been documented in this manner so as to set an example.
> - There is now a greater tendency for staff to move data to a common-use area than previously.
> - CRIA staff are examining data management issues, however, a structure has yet to be developed and implemented.
>
> **Comments:** If nothing else is done, I would strongly recommend the adoption of the standard "README' file for <u>every</u> data, or data-related-, directory. This is a key issue and should be standard data management practice.

Report 2b: Data Management b. Metadata (See Appendix C, D)

Metadata is information that describes datasets. If data is documented following good metadata standards it provides a consistent approach to the storage and retrieval of information. Spatial metadata standards have been developed for a number of countries and are used for documenting data storage and for use in accessing data through automated technologies on the internet using distributed search and retrieval.

Although a number of cases have been made for the development of a metadata standard for Brazil (beginning as early as 1996), little seems to have been done.

Australia and Brazil have similar environments, and probably a similar number and type of environmental datasets. In order for those datasets to be discoverable and useable, they need to be documented following agreed standards. Recently the ISO Technical Committee for Geographic Information/Geomatics (ISO/TC 211) released an International Standard for Metadata (ISO 19115), and this should form the basis for the development of any standards for use in Brazil.

The Australian Government, in 1996, developed a geospatial metadata standard that has been used extensively to document environmental (and other) data in Australia for the past 8 years. I believe that this standard could form the basis of a similar standard for use in Brazil.

Method: A brief examination of a number of existing metadata standards was made, and an examination also made of a number of on-line environmental data discovery tools.

Results: A report (Appendix C) and presented to CRIA in January 2004. In addition, a copy of guidelines on documenting species and vegetation data using the Australian spatial data standard (Appendix D) which I prepared for Environment Australia in 1998 (Chapman 1998), were supplied to CRIA staff.

Key issues include:

- Good metadata documentation is essential. To date, very little environmental data in Brazil is adequately documented following agreed metadata standards.
- An environmental data discovery system is required in Brazil, and is urgently needed by CRIA to discover what environmental data exists within the State and elsewhere and to be able to access the data where available.
- Australia has developed a good, robust metadata standard, and data discovery tools that could easily be adapted for use in Brazil and by CRIA.

Comments: I believe that there is an urgent need for Brazil to adopt a standard for documenting environmental metadata, and for the development of environmental data discovery tools. Any standard that is developed should conform with the new ISO standard on metadata – ISO 19115. My personal belief is that Brazil could do worse than to adopt and modify the Australian standard, along with its data discovery tools, used for the Australian Spatial Data Directory.

Report 3 – Climate Surfaces for South America (Appendices E, F)

Climate surfaces are a key basis for environmental modelling, and especially species modelling. When I first arrived in Brazil in March 2003, much of the modelling was being carried out using climate surfaces at 0.5 degree resolution (about 50-60 km). Modelling at this scale is insufficient to delineate environmental niches at a scale in which to be able to make environmental decisions, or for use in determining conservation priorities, etc.

> **Method:** I carried out a search for datasets at a better scale than the 0.5 degree being used, and at the same time, wrote a report setting out a case for the development of a 3 arc-minute climate surface for South America (Appendix E). Preliminary discussions were held with a number of people in an attempt to ascertain the availability of data for the development of such a surface.
>
> **Results:** In March of 2003, a report was prepared setting out the Case for a 3-minute Climate Surface for South America (Appendix E).
>
> A number of useful contacts were made, including within Brazil, Canada, Australia, the USA, Columbia and Venezuela, with the aim of obtaining suitable data for development of such a surface, and to elicit possible collaboration for its development.
>
> In July of 2003, I was able to obtain a number of climate surfaces at a scale of 10 arc-minutes (about 18 km) from the Centro Internacional de Agricultura Tropical (CIAT) in Columbia. These layers were a considerable improvement on those previously available (a factor of 25 times), although the method of preparation led to there being a number of undesirable artefacts within the data. The data also required considerable modification in order to prepare meaningful layers for use in environmental modelling (see under Report 4 and Appendix E, below).
>
> In December of 2003, I was made aware of a project in Guyana that led to the developing a 1 km climate surface for that country. On contacting the people involved, I became aware of a project being jointly conducted by organizations in Australia (Tropical Rainforest CRC) and the USA (University of Berkeley) with some assistance from CIAT in Columbia with the aim of developing global 30 arc-second climate surfaces. On contacting a previous colleague who was involved, I found that development of the surfaces was well advanced.
>
> In January, 2004, I was able to obtain access to the data (in Beta format) and download the datasets covering South America on a trial basis, and on the condition that I provide feedback to the developers. The dataset is not due for public release until March or April 2004. This data is at 30 arc-seconds (about 1 km resolution) which makes it 400 times the resolution of the 10 minute data-set, or 3600 times the resolution of the 0.5 degree datasets being used when I first arrived at CRIA. This is a major advancement. In reality, this data may be too fine, however the developers are preparing a 2.5 arc-minute (c. 5 km) dataset based on these layers which is the scale at which I recommend modelling be carried out at, as it is the scale most consistent with the scale of the biological data (museum and herbarium data) being used in the models.
>
> The development of these layers using the methodologies I advocated in Appendix E, now makes the development of a separate 3-minute surface for South America redundant. An addendum to the previous report (Appendix E) has been written in the knowledge of this latest information.

Significant Achievements:

- The acquisition of climate surfaces for use in modelling at 30 second, or 1 km, resolution which is an improvement in scale of 3600 times on that being used when I first began the project.
- Collaborative arrangements with key data custodians in Venezuela, Australia and Columbia, as well as developers of the climate surfaces in Australia, the USA and Columbia.
- A major outcome will be the development of species models at a scale that will be valuable for environmental decision making, conservation planning and development, climate change studies, disease prevention, etc.

Report 4 – Environmental Modelling (Appendix G)

Environmental, and especially species' modelling has become an important issue within the SinBiota and SpeciesLink projects. CRIA has taken this role seriously and in the past has been at the forefront of developments of modelling methodologies such as with Desktop GARP and Lifemapper, as well as having been involved in the running of many species models. A number of papers have been published on environmental modelling by CRIA staff during the past two years, and others are in preparation. Environmental modelling will continue to be a major issue in CRIA, and its importance will increase as more and better data for use in environmental models becomes available.

In a country as large as Brazil, or even as large as the State of São Paulo, no amount of environmental survey will ever provide a sufficient coverage for detailed environmental decisions to be made, or for state, or nation-wide, conservation planning to be carried out effectively. The use of environmental modelling for the planning of future surveys, for the filling of data gaps, for providing information for environmental decisions, and for conservation planning is therefore essential.

To date, CRIA has put most of its effort into modelling using GARP (Genetic Algorithms for Rule-set Production). This is an excellent methodology and has produced some valuable outputs, however it may not always be the best method in all cases, and this should be borne in mind. CRIA staff are now developing, in conjunction with a number of international collaborators, a broader Modelling Framework that will allow access to a broader range of modelling techniques and algorithms.

Methods: An examination of modelling in CRIA was conducted, in order to provide comments and ideas, to identify common pitfalls, to comment on issues of data quality, to examine different methods of modelling species' distributions, to examine some of the strengths and weaknesses of these different methods and to make recommendations on possible changes to GARP and future modelling projects that may be carried out within CRIA.

Results: A draft report on Environmental Modelling was presented to CRIA in June 2003, and modified a number of times (Appendix G).

Prior to modelling a species, one needs to consider the reason one is modelling and to what purpose the model is to be used. It would seem to me that a lot (but not all) of the modelling done so far in CRIA has been for the purpose of testing GARP and for developing the modelling algorithms. There is a need for this to continue, as there are many improvements that can still be made. However, consideration also needs to be given to other uses for the models. It is these possible uses that should drive the future developments of the GARP algorithms.

"The right model in the right place" is an important consideration in modelling, and is important not to expect that any one model or modelling method, will provide all the answers. The report makes some recommendations along the lines of not placing full emphasis on one modelling methodology, but to examine a range of modelling methods with the aim of selecting the most appropriate for the data available and for providing the results one is looking for.

A lot of effort to date has been put into the development of suitable species datasets for use in modelling, and hence the *species*Link project. The equivalent effort has not, however, been given to the development of suitable environmental datasets. A number or recommendations are made with respect to modifying the data layers being used in the GARP modelling tool, as well on the scale of modelling being conducted. In addition, time was spent acquiring improved datasets for use in the environmental modelling being conducted at CRIA.

The report makes a number of suggestions for future modelling research, as well as modifications to the software GARP.

Key issues include:

- The need to examine the purpose for which a model is being run
- Not all modelling software works in the same way and will thus produce different results. There is a need to determine the best method for use in each case.
- There is a need to examine the environmental layers being used in the model and choose the best available.
- The issue of scale should be looked at. Is the model being run at an appropriate scale for the data and the results required?\

Significant Achievements:
Following the presentation of the first draft of the report to CRIA, a number of the suggestions have already been either implemented or are planned, including to the methods used for modelling in CRIA (and elsewhere in South and North America). Some of the major outcomes of the project have been:

- Improved resolution of climate layers available for use in GARP and other models (see Report 3, above)
- The implementation of environmental layers that have more relevance to the environment (e.g. mean temperature of wettest and driest quarters and rainfall of warmest and coolest quarters rather than rainfall and temperature in January and July).
- Modifications to the methods used to prepare and resample grid layers being used in GARP, such as the use of Bilinear Interpretation and Cubic Convolution rather than the Nearest Neighbour method previously used.
- Modifications to the way environmental layers of different scales are combined.
- Planned modifications include possible extension of the methodologies for use in marine and aquatic environments; improved methods of validating the results; incorporation of a probability surface in the output and improved visualisation of outputs.
- In addition, a project has been commenced in SinBiota to develop a broader modelling framework, including the integration of a number of modelling methods, scales and data layers accessible in a user definable way.

Report 5 - Data Quality – a Discussion Paper (Appendix H)

A broad background paper was prepared covering issues of data quality. This paper is a precursor to Report 6, and the issues are covered in that discussion.

Report 6 – Data Cleaning Tools (Appendix I, J)

Museums and herbaria throughout the State of São Paulo and elsewhere in Brazil have begun to database their collections. Some of these, especially in the State of São Paulo are being carried out as part of the FAPESP/Biota *specie*sLink project being managed through CRIA. The main goal of the *species*Link project is to implement a distributed information system to retrieve primary biodiversity data from collections throughout the State. Twelve collections (3 herbaria, 2 acari, 3 fish, 1 algae and 3 microorganism collections) are already engaged in the first phase of the project. Others will join the project from time to time.

Errors in data are common, but a good understanding of errors and error propagation can lead to active quality control and managed improvement in the overall data quality. Errors in species' data are particularly common with errors in spatial position (geocoding) and in taxonomic circumscription two of the most common errors found in specimen databases. These errors can cause major problems in modelling and biogeographic studies. Assessment of the accuracy of input data is essential otherwise the results of any modelling will be meaningless

A large proportion of my time at CRIA was spent examining issues of data quality and data cleaning. This is particularly relevant, as I have spent a lot of the past 25 years examining these issues and developing methodologies to assist users in cleaning their data.

> **Methods:** Existing tools and methodologies for use in testing, cleaning and validating species data were examined; The feasibility of developing and producing a tool kit for data cleaning, along with the feasibility of developing guidelines for best practice were also examined. A number of talks were given to institutions on data cleaning and validation.
>
> **Results:** A draft report on Environmental Data Cleaning tools was presented to CRIA in June 2003. A subsequent first draft Best Practice manual was developed from this for possible use by the Global Biodiversity Information Facility, and a joint paper with two American researchers prepared for presentation at the Sixth International Symposium on Spatial Accuracy to be held in Portland, Maine in June 2004.
>
> The report examines a number of existing methods and guidelines for identifying errors in taxonomic data as well as spatial data, and makes a number of recommendations. The report recommends, among other things, the preparation of a Data Cleaning Toolkit on CD for distribution to museums, herbaria and other institutions. The Toolkit would include some publicly available software, guidelines to methodologies, links to on-line resources, and some universal datasets (such as species names for use in pick lists, etc.).
>
> Key issues include:
>
> - Recommendation for development of Data Cleaning toolkit
> - Examination of a range of available methods for data cleaning and validation
>
> **Significant Achievements:**
>
> - Development of Guidelines on Nomenclature (Appendix J)

- Development of an on-line method for detecting outliers in Latitude, Longitude and Altitude - spOutlier(**http://splink.cria.org.br/tools/**).
- Development of an on-line methodology for assigning latitude and longitude (localidade) at a distance and direction from a gazetted point –geoLoc-CRAIA (**http://splink.cria.org.br/tools/**).
- Presentation of a joint paper on Data Validation to the Sixth International Symposium on Spatial Accuracy.
- Preparation of paper on the new on-line tools.
- Request to produce a jointly authored best Practice document for the Global Biodiversity Information Facility on data cleaning.
- Agreement to include some software, some documents and data on a Data Cleaning Toolkit CD if it goes ahead.

Report 7 – Lifemapper (Appendix K)

Lifemapper (University of Kansas 2003a) was developed in the 1990s and is an on-line computer program managed by the University of Kansas. It is an attempt to test the possibilities of running distributed environmental models on the internet using data collected from a range of distributed sources. CRIA has been involved in assisting in the development of methods (used in the *species*Link project and in the Species Analyst used with Lifemapper) for accessing data from a range of distributed museums, and in developing the GARP software used in Lifemapper. CRIA and the University of Kansas have entered into a collaborative arrangement to work on these and other issues.

The first modelling of species on the internet was developed by myself and a number of colleagues at the Environmental Resources Information Network (ERIN) in Australia in 1994. From that early internet modelling arose GARP (Genetic Algorithm for Rule-set Production). More recently, a desktop version of GARP has been written by Ricardo Scachetti-Pereira at CRIA, and this forms the basis of the modelling behind Lifemapper.

Because of the relationship between CRIA and the University of Kansas, and the role CRIA has played in the development of Lifemapper, it was decided to examine the system and report on it. The resultant report was presented to CRIA in July 2003, and a copy forwarded to the University of Kansas

Methods: An examination of the Lifemapper system was conducted from a users point of view and a number of recommendations made.

Results: A report on Lifemapper was presented to CRIA in July 2003 and subsequently a copy was forwarded to Jim Beach at the University of Kansas.

Key issues include:

- The need to examine the environmental layers being used for the modelling (similar issues to that mentioned under the Modelling Report No. 4 – Appendix G).
- The need to look at the scale of modelling being carried out.
- The need to include caveats on the input data (because the data shown does not always include all possible data available, and often does not cover the entire known range of the species).
- The issue of using atmospheric climate for the modelling of aquatic species such as fish.

Significant Achievements:

- I have been informed that the Kansas staff responsible for Lifemapper will be implementing most of the recommendations in the report in the near future.

Report 8 – BioLink 2.0 – a preliminary evaluation (Appendix L)

BioLink (Shattuck and Fitzsimmons 2000) is software developed by the Australian National Insect Collection in Australia for databasing museum and herbarium information. There are a number of other programs developed for similar purposes and a range of these are being used by institutions involved in supplying data under the *species*Link project. Other programs include Biota (Colwell 2002), BRAHMS (University of Oxford 2003) and Specify (University of Kansas 2003b).

In June, a new version (Version 2.0) was released and it was thought valuable to do a brief evaluation of the software in order to determine its suitability for use by any institution involved with *species*Link who may wish to do so.

Methods: An examination of the BioLink software was conducted, looking at its usability from a user point of view as well as special aspects of the software that may not be available in the alternatives. No comparison was made with the other software choices available, and no actual data was entered into the database.

Results: A report on BioLink was presented to CRIA in September 2003.

Key issues include:

- The software has good import and export routines using XML
- Some problems were found with the use of the included software EGaz outside Australia, and the developers were notified of these.
- The software has a strong bias toward entomological data. For this reason, it would be an ideal choice for databasing of entomological or related collections. Its use with botanical data is not recommended as there are better choices available.
- The software has good links to export routines for linking to character-based databases such as DELTA and LUCID.

Significant Achievements:

- While investigating this software, I became aware of a number of good gazetteers covering South and Central America, and copies these were imported for use in CRIA, including with the on-line localidade system mentioned above.

Other issues covered

While I was at CRIA, I was asked to use my expertise to assist staff with a number of other issues. In addition, my broad contact and collaborations were used to help develop long-term collaborations with a number of international scientists. These included:

- Assisting with the preparation of a questionnaire and subsequent report on Data Sharing with Countries of Origin for presentation to the Global Biodiversity Information Facility as part of a contract on data repatriation and data sharing.

- Assisting with the planning and running of an Inter-American Workshop on Data Access run under the sponsorship of the Committee on Data for Science and Technology (CODATA).
- Assisting staff, students, and others with writing of scientific papers and reports, and reviewing a number of papers before submission for publication.
- Carrying out many discussions with staff, students and researchers across the State on a broad range of technical and scientific issues related to data management, environmental management, conservation planning and use of data in a policy environment.
- Writing a number of scientific papers in conjunction with staff at CRIA, as well as with international researchers.
- Presenting a number of seminars on data management, data quality, environmental modelling and decision support systems.
- As part of my membership of the Scientific Advisory Committee to the FAPESP-Biota program, carrying out the 4th Annual evaluation of the program in association with other members of the SAC.

7. Conclusion and Recommendations

I believe the project was a worthwhile project, and will lead to a much higher level of data management, data quality and standardisation of biodiversity data within the State, as well as allowing for an improved quality and level of environmental modelling, to a level which will be of value in environmental management, decision making and conservation planning.

There are a number of other topics that were not covered by this project. I recommend that FAPESP and CRIA give consideration to a further project to look at these issues, along with examining how the implementation of the recommendations made in this report have progressed. Issues not addressed that I believe should be considered are:

- Climate-change modelling. This has aspects of modelling that have not been considered within this report. They include the scale at which modelling is reasonable and practical, the use of Global Climate Change Models (GCMs) versus Regional Climate-Change Models (RCMs), how climate-change algorithms are best included into existing modelling methods, and the development and use of outputs for environmental planning and decision-making.
- Reserve and conservation-selection methodologies. A number of different methods exist around the world, and there needs to be an examination of which methods may be most appropriate for use within the State of São Paulo given the data available, as well as aspects of training and development.
- Decision-support tools. The time is rapidly approaching when data and tools arising out of the FAPESP-Biota Program will begin do be used for environmental decision making, policy formulation, conservation assessment, etc. To best carry this out, on-line decision-support tools will need to be developed. An examination of available options needs to be conducted, along with a user-needs assessment.
- Environmental Risk. As more and more environmental decisions are based on the data being made available through speciesLink, and SinBiota, etc., then the more the issue of Risk and Risk Assessment come into the equation. Risk Assessment in the ‘species’ area has been seldom touched on around the world, but if costly decisions are being made on the location of an endangered species, for example, or on a modelled distribution from GARP on the likely spread of a weed or disease, then methods for assessing the risk inherent in those decisions must be considered. An examination of available options and methodologies and recommendations on future considerations in this area made.

I believe that FAPESP and CRIA received more than value for money from the project, and would strongly endorse the process whereby external researchers are brought into the program from time to time to make their expertise available to researchers across the State. There have been several successful examples of this being done at CRIA in the past, and it is one of the key aspects to CRIA being able to maintain its place as a world-leading innovator in the field of Biodiversity Informatics. This has resulted in providing FAPESP and the Biota program with an effective and efficient means of making the results of their valuable research projects available to researchers and the public alike, not only within the State, but nationally and internationally.

I would recommend to FAPESP, that projects like this one be continued and encouraged, along with projects that send young researchers overseas for out-posting in relevant overseas institutions for short periods.

I thank FAPESP for the opportunity of carrying out this project and the staff and directors of CRIA, and the many researchers and students of the Biota program for making my stay in Campinas a pleasant and memorable one, and one that was extremely productive.

8. References

Belbin, L., Austin, M.P., Margules, C.R., Creswell, I.D. and Thackway, R. (1994). *Modelling of Landscape Patterns and Processes Using Biological Data. Sub Project 1: Data Suitability*. Report to Australian National Parks & Wildlife Service, Canberra.

Burbidge, A.A. (1991). Cost constraints on surveys for nature conservation. **In** C.R.Margules and M.P.Austin (eds), *Nature Conservation: Cost Effective Biological Surveys and Data Analysis.* CSIRO, Melbourne, pp. 3-6

Chapman, A.D. (1998). *Guidelines for documentation of species and vegetation data*. Canberra: Environment Australia. http://www.deh.gov.au/erin/documentation/pubs/metadata.doc [Accessed 11 Feb 2004].

Ferrier, S. and Watson, G. (1997). *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity*. Canberra: Environment Australia. [Also available electronically at http://ea.gov.au/biodiversity/publications/technical/surrogates/].

NLWRA (2004). *Natural Resources Information Management Toolkit. 2. Data Management Principles.* Canberra: NLWRA. [Available on-line at http://www.nlwra.gov.au/toolkit/2/2-3.html#236].

Shattuck, S.O. and Fitzsimmons, N. (2000). *BioLink, The Biodiversity Information Management System*. Melbourne, Australia: CSIRO Publishing. http://www.biolink.csiro.au/.

University of Kansas (2003a). *LifeMapper.* Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://www.lifemapper.org/

University of Kansas (2003b). *Specify*. Biological Collections Management. Lawrence, Kansas: University of Kansas http://usobi.org/specify/.

University of Oxford (2003). *BRAHMS. Botanical Research and Herbarium Management System*. Oxford, UK: University of Oxford http://storage.plants.ox.ac.uk/brahms/.