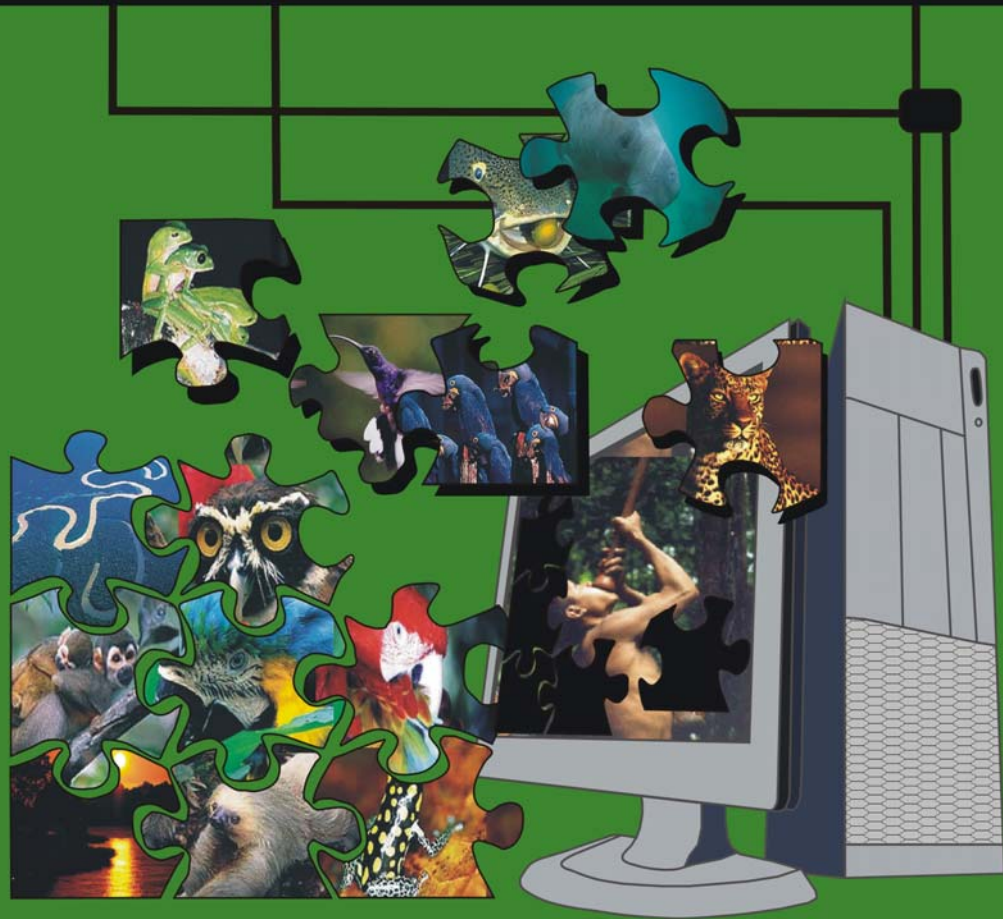


J.L. Campos dos Santos

A BIODIVERSITY INFORMATION SYSTEM IN AN OPEN DATA/METADATABASE ARCHITECTURE

A Biodiversity Information System in an Open Data/Metadatabase Architecture



J.L. Campos dos Santos

**A BIODIVERSITY INFORMATION
SYSTEM IN AN OPEN
DATA/METADATABASE ARCHITECTURE**

José Laurindo Campos dos Santos

juni 2003



INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE
AND EARTH OBSERVATION
ENSCHEDA, THE NETHERLANDS

ITC Dissertation number 100
ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands

Samenstelling van de promotiecommissie

Promotoren: Prof. dr. Peter M. G. Apers
Assistent-promotor: Dr. ir. Rolf A. de By
Leden: Prof. dr. ir. A. Nijholt (UT, EWI)
Dr. ir. H. M. Blanken (UT, EWI)
Prof. dr. ir. M. Molenaar (ITC)
Prof. dr. E. J. de Bruijn (UT, BBT)
Dr. C. U. Magalhães Filho (INPA, Brazilië)



CTIT Ph.D. thesis series number 03-52
CTIT, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISSN 1381-3617

ISBN 90-6164-214-0

Cover designed by Tomislav Hengl

Printed by ITC Printing Department, Enschede, The Netherlands

Copyright © 2003 by J. L. Campos dos Santos

All right reserved. No part of this publication apart from bibliographic data may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photo-copying, recording, or otherwise, without the prior written permission of the author, Hengelosestraat 99, P.O. Box 6, 7500 AA Enschede, The Netherlands.

A BIODIVERSITY INFORMATION
SYSTEM IN AN OPEN
DATA/METADATABASE ARCHITECTURE

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F. A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op woensdag 4 juni 2003 te 15:00 uur

door

José Laurindo Campos dos Santos

geboren te Manaus, Amazonas, Brazilië

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. Peter M. G. Apers

Assistent Promotor:

Dr. ir. Rolf A. de By

To My Two Graces,
Sidínea and Adriana

“Nature does not proceed by leaps.”

Carolus Linnaeus (also Carl von Linné),
(Swedish botanist, 1707 – 1778, father of taxonomy)

Nemesis Divina,
University Press of America, 2002.

Preface

After years of studying Computer Science, I concentrated my attention to data models, database design and implementations according to software engineering standards. When I joined INPA, I was eager to apply my skills to develop solutions to help scientists in their work. In reality, INPA, with its multi-disciplinary environment, had too many problems in data and information management to be addressed and too few experts to face them. To aggravate that, this would be expected to be the picture in similar institutions all across the Brazilian Amazon region. I saw this as a challenge and a lifetime opportunity to contribute, but I knew I had to cross a bridge towards bio-science, which would require me to open my mind to new disciplines in science with no way back.

After some successful solutions applied at INPA and elsewhere, our efforts paid dividends, and in 1997, I was invited to join the Pre-LBA Data Set Initiative and following, the LBA-DIS Working Group and NASA LBA-Ecology. During that time, we discussed about the overall LBA-DIS data issues and the positive impact on bio-science. The issues were divided into five categories: data policies, data and information standards, satellite data, DIS implementation, and management and communications. LBA-DIS today, as a result, is becoming a major integrator of science for the LBA project. In parallel to that, we participated intensively in major experiments in the Brazilian Amazon region, particularly the strengthening of the Scientific Collection Program at INPA and other institutes. From these incursions I became familiar with some bio-science problems regarding data and information analysis, and dissemination. I gathered sufficient evidence on how certain problems could be approached. Particularly, the design of conceptual models for database implementation was needed and should be materialised as a robust and state-of-the-art information system with a well-balanced dose of computer science theory and engineering delivery.

This thesis provides my attempt to addresses the problem of computer infrastructure functionality, database conceptual schema representation and data/metadata management of biological data. Additionally, it addresses the problem of legacy data by applying a georeferencing method integrated with the database as well as a proposal for an automatic protocol to resolve taxonomic differences.

To be honest, as an optimistic, I envisaged to go beyond my research plan. There still is a lot to be done. I hope to have a more realistic project to do it in a better way. But I did succeed, not only to cross the bridge, which gave me satisfaction,

but to understand a fraction of the spectrum of the biological data problems. I am convinced now that these problems are huge and extremely complex. This confirms that computer science techniques can and will deliver the important tools that may accelerate the production of scientific information in the bio-science domain.

José Laurindo Campos dos Santos
June 4 2003

Summary

Species and genetic diversity in different ecosystems are important components of biodiversity. To find answers to crucial questions concerning how they function, millions of organisms collected from the tropical rainforest and its water bodies have been and continue to be deposited in biological collections. Large datasets were collected and compiled during many unrelated and independent studies across the Brazilian Amazon region since the last century. Individual researchers are unable to fully comprehend these data and information pools, as their current and newly arising questions may depend on multi-disciplinary contexts and on well-documented data. Since some institute activities are related, but others are not interrelated at all, the majority of adopted solutions of data ordinance at this moment suffer from redundancy, data inconsistency, and interpretation gaps, leading to high costs in labour, processing and infrastructure, and consequently have less than optimal scientific results.

Computer technology has been a fundamental resource applied for bio-information management. For successful use of this technology, there are a number of requirements: an accurate information model, formal data and metadata management, as well as methods to integrate and revive legacy data, amongst others, by adding geographic information and analysis capability.

This thesis presents an overview of biological collections, their complexity and related bio-information management activities in the main institutes in the Amazon region. The adopted functional and system analysis was the result of interactions by interviews, information requirement analysis, data flows and evaluation of descriptions, with the participation of researchers as users, and curators as information managers and data providers.

A conceptual database schema, CLOSi (Clustered Object Schema for INPA's biodiversity data collections), was conceived to facilitate and stimulate the development of biological collection databases.

Usually, researchers refer to their data as raw data, which are structured in rows and columns of numeric or encoded sampling observations.

The usefulness of such data can only be assessed when they are associated with either a theoretical or conceptual interpretation model. This requires understanding of the types of variables, the units adopted, the potential biases in the measurements, sampling methods and a number of facts that are not represented in the raw data, but rather in the metadata. As a fact, information can be lost through degradation of the raw data or through lack of metadata. Metadata provides data users with the ability to locate and understand data through time. Data and metadata combined within a conceptual framework improves information production. An XML-based solution for the management of metadata biological profiles via the web was implemented. The presented solution uses the FGDC metadata standard, which incorporates the biological data profile, and which is represented as an XML schema and storage in a web-based XML repository.

Further, the design and implementation of a database, a web interface for biological collections and their integration with a retrospective georeferencing tool, with the advantage to contribute to collaborative gazetteer initiatives is detailed. The database development is based on the CLOSi schema as well as on building a web interface to access the database, supported by a three-tiered system architecture. The architecture consists of a client user interface over the web, an application server and a database management system.

This structure was produced from in-depth research of the needs of those who use biological collection data. The solutions made available can benefit institutes similar to INPA tremendously. The problem of *ad hoc* system development can be reduced considerably and resources can be diverted to the utilisation of the presented database design and implementation. An implemented database system will pave the way for much faster data digitising of biological data. With these two in place, data exploration and information dissemination can be allowed through additional functionality or complementary application implementation.

For managing XML-based metadata, a client-server set-up was used that allows metadata and data dissemination for users at the global scale via the web. This work is particularly suitable for less developed countries since the system developed and tools used are either public domain or free open source, ensuring low cost for a robust solution to any particular institutional environment.

The integration with the georeferencing application was found to be a successful and valuable addition to the web system, providing great benefit to researchers of biological collections that need geospatial analysis results. The advantages of this process include: increasing speed in georeferencing, maximising consistency between users, allowing the incorporation of interpretation standards established by researchers, especially curators, and quantifying textual locality vagueness.

A taxonomic reference system is extremely important for any biological scientific activity. Systematics is the research field within biology

that attempts to unfold the (evolutionary) tree of life, i.e., the biological taxonomy. Biologists do not know that tree, and are in the process of slowly uncovering its secrets in which they differ in their beliefs. To deal with this problem, we propose a negotiation protocol framework that would help automatic systems do this. We abstracted away, however, from certain complexities of taxonomic practice, such as the presence of synonyms and misspellings, and assume error- and redundancy-free taxonomic data sets on both ends, which may (and will) however differ in structure and completeness.

Thus, we could be providing a useful solution for the exchange and communication of biological information to the wide biological research audience, aiming to facilitate the research on our biological resources, satisfying our growing need for answers on their importance for our biosphere.

Summary

Acknowledgements

I believed to be born and live in the Amazon region was hard in all senses, but now I know that it was just a training camp. To embrace science and take a journey into a PhD research requires more than my survival kit could provide. Like in the jungle, the survival depends on the strength and harmony we build around us. My strength to reach this point came from people. Without their help, encouragement, appreciation, friendly contributions and advices this work would not have been possible and my life not so enjoyable.

First and foremost, I want to express my sincere gratitude to my promoter Prof. Dr. Peter Apers for his contribution and strong support throughout the work. It was a privilege to have Dr. Rolf de By onboard as my supervisor. Rolf provided me an outstanding professional guidance and constant support, feedback and advices. I am pleased for his personal views and to be there for all his students not only as a guide but as a friend.

I would like to thank the National Institute for Amazon Research, INPA, for seeing the importance of my studies and for offering me a fertile ground to implement these ideas and to the International Institute for Geo-Information Science and Earth Observation, ITC, for partially financing this research project.

At ITC, I want to express thanks to the former research coordinators Dr. Liesbeth Kusters and the current coordinator Prof. Dr. Martin Hale as well as Loes Colenbrander (piloting the PhD help desk), for all the support. To the secretaries Marijke Smit and Saskia Tempelman. To the colleagues from GIP, especially Martin Ellis, Irberto de Sousa, Rob Lemmens and Yuxian Sun. In the Educational Affairs Office, André Klijnstra, Bettine Geerdink, Marie Metz and Teresa van den Boogaard. To ITC librarians Marga Koelen, Carla Gerritsen, and Petry Maas-Prijs, excellent book and paper trackers. The staff at the IT department, Ard Blenke, Ard Kusters and Cecille Plomp. To the reception officers Roelof Schoppers, Hans Verdam and Tom Busscher – who can miss the announcements before the closing time? I did, but just once. To the ITC Hotel officers, Saskia Groenendijk and Marjolein Woerlee. Thank you all!

I want to thank researchers, curators, and technical staff who work with biological collection and related disciplines in the Amazon region, especially Dr. Angela Varella, Dr. Célio Magalhães, Dr. Charles Clemment, Dr. Efrem Ferreira, Dr. Fernando Rosas, Dr. Isolde Ferraz, Dr. Jansen Zuanon, Dr. João Ferraz, Dr. José Celso

Acknowledgements

Malta, Dr. José Rafael, Dr. Lúcia Rapp, Dr. Mario Cohn-Haft, Dr. Milton Nakashiro and Dr. Ronaldo Barten.

Thanks to Dr. Ozório Menezes Fonseca, the former Director of INPA, for entrusting me with the project and implementation of INPA Network and for supporting this project since the very beginning. To Jurg Sonderegger, Volker Duzat and Ricardo Rios for joining me during the INPANet Project. I had promised you fun in bringing technology into the jungle. I forgot to say that it would not always be like that.

To Marguerite Reder and Eric Marcon for the good time spent in Silvolab, Kouru, in French Guyana, discussing solutions for the Oiapoque project and fixing bugs to make data dissemination possible.

I would like to thank Dr. Cláudio Ruy da Fonseca, Dr. Júlia Ignez Salem and Dr. Peter Weigel for providing insightful views in the BIOAMAZÔNIA 2002 Report.

I am in debt with my good friends and colleagues Edwin Keizer, Etien Koua, Dr. Patrick Ogao and Wim Bakker for the reading, comments on drafts of papers and chapters of the thesis. I appreciate your help very much. As you can notice here I did follow your tips. To Dr. Lorena Montoya for the great help with the Spanish translation – muchas gracias!

I want to thank all PhD colleagues and friends for their company, help and discussions on scientific and social topics, either personally or via the ITC PhD committee. I want to address special thanks to Ivan Bacic, Giancarlo and Renata Guizzard, Dr. Mohammed Said, Dr. Liu Xuehua, Dr. Liu Yaolin, Zhang Qingming, Tomislav Hengl, Jelle Ferwerda, Arko Lucieer, Javier Morales, Dr. Citlalli Lopez, Narciso Basols, Dr. Cheng Jianquan, Dr. Huang Zhengdong, Masoud Kheirkhah, Alfred Duker, Arta Dilo, Nirvana Meratnia, Grace Nangendo Uday Nidumolu, Richard Onchaga and Pravesh Debba.

Life in ITC is not always hard work, the social and sport life make the difference to keep us fit. Thanks to the Dean, Jan de Rooter and our football team coach Jaap Duim and Johan Weggen.

These are indeed special people that had joined my life forever, Affonso Magalhães (in memory), Blanca Perez, Claudia Pittiglio, Dorit Gross, Ellen Steur, Dr. José Alberto da Costa Machado, Gabi Zimmermann, Katrin Möenter, Michael Bisset, Maura Gebaska, Monija Ivankovic, Dr. Norman Kerle, Sara Ruto, Sieberen Bosch and Silvia Giada. For you a big THANKS is not enough, you deserve all my gratitude.

I would finally like to thank my wife Sidíneia Amadio and my daughter Adriana Amadio Santos, my front line team, for their love, patience and sacrifice. Dr. Sidíneia had to slow down in science and undertake the whole burden of our family. My little princess Adriana took the writing management from Manaus. I had to report to her every new chapter written. By no means I could ever repay you. This work is yours and you should know that it could have never come into, without you being my source of inspiration and encouragement and without your love. I am always enchanted by the light you both brought to me. Thanks to my sisters Ana and Liliana and my brother Mário (the Campos dos Santos tribe) for their long-distance encouragement, and for keeping me abreast of news from home. For the loving memory of my mother and father. I believe they are happy today, just the way I am.

Contents

Summary	i
Acknowledgements	v
1 Biodiversity Scenario	1
1.1 Background	1
1.2 Experiments in the Brazilian Amazon	3
1.3 Overview of this Research	12
1.4 Outline of the Thesis	15
2 X-Raying Amazonian Biological Collections	19
2.1 Introduction	19
2.2 Biological collections in the Amazon region	20
2.2.1 Institutes and their collections	20
2.3 Problems with biological information	34
2.4 Bridging gaps towards BIS	36
2.4.1 A close look at INPA's initiatives	38
2.4.2 A toolkit approach to manage collection data	40
2.4.3 Technological influences for information dissemination	42
2.5 Concluding Remarks	44
3 Requirements and Functionality: Users and Systems	47
3.1 Introduction	47
3.2 Functional Requirements	48
3.2.1 Research	48
3.2.2 Conservation	49
3.2.3 Education and Capacity	49
3.2.4 Collection Management	50
3.3 System Requirements	54
3.3.1 Processing requirements	54
3.3.2 Data Types, Volumes, and Usage	56
3.3.3 Users	59
3.3.4 System and Data Security	63
3.3.5 Maintenance and System Flexibility	64

3.3.6	System Architecture Constraints	65
3.4	Summary	67
4	Clustered Object Schema	69
4.1	Introduction	69
4.2	Schema to Represent Biological Data	70
4.2.1	The strategy	70
4.2.2	The components of the schema	74
4.3	A CLOSi Schema for Biological Collection	80
4.3.1	Cluster Collection Management	80
4.3.2	Cluster Collecting Event of Collection	82
4.3.3	Cluster Locality of Biodiversity Data	83
4.3.4	Cluster Taxonomy	83
4.3.5	Cluster Agent of Collection	84
4.3.6	Cluster Reference	84
4.4	Reflection on the Design Work	86
4.5	Strategy to Compare CLOSi Effectiveness	87
4.6	Concluding Remarks	90
5	An XML-based Solution for Bio-Metadata	93
5.1	Introduction	93
5.2	Metadata in Brief	95
5.2.1	Scope	95
5.2.2	Content	95
5.2.3	Why Metadata?	97
5.2.4	Standards	98
5.2.5	The Biological Data Profile	100
5.3	From Biological Profile to XML Schema	102
5.3.1	XML criteria and the mapping process	102
5.4	Alternatives for Metadata Management	103
5.4.1	Application deployment and management from a clearinghouse	103
5.4.2	Metadata description and management supported by web application	105
5.5	Implementation: Three-tier Architecture	106
5.5.1	Client Side	108
5.5.2	Web Server	108
5.5.3	BiOME XML Components	109
5.5.4	XML (Database) Server: Metadata Repository	109
5.6	Discussion of the Implemented Solution	111
5.7	Concluding Remarks and Future Work	112
5.7.1	An Active Node Approach	113

6	Implementing Biological Data on the Web	117
6.1	Introduction	117
6.2	Tools for Website Development	119
6.2.1	Open Source Systems	119
6.2.2	The Web Server	119
6.2.3	The Database Management System	121
6.2.4	The Server-side Script	124
6.3	Prototype Implementation	127
6.4	Conclusions and What Comes Next	137
7	Georeferencing Amazonian Bio-Data	141
7.1	Introduction	141
7.2	Georeferencing Process	144
7.2.1	Expressing Latitude and Longitude	144
7.2.2	Dealing with distance and direction uncertainties	146
7.3	Understanding Vagueness and Uncertainties in Distance and Direction	147
7.3.1	Example of distance imprecision	152
7.4	Experiencing Retrospective Georeferencing for Amazonian Data	153
7.5	Potential for Collaborative Gazetteer	159
7.6	Appraisal of the Pros and Cons	161
7.6.1	Improvements from previous methods	162
7.6.2	Foreseen future	162
7.7	Conclusions	163
8	Automatic Reconciliation of Taxonomic Belief Differences	165
8.1	Introduction	165
8.2	Systematics in brief	166
8.2.1	Biological classification	166
8.2.2	Some basic terminology	170
8.2.3	The Linnaean TRS	170
8.2.4	The Phylogenetic TRS	171
8.3	Integration of taxonomic data sets	172
8.3.1	Why do taxonomic beliefs differ?	172
8.3.2	Components of the problem of communicating over taxonomy	172
8.3.3	Problem definition	173
8.4	Framework for Automatic Negotiation	174
8.4.1	Requirements	174
8.4.2	A communication protocol	176
8.4.3	Rules of negotiation	177
8.4.4	To express rules and negotiation proposals	178
8.5	Conclusions	179
9	Conclusions and Future Research	181
9.1	Summary	181
9.2	Achievements	183
9.3	Topics not addressed	184
9.4	Future research	184

References	187
Author's Bibliography	199
Abbreviations	203
Appendix A: The Syntactic Definition for CLOSi Schemas	207
Appendix B: CLOSi Controlled Value Classes	213
Appendix C: Examples of CLOSi Instantiated Values	223
Appendix D: Metadata of a Crustacean Collection	229
List of ITC Ph.D. Thesis	237
Samenvatting	241
Sumário	245
Resumen	249
Curriculum vitae	253

List of Figures

1.1	Part of the South America map showing the Pan Amazonian region and the fieldwork area coverage.	4
1.2	The Brazilian Amazon region.	5
1.3	West Europe in relation to the Brazilian Amazon.	6
1.4	Outline of the thesis.	17
2.1	Holotype, identified and total number of specimens in EMBRAPA CPAA and CPATU biological collections.	24
2.2	Holotype, identified and total number of specimens in IEPA collections.	26
2.3	Biological Collections at INPA.	28
2.4	Holotype, identified and total number of specimens in the MPEG biological collections.	31
2.5	Holotype, identified and total number of specimens in UFAM biological collections.	33
2.6	The traditional information production method showing the absence of metadata.	35
2.7	Sources of biodiversity data and information.	36
2.8	Degradation of information.	37
2.9	Interactions amongst functions and collections.	40
2.10	INPA's current organisational structure of the Scientific Collections Program.	41
2.11	Strategy for information delivery.	43
4.1	Clusters and relationships structure of CLOSi.	74
4.2	Object class Agent, its attributes and relationships.	76
4.3	Object class Biological Collection, its relationships and attributes.	77
4.4	Object class Line its relationships and attributes.	78
4.5	CLOSi Notation.	81
4.6	Cluster Collection Management, its classes and relationships.	82
4.7	Cluster Collecting Event of Collection with its class, attributes and relationships.	83
4.8	Object class Locality, its relationships and attributes.	84
4.9	Cluster Taxonomy its classes and relationships.	85
4.10	Cluster Agent_of_Collection, its classes and relationships.	85

4.11 Cluster Reference_Work, its classes and relationships.	86
5.1 A simple understanding about metadata.	96
5.2 Sections of the CDSGM standard.	99
5.3 Biological data profile extended elements.	101
5.4 Method for mapping FGDC to XML schema and template.	103
5.5 FGDC metadata BiOME schema - root level.	104
5.6 Segment of a metadata with description for citation information.	105
5.7 The implemented three-tier architecture.	106
5.8 Screenshot of the BioME portal: a way to deploy metadata components.	110
5.9 Example of the proposed web active node approach.	115
6.1 Running a server-side script: The three-tier process for dynamic web databases.	125
6.2 Web interface mapping hierarchy.	129
6.3 Screenshot of the biological collections database search page.	131
6.4 Screenshot of specified search page for collecting events.	132
6.5 The database search sequence.	133
6.6 Screenshot of the Login page.	134
6.7 Authentication result page.	135
6.8 Data entry page.	135
6.9 Data entry result and proceed prompt.	135
6.10 Georeferenced object data entry.	136
6.11 Georeferenced object data entry result page.	136
7.1 Example of INPA's fish expedition note.	142
7.2 Sources of uncertainty.	146
7.3 Uncertainty in distance precision.	148
7.4 Combinations: distance imprecision.	150
7.5 Combinations: distance imprecision with named placed.	151
7.6 Combinations: distance and direction.	151
7.7 Combinations: The sum of distance and direction	152
7.8 Combinations: generalisation distance and direction.	153
7.9 Georeferencing process and collaborative gazetteer within the CLOSi web context.	157
7.10 CLOSi web interface with georeferencing tool.	158
7.11 CLOSi data exported to the CAS system.	159
7.12 View based on Amazonian shape files.	160
7.13 Localities georeferenced in Presidente Figueiredo, Amazonas, Brazil displaying the area where fish were catch.	161
8.1 Example of a taxonomic subtree indicating evolutionary history of some new world flycatchers	169
8.2 Framework to negotiation participants.	175
8.3 Protocol for taxonomic disambiguation between two autonomous taxo- nomic information services <i>R</i> and <i>P</i>	177

List of Tables

2.1	Biological collection held by institutes in the Amazon region - (<i>n</i>) indicates number different collections of the same type.	21
2.2	Agronomic collections and plant nurseries in Amazonian institutes. . .	22
2.3	Motivation activities that led to the formation of EMBRAPA collections. . .	23
2.4	Current use of EMBRAPA collections.	23
2.5	Application context of collections at IEPA.	25
2.6	Current use of IEPA collections.	25
2.7	Some important collections at INPA.	27
2.8	Application context of INPA collections.	27
2.9	Current use of INPA collections.	29
2.10	Application context of MPEG collections.	30
2.11	Current use of MPEG collections.	30
2.12	Application context of UFAM collections.	32
2.13	Current use of UFAM collections.	32
2.14	Collections and software diversity.	41
3.1	The MVZ preliminary five year estimation of disk space needed for alphanumeric data.	57
3.2	Number of the potential simultaneous SCP users.	62
4.1	Entomological collection information; R indicates information recorded by the interviewed researchers	72
5.1	HTTP requests of XYZFind	111
8.1	Language feature comparison	179

List of Tables

Chapter 1

Biodiversity Scenario

1.1 Background

As human populations continue to grow, the exploitation of the world's natural resources increases. The misuse of some natural resources have made them unsustainable and the world's biodiversity is being lost at an alarming rate. Primack (1993), declares that 25% of the planet's species will become extinct over the next decade, with a figure of 25,000 biological species per year. Concern about the negative impact of human actions on the environment is obligatory. As a result, an increasing number of conservation initiatives are being implemented by local communities and national governments. Additionally, the commercial sector is responding to pressure to reduce the negative impact of its activities on the environment. At all levels — local, regional, national, and international — conservation organisations have become effective advocates for programmes and policies designed to conserve and sustainably use the Earth's natural resources (International Conservation Organizations Consortium, 1998).

Many nations have confirmed their commitment to the principles of Agenda 21 by ratifying the Convention on Biological Diversity (CBD), Rio de Janeiro (1992) and other treaties related to biodiversity conservation. Worth mentioning are Conventions on International Trade in Endangered Species of Wild Fauna and Flora (CITES)(Wijnstekers, 1992), the Convention on Migratory Species of Wild Animals (CMS), Bonn (1979), Convention on Wetlands of International Importance, especially as Waterfowl Habitat (Ramsar), Iran (1971) and the Convention Concerning the Protection of the World Cultural and Natural Heritage (WCNH), Paris (1972). They have also ratified treaties relating to broader environmental issues such as the Montreal Protocol, Montreal (1987) and the United Nations Convention to Combat Desertification (CCD), Rio de Janeiro(1992) as well as the United Nations Framework Convention on Climate Change (FCCC), Rio de Janeiro (1992) (World Conservation Monitoring Centre, 1999; The United Nations Program of Action from Rio, 1992).

All treaties agree on the need for comprehensive, high quality information on the status of biodiversity. Some important conventions, such as CITES and Ramsar,

1.1. BACKGROUND

as well as a number of regional conservation initiatives, present specific information requirements. Especially the CBD, which has adopted them, has also convinced parties to develop a complex level of biodiversity conservation knowledge, reports and actions.

Sources of biodiversity data are well known for being abundant, but at the same time, are incompatible (in many ways) and are dispersed over hundreds of organisations, governmental and individual agencies. In such a framework, without proper attention, problems with data management are almost inevitable. This situation contributes to the 'data crisis' which cannot be ignored, and additionally, we increasingly, face spontaneous computer application development (Brackett, 1997).

Further, there is no catalog of what data are available, thus, access is often on an ad hoc basis, scales and contents of data sets are not consistent, and neither are their history or metadata definitions. A more complex problem lies in the quality of available data, as it is often unknown and no standard mechanism has been put in place to qualify, let alone quantify it. Thus, a common language, a system platform for integration, information exchange and management, and analytical tools are needed urgently (World Conservation Monitoring Centre, 1999).

The scientific community agrees in large that Biodiversity Information Systems (BISs) can be important tools. In fact, they may become crucial mechanisms in the decision making process for worldwide policy development of environmental preservation and biodiversity conservation. For centuries, ecological data have been collected, primarily by single or small groups of researchers in small areas over relatively short periods of time (Karaiva & Anderson, 1988; Brown & Roughgarden, 1990; Michener *et al.*, 1997; BCDAM-MMA, 1998).

The world situation brings into light questions to ecologists 'on how ecological pattern and process vary in time and space, and to understand the causes and consequences of this variability' (Levi, 1992). For such questions to be properly answered, far more data are required than could feasibly be collected, managed, and analysed under the auspices of a single individual or group of researchers (Michener *et al.*, 1997). The strategy used to find the answer to a certain extent, is to use data that have been collected by other research teams for varying purposes. Also, funding agencies' mandates have focused increasing attention on preserving, sharing, and promoting the understanding of valuable data sets.

A common practice among scientific interdisciplinary teams is to share data with expert colleagues to address specific questions. Biological data sets are neither perfect nor intuitive. The sharing mode of data collected more recently has been kept close to the data producer (within the same institute or project), who knows the subject. Little additional information is needed to use and interpret such data sets. Research teams outside of the specific subject area need highly detailed instructions or documentation to accurately interpret and analyse historic or long-term data sets, and data from complex experiments (Michener *et al.*, 1997). Usually, researchers refer to their data as raw data, which are structured in rows and columns of numeric or encoded sampling observations. The usefulness of such data can only be assessed when they are associated to some model of interpretation. This requires understanding of the types of variable, the units adopted, potential biases in the measurements, sampling methodology, and a series of facts that are not represented

in the raw data, but rather in the metadata. Data and metadata combined within a conceptual framework produces the required information (Michener *et al.*, 1997; LBA Project, 1997).

For existing information, there is the risk of loss through degradation of the raw data or the metadata. Such loss, at a glance, seems to be unavoidable, since there are many processes that can cause it over time. The loss processes can be classified as: gradual degradation of storage media, discrete events, like retirement or death of researchers involved in the acquisition of the data or obsolescence of storage technology, and catastrophic events. The loss of metadata can occur throughout the period of data collection while the rate of loss can increase after research results have been published or the experiment has ended. Specific details are most likely to be lost because of the abandonment of data forms and field notes. Over a longer period of time, degradations and history loss can reduce the information about relevant details not covered by publications (Michener *et al.*, 1997). Kirchner (1994), points out that there is another factor that causes information loss, namely the loss of the conceptual model that was used as a framework and interpreter of the data. The models may be considered simple and can be expressed using statistical techniques to represent relationships among variables. Data sets associated to time series, in general, form the basis of simulation models. Thus, preservation of the information about a data set may also involve preservation of the simulation model and its associated input and output files.

Initiatives resulting from biological conventions alone do not guarantee the solution of problems faced by institutes, which have gathered data from experiments for a long time. Experiments can be vulnerable if data and metadata problems are overlooked or ignored. Our research attempts to address some questions regarding biological data and metadata representation, management and dissemination. In the following, we summarise the currently most important experiments in the Brazilian Amazon region, thus, experiments that produce large volumes of data, and identify potential users of solutions that we can incorporate into BISs and other fields such as environmental informatics (Alcama, 2002) or top level bio-ontology (Gangeni, 2002; Staab, 2002).

1.2 Experiments in the Brazilian Amazon

The Pan Amazon ecosystem encompasses the tropical rain forest of Bolivia, Brazil, Colombia, Equador, Guyana, Peru, Suriname and Venezuela, with an area of approximately 8.6 million square kilometers. The region is one of the most complex biomes on earth and contains the largest biodiversity on the planet. Figure 1.1 presents the South America map showing the Pan Amazon countries along the equator.

The Brazilian Amazon¹ comprises the states of Acre, Amapá, Amazonas, Mato Grosso, Pará, Rondônia, Roraima, Tocantins and portions of Maranhão and Goiás (See Figure 1.2). This region alone covers a vast area of 5.8 million square kilometers, contains almost one half of the world's undisturbed tropical evergreen forest

¹The Brazilian Amazon is also cited in the literature as the Legal Amazonian (Benchimol, 1996).

1.2. EXPERIMENTS IN THE BRAZILIAN AMAZON



Figure 1.1: Part of the South America map showing the Pan Amazonian region and the fieldwork area coverage (Source: CIA World Facts Book, USA, 2001.)

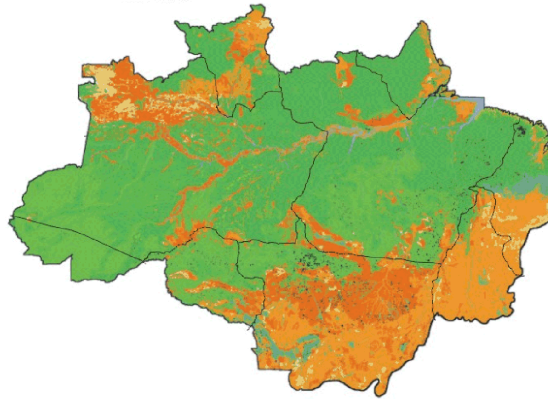


Figure 1.2: The Brazilian Amazon region (Source: Global Resource Information Database (INPE, São José dos Campos - Brazil, 2002.))

and a large area of tropical savanna. The basin is very important in the metabolism of the Earth, responsible for approximately 10 percent of the terrestrial net primary production, and it is a region of high biodiversity that the world ought to know about and cannot afford to lose (Pandolfo, 1991; Terborgh, 1992; Benchimol, 1992; Benchimol, 1996; da Silva & de Melo, 2001).

The Brazilian Amazon region corresponds to 60% of the total area of Brazil, large enough to accommodate West Europe (See Figure 1.3). Because of its importance, the Brazilian Government has prioritised its biodiversity, allocating large investments, initiatives and partnerships with international organisations to facilitate the access to and guarantee management of data collected in the region for the last one hundred years.

The 'Agenda 21,' Chapter 40, entitled 'Information to the Decision Makers,' states that information is needed at all institutional levels, that is, from local government to international councils, aiming to guarantee the sustainable development of the region. The current problems dealt with in this scope include: ignorance about the collections and their activities, redundancy of work inside the same scientific field, geographical area and period in time; lack of information about critical subjects and incompatibility of the methods and software platforms adopted (The United Nations Program of Action from Rio, 1992).

Responding to the global interest in conservation and sustainable development of the Amazon region, several projects are under development. Some of these projects are: Biomass and Nutrients (BIONTE), SIVAM (Sistema de Vigilância da Amazônia), Biodiversity Database Collection of INPA (Instituto Nacional de Pesquisas da Amazônia), (INPA-BioDB), BCDAM (Banco de Dados Compartilhado da Amazônia), Amazon Reference Database (BRISA), Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA), SHIFT (Studies of Human Impact on Floodplains and Forests in the Tropics) and various others, with more specific aims.



Figure 1.3: West Europe in relation to the Brazilian Amazon (Source: Global Resource Information Database (INPE, São José dos Campos - Brazil, 2002.))

In the following, we present general information about large scale initiatives, such as BCDAM, INPA-BioDB, LBA and SIVAM. We concentrate on the main objectives, their components, and levels of importance and investments.

BCDAM: Amazonian Shared Databases

To make recent and reliable data available is a top priority for organisations working on biodiversity. They are dedicated to monitor the region under human occupation, and the economical exploitation within the region as well as the enforcement of politics and plans, programs and projects for the sustainable development of the entire region. In Brazil, several state and federal institutes are working to gather data about Amazonian biodiversity. They aim at processing and generating detailed information about the social, economic and environmental situation of the Amazonian region and its population of 20 million people living in the area.

The BCDAM project is a solution that has been adopted to address some of the most critical problems. The project is an implementation of a cooperative system in which decentralised databases can be accessed by all participating institutes. The main aim of this system is to go beyond data access and data exchange, and to address data integration, processing and information dissemination among institutes as well.

Today, 66 institutes are participating actively and the data gathering effort that has usually been done in isolated ways now starting to focus on a more collaborative

format. However, biodiversity data maintained in several collections in Brazil are still unknown to others, unnecessary resources are spent on redundant experiments, there exists a lack of complete and current information about important themes, as well as incompatibility amongst methods and software. There is a large variety of software packages, which are mutually incompatible, making the integration and exchange of data amongst researchers difficult. This is also reflected in a lack of integration at the political level. The problem in a broader sense is not only due to lack of information. A number of institutes have acquired up-to-date information technology allowing them to implement solutions for their specific problems. Several databases have been implemented: geographical, statistical, bibliographical, legislative, project management, and expert systems. So, there is a proliferation of different databases and platforms. The problem lies in finding the information the system developers need to implement those vast systems. As a consequence, one can identify redundancy throughout these systems, though they still have gaps in important information.

The BCDAM project has expanded its objectives and focused:

- To promote and ensure biodiversity data exchange among institutes.
- To promote more rationality in the process of sampling, gathering, quality assurance, and information dissemination about the Brazilian Amazon, avoiding data and process redundancy as well as wasting financial resources.
- To promote data and information compatibility and complementarity.
- To promote project development and shorter on data gaps that may exist.

BCDAM also provides tools such as BRISA, which stores information about the Brazilian Amazon and keeps descriptions of the content and several access ways to the databases available within BCDAM. This database is available via the web and works as a central index. A user can check which database has the information s/he needs and what the data quality is. The information available in BRISA comprises data from the institutes that own the data, information about the data itself, and descriptions of the contents of these databases (the descriptions, however, do not follow a defined standard). There is also a tool available to manage links to other databases distributed in the region (BCDAM-MMA, 1998).

INPA-BioDB: Biodiversity Information System

INPA has more than 40 years of tradition in studying the Brazilian Amazon, the biggest region of biodiversity on our planet. Millions of organisms, collected from the rainforest and its water bodies, have been deposited in the biological collections of the institute. They are divided into six botanical, four micro-organismal and nineteen zoological collections, made up from vertebrates and invertebrates.

In 1997, the institute started a project that included the development of its own information system for the management of INPA's collection data. The restricted resources did not allow the development of a complex system in one single step, so it was decided to start with one part of the collections, the entomological collection (insect collection). This is the biggest collection at INPA and represents also a big

diversity of information as there are various researchers and departments working with insects.

The Entomological Database Model project (EDMI), started with an evaluation phase during which the requirements of a collection information system and publications about such projects, data models and systems were defined and studied. The system used by the Museum of Vertebrate Zoology (MVZ) in Berkeley, Calif, was used as reference. Later on, the models of the Association of Systematic Collections (ASC) and the Ohio State University Insect Collection (OSUIC) (Johnson, 1997) were also considered. The effective data requirements at INPA have been evaluated with the help of various researchers at the institute who are working with insects.

For the modelling itself, a method called Object Protocol Model (OPM) of which the representation had been modified (Chen & Markowitz, 1995; Chen & Markowitz, 1996; Sonderegger *et al.*, 1998) was adopted. The Prototype of INPA-BioDB was based on the model EDM; a first prototype was developed consisting of a database and the interfaces for the data input and for accessing the system via the world wide web (Oliveira, 1999). In Chapter 2, we present a detailed diagnostic of INPA's collections in the context of the Amazonian institutes.

LBA: Large-Scale Biosphere-Atmosphere Experiment in Amazonia

The LBA experiment is an international research initiative led by Brazil in conjunction with NASA LBA Ecology, which aims to design and create the knowledge needed to better understand the climatological, ecological, biogeochemical and hydrological processes that occur in the Brazilian Amazon. Also included are the impact of land use and land cover, and the inter-relationships between the Amazon ecosystem and the Earth system (Nobre *et al.*, 1996).

The data that have been gathered by the contributors are voluminous and complex. A data management strategy is needed for the study in which the database will act as repository for all scientific data, including spaceborne data, airborne, and ground-based data. The large volume of data to be stored, maintained, retrieved and analysed, and the distances between sites that produce and consume data had to be taken into account. There is a need for integrating data that differ syntactically and semantically, generating derived data to be incorporated dynamically in the repository and incorporating the many semantic levels that the repository may have.

The system for this task has to be a flexible computational environment that operates on the basis of independent and integrated subsystems comprising three levels of abstractions: the component, analytical and distribution levels. The component level deals with collected data; the analytical level acts as a scientific notebook (with a set of scientific functions for data analysis) and the distribution level is used for storage, analysis and information dissemination to the community outside of the experiments.

Information technologists and application domain experts from South America, Europe, and the United States attended a NASA LBA Ecology workshop in College

Park, Ma, to discuss past experiences, lessons learnt and potential options for implementing an information system for the LBA project. This international research community aimed to design and create the knowledge needed to address the objectives of LBA (Nobre *et al.*, 1996; LBA Project, 1997).

To understand how the Amazonian ecosystem functions, it is necessary to quantify, understand and represent the physical, energy, water, carbon, trace gas, and nutrient process cycles found within the area. It will help to determine how these relate to the global atmosphere and respond to deforestation, agricultural practices and other land use changes, and how these responses are influenced by climate.

Discussions during the workshop regarding the management, distribution, and archival of LBA data clarified the idea that these activities would be different from those of previous experiments such as FIFE² and BOREAS.³

The LBA-DIS is a component of the LBA that deals with the global repository of data produced by the research of the LBA project. Although many similar systems have been developed in the past few years (Kobler *et al.*, 1995; Koschel *et al.*, 1996), some of the main characteristics of the LBA demand the development of software solutions that involve state-of-art computer technology. For example, client/server architecture and semi-structured data management.

The LBA-DIS was based on three project phases identified as: (a) Primary Investigator to Primary Investigator (PI to PI), science data acquisition and preliminary data analysis; (b) PI to DIS, science data storage, data quality assessment analysis, processing and valued added product generation, and (c) Data to Archive - archival and distribution. This approach is analogous to the publication of a scientific paper where the scientist (a) collects and analyses data and writes a paper; (b) the paper is submitted for peer review and edited to conform to publication standards; and (c) the document is published and placed in a library for access (LBA Project, 1997).

LBA-DIS output may be considered a computational environment that can operate as a tool to help the LBA scientific teams to produce scientific results instead of a simple database system. The system must be more than a database system that manages a large volume of data generated by different scientific teams, as it must also have features that will help in the effective analysis of data produced.

A consensus among the LBA-DIS working group is that the activities to manage the wide range of data need to be distributed across the different LBA organisations involved, implying the existence of multiple DIS nodes that are responsible for different sets of the LBA data. The idea of this DIS node architecture brings

²The First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE) was a large-scale climatology project conducted on the Konza Prairie of central Kansas from 1987 through 1989. This project was designed to improve understanding of carbon and water cycles; to coordinate data collected by satellites, aircraft, and ground instruments; and to use satellites to measure these cycles (Oak Ridge National Laboratory, 1999b).

³The Boreal Ecosystem-Atmosphere Study (BOREAS) was a large-scale experiment initiated in 1990 aiming to investigate interactions between the boreal forest biome and the atmosphere. Surface, airborne, and satellite-based observations were collected to study the biological and physical processes and conditions that govern the exchanges of radiative energy, water, heat, carbon, and trace gases between boreal forest ecosystems and the atmosphere, particularly those processes that may be sensitive to global change (Oak Ridge National Laboratory, 1999a).

home an important concept, the metadata, or documentation describing pertinent aspects of actual data. Metadata provide the ability to organize and maintain an organisation's investments in its data. It can ensure the users' ability to locate and understand data in the present and in the future (LBA Project, 1997).

The LBA-DIS supports several functions. These functions are divided into mandatory functions, comprised of information ingestion, management, distribution, integration and long-term archiving; and support functions, that is, software tool development, data registry, metadata preparation, data quality assessment and data documentation.

The LBA project will generate many data sets, either by sampling or by further generation (i.e., by corrections, extrapolations, modelling, etc.), that can be classified, according to the project's planning, in five main components that have a close interaction. These components are Physical Climate, Carbon Storage and Exchange, Biogeochemistry, Atmospheric Chemistry, Hydrology, and Land Use and Land Cover. The project has reached an advanced stage of development and many data/metadata problems have been addressed. However, an evaluation of the solutions currently applied and further development of new solutions is still necessary.

SIVAM: Amazon Surveillance System

Governmental organisations are acting in the region in an isolated way, sometimes carrying out similar kinds of activity without sharing the obtained knowledge and without optimising the use of resources. As a result, we know little about the vast region and there is no effective control over criminal actions (e.g., illegal logging, unauthorised flights, deforestation and mining). SIVAM was implemented to establish better control in the region and is a powerful network for the collection and processing of information.

The system gathers information obtained by several governmental organisations that work in the Brazilian Amazon, processes and integrates this information in large databases, allowing organisations to share data and knowledge. In this way, the waste of effort that exists today can be minimised and can be adapted to the correct use of procedures and resources available for the development of these activities, taking into account the competence of each organisation involved.

The 1980s witnessed a number of unlawful activities, such as drug production, disarray in the occupation process, invasion of Indian areas, smuggling, predatorial actions, mainly from lumber companies and illegal mining, and the occurrence of a series of other crimes. In fact, with the difficulties of communication and control in the region, it was very difficult for the Brazilian government to measure the scale of illegal activities happening in the Brazilian Amazon. Over the years, the situation has worsened and countermeasures had to be taken.

SIVAM is recognised as a strategic project for the country because:

- there is unanimity among the regional governments in the recognition of its importance for the integration and the sustained development of the Amazon region;
- there will be effective control over the Amazon region, not only of its air space, but, even more importantly, of the use of its water resources, its biodiversity,

over the occurrence of deforestation and forest fires, of the settlement and movements of the Indian population, in the surveillance of border areas, in support of the repression of smuggling, drug dealing, and illegal mining;

- the Brazilian Government has interest in its full implementation, for it represents a possible solution for some of the region's problems, which have global consequences;
- in the near future, will promote the concrete participation of other countries in the region, supporting to Brazil's integration, in a process of collaboration for regional development at the international level.
- the resources for its complete development are guaranteed, with a financing contract split between Raytheon, a North-American company, Atech Foundation, and Embraer, both Brazilian companies; and
- the full operation of the system tends to be self-sustained, through the collection of taxes for services rendered;

SIVAM has a common and integrated infrastructure of technical procedures destined to be used in data acquisition and management and for the visualisation and dissemination of images and information. These procedures include remote sensing, environmental and weather monitoring, communication activities, radar surveillance, computer resources and telecommunication media. The application of these technical procedures, and the association of the data obtained from the different sensors, provides the particular type of information for the operational needs of each user.

The system has divided the Brazilian Amazon into three areas: Manaus, Belém, and Porto Velho. It was created by Regional Surveillance Centers of the Amazon (CRV) in these three provincial capitals. These CRVs have their work coordinated by the General Coordination Center (CCG) in Brasília.

The CCG is the place where all the information obtained from the data processed by SIVAM, as well as the resources of the organisations that take part in the system are centralised and made available. The complete data set allows the CCG to make knowledge available to the organisations so as to permit the elaboration of a strategic action plan for the Brazilian Amazon.

The CRVs are connected with each other and with the CCG, functionally and operationally. The CRVs are concerned with the concentration, processing, and dissemination of data in its area of operation, providing the necessary knowledge for the work of the participants of SIVAM, and enabling, at the same time, a better capability in communication. The CRVs have telecommunication, data management and visualisation resources, remote sensing by satellite, radar surveillance, weather information, communication monitoring, and general information for the coordinate action of the participant organisations of SIVAM.

The remote organisations, as a connected part of the system, are responsible for the collection and expedition of information to the corresponding CRV. The remote organisations provide support to local actions, through the telecommunication procedures that are implemented. They are now connected to the CRVs through the National Telecommunications Service, or by satellite, using stations assembled by SIVAM (Secretaria de Assuntos Estratégicos, 1993). Despite the importance of

the project, strong criticism from different sectors of society has been made and also questions have been raised about the way that SIVAM has been conceived and implemented, regarding political interference and financial misconduct and fraud allegations (Martinelli, 2001; Bortoni & de Moura, 2002; Brigadão, 1996).

1.3 Overview of this Research

Spotting Problems

The integration of biodiversity databases for the purpose of sharing data and analysis amongst scientific teams, can be a useful approach for understanding more about the Earth's ecosystem. After having been involved in several environmental projects in the Amazon region, we observed that some Environmental Information Systems (EIS) address several problems, ranging from user facilities to management data and metadata, mechanisms to infer patterns and non-explicit relationships from data, and integration of databases and GIS facilities. These problems are present in all EIS (Gunther, 1998). Other observations we made were related to the common problems of development of projects in the region. They included:

- The common occurrence of data production, not only by sampling, but also as a result of applying modelling functions.
- The presence at different scales of geo-spatially and temporally referenced data.
- The intention to use GIS systems and tools.
- The need to input and/or output data/information from/to other projects.
- The expectation that volume of data will be large.
- The need to manage metadata and/or context data.

We enumerate some general issues, which we targeted to investigate and contribute to:

1. The need to make environmental information available for researchers and the general public, as well as helping decision-makers to implement safe and sustainable development. The web technology is probably the best way to disseminate information on a global scale (Crowder & Crowder, 2000; Deitel *et al.*, 2001).
2. The questions about environmental information involve a number of different subjects (e.g., physical climate, carbon storage and exchange, biogeochemistry, hydrology, land use and cover, biological collections, etc). There is need for data and functionality to be made available in an integrated manner. The alternative solution aims to minimize problems of integrating specific 'problem-oriented' systems, that generate information about one specific subject. Users from research teams usually cannot fully use the data gathered by other teams

due to lack of information (poor metadata and context information), and hardware and software incompatibility. Consequently, data sharing becomes a research team effort, delaying any multidisciplinary data analysis. For the purpose of monitoring environmental behaviour or events, such delays are unacceptable.

3. A shortage of financial resources for environmental measures indicates the need for decision-makers to know '*where to best spend financial resources,*' or '*which policy should be implemented first regarding conservation.*' (Delbaere, 1998). It is necessary to generate strategic information from appropriate analysis and to prepare intuitive presentation for users. In this scenario, Geographical Information Systems (GIS) are required, since they can provide geographic data, and maps for environmental modelling and statistical analysis. Distribution of biodiversity has strong requests in projects like SIVAM and INPA's Bio-DB (Secretaria de Assuntos Estratégicos, 1993; Sonderegger *et al.*, 1998). In GIS, the basic concepts are location, spatial distribution and relationships, in which the basic elements are spatial objects (Fedra, 1994). In some networked projects, GIS products like ArcView⁴, ERDAS⁵, ILWIS⁶, etc, are used at the client side machine to display maps. This method avoids processing data at the server machine, but has a disadvantage that the GIS applications are needed for every client. The ideal alternative for this would be a system that can be accessed via web browsers, and that is capable of providing data and functionality to all clients, eliminating the software cost for sizeable institutes like INPA and MPEG (Museu Paraense Emilio Goeldi).
4. More than 95% of all the available biological collections and the majority of environmental experiment data do not have geo-referenced information, and consequently cannot identify the spatial position from the site where they were collected (Kerschberg *et al.*, 1996). Geographical data is, however, a critical feature of biodiversity. Thus, the integration of GIS databases in a database environment will provide good facilities for scientists to geo-reference their data sets (Kerschberg *et al.*, 1996).
5. Among new trends in database technology, data mining or knowledge discovery in databases has empowered data exploration. In providing an integrated database with modelling, geo-processing and aggregation functions, and storing data from several sources and different platforms, it would allow scientific teams to explore the integrated database. Such multidisciplinary exploration and data analysis from different data sets might give answers to questions of essential environmental hypotheses. (Piatetsky-Shapiro & Frawley, 1992).

⁴ArcView is a trademark of ESRI in the USA, European Community, or certain other jurisdictions.

⁵ERDAS is a service mark of ERDAS LLC.

⁶ILWIS is a trademark of the International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands.

Research Motivations

This research addresses a number of deficiencies present in current BIS environments, particularly systems that support large environmental experiments in the Brazilian Amazon. There is an urgent need to survey the status of biological collections in institutes and to describe their functional and systems requirements. We investigate computer technology, as infrastructure, to overcome some of the deficiencies already highlighted. In these projects, there is a clear need for a computational environment to support biological data representation and management, integration, analysis and information dissemination on a global scale.

We believe that biodiversity analysis is an important candidate application area, because of its strategic value to developing countries, for the world's environmental condition, the inherent complexities of understanding and modelling ecosystems, and the promises that current software and database technology seem to hold to extend their analysis capabilities. These technologies are available but in a non-integrated form, and the idea of an integrated computational environment seems to be promising.

This research can provide the means for a collaboration between a number of disciplines, and to accommodate the users with tools. This requires, at least partially, the understanding of biodiversity issues and census methods, and the active involvement of the mentioned disciplines.

Another point to emphasise is the use of distribution maps at a local scale, while at the national scale the larger ecological complexity is typically too poorly understood to support it with biodiversity analyses. Moreover, advances in automated biodiversity analysis also hold promises for biologists, who typically operate at the regional level. One may, for example, consider the study of geographic distribution of genes in a set of sibling species, and its evolutionary explanation. Also, data analysis and simulations can help environmental scientists to prove new hypotheses, generating scientific knowledge much faster and consequently, reaching well-founded predictions earlier which would allow intervention to prevent environmental disaster. In the political arena, decisions affecting natural resources are taken at various governmental levels, and in the long run one would hope that biodiversity analysis folders will become an accepted means of argumentation at all these levels. It seems reasonable to assume, however, that the added value of biodiversity techniques will be most prominent at regional levels: especially in conservation campaigns to protect important natural habitat, these folders may be crucial.

The government's conservation and sustainable development programs are thus the most obvious underlying policy objective of the research. In countries located in the tropical zone, a better understanding of the nation's ecosystems, and their subsequent conservation has, in fact, also led to important forms of economic growth, especially with eco-tourism, that now forms the basis for conservation policies. This can be observed in countries like Brazil, Belize, Chile, Costa Rica, Ecuador, and Peru, which are committed to increase investment and reinforce their environmental policies toward conservation.

Furthermore, there is a need to provide methods for geo-referencing of legacy data sets and offer GIS capabilities in an integrated platform. At the moment, there are no facilities that allow researchers to share data and metadata, to understand,

and analyse them. We point out that spatial data management is much needed today, in real geo-information applications, where such management functionality is lacking or only partly supported by today's GIS. There is a strong commitment towards the use of state of the art software, database design and implementation techniques, and the correctness of functions developed, and their usefulness for real applications. Additionally, we are committed to evaluate and define suitable schema representations for an open computational environment within the spatial database technology theme. Also, this research is compatible with the interests of environmental research institutes in Brazil, specially those based in the Amazon region, like INPA, MPEG, and EMBRAPA, while at the same time, it supports the research interests about applications of spatial database systems at University of Twente and ITC.

We foresee that this work, in the long run, can create an atmosphere in which multidisciplinary research teams can explore each other's data and metadata using an open computational environment that provides benefits for conservation issues, which translates itself into interesting statements of geo-information problems, and accommodates their solution.

Objectives

The overall objectives of this study fall into seven categories:

1. A description of the biological collection scenario in the Amazon's main institutes, by presenting diagnostics of its conditions and potential for education and research purposes (Chapter 2).
2. A description of the requirements analysis; functional and system requirements based on INPA as case study (Chapter 3).
3. A definition of a schema representation of biological collection data (Chapter 4).
4. A bio-metadata solution integrated to a database for data and metadata dissemination (Chapter 5).
5. A definition of a computer infrastructure to manage and disseminate data and metadata (Chapter 6).
6. A method to geo-referencing Amazonian biological legacy data (Chapter 7).
7. A negotiation protocol for reconciliation of taxonomic belief (Chapter 8).

1.4 Outline of the Thesis

The outlines of the thesis is shown in Figure 1.4. The book consists of nine chapters and the arrows indicate the information flow for reading orientation.

Chapter 1, as a general introduction, highlights, as background information, the main concerns of government and environmental institutes towards scientific data and metadata information. Also, it identifies the main scientific experiments

that are underway in the Brazilian Amazon and it presents an overview of this research, focusing on the problems, research motivations, and objectives.

Chapter 2 presents an overview of the condition of biological collections in the Brazilian Amazon and of the institutes in which they are based. It also presents their motivation activities and effort regarding management and information sharing. The chapter also describes the problems and drawbacks faced by that and the efforts to provide ready-to-use information. It discusses detailed initiatives carried out by INPA, trying to integrate its biological collection program to activities related to research. Collection management via a toolkit approach and the influence of computer technology are also addressed. The chapter closes with our recommendations for necessary actions, which can provide advancement in collection management and strengthen partnerships.

Chapter 3 focusses on the requirements analysis, (functional and systems) performed at INPA's Scientific Collections Program (SCP), and aims to provide the necessary information for the development and implementation of an automated system to manage collection data.

Chapter 4 describes the Clustered Object Schema representation of biological collections aiming at database implementation. The description includes the elements defined for clusters, a graphical notation and their syntactic definitions.

Chapter 5 presents a solution for metadata management. The solution includes the use of Extended Markup Language (XML) technology to describe biological metadata over the web. The chapter also presents metadata issues regarding content and standards, and describes a method adopted to implement it in a networked client/server platform as well as an approach to implement it in an online nodes environment for harvesting metadata information.

Chapter 6 describes a prototype implementation for managing biological data on the web. Evidence is given to support the use of an open source system all along to the three-tiered architecture; web servers, Database Management System (DBMS), and server-side script resource. The prototype implements partially a CLOSi-based database and its web interface.

Chapter 7 demonstrates a retrospective georeferencing method applied to a biological legacy data from a CLOSi-based database. Georeferencing issues, such as expressing latitude, longitude, and the calculation of the maximum error distance originated from uncertainties are also discussed. It shows an option to contribute to a collaborative gazetteer (directory of information about locality) by integrating the newly georeferenced data set to distribution on the web.

Chapter 8 presents a proposal for automatic negotiation protocol that can resolve taxonomic belief differences. Fundamental problems present in biological classification and the basic terminology and taxonomic reference system (linnaean and phylogenetics) is described. The problems around integration of taxonomic data sets is detailed as components of the problem of communicating over taxonomy. Followed, we present the framework protocol for an automatic negotiation, which would process protocol specifications, its rule and negotiation strategy and highlight features that a computer language need to have during an eventual implementation.

The book closes with Chapter 9, which draws conclusions and formulates some recommendations for future development of this research.

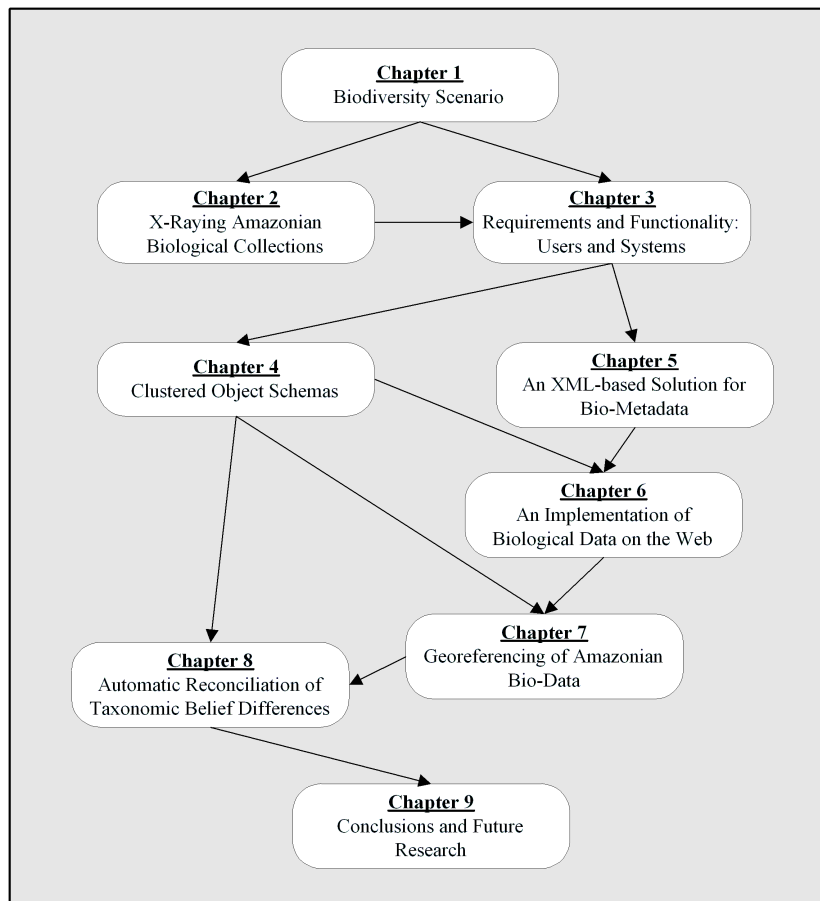


Figure 1.4: Outline of the thesis.

1.4. OUTLINE OF THE THESIS

Chapter 2

X-Raying Amazonian Biological Collections *

2.1 Introduction

In this chapter, in Section 2.2, we present an overview of the current situation of biological collections in the Amazon region. It identifies the main institutes and their collections, their motivation, activities, and efforts regarding management and information sharing. Following this, in Section 2.3, we describe the problems concerning biological information, which include information production methods, the sources, its degradation processes and data and information representation. Next, in Section 2.4, we indicate initiatives underway to provide ready-to-use information on a large scale as well as the major drawbacks of solutions based on normal file systems. We then detail some initiatives carried out by INPA, which tries to integrate its biological collection programme to activities related to research. Collection management via a toolkit approach and the computer technology influences are also addressed. With Section 2.5 we finalize the chapter by presenting our conclusions and recommendations about the collections, actions that would provide advancement in collection management, the importance of scientific partnerships and the foreseen role of the Brazilian national scientific body (MCT), mainly implementing policies, and sponsoring initiatives.

*This chapter is based on:

- (1) Campos dos Santos, J. L., de By, R. and Magalhães, C. (2000). A case study of INPA's Bio-DB and an approach to provide an open analytical database environment. In *International Archives of Photogrammetry and Remote Sensing*, 33(B4): 155—163.
- (2) Magalhães, C., Campos dos Santos, J. L. and Salem, J. I. (2001). Automation of biological collections and information about the Amazonian biodiversity. In the *Strategic Partnership*; Ministry of Science and Technology, Brasília, DF, Brazil, 12: 294—312. (In Portuguese).

2.2 Biological collections in the Amazon region

2.2.1 Institutes and their collections

The first part of our research included visits to several universities and research institutes in the Amazon region. A survey was conducted to identify procedures adopted in the management of biological data, possible problems at institutional and inter-institutional levels, biological collection availability (including data sets) and additional factors that could have an impact on scientific information generation, its management, analysis and dissemination. The data gathering was essential to our analysis and provided the elements that guided solutions towards data and metadata management, which are described in the next chapters.

The missions took place from November 1999 to August 2001, and the institutes visited were:

- INPA (Instituto Nacional de Pesquisas da Amazônia), UFAM (Universidade Federal do Amazonas), and CPAA (Centro de Pesquisas Agroflorestal da Amazônia Ocidental) — Manaus, Amazonas — November 1999;
- MPEG (Museu Paraense Emílio Goeldi), CPATU (Centro de Pesquisas Agroflorestal da Amazônia Oriental) — Belém, Pará — June 2000 and August 2001;
- MCT (Ministério da Ciência e Tecnologia) and MMA (Ministério do Meio Ambiente) — Brasília, DF — July 2000 and June 2001;
- Silvolab-Guyane (Groupment d'Intérêt Scientifique) — Kourou, French Guyana — December 2000;
- IEPA (Instituto de Pesquisas Científicas e Tecnológicas do Estado do Amapá)— Macapá, Amapá — May 2001.

Apart from IEPA, all institutes have a more or less established Scientific Collection Program or some type of initiative, with objectives that match, in a certain way, those described in more advanced SCPs. Also, these organisations are investing continuously in computational infrastructure for scientific data management. Despite this, they have insufficient resources to develop a networked contribution amongst organisations. Additionally, it required extra effort from our side to establish the necessary links for accomplishing the planned fieldwork missions. The main reason for this is that representing scientific data aimed at global dissemination and implementation of functionality for sharing is threatening. Sometimes institutes lacked a clear data policy. In certain domains, for instance, botanic economics and tropical diseases, there were more difficulties for data description and acquisition due to the sensitivity of the data itself, which are considered to be of strategic value with potential economic interests, contrary to government interest, therefore, constituting a threat for some sectors. Due to official impediments regarding data policy, as stated in the Oiapoque project (Marcon, 1999), Silvolab-Guyane biological data have been left out, until further agreement which should be reached before the second phase of the Oiapoque project.

All institutes have been collecting species information throughout their existence and this information has been deposited into biological collections. The formation of

Table 2.1: Biological collection held by institutes in the Amazon region - (*n*) indicates number different collections of the same type.

CPAA	CPATU	IEPA	INPA	MPEG	UFAM
Insects	Herbaria	Birds	Bacteria (2)	Birds	Fish
Microbiological	Insects	Carcinology	Birds	Fish	Herbaria
	Trees	Fish	Fish	Herbaria	Herpetological
		Herbaria	Fruits	Herpetological	Insects (2)
		Insects	Fungi (2)	Histological	Microbiological (3)
			Herbaria	Insects	Plant Tissue
			Herpetological	Mammal	
			Invertebrates (14)	Pollen	
			Mammal	Trees	
			Pollen/Seed		
			Trees		

biological collection happened, in most of the cases, as a natural process, without a national or institutional policy or guidelines; it was implemented by the researchers themselves, who aimed to preserve the collected specimens or to fulfill requirements of specific research projects. The result of these initiatives is a huge volume of data and information about the biota, which continues to grow.

The Figure 1.1 presents a map identifying the area covered during our fieldwork. Manaus was selected to be the base office due to logistic facilities provided by INPA. Central offices of MCT and MMA are located in Brasilia and deal with all legal aspects of scientific data, policies, and programmes for development at the federal level. The visit to these two institutes was particularly interesting. It was possible to perceive how the complexity of legal issues and the number of open questions makes the implementation of a national programme for protection and sustainable development in the region a sizeable problem.

To manage biological collections in places with megadiversity, like the tropical Amazon region, management rely on automated systems, that are problem-driven applications or database implementations. Despite a clear lack of importance regarding the utilisation of biological collections, they are fundamentally important when associated to economical values, identification of fragile ecosystems, land demarcation based on endemism, and endangered species. We observed that institutes differ in the way they measure the degree of importance of their biological assets. For instance, research institutes associate their importance with the number of scientific papers produced, whereas for universities their teaching activities are more important.

Given the size of the Amazonian area, the number of existing biological collections is small. This is due to the limited capacity the institutes have which forces them to concentrate their effort only on a few types of collection. In Table 2.1, we present an overview of biological collections kept in the main Amazonian institutes, while in Table 2.2 we show the agronomic collections and nurseries.

2.2. BIOLOGICAL COLLECTIONS IN THE AMAZON REGION

Table 2.2: Agronomic collections and plant nurseries in Amazonian institutes.

COLLECTIONS	ORGANISATIONS			
	CPAA	CPATU	INPA	UFAM
Germplasm Bank	Cupuaçu	Ananás	Aromatic Plants	Araça-boi
	Dendê	Bauhinia	Forest essences	
	Forest species	Cupuaçu	Forest species	
	Guaraná	Ipecacuanha	Medicinal Plants	
	Mandioca	Jaborandi	Tropical Fruits	
	Pau Rosa	Pimenta do Reino		
	Seringueira	Urucum		
Nurseries	Medicinal Plants	Guaraná		Fruits Medicinal Plants
Other Collections		Tropical Fruits		

The description given here shows that collections collected so far are small compared to the diversity present in the region. It is necessary then to consolidate the existing collections and propose the formation of new ones. Recently, the Brazilian government has taken an important step towards the implementation of this idea by establishing a council to deal with management of genetic material at the national level. The mission of this body, among others is to establish mechanisms to access, disseminate, and protect the nation's genetic resources (Magalhães *et al.*, 2001).

EMBRAPA: CPAA and CPATU

EMBRAPA is a federal institute that develops research in the northern region of Brazil through its CPAA (West Amazonia) and CPATU (East Amazonia) research centres. EMBRAPA promotes research in the field of agronomy, developing new methods to develop the economic use of vegetal species. The institute assumes it is mandatory to form and maintain collections *ex situ* (e.g., insects and herbaria collections) as well as the conservation of genetic resources in a germplasm bank. Both collections (insects and herbaria), were formed by researcher initiative, taking into account social demands. Today, they represent an important asset for EMBRAPA.

The insect collection is relatively small and lacks resources for maintenance and development. The herbaria collection, due to its importance, receives partial support from the institute. The number of species available in the germplasm bank is small when compared to the existing species number in nature and their economic potential. Despite this, there are fourteen germplasm banks. Table 2.3 and Table 2.4 present the motivation associated with the formation of these collections and their current use.

Figure 2.1 presents the numbers of holotype specimens, identified specimens and total specimens held in the collection. A holotype specimen is the single specimen designated by an author as the type of a species or lesser taxon at the time of establishing the group. From the same figure, it is possible to notice the natural interest of both institutes, which focus interest mainly on the herbaria collection, placing the

Table 2.3: Motivation activities that led to the formation of EMBRAPA collections.

ACTIVITIES	COLLECTIONS		
	Insects CPAA	Insects CPATU	Herbaria CPATU
Biotechnology Projects	no	no	no
Donations	no	no	no
Education Activities	no	no	no
Institutional Policy	yes	no	no
Personal Initiative	yes	no	no
Research Activities	yes	yes	yes
Taxon Identification	yes	no	no

Table 2.4: Current use of EMBRAPA collections.

ACTIVITIES	COLLECTIONS		
	Insects CPAA	Insects CPATU	Herbaria CPATU
Biogeography	no	no	no
Biology/Ecology Studies	no	no	no
Demarcation Protected Areas	no	no	no
Evolution Studies	no	no	no
Micropropagation Studies	no	no	no
Sampling Policy	no	no	yes
Taxonomy Verification	yes	yes	yes
Witness of Biodiversity	yes	yes	yes

2.2. BIOLOGICAL COLLECTIONS IN THE AMAZON REGION

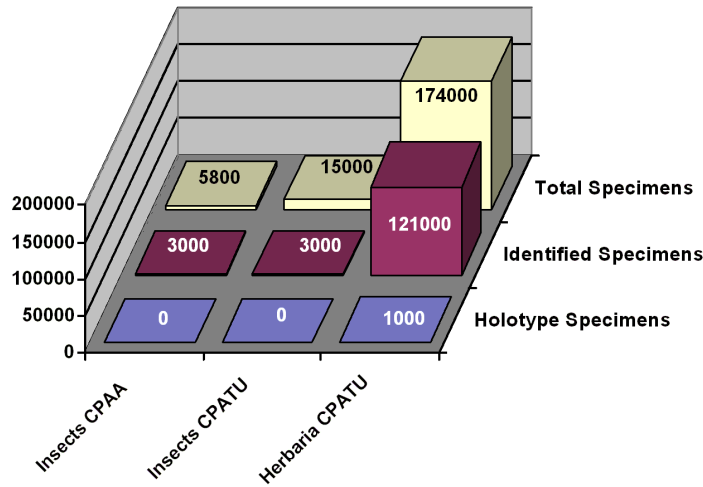


Figure 2.1: Holotype, identified and total number of specimens in EMBRAPA CPAA and CPATU biological collections (Source: da Fonseca *et al.*, 2002).

rest at a lower level of importance, functioning as accessory tools related to specific research questions.

From our observation, the research activities related to the collections in both research centres are few. For example, at CPAA, only one research activity is associated with the insect collection and at CPATU there are two activities, one is being carried out directly in the collection while the other is associated with the collection. None of the collections have activities associated to biotechnology. It is evident that EMBRAPA needs to invest in collection expansion, management and increase of taxonomic identification. Those actions can guarantee new opportunities and more knowledge generation and dissemination.

IEPA

IEPA collections were formed from donations from the state Natural History Museum (Ángelo Moreira da Costa Lima) and the Medicinal Plants Herbaria (Waldomiro de Oliveira Gomes), aiming mainly at supporting research activities. The institute recognises their importance but is not able to derive programmes to utilise them fully. The role of these collections has not been defined in the context of the institutes. Because of this, the collections cannot contribute fully to social and scientific demands. Additionally, this has contributed to the lack of financial resource for its expansion and modernisation.

Da Fonseca *et al.*, (2002), called attention to the fact that three of the IEPA

Table 2.5: Application context of collections at IEPA.

ACTIVITIES	COLLECTIONS				
	Carcinology	Insects	Herbaria	Fish	Birds
Biotechnology Projects	no	no	no	no	no
Donations	no	no	no	no	no
Education Activities	no	no	no	no	no
Institutional Policy	yes	no	yes	yes	no
Personal Initiative	no	yes	yes	no	no
Research Activities	yes	yes	yes	yes	yes
Taxon Identification	no	no	no	yes	no

Table 2.6: Current use of IEPA collections.

ACTIVITIES	COLLECTIONS				
	Carcinology	Insects	Herbaria	Fish	Birds
Biogeography	no	yes	no	yes	no
Demarcation Protected Areas	no	no	yes	yes	no
Evolution Studies	yes	no	no	yes	no
Sampling Policy	yes	no	no	yes	no
Taxonomy Verification	yes	yes	no	yes	no
Witness of Biodiversity	yes	yes	yes	yes	yes

collections (See Table 2.1) were formed as a result of an institutional programme. These programs include biotechnology projects, educational activities, institutional policies, researchers personal initiatives, requests for specimen identification, and research activities. However, today, collections are utilised merely as a witness of biodiversity, ignoring the current use related to these collections, such as demarcation of protected areas, sampling policies, taxonomic verification, biogeography, and evolution studies. Table 2.5 presents the activities in each of the five collections during their formation, while Table 2.6 lists the current activities in which the collections are involved.

A common problem present in the collections is the discrepancy between the total number of specimens in a collection and the number of specimens classified and identified. The main reason for this is due to shortage of qualified personnel to manage the collection, and the lack of inter-institutional cooperation. Figure 2.2 presents the numbers for holotype, identified specimens and total number of specimens in the IEPA collection.

Having said this, we can conclude that IEPA's situation is far from being satisfactory and reflects well the need for integration, and collaboration with other institutes whose biological collections are better structured, to facilitate the institute to provide information for productive activities.

It is worth mentioning that IEPA collection has great potential. It is located in a region with a unique biological richness and therefore can be classified as a strategic institute. Unfortunately, this asset has not been prioritised and not a single action

2.2. BIOLOGICAL COLLECTIONS IN THE AMAZON REGION

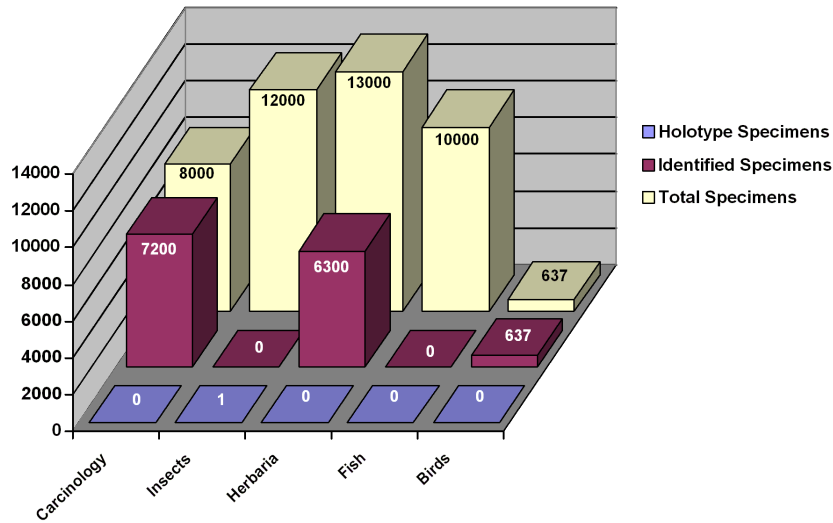


Figure 2.2: Holotype, identified and total number of specimens in IEPA collections (Source: da Fonseca *et al.*, 2002).

has been taken towards the use of an automated system to manage existing data about the collection.

INPA

INPA has more than 40 years of tradition in studying the Amazon ecosystems. During this time, millions of organisms, collected from the rainforest and its water bodies, were deposited in the collections of the Institute. They are divided into six botanical, four micro-organismal, and nineteen zoological collections, consisting of vertebrates and invertebrates. Figure 2.3 presents the scope of the biological collections at INPA.

Despite the lack of a clear protocol guiding INPA to concentrate collecting effort in certain areas, and by not establishing taxonomic groups as a high priority, INPA is still considered the most prominent institute in taxonomic research in the western Amazonia. INPA has also been affected by the lack of taxonomists available. It has been reported that the number of taxonomists available is not sufficient to deal with the demands of today. This phenomena, also known as 'biodiversity crisis' is affecting global wide initiatives (Wilson & Peters, 1988; Gaston & May, 1992; Feldmann & Manning, 1992; Magalhães *et al.*, 2001). Fonseca and Ferreira (1998) indicates that only 16% of INPA's researchers work with taxonomy, a number that cannot meet the demand.

In 1997 the institute started a project that included the development of its own

Table 2.7: Some important collections at INPA.

COLLECTIONS	RECORDS		
	Total	Computerised	Percentage
Amphibians	8.074	7.794	96.53
Birds	500	0	0
Crustacean	718	533	74.2
Fish	40.000	14.464	36.16
Herbaria	203.450	32.484	15.9
Invertebrates	2.165.000	10.115	0.46
Mammal	3.000	2.714	90.46
Mollusca	261	261	100
Pollen	24.200	0	0
Reptiles	2.225	1.187	53.34
Trees	10.200	200	1.9

Table 2.8: Application context of INPA collections.

ACTIVITIES	COLLECTIONS										
	A	B	C	D	E	F	G	H	I	J	K
Biotechnology Projects	no	no	no	yes	yes	no	no	no	no	no	no
Donations	no	no	yes	yes	yes	no	no	no	no	no	no
Education Activities	yes	yes	yes	yes	yes	yes	yes	no	no	no	yes
Institutional Policy	yes	yes	yes	yes	yes	yes	yes	no	yes	no	yes
Personal Initiative	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no
Research Activities	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	yes
Taxon Identification	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	yes

Note: A=Amphibians; B=Birds; C=Crustacean; D=Fish; E=Herbaria; F=Insects; G=Mammals; H=Mollusca; I=Pollen; J=Reptiles; K=Trees

information system for the management of INPA collections. Unfortunately, restricted resources did not allow the development of such a complex system in a single step, so it was decided to start off with only one collection, the insect collection. This collection is the biggest of the institute, and also represents a large diversity of information as there are many researchers from various departments working with this animal group.

Table 2.8 presents the application context of INPA collections and Table 2.9 the current use of collections.

MPEG

MPEG has a clear objective, as far as collections are concerned, since it is a museum, aiming to form and keep biological collections that cover most the scope of its activities. MPEG has the most important biological collections in the region which dates back one hundred years. The majority of their collections have international recog-

2.2. BIOLOGICAL COLLECTIONS IN THE AMAZON REGION

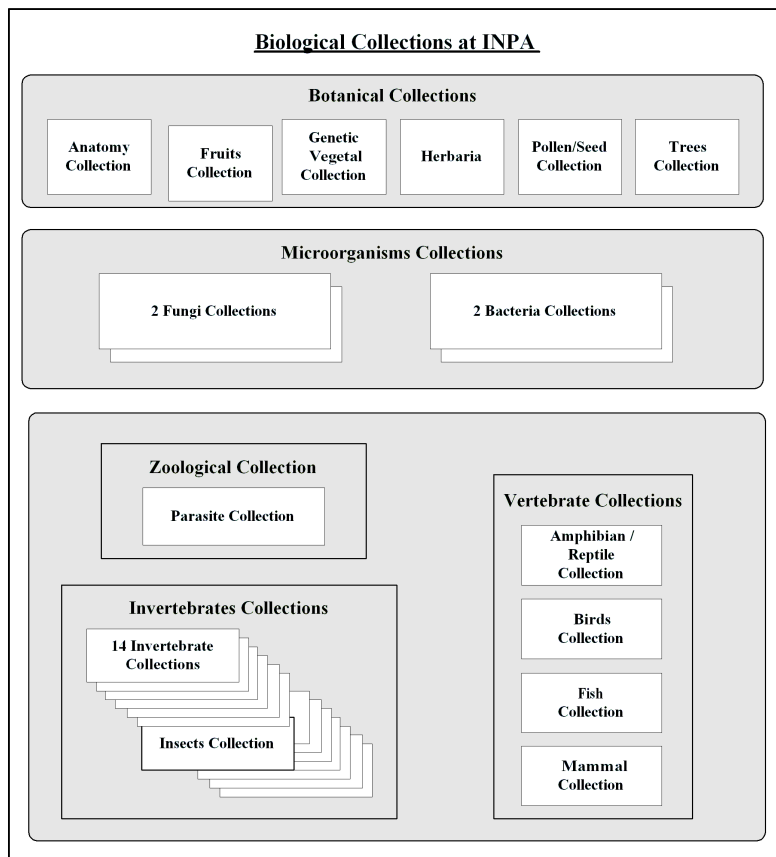


Figure 2.3: Biological Collections at INPA.

Table 2.9: Current use of INPA collections.

ACTIVITIES	COLLECTIONS										
	A	B	C	D	E	F	G	H	I	J	K
Biogeography	no	yes	no	yes	yes	yes	no	no	yes	yes	yes
Biotechnology Projects	no	no	no	no	no	no	no	no	no	no	no
Demarcation Protected Areas	no	no	no	yes	no	no	no	no	no	no	yes
Evolution Studies	yes	no	yes	yes	yes	yes	yes	no	no	no	yes
Micropropagation Studies	yes	no	no	no	no	no	no	no	no	no	no
Sampling Policy	yes	no	no	yes	no	no	yes	no	no	no	yes
Taxonomy Verification	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Witness of Biodiversity	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

Note: A=Amphibians; B=Birds; C=Crustacean; D=Fish; E=Herbaria; F=Insects; G=Mammals; H=Mollusca; I=Pollen; J=Reptiles; K=Trees

dition, for instance, the herbaria; when the subject is tropical diversity the use of its material is obligatory. Despite their visibility and the potential of its collections for research, they are not considered essential for the development of policies that can help decision making or act in the social productive sector. The reason for that seems to be rooted in the researchers' behaviour which is centred on their research activities. They tend to isolate themselves to produce scientific results. Furthermore, culture features can make the task of convincing researchers of the advantages of the interface with other sectors of society or interdisciplinary domains, extremely difficult.

The quality in human resources available has been responsible for the advances of MPEG collections. Such resources allow for continued operation, which is supported by the partnership of the Federal University of Pará. This was the first observed case in which the formation of a biological collection was a direct result of institutional policy. Other collections, such as trees and histological, are considered to be complementary, and are directly associated to the herbaria, were formed to support research projects, and, in exceptional cases, by personal initiatives. Table 2.10 presents the reasons that led to the formation of the collections and Table 2.11 presents the activities supported by the collections.

Six collections (Herbaria, Trees, Hystological, Herpetological, Fish and Mammal) are intensively used for taxonomic research, as indicated in Figure 2.4. It can be noticed that the number of identified specimens across the collections are high and indicates the relevance of these collections for science.

UFAM

UFAM was the first university created in Brazil (1909), under the name of Free University School of Manáos. With the collapse of the rubber era, in the 1920s, a considerable volume of information was lost, mainly components from vegetal species considered medicinal. After the 1960s the biological collection activities res-

2.2. BIOLOGICAL COLLECTIONS IN THE AMAZON REGION

Table 2.10: Application context of MPEG collections.

ACTIVITIES	COLLECTIONS								
	A	B	C	D	E	F	G	H	I
Biotechnology Projects	no	no	no	no	no	no	no	no	no
Donations	no	no	no	no	no	no	no	no	no
Education Activities	no	no	no	no	no	no	no	no	no
Institutional Policy	yes	yes	no	no	no	yes	yes	yes	yes
Personal Initiative	yes	no	no	yes	yes	no	no	no	no
Research Activities	no	yes	yes	yes	yes	yes	yes	no	yes
Taxon Identification	no	no	no	no	yes	yes	no	no	yes
Note: A=Insects; B=Herbaria; C=Trees; D=Hystological; E=Pollen; F=Herpetological; G=Fish H=Mammal; I=Birds									

Table 2.11: Current use of MPEG collections.

ACTIVITIES	COLLECTIONS								
	A	B	C	D	E	F	G	H	I
Biogeography	yes	yes	yes	no	yes	yes	yes	yes	yes
Biotechnology Projects	no	no	no	no	no	no	no	no	no
Demarcation Protected Areas	no	no	no	no	no	no	no	yes	yes
Evolution Studies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Micropopagation Studies	no	no	no	no	no		no	no	no
Sampling Policy	no	yes	no	no	no	yes	yes	yes	yes
Taxonomy Verification	yes	yes	yes	yes	yes	yes	yes	yes	yes
Witness of Biodiversity	yes	yes	yes	no	yes	yes	yes	yes	yes
Note: A=Insects; B=Herbaria; C=Trees; D=Hystological; E=Pollen; F=Herpetological; G=Fish H=Mammal; I=Birds									

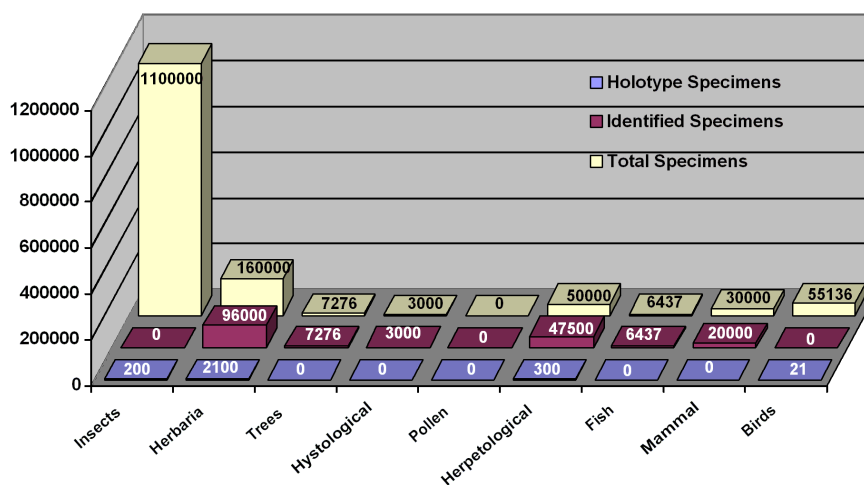


Figure 2.4: Holotype, identified and total number of specimens in the MPEG biological collections (Source: da Fonseca *et al.*, 2002).

tarted but up to the present time, the institution does not have a programme nor mechanisms for the creation and maintenance of its collections. In not doing so, the institution is not obliged to finance such activities and as a consequence the collections are neglected. Despite all adversities, researchers manage to support the collection with financial resources from individual research projects with a strong motivation towards education. Occasionally, the researchers salary are diverted to finance collection activities. Table 2.12 presents the activities responsible for the formation of collections at UFAM and Table 2.13 shows the current activities related to collections.

The total number of items in the collections is small but impressive when taken into account that they are managed in an informal way. The majority of collections is still in an embryonic stage and needs support of all kinds. Today, 13 research projects under development have a relation to biological collections and 21 projects are directly associated to the collections, being an average of 1.5 project per researcher. Human and financial resources available to the collections come from such projects. Figure 2.5 presents the total number of specimens in UFAM collections.

A comparative analysis amongst institutes and the several types of collections has recently been carried out by da Fonseca *et al.* (2002). The analysis gives emphasis to the number of families and genera each collection has, and raises questions concerning exchange and maintenance, environmental/ecological relationships and automatation.

2.2. BIOLOGICAL COLLECTIONS IN THE AMAZON REGION

Table 2.12: Application context of UFAM collections.

ACTIVITIES	COLLECTIONS					
	A	B	C	D	E	F
Biotechnology Projects	no	no	no	no	no	no
Donations	no	no	no	no	no	yes
Education Activities	no	no	no	yes	no	yes
Institutional Policy	no	no	no	no	no	no
Personal Initiative	yes	yes	yes	yes	yes	yes
Research Activities	yes	yes	yes	yes	yes	yes
Taxon Identification	no	yes	no	yes	no	no
Note:	A =Plant Tissue Culture; B =Insects; C =Herbaria; D =Herpetological; E =Fish; F =Microbiology					

Table 2.13: Current use of UFAM collections.

ACTIVITIES	COLLECTIONS					
	A	B	C	D	E	F
Biogeography	no	yes	yes	yes	yes	yes
Biotechnology Projects	no	no	no	no	no	yes
Demarcation Protected Areas	no	no	no	no	yes	no
Evolution Studies	no	yes	yes	yes	yes	no
Micropropagation Studies	yes	no	no	no	no	no
Sampling Policy	no	no	no	no	yes	no
Taxonomy Verification	no	yes	yes	yes	yes	yes
Witness of Biodiversity	no	yes	yes	yes	yes	yes
Note:	A =Plant Tissue Culture; B =Insects; C =Herbaria; D =Herpetological; E =Fish; F =Microbiology					

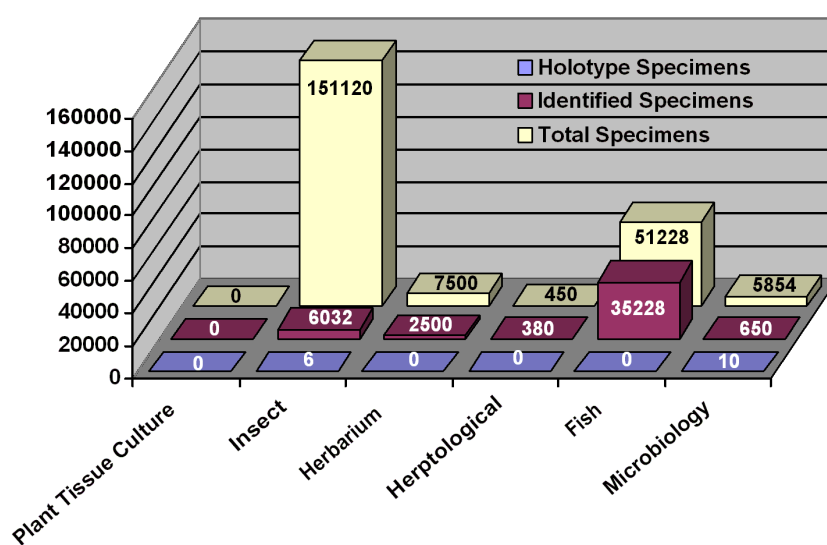


Figure 2.5: Holotype, identified and total number of specimens in UFAM biological collections (Source: da Fonseca *et al.*, 2002).

2.3 Problems with biological information

We have identified four classes of problems that can directly interfere with biological information management. The problems include:

(a) Information production method

Information is produced from a process that involves four elements: researchers, data, metadata, and a theoretical conceptual framework. There are two conditions that affect the way in which information can be produced:

1. Poor quality of data (raw data, table structure, numeric or encoded sampling observations) and inadequate storage devices. This is due to the large quantity of incomplete data, which are still kept in field work notes or in non-scalable media devices.
2. Non-existence of metadata.

To produce information, the traditional method comprises user (researcher), data (stored in media devices or paper notes and being of any type), metadata descriptions, and a theoretical conceptual framework. Data can also be stored in devices connected to the Internet. The lack of metadata can jeopardise the process of information production. To have a complete framework, it is necessary to provide improved ways to better describe data sets and to implement an intensive effort to digitise data. The dissemination of a good data description on a global scale is important and will facilitate the understanding of a scientist's hypothesis. The method is presented in Figure 2.6.

(b) Source of biodiversity data and information

We have found that scientific data sets are abundant, but incompatible and dispersed. This has led to the development of spontaneous systems, in which applications are driven by specific problems without taking into account the need of data integration or their dissemination.

Also, there are no catalogues available that indicate the existence of such data and access to the few data sets can only be done on an *ad hoc* basis. Usually, the scale and data content are inconsistent, with no history, unknown quality and no method to qualify it. Figure 2.7 portrays the scenario of the source of biodiversity data and information in the Amazonian institutes.

(c) Information degradation

Information content of data and metadata can undergo severe degradation over time. When data are acquired, the processes of validation and calibration lead to good data management until the time of publication of research results. After this, specific details about problems with individual data items, a specific data set or general details are rapidly lost.

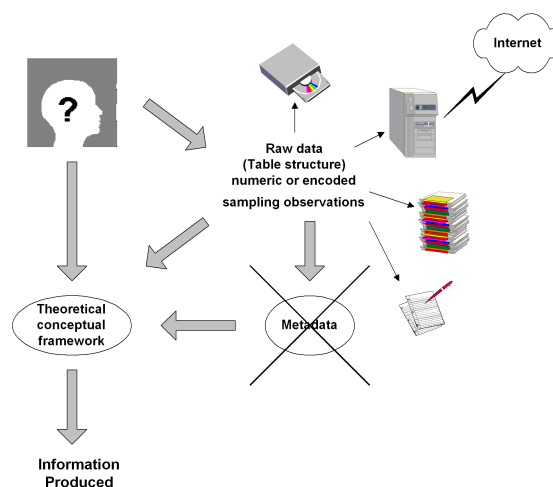


Figure 2.6: The traditional information production method showing the absence of metadata.

Also, a researcher's retirement or career change commonly causes access to information details to become difficult or unlikely. Accidents or changes in storage technology may eliminate access to remaining raw data and metadata at any time.

Figure 2.8 presents an example of the normal degradation of information content associated with data and metadata over time (information entropy). The dashed line means a sudden change in storage technology or any accident that may occur.

To overcome some of these problems, it is important to implement policies to control the information chaos and postpone the natural degradation. It is necessary to develop, use or extend a metadata standard across scientific research teams and to have a system to register and distribute data and metadata of high quality.

(d) Inherent biodata modelling

Traditional research institutes in the Amazon region date back over one hundred years, consequently, legacy data is a major issue. Spatial references are vague or non-existent and temporal references are uncertain.

Additionally, when scientific data sets represent biological species, the problem of taxonomic beliefs adopted across the data sets becomes apparent. Biological data representation must express a multi-taxonomic system. Another fact is that obtaining interoperability amongst biological data sets is difficult.

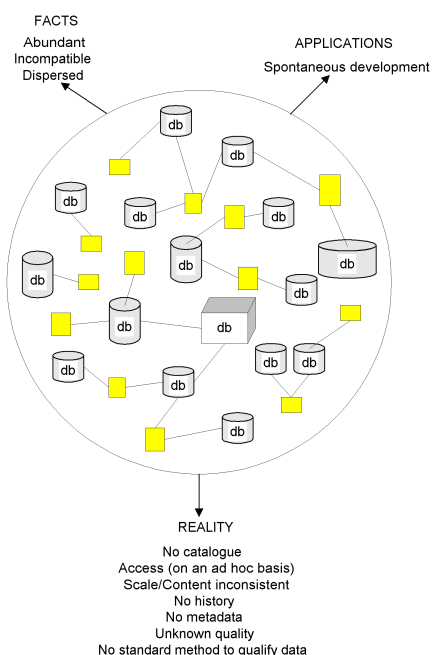


Figure 2.7: Sources of biodiversity data and information.

2.4 Bridging gaps towards BIS

The demand for information to assess environmental issues, like human impact on protected areas and endangered species, recovering from environmental degradation, and bioprospecting is constantly increasing. For most of these issues, the information may exist, but the problem is how to obtain it. Scientific material that has been published and made available, in general does not easily convey the full information needed.

It is in this context that biological collections can play an important role in meeting demands and answering questions, since collections incorporate intensive efforts and years of investigation about the fauna, flora, and microbiota.

There are efforts under way to provide ready-to-use information. Bisby (2000), and Edwards *et al.* (2000), mention several initiatives and describe projects and applications under development in the field of bioinformatics. Such applications focus on global solutions that allow interoperability and synthesis of information amongst local and remote systems that deal with information about biodiversity.

Magalhães *et al.* (2001) summarise some initiatives that are based on automated systems for scientific data management. Amongst them are worth mentio-

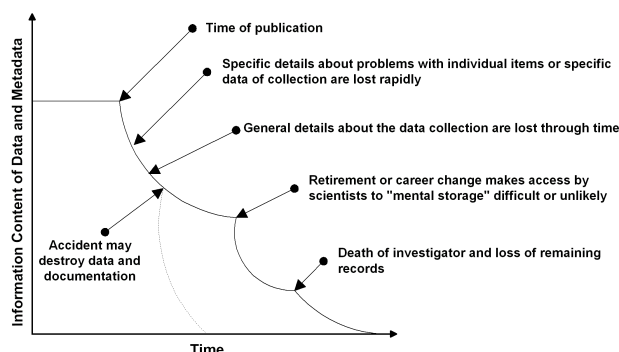


Figure 2.8: Degradation of information (Source: Michener *et al.*, 1997).

ning: BIN21 (The Biodiversity Information Network), Species 2000, ABI (Association for Biodiversity Information), IABIN (The Inter-American Biodiversity Information Network), NEODAT (The Inter-Institutional Database of Fish Biodiversity in the Neotropics), IPGRI (The International Plant Genetic Resources Institute), WDCM (The World Data Centre for Microorganisms), CONABIO (Comisión Nacional de Biodiversidad), INBio (Instituto Nacional de Biodiversidad), NBII (The National Biological Information Infrastructure), GBIF (Global Information Facility), BDT (Base de Dados Tropicais), BINbr (Rede de Informação em Biodiversidade Brasil), BIOTA-FAPESP (Programa de Pesquisas em Conservação Sustentável da Biodiversidade do Estado de São Paulo), BCDAM (Base Compartilhada da Amazônia) and SIVAM (Sistema de Vigilância da Amazônia). These global efforts require the use of database systems to manage and distribute data.

Institutes like INPA are getting prepared to better manage their collections by improving data quality and replacing register card archives with digital data.

Several implementation are based on common file systems. The major drawbacks include:

- Inter-related data are stored in separate files.
- Lack of automated control. System analysts and computer programmers must determine which parts are needed and must decide how the files are related, and coordinate file processing.
- Existence of data redundancy.
- Lack of data integrity, thus producing inconsistent results.
- Application program dependency. Formats of files and records are part of the application code. Any change in format must be reflected in the code. This is a time-consuming and error-prone task.
- File incompatibility. Files written by different programming languages cannot be readily combined or compared.

- Difficulty in representing data and relationships to users.

Database systems, in contrast, can accommodate the majority of user requests and change the way researchers work and produce information. The main characteristics of database systems are:

- They support data integration — all the application data are stored in a database and programmers are not responsible for maintaining files.
- They help to reduce data duplication — data are stored in a single place, meaning fewer data integrity problems occur.
- They improve program/data independence — record formats are stored in the database itself; they are accessed by the DBMS, not by the application program, which minimizes the impact of data format changes on application programs.
- They allow for representation of different views/perspectives.

Database technology is an important ally to researchers in biological science. The technology embedded in these tools makes the management of collection data much easier. In a database design, a consistent schema representation is essential, as without it the database technology may be unwisely or at least inefficiently used. Data from digital databases can be accessed over fast computer networks. Regarding information dissemination, the Internet represents another ally and is the front runner vehicle for worldwide access, while at the same time, allowing satisfactory restrictions to sensitive information.

2.4.1 A close look at INPA's initiatives

Allkin (1988) classified INPA's information within collections. The classification was an attempt to identify the type of data and the individuals responsible for generating and managing data, as well as who the potential users are and what the important delivery methods are. This classification did not consider the use of imagery as an information source. The content of this classification can be summarised as follows:

- **Curatorial records** They are identified by a unique number, have an entry date, loan records, supplementary collections made, where stored, etc. The curatorial staff and INPA managers are those who generate, manage, and use published material or Intranet applications.
- **Specimens** This information is based on visual inspection of individual specimens, carried out by research staff and visiting specialists or through loans. Specimen specialists are responsible for managing this type of information; the main users are taxonomists and INPA's research teams. The information and collection materials can be accessed by staff, by loan procedures, or gift duplicates, usually via prior identification.
- **Identification** It represents the identity of the material — on the original label or on a subsequent determination label. This type of information is generated by collectors or systematic specialists and is managed by specialists and curatorial staff. The users range from naive to expert and the information might be delivered to them as checklists or through other digital mechanisms.

- **Label data** This represents additional information written on labels by collectors. The scope of the information encompasses: geographical data, habitat description, morphological details, etc. This information is generated exclusively by collectors and is managed by the curatorial staff. The users are systematicists, ecologists, GIS specialists, and conservationists.
- **Aggregated information** Derived from specimens, this information is obtained from accumulation, comparison, and analysis of data from many specimens. Such information is generated and managed by research staff, systematic specialists and analysts. The users who consume this type of information are: researchers, agronomists, foresters, conservationists, and popular audiences. The delivery mechanisms include field manuals, monographs, and theses.

INPA promoted the first work to assess the basic requirements for its botanical collection database. Goycochea (1998) suggested that INPA adopt a common approach to the collection of information across all its collections. The suggestion was considered feasible, manageable and would be satisfactory for most of the curatorial needs. Others suggested including the use of standard terminology (for data and metadata, independently of software platforms) and a much deeper analysis of functional and system requirements. This would improve the usefulness and accessibility of information stored and would improve communication amongst research teams, as well as ensure the collection inter-relationships.

Sonderegger *et al.* (1998), and Oliveira (1999), took a similar approach, emphasising the need for identifying the structural requirements of a more general BIS, and recommended that INPA include collection records and research information from all its scientific fields. Allkin (1998) notes that initiatives of this kind have also been thought of in Europe, the USA, and Asia, but so far without much impact. These systems often failed and did not meet the institutes' expectations. The reasons lay mainly in the lack of a comprehensive method, poor design, and poor project management, rather than in technology gaps or financial limitations.

It has been identified from previous work that activities of interaction do exist at INPA (Sonderegger *et al.*, 1998). Figure 2.9 presents the interactions between collections and activities.

The SCP is responsible for the management activities and procedures of physical collections and to automatise them using a database (dashed area in the figure). There are relationships between the SCP and research via collections material that is utilised in research (dashed lines). Research provides the input to collections (e.g., items, species identification) and databases and uses the data and information for analysis and publications. The SCP and research also interact with other activities such as: education, public interest, and government strategic actions and by producing output to scientific contribution for global conservation. This structure seems to be suitable for data from different collections which needs to be integrated and distributed amongst the various existing functions at INPA.

The SCP implementation followed the progress of the MVZ at Berkeley, especially the mechanism of defining functional and system requirements (Blum *et al.*, 1995). Another work considered was the ASC, a biological collection model (Association of Systematic Collections, 1997). ASC proposed a framework for global biological database connection, suggesting that integration of data from several institutes

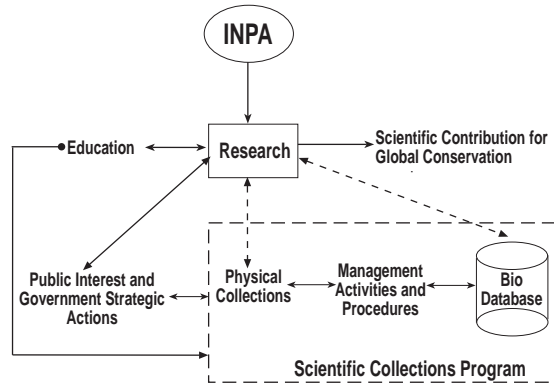


Figure 2.9: Interactions amongst functions and collections.

requires the existence and the enforcement of standards for data sharing. The work carried out by the OSU database model was also taken into account. The model is a representation of an insects collection and contains general attributes for biological collections as well as specific attributes of entomological data (Johnson, 1997).

Figure 2.10 presents the current organisational structure of INPA detailing how the SCP has structured the curatorships of biological collections. Under the research coordination, SCP and the library collections program have a relationship and complementary activities. Collections are managed by curators and large collection have subcuratorships. This structure has provided INPA with flexibility and independence at collections level and an ideal opportunity for development of a scientific data policy for sharing and dissemination of information across collections. INPA expects that partner institutes can also profit from its experience and adopt components of the initiatives tested and recommended.

2.4.2 A toolkit approach to manage collection data

Over the last decade, INPA's curatorial staff has put much effort in collection data management. Due to the considerable number of collections and the lack of a consistent overall plan, which would cover the majority of problems in representing collection data, the solution adopted includes a variety of software packages running on heterogeneous platforms. Table 2.14 illustrates the proliferation of systems that have been adopted across INPA collections.

It is well-known that such an approach causes problems regarding data integration. Nevertheless, because of the technology and resources availability at the time, the approach taken was to identify systems and apply them for managing collection data. This can be referred to as a toolkit approach. Given the software variety, the functionality of data exchange amongst different applications was much emphasised. Allkin (1998) stressed that the fundamental requirements for data exchange

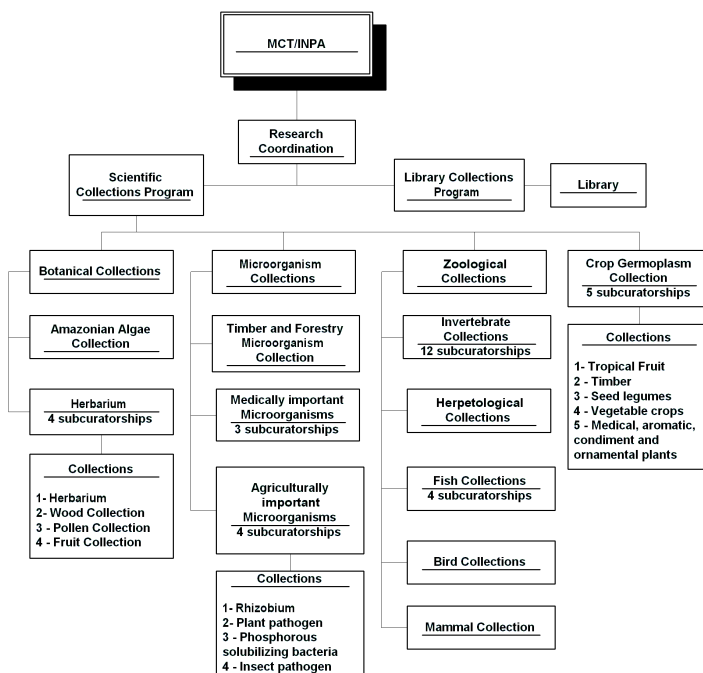


Figure 2.10: INPA's current organisational structure of the Scientific Collections Program.

Table 2.14: Collections and software diversity.

COLLECTIONS	SYSTEM IN USE	OBSERVATIONS
Amphibians	MS-Access	Table structure
Birds	none	none
Crustacean	SGC	SGC file format
Fish	MUSE	Based on Xtrieve
Herbaria	Foxpro and Brahms	Flora-Ducke and Brahms file format
Insects	Dbase, MS-Excell	Table structure
Mammal	MS-Excell	Access: Online database
Mollusca	SGC	SGC file format
Pollen	none	none
Reptiles	MS-Excell	Table structure
Trees	MS-Access	Table structure

are primarily the definitions and adherence to appropriate data standards, and secondly, transfer mechanisms between software products.

Some collection data management solutions have relied on the decision of the curatorial staff instead of on standardised criteria for adopting ideal software solutions. Obviously, this has caused major drawbacks for data integration. Some solutions at INPA emerged after the experience of the Plantas do Nordeste Project (PNE). The project is based in the northeast of Brazil and has been running since 1992. More than thirty Brazilian institutes participate in the project, federal or state research agencies, universities, non-governmental organisations (NGOs), and grassroots organisations involved in alternative agriculture, forestry, and community development. To maximize information dissemination and to identify priority audiences, it created the Plant Information Center, which utilizes the following criteria for adopting software solutions: generic software packages that features data independence, software packages that support data quality and integrity maintenance, developed products, softwares with functionality for data exchange, and data security (PNE Project, 2001b).

Three biological data management systems are currently being used by PNE: Alice, BRAHMS and DELTA (Campos dos Santos *et al.*, 2000; PNE Project, 2001a). These systems have complementary functions and are partially compatible, allowing data migration. Alice is used for merging, coordinating, and disseminating PNE's information about species (nomenclature, distribution, ecology, morphology). BRAHMS is used to manage herbaria and their specimen collection records. The morphological descriptions stored in Alice can be exported into the DELTA system. In this way, it is possible to produce identification guides (possibly electronically) from pre-processed species descriptions (PNE Project, 2001a).

Within PNE, a number of terminological standards have been established to resolve incompatibilities in areas of common interest. Descriptive standards have been developed for the vegetation types of the region, habitats, and use properties of forage plants and use properties for medicinal plants (PNE Project, 2001c).

2.4.3 Technological influences for information dissemination

The interest of INPA is to distribute and disseminate information from its collections to a global audience. Since most users are specialists, the analysis and summary of information contained in the collection will be performed by them. This strategy is the result of experience from the successful Ducke Reserve Project, which implemented the information delivery wanted by INPA (Hopkins, 1999). Figure 2.11 presents the strategy adopted in the Ducke Reserve Project. There are five processes that lead towards information delivery: coordination, integration, analysis, presentation, and delivery. The way these data were treated in the process before delivery depended largely on the number of steps or agents involved in the process.

Once data are digitised and stored, the access to the information presumably will help research, since it becomes possible to solve problems using data from multiple sources. Databases like Genome, Microbial Germplasm, Natural Fungi Collection and Nature Conservancy were developed using relational database technology

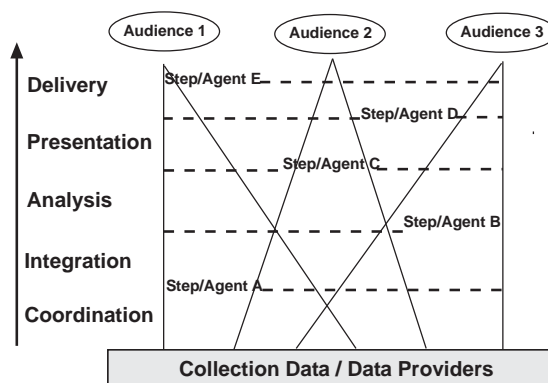


Figure 2.11: Strategy for information delivery.

(Newsome *et al.*, 1995). The users of these databases are more likely to be scientists, not familiar in accessing and querying database via command line interface. The problems range from the syntax of a query language, the issue of expressing what they want to extract from your database to the way the language is used during the process of interaction, which usually comprises command line interfaces. These related problems will cause errors and continuous drawbacks, as well as lack of feedback during the query process (Jarke & Vassiliou, 1985). The set of problems known in accessing data by naive users can limit the dissemination of information to a certain extent.

Tool developers have built a large variety of applications to construct browser-based interfaces to access these databases. These facilities can be categorised into one of three classes:

- Those that use an interoperability language to interface web browsers to remote databases (Cleary *et al.*, 1999).
- Those that are based on a schema by embedding directives into markup languages, for example XML (Bourret, 1999).
- Those that use tools designed to assist programmers in furnishing web access to comply with Open Database Connectivity.

The technology already in use can influence and ensure data access and guide the development of user interfaces. For instance, the use of the JAVA language seems to be appropriate, since with this language one can enhance the overall performance of the system by transferring some processing tasks to the front-end side of clients, to be achieved by JAVA applets. Another important aspect about JAVA is its portability. JAVA is an architecturally-neutral language, its compiler generates a JAVA virtual machine code instead of a machine code specific to the computer system one is using. Unlike C and C++, there are no implementation-dependent aspects of the code. Also, the libraries that are a part of the system define portable interfaces. The JAVA

system itself is portable, the new compiler is written in JAVA and the runtime is written in C with a clean portability boundary.

There is the need to make information available in a more flexible form, for instance, to have an integration-driven database schema, where schemas of existing or proposed database to be integrated into a global unified schema (Batini *et al.*, 1986). When integration is implemented, it can be customised and used.

2.5 Concluding Remarks

About the collections

The research activities within the collections show how underused and untouched the collections are. Collections have no value unless they are made available for inspection. Institutes have to commit themselves to find mechanisms to overcome their difficulties, and to meet their objectives, which include allowing full access to their collections.

It is evident that from all collections, only the herbaria have obtained a visible place, calling for attention and investments at both national and international levels. It is also evident that all herbaria visited are today included in the strategic plans of institutes, deserving funding for their digitalisation and management via database management systems.

There are opinions and different confrontations between decision makers and scientists. On one side, the decision makers request scientific evidence about the risks and the mismanagement of the biological collection to free up financial resources for their maintenance. On the other side, the scientists need all possible resources to succeed in their investigations and provide the much needed evidence.

What would make the difference?

To bring the participants to consensus, we envisage that the following actions would reduce the gap between scientists and the rest of society, they are:

- To develop an inter-institutional policy to prevent item loss from existing collections.
- To develop an open and advanced policy for use, formation, and sampling method to standardise events of collections, procedures of collection management, descriptions of geo-spatial components, and reference material related to the collections to make the data interoperable.
- To promote financial investment in collections and development of human resources in the field of environmental informatics and bioinformatics. The combined skills can be an important tool for information management focusing on the associated social aspects.
- To pursue partnerships for cooperation, aiming at the expansion of collections and at the use of technology for collection management.

- To develop computer infrastructure, integrated databases, and metadata with environmental and ecological information.
- To develop computer systems for education purposes, especially for environmental education.
- To develop application software to express and disseminate the Amazonian biodiversity based on collection information.

Our work is focusing on biological data and metadata management and contributes on biological collection development. It will provide development and tests of computer infrastructure, application requirements and conceptual representation of Amazonian biological data for database implementation. Further, attention is given to improve legacy data by incorporating geospatial information for use in advanced analytical environment.

Benefits of scientific partnerships

INPA and MPEG have a considerable competence in species taxonomy and also have the largest collections of Amazonian fauna and flora. It would be beneficial if institutes took action to establish partnerships for extended studies on inventories and use of the biological resources, enabling the integration and distribution of data sets amongst partner institutes. This scenario would require a considerable effort for negotiation, planning, coordination, and obviously financial resources. However, the synergy would bring more benefits for the country than isolated action. Moreover, Brazilian universities and museums could collaborate with international institutes for integration or repatriation of information.

Data integration could take advantage from the experiences of BCDAM (BCDAM-MMA, 1998), SIVAM (Secretaria de Assuntos Estratégicos, 1993) and PROBEM projects (Ministério de Ciência e Tecnologia, 2002). Other initiatives under development, for instance BIOAMAZÔNIA (Silva, 2001), depend on partnerships and the collaborative use of databases (data sets). Furthermore, geo-spatial components of biological data are essential. For this, INPE (Instituto Nacional de Pesquisas Espaciais) has all the necessary expertise to contribute.

In this context, it is important to emphasise and understand the interconnections amongst collections, as proposed by Lane (1996), which is applicable to this scenario. Additionally, given the small number of taxonomists in the region, partnerships are the only plausible way to advance. This is seen with INPA's collections that are getting the benefits from a long term partnership with the Max-Planck-Institute für Limnologie (Germany), the Institut de Recherche pour le Développement (France), the Smithsonian Institute (USA), and the Department for International Development DFID (United Kingdom).

MCT: strengthening partnership

INPA is under the administrative structure of the MCT and their activities have complementary relationships. Regarding the subject discussed in this chapter, we

2.5. CONCLUDING REMARKS

observed three major roles for MCT that can facilitate advancements in the initiatives that INPA is pursuing. Our observation should not be considered as recommendations towards INPA's management, but an additional opportunity to point out important roles that can bring benefits for the institutes involved. Given the characteristics of institutes and their activities developed, the observations hold for other institutes in the region. The three roles are:

1. Policies formulator, coordinator and sponsor of initiatives. MCT has a national mission and by having that, should enhance the national strategic programme for biodiversity in development by MMA, to formulate a scientific policy for inventories of the biodiversity in the region (Wheeler & Cracraft, 1997). This policy would encompass definitions of priorities of study, human resources development and investment in infra-structure. MCT could also promote and coordinate, through its institutes in the region, the definition of guidelines, objectives and targets for its accomplishment. The CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) could be responsible for the implementation via its programme of scholarships and research projects funding;
2. Policy definition for intellectual property and information access. This is a complex subject, specially because there are several participants with undefined profiles, both legal and juridical;
3. To develop and finance the implementation of integrated BIS. This could be done within the PPA (Plano Plurianual) of MCT via its Management Program for Ecosystems.

Brazil should consider the Amazonian biodiversity issue as high priority, due to its importance and influence in the regional and global processes, and its potential to promote economic wealth. To know, understand, maintain and use the biodiversity, it is important to have high quality information available, that in this context, can be considered an important asset. This can be achieved by automated systems that must be able to manage, analyse and propagate the information held by the institutes. The systems can become a valuable tool to support scientific research and education, and consequently be able to expand the knowledge about the biodiversity, fundamental for the region's future.

Chapter 3

Requirements and Functionality: Users and Systems

3.1 Introduction

In 1997, INPA started to automate its collection information. It adopted a tool kit approach, such that every curatorial staff member could identify, develop or use a system that could provide functionality for data and information management. This resulted in a proliferation of different system platforms (spontaneous development) as presented in Table 2.14 (Chapter 2). In 1998, the Scientific Collections Program (SCP) signalled the need for a new system, capable of integrating data from different platforms and using up-to-date and scalable information technology, methodology, hardware and software. In parallel, INPA had deployed computer infrastructure including personal computers, network connections, software for document preparation, statistical analysis, e-mail, Internet, database management, etc.

From 1999, we initiated a three year project to provide information for guiding system replacement or alternative procedures which could enhance the management of collection information in the Amazon region. To achieve this, the project was developed in stepwise phases, which were: survey of the biological collection phase, analysis and design phase (requirement analysis, information modelling and interface design), programming and test phase, data migration/capture phase, deployment and maintenance phase. Depending on user and system requirements, certain applications could be developed as components, allowing high priority applications to be made available before the entire system was completed.

This chapter aims to detail the initial analysis and design activities for INPA's biological collections. We adopted to a certain extent the method used in the Museum of Vertebrate Zoology (MVZ) — University of California, Berkeley, in the scope

of the Collection Information System re-engineering project. Due to domain similarities the method also identifies the functional and system requirements at INPA's SCP (Blum *et al.*, 1995). We looked into work carried out by Allkin (1997 and 1998) and Sonderegger (1998) regarding recommendations for automated collection systems. INPA had been selected as a test case due to its more advanced status on SCP in Amazonia, as portrayed in Chapter 2. Due to the recognised awareness of how important collection information is, together with INPA's commitment to implement a comprehensive BIS to manage biological data, INPA became an ideal institute to identify the function and system requirements. The analysis comprised of functional requirements, which determined the needs for specified information processing technology, and system requirements, which described the essential characteristics of the hardware and software.

3.2 Functional Requirements

The collections provide support to research, education, public interest and government strategic actions (see Figure 2.9 in Chapter 2). The management activities and procedures, that are developed by the SCP, are the result of the fast growth of collection data and the need to study and communicate this data in new ways.

3.2.1 Research

Information-related activities within the research area are grouped in four categories, as follows:

1. **Data Gathering** — The best way to gather data is by using electronic means and perform it as close as possible to the source (e.g., during a collecting event). Electronic data acquisition can increase efficiency by eliminating data entry and enforcing consistency in data collection procedures. Although it is practical to record data into computer devices, such as laptops, personal digital assistants (PDAs), global positioning system (GPS), etc, by installing them on field work locations or in laboratories, in the Amazon region this can also constitute a problem. The climate is extremely hostile to fragile equipment and it can restrict its utilisation tremendously.
2. **Data Management and Analysis** — For researchers, the most important requirement for an automated information system is to provide access to and views of collection data from any geographical location within reasonable limits. To achieve this, it is necessary to establish network functions together with a client/server architecture and interface components to query/browse/print results. These requirements can contribute to the integration of data, visualisation and application for analysis (e.g., GIS), and consequently, the enhancement of data quality. More frequent use of collection data increases the researchers' knowledge about critical issues related to the collection data. For instance, accuracy, completeness, consistency, etc., which can be controlled by more efficient collection management protocols. Some collections produce very large data sets. The challenge is to provide structure for to these data sets and

the subsequent long term use of them. To achieve this, INPA has to rely on database technology.

3. **Publication** — The basic infrastructure for research output, data and information dissemination is INPA's web server. This technology is now commonly used to publish biological collection information from data, metadata, digital printed versions, such as relational data structure, images, and texts. The SCP is already on INPA's web portal, though not as dynamic as it should be. INPA has to provide facilities for a more efficient update mechanism through the web. The front-end should contain advanced query interfaces to all online catalogs, digital images of collection items, curatorial personnel records, descriptions of research programme, and research activities developed in relation to each collection.
4. **Communications** — The electronic-based communication between researchers, their colleagues outside INPA, and the general public, is carried out through the local INPA network in the campi and internetworking facilities. Applications such as mail programs, web browsers, ftp clients are installed in each desktop computer. Services run without interruptions, with access restrictions as defined by the SCP policy.

3.2.2 Conservation

INPA can be considered as an important information base for conservation initiatives throughout the Amazon region. For optimal use, linkages are required with other information sources and expertise outside the SCP mission. The expertise required includes: remote sensing, satellite imagery, demographic information and land-use and cover data. To make collection data useful in the context of GIS outputs, collecting localities must be recorded and georeferenced to a specified coordinate system and spatial resolution. In INPA's collections, 98% of the data has no exact georeference. The georeferencing process deserves high priority and must be carried out with a combination of query mechanisms and data manipulation functions embedded in database systems, electronic gazetteer and tools that enable users to digitise spatial data from maps.

3.2.3 Education and Capacity

Education and information reinforce one another. For the successful use of BIS, adapted capacity is required. Educational systems are fulfilling a leading role and continuing advancements are necessary.

INPA in this respect, has a well-known, high level education programme. Its postgraduate disciplines, seminars, workshops and the interactions with most of the academic institutes in the region makes education a central point in the whole Pan Amazon. High level education in this respect provides all the advantages of the technology already made available, more specifically, dissemination of information, which can be accessed by the student community. Information here consists of a variety of printed documents, research results, bibliography, etc. Students have network connection, computer facilities and software to support their research and the

main vehicle of communication is the web. There is a great potential for the provision of education through distance learning and INPA is already planning to provide tools for web based academic material. Even though the technology is already available in the market, INPA lacks the necessary resources to develop distance learning educational materials. The issues mentioned indicate that the provision of education and the capacity to develop this also requires considerations regarding functional and systems requirements.

3.2.4 Collection Management

The collections

The scope of the information to be managed by a BIS requires a standardised information model or schematic representation and an adequate system design. Detailed information about the diversity of INPA's collections can be found in Chapter 2.

In the following, we list the subjects that describe a collection and that must be included in a BIS. The list is not a comprehensive one, it will be elaborated upon in the next chapter when describing a schematic representation for biological collections.

At this point we provide an overview of the diversity of attributes that a BIS has to manage. Cluster (or categories), class, object and entity information include:

- **Collection items** — Accession number (an index that represents the number in the collection), type, receiving status of items, lot description (a lot is a group of specimens associated at some taxonomic level and from the same collection locality), specimen, parts, situation.
- **Taxonomy** — Taxon rank, parent taxon, synonym, taxon relation, classification, taxon origin reference, determination, description.
- **Collecting event** — Date and time of the event, place, collecting method, collectors, expedition description.
- **Locality** — Place names, habitat description, cartographic reference, spatial coverage, geo-reference
- **Individual user** — Name, address, affiliations, skill description, organisation position, relationship to organisation.
- **Organisation** — Name, address, acronym, relationship to individual users.
- **Reference** — Book, technical report, thesis, article, web publication, publication in proceedings; authors, title, keywords, editors, pages, volume; relationship to collection items, events, locality, taxon identification and classification.
- **Transactions** — Loans, invoice number and date, transfer term of material, shipment, loan/borrow period, return date, transport information.

A more in-depth description of what biological collection management is, the purposes and processes involved is given in Chapter 4.

Collections management functions

Traditional collection management includes the following basic functions: creation and maintenance of physical specimens, catalog records, transaction management, and information retrieval and reporting.

Creation and maintenance of catalog records

1. **Cataloging** — The information-based objective of cataloging is to support the systematic (re-)traceability of collection items and associated information. To do so, we must be able to determine a unique identifier to collection items; to record and maintain descriptions about items; and record their relationships to other items. Two major activity groups are involved in cataloging: (1) Data entry and validation, and card labelling, and (2) Tag generation.

For the first group, the systems must have the following capabilities:

- (a) To enter records for new items and for items that are already cataloged on paper.
- (b) To reduce redundancy during data entry by helping the flow of information from the accession process to the cataloging process.
- (c) To permit flow of information from the access process to the cataloging process.
- (d) To allow linkage of many specimens to the location of a single collection event.
- (e) To import field data that has been captured in devices used in the field or installed in laboratories.
- (f) The system must support a validating and correction process for new records, that is, listing new records, allow records to be edited and validated data to be inserted into the database.

The process of tag generation must provide print-outs of cards, labels and tags identification, either individually or in batch-mode.

2. **Catalog maintenance** The system must allow retrieval and update of existing data as well as to print identification for new items entered or replaced. Permission for updates must be imposed according to the SCP collections policy.
3. **Data quality management** The current structure of the collection data are non-normalised, and may contain redundancy across collections. The curatorial personnel are still not able to maintain consistency in implemented databases. There are two functions that must be pursued:
 - (a) Normalisation of data design.
 - (b) Examination of data content and improvement of consistency. Data consistency should be promoted within the system by using normalised data structure.

Transaction processing and management of transaction information

1. **Acquisitions** — The acquisition documentation describes the way in which an item ownership was established. To do that, a system should provide the following functions:
 - (a) Facilitate the flow of information from acquisition to the collection catalogs.
 - (b) Minimize the redundant data entry and produce cards, labels, tags, etc.
 - (c) Link acquisition to collecting events, and shipping authorisations.
 - (d) Link accessions to specimens.
2. **Loans** — The objectives of loan functions are to know the situation and location of any item of collection, and to allow curatorial personnel to report on the usage of the collection and its objects. A system must provide the following functions:
 - (a) Support login of requests for material.
 - (b) Query of and report on collection catalogs allowing query results to be utilised during the composition of a loan process.
 - (c) Produce invoices and shipments, and enforce loans policy.
 - (d) Identify the items of the collection on loan as well as all details about them.
 - (e) Track items on loan and keep a history of each specimen on loan.
 - (f) Monitor active loan period.
 - (g) Process the return or extension of a loan, either partially or in full against the original invoice by permitting updates during this process.
 - (h) Allow complete or partial loan transfer.
 - (i) Display statistics of loans by period.
3. **Borrowers** — The system must be able to manage the collection items borrowed by the SCP from other collection holders. The functions must include:
 - (a) Track of the borrowed collection at transaction level.
 - (b) Record all information regarding the borrowing process, such as the storage location of borrowed items.
 - (c) Print a transaction-level invoice that the lender can use to control the return of materials.
4. **Accession displacements** — It is imperative to record collection item movements. It ensures accountability for collection items and to ensure that collection personnel know what has happened to a specific item. In these circumstances, the general catalog should record the information whether or not the item is physically present. The automated system should support the following functions:
 - (a) Update catalog record by registering the displacement, indicating all information related to this transaction.

- (b) Link the accession displacement to a shipping record and authorisation information.
 - (c) Produce displacement summary by period.
5. **Shipping** — SCP needs to control all the outgoing shipments. The system should support this function by the following processes:
- (a) Printing elements for a complete invoice.
 - (b) Monitor shipments.
 - (c) Link to authorisation and possession.
 - (d) Produce summary reports of shipping activities.
 - (e) Provide online descriptions of carrier's details regarding cost, permits and insurance.

Information retrieval and reporting

Today, the access to information is provided by curatorial staff that utilize their expertise to locate and retrieve the requested information. SCP aims to make as much of the collection data available as possible to users at local and remote locations, taking into account mandatory access restrictions. Curatorial staff should be involved in information retrieval only when users lack access to the online material or when it concerns a specific problem. In the following, we describe tasks and capabilities that the system must have regarding retrieval of and reporting on the information.

1. **Query and reporting** — These functions will serve more users. The system must support the following functions:
 - (a) Form-based advanced queries — the interface should allow users to retrieve records as either formatted text or structured data, ideally to export to another system, allow browsing and visualisation of controlled vocabulary defined for the system, and should guide users during query/browse evaluation providing statistics and estimated results.
 - (b) Customisation of query and reporting functions.
 - (c) Standard query language (e.g., SQL, XQL).
2. **Standard reports** — The system should provide a set of standard reports in support of daily collection activities, including summary statistics that can be used in annual reports and project proposals. Important information to be contained in the reports includes:
 - (a) Summary of transaction activity.
 - (b) Summary of cataloging activity within a specified period.
 - (c) Number of accesses to items of collections.
 - (d) Summary of specimens prepared, by collection, by period.
 - (e) Report on the growth and usage of a collection.
 - (f) Summary of the collection related publications within a period and with statistics of researchers, authors, and additional bibliography.

3.3 System Requirements

The system requirements aim to provide a feasible specification of a system architecture that can respond to the immediate needs, taking into account scaling and expansion that may occur in the future. System requirements usually describe the components, and potential of a system that can fulfill the functional requirements (Hatley *et al.*, 2000; Maciaszek, 2001). We tried to be as detailed as possible, since INPA's experience can be viewed as part of a broader network of Amazonian initiatives, with existing and evolving facilities. Additionally, information technology trends incline toward openness and interoperability (Fairchild, 1996; Sventek & Coulson, 2000). System modules of today are developed as interchangeable components that allow a complete, automated system to be comprised of different components from different developers. Collection management activities are moving away from single package solutions designed to work in a specific environment, dealing with a specific problem, or even a reduced number of problems, to integrated solutions focusing in interoperability functions. This prevents an institute from adopting a system that will have no future and all the resources already utilised from being wasted. Also, it can facilitate the process of system requirements, where characteristics can be overlooked or underestimated, that can later be adjusted simply by updating components with low costs. Obviously, the current system components convention cannot be applied across institutes due to the high costs involved for such a sudden change.

The system requirement analysis reviews the categories of processing that will be needed to meet the requirements of SCP. This is followed by our review on data types, data volume and usage, user classification and system and data security. Finally, the analysis enables us to report on the maintenance and system flexibility requirements, and to review existing constraints on the system architecture.

3.3.1 Processing requirements

1. **Database management capabilities** — INPA's collections hold approximately 2,500,000 items that have been described to a certain extent, by applying standard procedures. The information on cards, labels and tags presents consistency and data structure. This makes DBMS 'the eligible' candidate for data management. A DBMS can evaluate and operate on data within their proper context (Stonebreaker *et al.*, 1999; Mannino, 2000). For an appropriate data model, the development should consider the (object) relational technology because it offers flexibility and support for non-predefined queries, a common need in scientific data management, and also because these products dominate the data management market. Information models developed for museum data indicate that collection information is more complex than was previously thought (Blum *et al.*, 1995). This complexity requires a sizeable database, for instance with a large number of many-to-many and recursive relationships, and consequently placing the DMBS into a client/server class (Hatley *et al.*, 2000).
2. **Application development, query, and reporting tools** — Following Hat-

ley's arguments, database servers offer robust management capabilities with security enforcement. However, database servers have no facilities to help the access of information for naive users. Instead, they receive requests and respond to external applications for presentation. To speed-up development, SCP will need to count on application development tools to create a well-designed user front-end application. Additionally, an *ad hoc* query and report generator tool will also be necessary to provide mechanisms to extract and format data from the database.

3. **Text and document management** — There appears to be no urgent need for document management via an institutional information retrieval system, like File Magic, the Hummingbird DM or Lotus Domino. Such systems capture, store and organise documents and information with secure access to/and the retrieval of documents, making important information readily available in a network environment. In spite of the large numbers of document produced by INPA, adopting a retrieval system is viewed today as a potential long-term need.
4. **Geographic information processing capabilities** — All collection items were collected in a certain geographical place. Spatial data thus is an important component of collection. Only the very recent items collected (about 1% of the whole collection) are available for analysis by GIS systems. SCP needs geographic information capabilities for georeferencing legacy data, map generation and other analytical outputs from collection information and georeferenced data. The process of georeferencing is time-consuming and should rely on tools/equipment that enables a user to digitise geographic attributes, speeding up the process. The analysis and display functions still need to be addressed.
As a starting point, INPA can invest in building-up complementary expertise by setting-up a bio-geo-informatics laboratory and by collaborative works and inter-institutional relationships in those areas. Partnership amongst institutes is vital. The Instituto Nacional de Pesquisas Espaciais (INPE), the Instituto Brasileiro de Geografia e Estatística (IBGE) and the International Institute for Geo-Information Science and Earth Observation (ITC), are the type of organisation, which concentrate expertise in those areas and links for collaborations should be looked into.
5. **Image processing capabilities** — Requirements for capturing, manipulating, and displaying images appear in several areas. There are three areas in which image processing is considered, and the related tasks should be undertaken by collections personnel:
 - (a) Collections where type specimens can not be lent and researchers must examine on site of collection (e.g., mammals, birds). It would be important to make available over the Internet a series of good quality photographs of birds and mammals which could enable researchers to determine if on-site examination is required.
 - (b) Media publications (Internet, newspaper, books). Historical material is particularly important as habitats are changing fast.
 - (c) Field notes scanned images.

3.3. SYSTEM REQUIREMENTS

From the use of images on web documents, small scale needs will arise. In such document types information can be summarised (e.g., essays) and composed with a small number of images. In each, the context image capabilities go into basic capture, storage and delivery functions. Even though GIS and/or mapping applications, include digital image processing (DIP) software, and provide functions for image processing, high resolution printers and scanners are needed to serve all image services.

6. **Audio** — The requirement for digital audio comes from the mammal and bird collections. Audio signal processing requirements include: analog to digital conversion; mass storage of digital audio files, digital audio playback; and the generation of sonograms from several spectrographic analyse. Today very little has been done by the SCP to allow these capabilities. Although digital audio files are large, their provision could preserve the collection without loss of quality, and make the collection accessible to online users.
7. **Video** — The demand for digital video is unknown, as manifested in all collections, it is of interest for studies in animal behaviour (e.g., mammal, reptile and amphibian collections). Also, video material could serve as a more permanent way of archiving.

3.3.2 Data Types, Volumes, and Usage

In general, information to be produced requires data of different types (e.g., alphanumeric, image, audio, video). Consequently, this imposes rules for system input and output, that are reflected in the information volume and usage protocol. These characteristics and circumstances are responsible for defining system design issues (i.e., storage mechanisms, backup policy, data structure and system performance). In this section, we describe the characteristics for the different data types regarding input/output, volume and data usage.

Alphanumeric

- **Input** — Keyboard, external data files (being authority files), scanned paper-based information.
- **Output** — Display monitors, laser printers, digital files, system export functions (via network and to portable media.)

Beyond doubt, concerning collection management, output printing constitutes a problem to be resolved. Most specimen catalog cards, tags, and labels are still being prepared manually (India ink and special type of paper). Box (or dry) labels are in the process of substitution by labels produced by laser printers. This substitution will be extended to produce labels to be used in fluid collections (alcohol, formalin, and glycerin). Tests performed at INPA's fish laboratory indicate that the fixation of the laser printer ink to labels is satisfactory and they are now in use. Tag identification presents a special problem: tags may have different sizes and some come with fastening strings pre-attached, moreover, information may be written on both sides. An option to use an automated

Table 3.1: The MVZ five year estimation of disk space needed for alphanumeric data (Source: (Blum *et al.*, 1995)).

SOURCE	VOLUME (Mb)
Catalog	646
Transactions	12
Total Alphanumeric	658
Overhead Rate	× 2
Required Storage Space	1,316

printing system, should be considered to print information onto adhesive label paper and then affix labels to the tag. Tests carried out in collections at INPA indicated adhesives degradation and reduction in adhesion to the tags. For this problem no solution has been found yet. Regarding jar labels, they contain summary information, specially the taxon name and the location where the specimen was captured. Besides labels and tags still being produced manually, in the future, SCP expects that all labels (wet and dry), catalog cards and specimen tags will be printed by an automated system.

- **Volume** — It was not possible to estimate the storage volume required for alphanumeric data, since we had no indication about the future data structure or the platform of the database management system. The only information available, in this respect, was the MVZ's projection based on collection size and annual growth rate for a period of five years (see Table 3.1). Total estimated alphanumeric data, based on collection catalogs, transaction management records, and DBMS overhead rate (table indexes, but not the temporary space required by the DBMS for the data manipulations) of 100%, is around 1.3 Gigabyte (Gb). There is no available information regarding the MVZ growth rate for the next five years. A study to quantify the SCP data volume has yet to be carried out.
- **Usage** — We were unable to predict the frequency of usage for creating, reading, updating and deleting data, because first, SCP has no consolidated history of all the requests received and records provided.

Image

Biological collection information needs image management. Potential candidates for digital images include: specimens and preparations, field notes, correspondence, photos, slides, illustrations and maps.

- **Input** — High resolution (gray-scale and colour) scanner, video camera with a frame-grabber device and digital camera. GIS software and products shall be

3.3. SYSTEM REQUIREMENTS

acquired via the network or on CD-ROM, either from commercial providers or academic institutes.

- **Output** — PC graphical monitor and desktop laser colour printers to attend a networked community of users.
- **Volume** — The storage space required depends on image size, colour depth, image content and file format. Even though we cannot estimate the disk space, we can guess that it will be substantial. Just to give an idea, tests reveal that a small primate skeleton comprised of four images will require 500KB. A field note page requires 25KB.
- **Usage** — Digital images should be created and made available via an online image resource. Images do not need updates by editing, at best an image will be replaced by a new one.

Audio

The need for audio data is derived from the bird, mammal and insect collections, which are comprised of vocalisations, recorded in the field. Vocalisations have to be analysed from the sonograms.

- **Input:** Magnetic tape in analog form. Basic devices to be used include: tape recorders, speakers, an amplifier. Facilities for converting analog to digital and interface to personal computer ports must be implemented.
- **Output** — Sound analysis workstation and spectrographs.
- **Volume** — No study was conducted to estimate the size and hard disk space that will be necessary to store the entire collection. However, the majority of recordings are bird vocalisations. The study carried out by the MVZ indicates that one complete vocalisation represented in digital form at 48 kHz will require 350 MB of digital data (Blum *et al.*, 1995).

A good collection would eventually have an average of 10 audio samples per species and each sample being on average 2 minutes (120 seconds). Lets consider that half of the samples are recorded in mono, half in stereo, adopting the well-compressed MP3¹ format. For other formats consider factor 4 of the MP3.

The file size after 10 seconds reading mono at 44 kHz (CD quality audio) is around 70Kb. The estimation of the space needed to store digital audio files of INPA collections can be calculated by the following:

$$A = T * N * C * L * F$$

where:

A is the audio estimation size

¹MP3 is the file extension for MPEG, audio layer 3. Layer 3 is one of three coding schemes (layer 1, layer 2 and layer 3) for the compression of audio signals. Layer 3 uses perceptual audio coding and psychoacoustic compression to remove all superfluous information (more specifically, the redundant and irrelevant parts of a sound signal).

T is the total number of species in the collection;

N is the number of audio samples per species;

C is the compression factor for MP3 mono/stereo;

L is the audio length time; and

F is the file size after reading 10 seconds mono/stereo.

Considering the INPA collections, (see information on Table 2.7), the number of species are: birds = 500; mammals = 3,000; specific insects = 2,000; frogs = 700. The total species is 6,200.

The audio estimation size (A) will be ≈ 124 Gb in MP3 or ≈ 620 Gb for other formats.

- **Usage** — The data usage patterns regarding creation of digital audio files include the digitising of all analog material onto more appropriated media for the benefit of inventory, monitoring and analysis and as the result of ongoing fieldwork and analysis of vocalisations still in analog form. Due to the size of the expected volume of data it is not recommended to maintain it online. The option would be to keep the samples in removable media (e.g., CD jukebox) and the process of load/unload would remain a manual one.

Video

INPA collection does not yet include digital video. Some material produced is not available for public use, and some belong to researchers and other are used privately. Due to the importance of video material (e.g., study of animal behaviour), the next step of SCP should also concentrate on acquiring video equipment for a multimedia laboratory to serve all collections.

3.3.3 Users

User classification

- **Curators** — are responsible for collection activities and relationships with other activities of the institute. The main information required is for use in reports and data analysis. Since the level of computer skill amongst curators varies (none of them are computer expert) the recommended approach should be via an easy-to-use interface. Such interface should be able to retrieve data and make use of pre-defined queries as well as to provide a flexible *ad-hoc* query tool. Curators also are responsible for research expeditions and for the creation and recording of the field notes collected, and specimen labels and tags. In some collections, the possibility of using electronic equipment for recording data (e.g., portable computers and PDAs) is discussed. This can have a positive impact on the quality of collection data by reducing the workload of transcription and interpretation. Also, this can allow the implementation of an electronic notebook, through which researchers can take collection material into the field and work with the newly collected material in the field. For material to be posted on the SCP web site, a curator may be requested to provide the

3.3. SYSTEM REQUIREMENTS

material; which can be teaching material or any specific information regarding a particular collection.

- **Subcurators** — are the more frequent users of the system. They can perform all information-related tasks, which encompass the activities of curators and assistants, as well as tasks for system administration (e.g., accounting and authorities, etc.). Also, subcurators are responsible for maintaining the quality of the information of collections, and for training users of the system.
- **Assistants** — perform the basic activities in the collections (e.g., cataloging, accessions, loans, borrowing, shipping, etc.). They are responsible for the data entry and monitor transactions within collections. A problem faced by the SCP is the shortage of assistant staff and thus, some activities float to curator level. Usually, SCP counts on short period apprentices (students) with a low level of expertise in both collection procedure and computer use. This situation is common across all institutes that are biological collection holders. Tools to be used by assistants should help them in the daily activities, and mainly prevent them from entering inconsistency to collection data.
- **Students, Research and Teaching Assistants, and Laboratory Technicians** — INPA has a diverse graduate program and students perform several activities related to collection information. They produce material out of their research or by specific activities within the SCP, under the close supervision of curators. Students are engaged in other activities including: assisting in teaching and preparation of course material and working as curatorial assistants.
- **Other Users** — Amongst other users there are multi-disciplinary researchers from outside institutes and the general public. These users require basic descriptions of the collections and the items and a system capable to extract data and describe requests precisely. This is desirable because it can make collections more useful while at the same time helping curatorial staff to serve a larger number of users.
- **Information Technology personnel** — The descriptions of the processing management indicates that to manage a large digital biological collection will require robust database server technology. Such a platform requires well-trained personnel to set it up and run it accordingly. In addition to that, customisable tools will be required to access collection databases, to develop and manage web server services; and provide capabilities for development of GIS and mapping products. These areas encompass a variety of computer technology skills, which are:
 - Server Manager** — manages systems security, physical resources, and performs operating system updates.
 - Database Administrator** — manages database physical resources, allocates and monitors database space; creates logins; configuration of database software; optimises database performance; and performs backups and restoration.
 - Database Designer** — creates and modifies database structure; manages indexes, monitors performance; manages integrity constraints and user authorities.

System Developer — activities are performed by programmers and analysts working with other technical personnel and users and database designer to implement database applications at levels of client and server.

Web server Developer/Administrator (Webmaster) — develops and maintains web server platform; works in conjunction with other technical personnel and web sites content to create and maintain information resources on the web server; develops and maintains a code that allows the web server to respond as a client to databases.

PC Manager — installs and maintains personal computers and devices; installs and upgrades hardware and software; configures and provides instruction for using PC-based applications.

Authority capabilities

The SCP policy requires user authority management. The system administrator should grant and/or revoke authority according to implemented guidelines. The aim of these authority guidelines is to protect the system from unauthorised access, to control data integrity, and to prevent access to sensitive or unverified data. All users, apart from the class of authorised users, will require to comply with a log-on and password to get access to the system. Independent of the classes of users, web material to be used in dissemination and teaching can be updated only by specific authorised users.

In the following, we present a proposal for authority classes:

- **Curators** — They must have full authority to read and modify all collection data. Curators must be aware of updates done by others. The task of data modifications shall be performed by subcurators, since they can update information following a paper record and following the SCP procedures.
- **Subcurators** — They must have the authority to modify authorities, create, read, update, and delete any record in the database.
- **Assistants** — Should be allowed to read and modify the data about collection items. The system should prevent this class of users from deleting an entire data set. Regarding modification, assistants should not change records which require expertise (e.g., taxonomic classification and determination). Also, they cannot modify authority files.
- **Students** — Should be allowed to read all public data. However, students engaged in more elaborated curatorial activities, can be granted the same status as subcurator by a curator.
- **Other Users** — Should have access to all public data.
- **Information Technology personnel** — This class of users must have full authority, i.e., to create, modify programs and data.

Locations

The number of users and their location determine the necessary topology of the communication infrastructure, application, and number of computers and workstations.

3.3. SYSTEM REQUIREMENTS

Table 3.2: Number of the potential simultaneous SCP users.

User Group	Number of Users
Curators	11
Subcurators	33
Assistants	35 – 45
Students	≈ 100
Other Users	≈ 20

INPA has provided, before any attempt of a BIS implementation, the ideal infrastructure to attend any request at Intranet and Internet level with client/server architectures. The system supports implementations of query and report generation via hyper-text transfer protocol (HTTP; i.e., a web server) integrated to database servers.

System availability

INPA's activities vary in range and time schedule. The requirements for system availability indicated by the community is year-round. At the moment, the infrastructure is trying meet these demands by having network and server redundancy. INPA cannot provide the personnel to cover extended working hours (weekends and public holidays). This will affect the system administration tremendously, increasing unpredicted down-time of servers. So far, no solution has been put forward and there is no alternative for fast recovery after major incidents, like hardware failure. Even though the system presents a certain degree of redundance, to replace hardware can take weeks due to unavoidable institutional procedures of acquiring new materials.

Number of potential users

Curators and subcurators are the users of the system. The student community numbers around 400 students per year. It is estimated that 25% of the students access collection items daily. Other users, mostly visitor researchers, are estimated to be around 20 simultaneous users. This number will tend to increase due to the SCP plans to made computer stations available at the central library for the general public to access collection information. Since the computer infrastructure is already in place, and said to be satisfactory, the combined demand it seems will not affect system performance. In case performance degrades a multi-task strategy must be considered. The number of potential users are presented in Table 3.2.

3.3.4 System and Data Security

A system should be protected from failure, human error, theft, and natural events of nature (e.g., fire, flood, etc.). To prevent any loss, redundancy is the most effective solution that a system can have. It is done through a backup procedure; it should maintain daily copies of the work. The backup protocol should cover all data and information stored in all computers used by the SCP system.

MVZ backup policy highlights important factors that a backup protocol should focus on, which are:

1. **Objective** — Backup procedures aim to ensure information recovery from a copy in the event of a disaster. The backup archive maintains the state of the information (a version) at specified intervals, allowing it to be recovered. In adopting a backup policy, it will be necessary that additional attention is paid to media life span, long term data format and media compatibility, and backup security off-site.
2. **Update rates** — Since not all data changes at the same rate, the frequency of updates should be in accordance with the effort required to re-create them. For instance, application program files are never modified and need to be copied only once. Application configuration files should be copied after each modification.
3. **Down-time**
 - (a) For backup — Some backup systems require the procedure to be performed with the repository off-line. Some, more sophisticated, systems allow backup to be performed with the system on-line. Incremental backups can copy only the changes made since the last full backup.
 - (b) Before recovery — Critical information should be backed up in ways to minimize the recovery time. For this situation, tape-based backup is not recommended.
4. **Undetected error** — Subtle error can go undetected in digital files, particularly large databases, for long periods. Backup strategy can be an effective resource to rectify such errors. This was the case that occurred during a data recovery of INPA's central library.
5. **Backup checks** — Backup itself can fail and data may end up not being copied correctly. The backup protocol should adopt automated verification at backup time. Also, the copies should be verified for readability in another similar hardware device.
6. **Backing up backups** — Backups, for obvious reasons, should not be kept at the same site where data and system are running. Also, they should be protected against humidity, especially in regions like Amazonia where the humidity can be high for long periods. Backups should also be protected against unreliable hardware and software upgrades. Such circumstances can be destructive.

Backup policy for database server

The creation and update rate of INPA's collections is expected to be large and should therefore be supported by a daily backup. The restore time is not critical as many transaction processing systems are, but minimising the down time and the number of backup versions should be a high priority. The archive protocol should guarantee the daily version to be recoverable up to a month, monthly versions to be recoverable up to a year, and semiannual versions to be recoverable for ever. In case any server software is upgraded or reconfigured, backups should be performed.

Backup policy for desktop computers

Curators share the opinion that SCP should provide backup services per equipment that collection staff use on a daily basis. SCP currently provides a backup service for the main server but some machines placed in specific collections may be unprotected. Propagating backup facilities across the SCP network will demand activities from the system administrator, consequently inducing more costs. The ideal solution (attending the curators request for backup of all information) would require large initial investments, not available at the moment.

3.3.5 Maintenance and System Flexibility

Hardware and software, in parallel with users needs, are dynamic and constantly evolving, providing the necessity for constant maintenance and flexible synchronisation of both. Current collection management is performed by several application solutions with no integration in mind. For an integrated solution a system must provide a wider set of functions. Those functions are common in client/server architecture, in which the data server requires procedures and triggers that ensure data integrity and database performance, and on the client side, enables users to enter and edit their data in a user-friendly way.

Any proposed system aiming to maximize flexibility and to minimize maintenance cost should consider the following factors:

1. **Scalability** — The computer infrastructure should be scalable to accommodate capacity increase by incrementing components rather than system replacement. For example, the expansions of storage capability, memory or CPUs to improve performance, network backbone throughput, etc.
2. **Extensibility** — It refers to a system's flexibility to incorporate new features (hardware and software technology) needed. For instance, the MVZ approach divides the extension of a system into minor, moderate and major. The minor ones are small extensions that have little effect in user perception of the system; moderate ones change the way a user performs an existing function and modifies the system to hold new information; and major ones refer to the development of new functions, and may imply the insertion of new software (Blum *et al.*, 1995). INPA consider two types of extension: light and heavy. The light extensions incorporate the minor and moderated extensions of the MVZ

and the heavy ones are similar to the MVZ's major extensions. These classifications only provide an idea of the kind of modification that is expected for a system. During our investigation, it was not possible to anticipate the full scope of maintenance before an actual system is implemented, since it is either from observing a running system that one can project the rate of progress of user needs or by using similar solutions adopted by institutes which manage similar types of information to anticipate the needs.

3. **Standard Components** — INPA has adopted commercial-off-the-shelf (COTS) hardware, used in its computer infrastructure, and some software (e.g., operation system, document preparation, database management, etc.). The COTS approach can bring two advantages:
 - (a) Regular product upgrades.
 - (b) Interoperability amongst applications and continuing opportunity to incorporate new functions in the system.

There are discussions going on about open source (free) software and how this strategy can deliver most of the needs. In the context of the Amazonian community these options are, at first, welcome, since they do not impose pressure on the constant lack of resources. However, there may be side-effects, yet not well publicised, regarding the advantages claimed by the open source approach. For instance, there is no guarantee the a system reaches a sufficient usable stage, there may be problems connected to intellectual property and it may be difficult to know or follow the current status of software.

4. **Documentation** — System documentation contributes to reduce the costs for system maintenance. Application design procedures and naming conventions should be defined, described and enforced to all application codes. It should be a practice to include complementary documentation in the code (low level detail); and additional documentation to describe the system components, how independent they are, and the strategy for reusability. We observed that most of the systems in use, that were developed 'in house', have poor documentation, in some cases none at all. This has caused concerns amongst curators and all agreed that it must change.

3.3.6 System Architecture Constraints

INPA, through PPG-7², has invested heavily in information technology, especially in computer infrastructure. A BIS should utilize the existing infrastructure and new investment should be used for upgrades and expansions to meet essential required functionalities.

²Pilot Program for the Protection of the Tropical Forests of Brazil PPG-7 (Programa Piloto para a Proteção das Florestas Tropicais do Brasil — PPG-7). This international programme involves Russia, Germany, the United States, France, the United Kingdom, Italy, Japan and Canada, all countries belonging to the G-7 along with the Brazilian Government. Part of the funds donated were used to set up the Rainforest Trust Fund — RTF, a multilateral fund with various donors, administered by The World Bank.

3.3. SYSTEM REQUIREMENTS

- **INPA network and the Internet** — INPA has three Campi (Aleixo I, Aleixo II and V8, interconnected via optical fiber (at 155 Mbps scalable to 622 Mbps) in the network backbone, the INPANetwork. The network adopts the Asynchronous Transfer Mode (ATM) technology and the sub local area networks are wired with category 5 unshielded twisted pair (UTP) cable to provide Ethernet data communication services. The networking uses Transmission Control Protocol / Internet Protocol (TCP/IP) and the dial-up connection to the INPANetwork is provided via a modem, with limited resources, which supports TCP/IP connection through Serial Line Internet Protocol (SLIP) and Point to Point Protocol (PPP). The demand commonly exceeds the number of lines available, and it troublesome to get a connection. The connection between the INPANetwork and the Internet is described in Chapter 4.
- **Collections Computer Facility database server** — The SCP computing resources are integrated with the technical group for informatics, known as GTI, resources which provide a variety of computing services to its users. Important resource capability includes: Itautec Info Servers, an Intel-based processor of 500MHz and up scalable to 4 processors and 13-80 Gb of disk space, and DB2, Oracle and MySQL DBMS. The database server is also furnished with DAT tape drive (24 Gb) for backing up the database. Apart from MySQL site licences for DBMSs are limited and the institute does not yet have an agenda regarding maintenance and training for either hardware or software. GTI tries to provide all the necessary administration procedures, that is, security, configuration management, and systems upgrades. There is now a database administrator available in charge of performing the necessary operations of database design and implementation.
- **Desktop environments and applications** — INPA has installed approximately 650 personal computers, of which 97% are Intel-based and 3% are Macintosh. Current operating systems, 65% Windows 9*, 25% Windows 2000 and XP, and 10% Linux and others. The Macintosh community is very small and concentrated in the field of molecular biology related to the Genome Project. The Intel-based users are those who develop curatorial and administrative activities as well as researchers and graduate students. Applications currently used on desktop computers include:
 1. Word Processing - MS-Word (Windows, Mac).
 2. Spreadsheet - MS-Excel (Windows, Mac)
 3. DBMS - MS-Access (Windows), FileMakerPro(Mac), dBase (DOS), Paradox (DOS, Windows), DB2 (Windows), Oracle (Windows) and the DBMS MySQL (Windows).
 4. Web Browser - Internet Explorer and Netscape.
 5. E-mail - Eudora (Windows, Mac) and MS-Outlook (Windows).
 6. Network communication - ftp and telnet.
 7. Illustration and Graphics - Powerpoint (Windows).
 8. GIS - ERDAS, ARC/View.
 9. Statistical Analysis and Graphing - Systat, Statistic, SAS, SPSS.

- **Unix environment** — INPA also has three Sun Ultra Sparc II workstations, with limited disk space, a 4 and 24 Gb backup tape drive and a CD-ROM drive. There is interest to increase the use of the Unix/Linux system across the institute.

3.4 Summary

The descriptions presented here aim to provide the ground information for describing data and information representation, software and hardware that can comprise a BIS solution for the biological collections at INPA and elsewhere. The analysis can also give the curatorial personnel a picture of a BIS and the skills that will be required to develop, implement and manage it.

In summary, a BIS should be built based on a multi-user, (object) relational database server technology to be the data repository. The database features should favor a variety of client applications with emphasis on flexible query and report tools, web server integration and GIS/mapping systems to enable collections data to be visualised together with spatially-referenced environmental data.

We aimed to make this requirements analysis as prospective as possible. We have noticed that during this work, research protocols and user requirements have been changing slightly, mainly due to the institutional strategic plan and re-direction of priorities. Today, a BIS is still an abstraction which INPA is focusing to bring into reality. We envisage that once the system is implemented, users and systems will have to go through a process of adjustment of requirements. This does not constitute a drawback in our analysis. We understand that this is part of a software engineering process and INPA should anticipate the extension of system maintenance.

3.4. SUMMARY

Chapter 4

Clustered Object Schema *

4.1 Introduction

In this chapter, the way in which biological collections can be represented schematically, aiming for database design and implementation is described. The schema is a result of a survey performed in scientific institutes in the Amazon region where information was structured to represent biological collection data and events that make use of the data. For biologists/curators when reading this chapter, we recommend to read it in conjunction with Chapters 3 and 6, since these chapters present details of the system, user requirements and schema implementation of a biological database. The components of such a schematic representation are comprised of clusters, object classes, relationships and attributes and a graphical notation that is presented in Section 4.2. We call this structure a CLOSi Schema. CLOSi has constructs related to the Semantic Data Model (SDM) (Hammer & McLeod, 1981) and Object-Protocol Model (Chen & Markowitz, 1996). We have used CLOSi and implemented a database prototype on top of MySQL, a relational database management system (Dubois & Widenius, 1999; Welling & Thomson, 2001). Details of this implementation are presented in Chapter 5. In Section 4.3, we describe the components of the schema by presenting the main clusters' graphical notations. Finally, in Section 4.6, we present a summary of the chapter with our concluding remarks. The BNF for CLOSi schemas is presented in Appendix A. The Appendix B presents a list of CLOSi controlled value classes and some examples of CLOSi instantiated values are presented in Appendix C.

*This chapter is based on Campos dos Santos, J. L., de By, R. A., Apers, P. M. G., and Magalhães, C. (2002). Clustered object schemas for INPA's biological collections data. In the Proceedings Volume I (Information System Development I) of the SCI 2002 (6th World Multiconference on Systemics, Cybernetics and Informatics), Orlando, Florida (USA), July 14-18, 2002, pp 38-44.

4.2 Schema to Represent Biological Data

4.2.1 The strategy

Justification

Several research institutions are setting up biological collection programs as part of their scientific strategic plan. The programs need computational solutions to help in the design and development of databases and integrated tools for multi-biological collections management that can be used across institutions in the Amazon region. The way one can achieve that is by understanding the physical complexity of biological collections. It also requires knowledge about the data and its objects, and the information that will eventually be stored in the database. The application of computer technology and related solutions in bio-science is welcome since it will promote advancements in both areas with positive outcome in data management and dissemination of information.

Our strategy towards the representation of collection data was centred essentially on surveying users' perspectives and needs. We involved collection users in the processes of analysing data and system requirements, since they play an important role, especially during the phase of data requirement analysis. For this survey, we concentrated on the scope of which research institutions could provide the essential information we needed, definition of a feasible schedule of field work missions and the content and the context of interviews and its evaluation.

The scope of the survey

The main objective of this survey was to investigate the real situation of the biological collections in the Amazon region, to understand the information flow within a collection and amongst collections, to describe the components, functions and events that manipulate collection information. Since no comprehensive compilation about biological collections had been gathered before, this survey became an important source of information that can be used by institutions and researchers. Additionally, the result of the survey would be a basis for a representation for database implementations. The geographical area, time and resources available were limiting factors of this survey. We were able to compromise those factors with the objectives and to ensure collecting the required information.

The selection of institutions

Since we were familiar with INPA's institutional structure, we started by identifying affiliated institutions that have the common practice of exchanging collection information and material within the Amazon region. We compiled a list of institutions to approach them and to present our research initiative. The SCP at INPA helped us to establish the contacts and identify the personnel in each collection of interest in the region. Through INPA's affiliated network we were able to select and reach an agreement for the survey with the following institutions: INPA, MPEG, IEP, EMBRAPA, Silvolab and MCT.

Field work

The field work started in November 1999 and lasted till August 2001, with a total of four trips of two months each, when we visited the collections of INPA, EMBRAPA, IEPA, MPEG, and Silvolab. The institutions provided us with logistics and personnel available during our visit and the activities were accomplished according to plans. Figure 1.1 illustrates the area covered during the field work missions and the locations where the selected institutions are based. More details regarding the fieldwork schedule can be found in Section 2.2.

Interviews and evaluation

We carried out investigations in the selected research institutes in two steps:

1. **Interviews** — We interviewed scientists who worked with biological collections, most of them being curators, and
2. **Evaluations** — We asked researchers from these institutes to validate the descriptions of object types and data flow in the collections.

Each participant in these processes was a specialist in a taxonomic group or a certain biological aspect of some taxonomic group. The interviews had an open format and researchers were asked the same general questions. This allowed us to develop a protocol to conduct the steps from field sampling to recording information.

The data collected consisted of two parts: a general one which holds the information that is normally collected in all studies (e.g., date, time, locality description), and a specific one that corresponds to the scientific interest of a study (e.g., the altitude of a locality or the moon-phase may be of interest in one study but not in another).

The process of interviewing scientists, who worked on different studies and in different institutes helped to differentiate between information that is common to all and that which is used by just a few scientists.

Classification of the information collected

An example of interviews carried out at the entomological collection of INPA is presented in Table 4.1. The details are follows: the Data column contains the attributes of the objects in the collection; Type determines the kind of value for the data (e.g., string, number, date, hour, binary and table – table is an association to an optional standard value), and the Entomological Fields, include information about taxonomic entities investigated.

We divided information into the groups Collections Information, Taxonomic and Identification Information, Collecting Event Information and Social and Ecological Information. The group Collection Information contains collection management information and attributes that describe the place and state of an object in the collection. The Access Number is assigned to objects of different collections if they were collected at the same event of collection. The Collection Number is the identifier of a specimen or a group in a collection. Fixation, Conservation and Preparation are descriptions which were required by nearly all scientists. Rack and Drawer describe

4.2. SCHEMA TO REPRESENT BIOLOGICAL DATA

Table 4.1: Entomological collection information; R indicates information recorded by the interviewed researchers (Aca = Acarina, Col = Coleoptera, Dip = Diptera, Eph = Ephemoptera, Hem = Hemiptera, Hym = Hymenoptera, Lep = Lepidoptera, Odo = Odonata, Ort = Ortoptera.)

	Data	Type	Entomological field								
			Aca	Col	Dip	Eph	Hem	Hym	Lep	Odo	Ort
Collection Information	Access Number	String									
	Collection Number	String									
	Fixation	Table	R	R	R	R		R	R	R	R
	Conservation	Table	R	R	R	R	R	R	R	R	R
	Preparation	Table	R	R	R	R		R	R	R	R
	Rack	Number						R	R	R	R
	Drawer	Number	R		R			R	R		
	Fitter	Table									
	Cataloguer (Digitiser)	Table									
	Loan	Binary									
	Availability	Binary									
Material Condition	String										
Taxonomic And Identification Information	Order	Table			R						
	Super-Family	Table									
	Family	Table	R	R	R		R	R	R	R	R
	Sub-Family	Table		R	R			R			R
	Genus	Table	R	R	R	R	R	R	R	R	R
	Sub-Genus	Table		R				R			
	Species	Table	R	R	R	R	R	R	R	R	R
	Sub-Species	Table						R			
	Status/Type	Table									
	Taxonomic History	Table									
	Identifier	String	R	R	R	R	R	R	R	R	R
Identification	Date	R	R	R	R	R	R	R	R	R	
Collecting Event Information	Field Number	String							R		
	Country	Table	R	R	R	R	R	R	R	R	R
	State	Table	R	R	R	R	R	R	R	R	R
	Community	Table	R	R	R	R	R	R	R	R	R
	Locality	Table	R	R	R	R	R	R	R	R	R
	Micro-Locality	Table	R	R	R	R	R	R	R	R	R
	Coordinates	Table		R	R	R	R	R	R	R	R
	Altitude	String		R	R						
	Date	Date	R	R	R	R	R	R	R	R	R
	Hour	Hour	R		R	R		R	R		R
	Clima/Conditions	String		R	R	R	R		R		R
	Collecting Method	Table	R	R	R	R	R	R	R	R	R
	Habitat	String	R	R	R	R	R	R	R	R	R
	Moon	String			R						
Collector	Table	R	R	R	R	R	R	R	R	R	
Social and Ecological Information	Sex	Binary	R		R		R	R	R	R	
	Parasite	Binary		R	R						
	Solitary	Binary						R			
	Social	Table		R				R			
	Mature/Immature	String		R	R	R		R		R	

the physical location of an object in the collection; which of them is used, depends on the kind of collection. The Fitter is the person who added an object into the collection. This object attribute and the attributes Cataloguer, Loan and Availability are important data for collection management.

The attribute Material Condition may help a curator to decide whether he should loan an object or not. The attributes not marked in this table are not of scientific interest but they are important for the management of the collection.

The group Taxonomic and Identification Information contains the identifiers of the biological classification as Order, Family, Genus, Superspecies, Species, Subspecies and nearly all of them were used. Status/Type describes the meaning of an object for the taxonomy, e.g., a holotype specimen is the one that has been used as the standard for the original description of a species. The Taxonomic History contains the information about the taxonomic names, synonyms, their authors and the date of the description.

The information recorded during a collecting event is included in the group Collecting Event Information. Field Number is an identification number for the event and is normally correlated with the field notes that a collector makes for the event. The remaining attributes define the collectors, the place and the time of an event. Generally, all entomologists use the same attributes. Exceptions are the altitude and moon phase that are just used in special studies.

Specific attributes of collection objects have been put into the group Social and Ecological Information. The attributes Sex and Mature/Immature are important in all research areas. It is even desirable that the latter attributes give a more precise impression than just the 'mature' or 'immature', it should contain the precise description of the life stage that the collection object is in. The attribute Parasite, Solitary and Social describe a social or ecological aspect of an animal and are not used by all scientist, nevertheless they must be considered in the model of the database.

The results of the interviews were grouped by functionality and clustered as object types, representing the items in collections. We named this representation CLOSi (Clustered Object Schema for INPA's) biological collections. CLOSi is a biological data representation that supports specifying database schemas in terms of objects and classes. Further details of CLOSi can be found in (Campos dos Santos & de By, 2000; Campos dos Santos *et al.*, 2002a). Simplified versions of the syntactic definition is presented later when we describe the elements of CLOSi. For all syntactic definitions developed for CLOSi, a `<text>` definition consists of a chain of strings with blank spaces between strings. A `<string>` definition consists of a chain of characters without any blank spaces between characters. The complete BNF Grammar for CLOSi can be found in Appendix A. Figure 4.1 presents the structure of interrelated groups of biological collection concepts (clusters) and Figure 4.5 presents the graphical notation we defined for CLOSi.

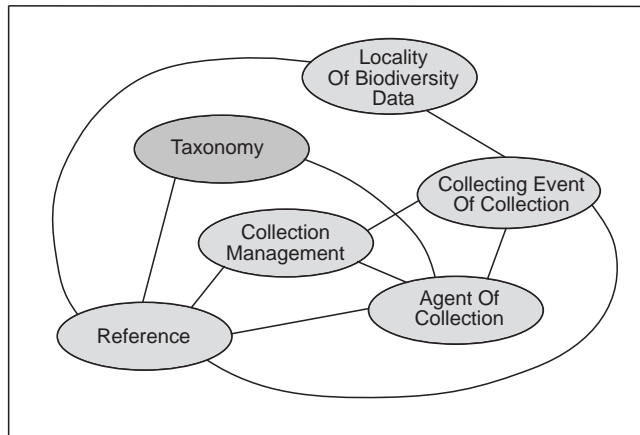


Figure 4.1: Clusters and relationships structure of CLOSi.

4.2.2 The components of the schema

Clusters

A cluster groups a set of inter-related object classes. The schema represents functional groups that are specified in terms of object classes and their relationships. A cluster can be viewed as a subsystem at the conceptual level. The clusters were organised as follows:

- **Collection Management** — Describes information about biodiversity collection data, particularly the aspects of information management of the collection components (e.g., object of collections, object situation, agents who deals with objects, the relationships with collection events, determination/classification aspects of object, etc.);
- **Collecting Event** — Describes information on collecting events, which includes: method(s) adopted during the event, site where the event took place, the collector's identification and the scientific area that the event covers,
- **Locality of Biodiversity Data** — Describes information about where the data was collected allowing details of named places, habitat, spatial domain coverage, cartographic reference and geo-reference description;
- **Taxonomy** — Describes information about the taxonomic classification, identification and the ecological relations of taxa;
- **Agent of Collection** — Describes persons and organisations dealing with the biological collection information;
- **Reference** — Describes publications related to biodiversity data within a collection.

Object Classes and Relationships

An object class describes a specific aspect of a collection; it is identified by a unique name. An object class can have a relationship to other classes of any cluster. An object is described by a class description and is associated with one or more attributes.

A class can be a specialised class and for that, the class relationship should be instantiated as *Is_a*. An object class must belong to a unique cluster and can be associated with an optional list of control value classes (set of ordered atomic values).

Is_a Object Classes

Specialisation is an abstraction method that allows the definition of object classes consisting of subsets of objects of another object class (the superclass). The inverse of specialisation is generalisation.

If the object class OC_i *Is_a* OC_j then all instances of OC_i are also instances of OC_j . Specialisation defines a transitive relationship between object classes. If OC_i is a direct specialisation of object class OC_j , and OC_j is a direct specialisation of object class OC_k , then OC_i is a transitive specialisation of OC_k . The specialisation OC forms a direct acyclic graph.

Each specialisation object class must satisfy generalisation's referential integrity constraints requiring each object in it to belong to all its superclasses. A specialisation object class, inherits all the attributes of its superclass. CLOSi does not allow multiple inheritance, i.e., direct multiple superclasses.

Attributes

An attribute is the basic unit of information on object class occurrence. It may be local to the object class or inherited by relationship. Attributes associated with object classes have constraints enforced by cardinality, which specifies the minimum and maximum number of values for attributes. In the absence of any specified cardinality constraint, it is by default [0,1] for single value attributes, and [0,] for set-of value or list-of valued attributes, where an unspecified maximum represents an unlimited number of values.

Also, referential integrity constraints are implied by a CLOSi schema and regarding attributes associated with controlled value classes; the value of such attributes must belong to their associated value classes.

Types of attributes

CLOSi supports four types of attributes: standard, composite, derivation and geo-attribute. They are detailed as follows:

1. **Standard** — This attribute describes the common property of an object class, that is, attribute name, cardinality, the attribute data type and the attribute description. The cardinality indicates how many values are minimally and maximally associated with an object occurrence. For example: A standard attribute called *CollectionCode* has cardinality [1,1] and the attribute type of

4.2. SCHEMA TO REPRESENT BIOLOGICAL DATA

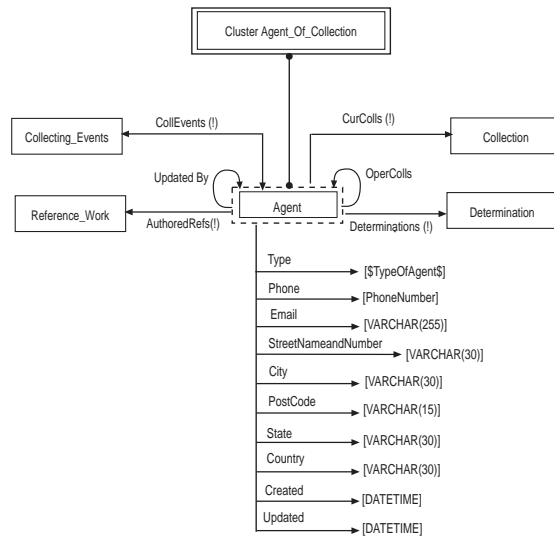


Figure 4.2: Object class Agent, its attributes and relationships.

VARCHAR(20).

An attribute type can also be identified by object class. For example: An attribute called OperColls (a staff personnel who operates the collection work activities) has a cardinality of [1,n] and the attribute type is the object class Agent. This implies that the information related to OperColls is described by all the attributes of the the object class Agent. Figure 4.2 presents the class Agent, its attributes and relationships. The complete CLOSi notation is describe below and presented in Figure 4.5.

2. **Composite** — A composite attribute is formed by the concatenation of one or more attribute members and an attribute description. For example: The composite attribute called CollectionID is described by a cardinality [1,1] and the attribute members [OrganisationAcronym + CollectionCode]. Figure 4.3 presents the class Biological_Collection with composite attribute Collection_ID.
3. **Derivation** — A derivation attribute represents a relationship between object classes and indicates that the attribute has an original object class (derivation_class_from) as type and can be derived as an attribute of another object class (derivation_class_to). For example: The attribute Objects represents the relationship between the classes Biological Collection and Collection.Object, meaning that a biological collection is comprised of a set of collection objects (See Figure 4.6).
4. **Geo attribute** — A geo_attribute describes geo-referential characteristics of an object. It is described by an attribute name, a cardinality and a coordinates

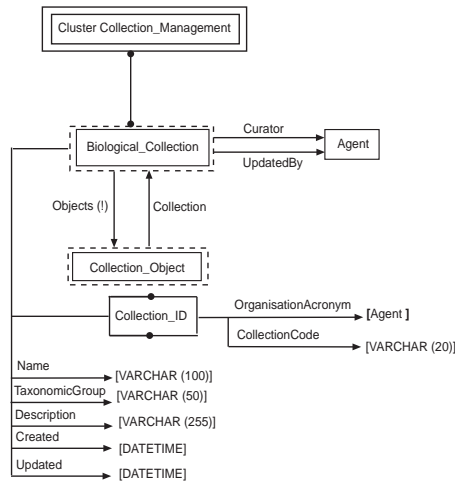


Figure 4.3: Object class Biological Collection, its relationships and attributes.

type or distance. Coordinates at latitude and longitude are each defined by four values: degree, minutes, seconds and hemisphere. The Coordinate type can be of geographic, rectangular or nodes type. The geographic type describes the latitude and longitude and is also described by four values: degree, minutes, seconds and the identification of the hemisphere (e.g., N, S, E or W). The rectangular type also describes the latitude and longitude, which are defined by two values: rectangular coordinates in meter units and the hemisphere.

Figure 4.4 presents the Cluster Locality_of_Biodiversity_Data with the class Geo_Reference_Object and the subclass Line with its relationships and attributes.

Controlled Value Classes

Each controlled value class defined in CLOSi has a unique class name, and can either be of string or numeric type. A controlled value class can be associated with an optional standard value, a class description, the name of a cluster and the class where the controlled value is described, and a list of declared-domain properties. Declared-domain properties are defined as tag-value pairs and will be used in application programs.

A string type controlled value class consists of a set of enumerated atomic values, which are strings. If a standard value is declared, then the standard value must be one of the enumerated atomic values.

4.2. SCHEMA TO REPRESENT BIOLOGICAL DATA

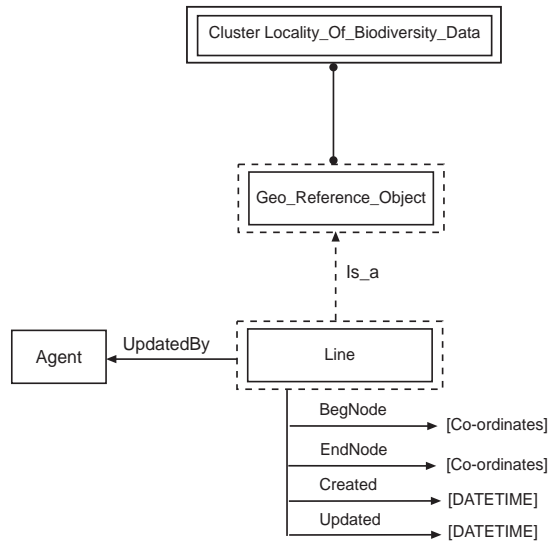


Figure 4.4: Object class Line its relationships and attributes.

For example, a string type controlled value class ProjectType is defined as follows:

```

CONTROLLED VALUE CLASS ProjectType
{"Overlap",
"Homologous",
"Single",
"Non Overlap",
"Unknown"}
STANDARD: "Unknown"
  
```

Each string value of a controlled value class is associated with a specific CODE_TYPE associated with a data type for this code. Data type can be a numeric type such as: INTEGER, SMALLINT, TINYINT, REAL, FLOAT, DECIMAL and NUMERIC, or a character string type such as: CHAR and VARCHAR. A code can consist of alphanumeric characters and special symbols (. : + - * / ! =). The symbol '?' is allowed only when the code starts with A-Z or a-z.

Each controlled value also has a value description indicating which cluster and class the controlled value class has been defined in, as well as a standard value and a description of the controlled value class. For example,

the controlled value class `ProjectType` can be defined as follows:

```
CONTROLLED VALUE CLASS ProjectType
{"Overlap", 1, "Overlap Project"},
("Homologous", 2, "Homologous Project"),
("Single", 3, "Single Project"),
("Non-Overlap", 4, "Non-Overlap Project"),
("Unknown", 0, "Unknown Project Type")}
CODE_TYPE: SMALLINT
DESCRIBED_IN: Cluster Collection Management in
Class Collection Object.
STANDARD: "Unknown"
DESCRIPTION: Control vocabulary for the ProjectType class.
```

A numeric type control value class consists of a set of ranges, where each range is either a number or an interval defined by a lower and an upper limit. For example, a numeric type controlled value class `INTERVALS` can be defined as a {1-10, 100, 200-300}. If a standard value is specified, then the standard value must be within ranges defined for this controlled value class. Numeric type controlled values or ranges can have value descriptions but cannot be associated with a `CODE_TYPE`. This example was partially discussed by Chen and Markowitz (1996). All the Controlled Value Classes defined so far for CLOSi are presented in Appendix B.

Graphical Notation

A cluster is graphically represented by a double solid line rectangle and its name placed in the middle of the internal rectangle.

There are two ways to graphically represent classes: one for classes that belong to a current cluster (part of a current cluster) and another for those belonging to an external cluster. The first is graphically represented by a double rectangle with a dashed line in the external part. The latter, is represented by a single rectangle with a solid line. In both cases, the name of the classes are placed in the middle of the internal rectangle.

Relationships may exist between a cluster and its classes (cluster association), between classes to specify derived relations, represented by a solid arrow, and between classes to define an *Is_a* relation, represented by a dashed arrow.

Classes have attributes that are graphically represented by a solid arrow. Class is comprised of attribute names and its data type, including controlled value class and coordinates type.

The composite attribute is represented by a rectangle marked by two

bullets attached to the attribute members. The attribute members are described as normal attributes.

Figure 4.5 presents the notation to express a CLOSi schema. The notation represents clusters, object classes, relationships and attributes.

4.3 A CLOSi Schema for Biological Collection

We argue that since the conceptual design of CLOSi was originated from multiple sources (collections and institutes policy) it is possible to accommodate multiple set of biological information requirements from institutes. The schema covers most of the general aspects of a biological collection. The involvement of a large community that deals with collection data during the process of survey had ensured the usefulness in multiple collections site.

CLOSi has initially been tested with a small part of the entomological collection at INPA. After improvements, we are now using it on the Crustaceous and Fish collections. In the following, we illustrate the cluster schemas using data from the Crustaceous collection. Figure 4.6 to Figure 4.11 present the object classes belonging to clusters as well as the relationships to other classes (internal or external). Examples of instance values of object classes and their attributes adopt information from INPA's Vertebrate collection and are illustrated in the Appendix C. To simplify the examples, we present only some part of the graphical representation of each object class and do not detail their attributes at this chapter. All object classes have three common attributes, that is, Created (date and time of creation of the object), Update (date and time of the last update of the object) and UpdatedBy (who performs the update).

4.3.1 Cluster Collection Management

This cluster describes information about biological collection data, particularly collection management activities. The cluster is comprised by the following object classes:

- **Biological_Collection** — Represents the assemblage of biological specimens that usually correspond to a series of catalogue numbers.
- **Collection_Object** — Describes biological items that are part of a collection.
- **Lot** — A subclass of the **Collection_Object** class and consists of one or more specimens that have been collected during the same collecting event.
- **Specimen** — A subclass of the class **Collection_Object**.

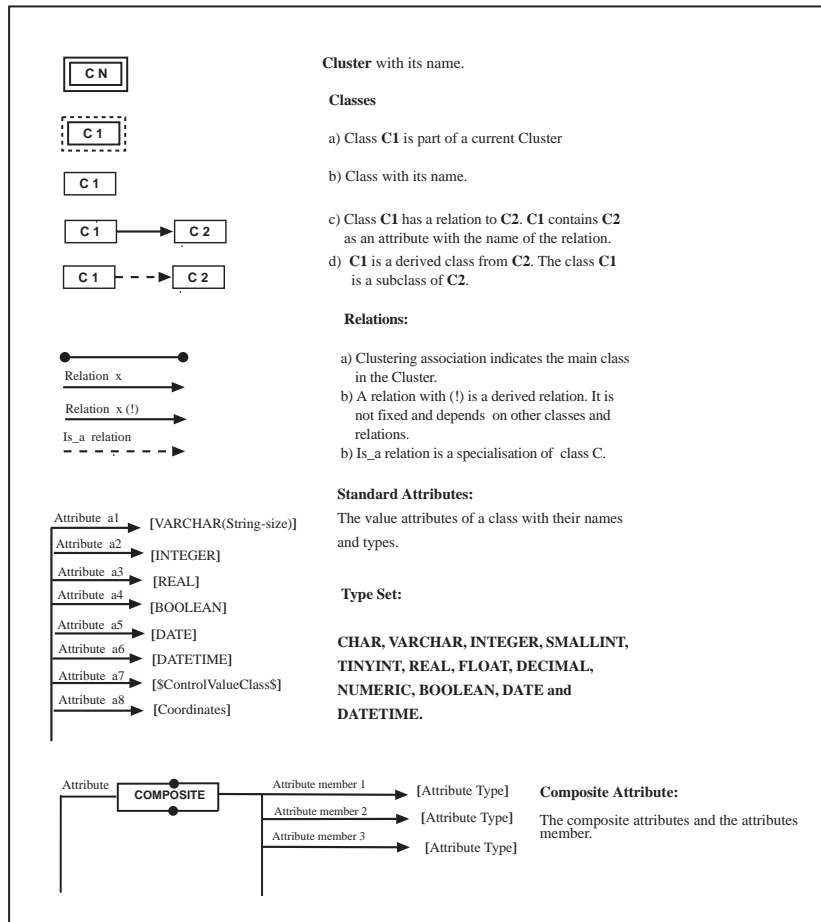


Figure 4.5: CLOSi Notation.

4.3. A CLOSI SCHEMA FOR BIOLOGICAL COLLECTION

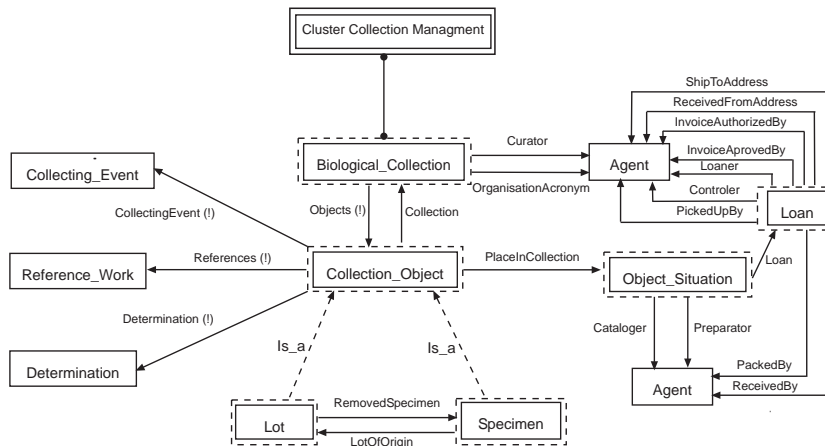


Figure 4.6: Cluster Collection Management, its classes and relationships.

- **Object.Situation** — Describes the state of an object, where it is located in the collection and whether it is loaned, as gift, or in exchange.
- **Loan** — Defines when, by whom and in which circumstance an object is loaned.

Some objects of this cluster also have relationships to external objects belonging to other clusters. For instance, the class `Collection_Object` has a relationship to the object class `Reference_Work`, indicating the published material in which the collection object is mentioned. The relationship `CollectingEvent` describes the event in which the object was collected and `Determination` describes the process of identification at the species level and provides a taxon name. The relationships between `Lot` and `Specimen` describe the removal of a specimen from a `Lot`. Figure 4.6 presents details of the cluster `Collection_Management` and its relationships.

4.3.2 Cluster Collecting Event of Collection

The `Collecting_Event` object class holds the information about collectors, the place and time of an event, the collector's identification and the subject that the event covers. This object class has relationships with external object classes. As can be observed in Figure 4.7, the relationship `Reference` describes the reference work in which the collecting event is cited. The `CollectObject` refers to the collection objects that have been collected during the event. The `PlaceOfEvent` identifies the locality where the event took place.

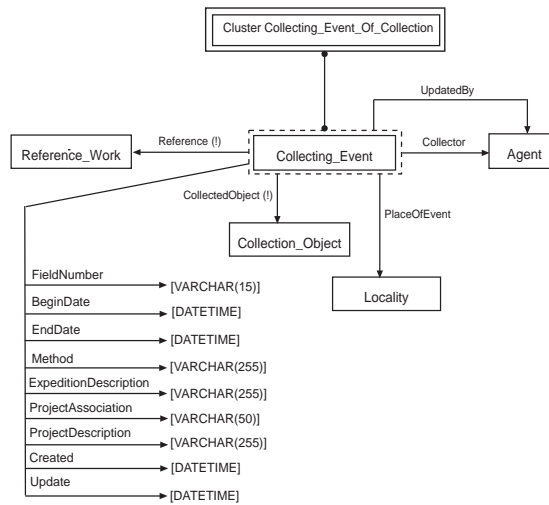


Figure 4.7: Cluster Collecting Event of Collection with its class, attributes and relationships.

The relationships Collector and UpdateBy indicate the list of collectors that participated in the collecting event and who updated information respectively.

4.3.3 Cluster Locality of Biodiversity Data

This cluster describes information about the Location where the data was collected. It defines classes Locality, Named_Place, Habitat, Spatial_Domain_Coverage, Cartographic_Reference, Geo_Reference_Object and the subclasses Chain, Line and Point (see Figure 4.8).

4.3.4 Cluster Taxonomy

The cluster describes information about the taxonomic classification, identification and ecological relations between the taxa. Figure 4.9 presents the cluster Taxonomy, which is formed by the classes Taxon_Name, Taxon_Relation, Classification and Determination (Identification). Classes within the cluster also have relationships to external classes; for instance, the class Taxon_Name has relationships with the external classes Reference_Work (TaxonOrigRef and Reference) and similarly class Agent with TaxAuthors, Repository and UpdatedBy. The class Determination has relationships to

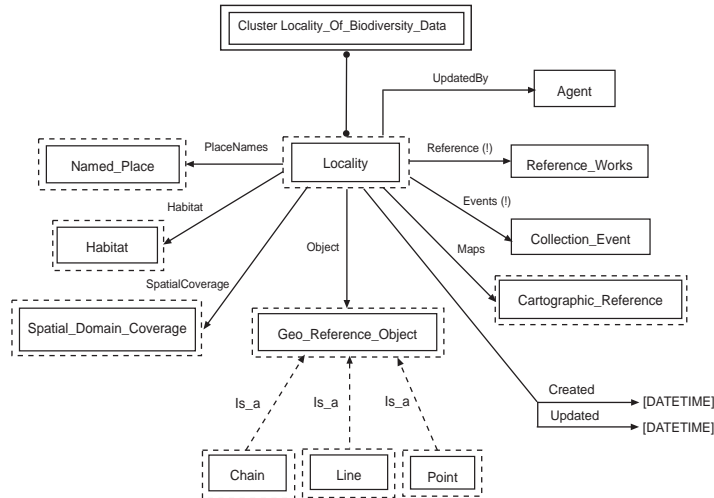


Figure 4.8: Object class Locality, its relationships and attributes.

the external classes Collection_Object CollObject and Agent (Determiner) and UpdatedBy.

4.3.5 Cluster Agent of Collection

The cluster describes persons and organisations, which deal with the biological collection information. The classes Agent (as superclass), Person and Organisation (as specialisations) form the cluster Agent_of_Collection. Figure 4.10 presents the cluster Agent with its object classes and relationships. The object class Agent interacts with all activities during a collecting event, collection management, in classification and determination of taxa, as well as reference work and all information that is updated in the system.

4.3.6 Cluster Reference

This cluster comprises the following object classes: Reference_Work, Book, Book Section, Technical_Report, Thesis, Article, Web_Publication and In_Proceedings. External object classes have relationships with Reference_Work via the relationship Reference and AuthoredRefs to object class Agent. Figure 4.11 presents the graphical representation of the cluster Reference and its relationships.

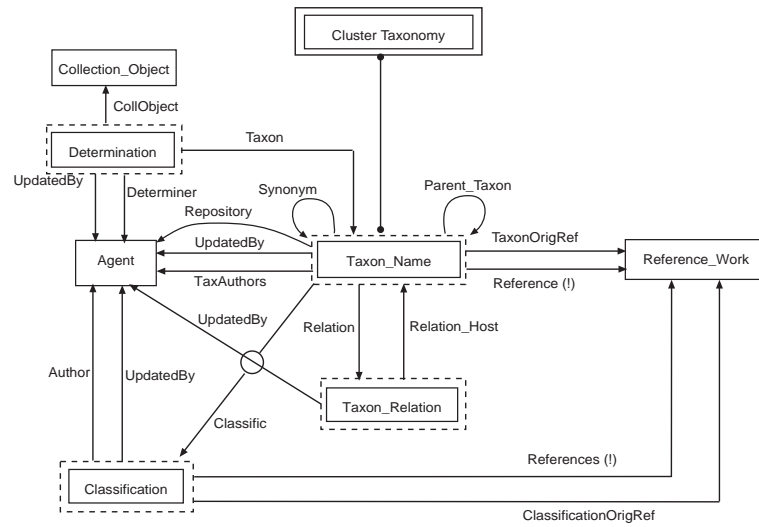


Figure 4.9: Cluster Taxonomy its classes and relationships.

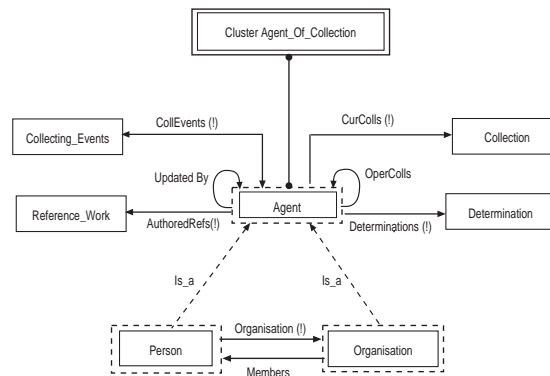


Figure 4.10: Cluster Agent_of_Collection, its classes and relationships.

4.4. REFLECTION ON THE DESIGN WORK

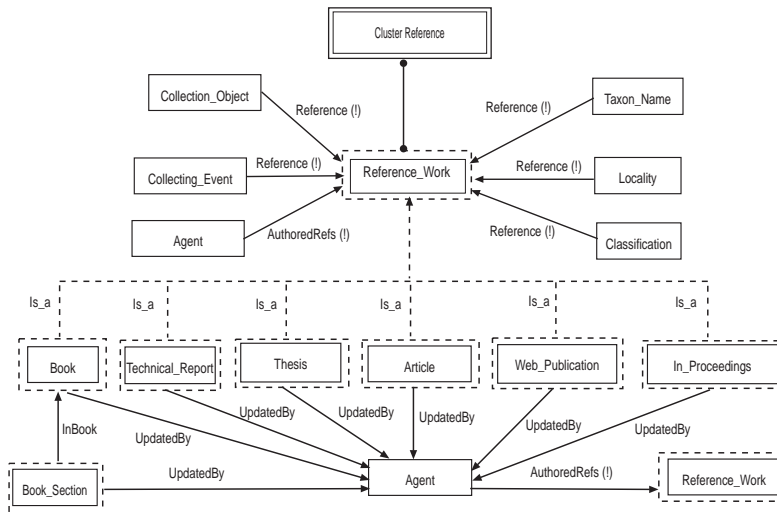


Figure 4.11: Cluster Reference_Work, its classes and relationships.

The cluster describes publications related to biodiversity data within the collection or literature (including non-conventional) that can be associated to a collection object. This would help to provide functionality where researchers can continuously contribute for the enrichment of collections. Today, the majority of collections do not keep track of the reference work that originated from the collection objects and the collection is treated as a mere repository of biological material. Elements of this cluster can be the vehicle for integration with library databases.

4.4 Reflection on the Design Work

The identification of objects, attributes, processes and events, related to biological data management was performed in a bottom-up approach of data analyses (from detail to a general level). The elements utilised were provided by the survey conducted (see Chapters 2 and 3). Relationships amongst these elements were also identified. From the complete structure produced, it was possible to define functional groups creating the basis for the clusters (see Figure 4.1). A graphical notation was produced to represent the structure using simple elements (rectangles, arrows), lines styles and labels (see Figure 4.5).

Since CLOSi tries to map a general view of biological collections, par-

ticularities of specific collections cannot be fully represented. For that, we envisage the extension of CLOSi with extensible micro-schemas associated.

We have tested CLOSi schemas on two of the invertebrate collections (entomological and crustacean). The personnel involved in the management of these collections reported that the documentation available for CLOSi was satisfactory and relatively easy to understand. The notation applied utilises an intuitive method for representing CLOSi concepts. Regarding to mapping to database implementation, designers expressed that mapping the schema to an ER diagram was performed without semantic loss. The schema contains all the necessary information to form a conceptual and physical design of biological collection databases. This evidently will result in speeding up the process of database implementation.

4.5 Strategy to Compare CLOSi Effectiveness

Given the focus of CLOSi as a conceptual level for a biological database implementation, a question that remains to be answered at this point is “how effective is the CLOSi schema compared with some conceptual options already implemented, which aim at the management of biological collections?”

To answer such a question, we must use the most prominent designs available elsewhere and compare them with CLOSi features. The comparison requires detailed information about the conceptual model behind each implementation.

Description of criteria to select system for comparison

We found out that the majority of applications designed for multi-disciplinary collections are not free software and consequently information available, particularly about the model adopted is insufficient for a comparison study. The way one can succeed in comparing models at conceptual levels is by a close scrutiny of the models within inter-institutional relationships framework and common goal initiatives towards integration or extension of models concepts. A partnership with targeted organisations would grant us access to the concepts implemented in those solutions as well as details about the information technology used. We have limited temporarily our attempt to compare CLOSi with other schema/models of similar nature. Instead, we tried to identify potential systems that match our criteria for a comparative study.

The systems selection was based on our own criteria, which include:

- The system must implement a model or a conceptualisation of a biological collection.

4.5. STRATEGY TO COMPARE CLOSI EFFECTIVENESS

- The implemented system must aim at specimen management.
- The system must be functional and in use by at least one collection.
- The system must be available for download (for free or as COTS.)
- The system must provide customisation functions.
- Technical support available, preferably online.

We present in the following a list, with a brief description, of systems designed for multiple collection types. The implemented systems are well-known, however, information about their conceptual models is scanty.

1. **BASIS** — A stand-alone relational database system for PCs written in FoxPro for Windows. BASIS was designed as a general database and catalogue program to help the compilation, integration, and retrieval of systematic, biological, and bibliographic information for arthropods at and below the family level on a world or regional basis.
Online information can be found at <http://www.sel.barc.usda.gov/>
2. **BioLink** — Designed to assist users working with taxon and specimen-based information. The main users are: taxonomists, ecologists, collection managers and biogeographers. The system is suitable for researchers, large museums or collections, or teams of global collaborators.
Online information can be found at <http://www.ento.csiro.au/biolink/>
3. **BIÓTICA** — Designed to manage curatorial, nomenclatural, geographical, bibliographical and ecological data. Aims to assist in the capture and updating of the data. Biótica was developed in a modular form, both in the structure of the database and in its system (programs), taking into account the needs of the biological taxonomists, curators, biogeographers, ecologists, ethnobiologists, etc.).
Online information can be found at <http://www.conabio.gob.mx/>
4. **BIOTA** — Designed to manage specimen-based biodiversity and collection data by providing an easy-to-use graphical interface to a relational database structure. It also provides tools for data input, update, maintenance, analysis, and reporting. Biota is used by ecologists, conservation biologists, reserve managers, biogeographers, taxonomists, systematists, and collection managers.
Online information can be found at <http://viceroy.eeb.uconn.edu/>
5. **KE EMu** — Provides collections management facilities for museums, herbaria and botanic gardens. KE EMu can manage very large collections and integrates a multi-disciplinary catalogue with interpretative information and multimedia resources.
Online information can be found at <http://www.kesoftware.com/emu/>

6. **MUSE** — Designed explicitly to manage natural history collection, and is based upon the experience of curators and collection managers. Built-in taxonomic dictionaries are also available for several disciplines. The system can manage loans, provide labels and reports. The database engine utilises Btrieve from Novell and is implemented in a relational structure.

Online information can be found at <http://www.biodiversity.uno.edu/>

7. **SAMPADA** — This is a natural science collections system. It aims to increase the level of utilisation of biodiversity data associated with museum specimens. It supports museum and herbaria for the capture, management, integration, dissemination and analysis of specimen data for research, education, resource management and policy making. SAMPADA is a platform independent application for the storage and retrieval of data from biological collections. This includes information about museum specimens, field information and notes, images, slides, tissues, sounds, video as well as associated information. In addition, it manages information about collection transactions, such as loans, accessions, deaccessions, exchanges, etc. SAMPADA includes an embedded reporting tool.

Online information can be found at <http://www.ncbi.org.in/sampada/>

8. **SPECIFY** — Designed to support data entry, queries and reports for single or multiple-taxonomic discipline collections. The application is based on a multi-disciplinary information model derived through a series of Natural Science Collections Alliance (NSCA) workshops in the 1990' s. The Specify database accommodates the many large and small differences among taxonomic disciplines in their approaches to collection information. Such historic conceptual differences, such as specimens vs. lot-based collecting paradigms, the recognition or absence of a “collecting event” as a field concept (e.g., ichthyology has it, botany does not), or discipline-specific conventions for loan and exchange and gift management, have a profound impact on the automation of specimen cataloging, curation and collection data analysis within the respective disciplines. Specify supports the union of data from various taxonomic disciplines within a single database.

Online information can be found at <http://usobi.org/specify/>

9. **TAXIS** — This is a DBMS designed for biologists. The main purpose of this software is to provide an interactive and user-friendly tool to facilitate processing of taxonomic information: keeping records of biological collections, studying collected specimens by recording characters, photographs, drawings, etc. TAXIS automatically serve as an interactive identification system that makes use of recorded characters, taxa and

4.6. CONCLUDING REMARKS

images. The structure of TAXIS database allows to apply it to any group of organisms.

Online information can be found at <http://bio-tools.tcn.ru/>

Outcome

The comparative study would provide us with information regarding the rank of CLOSi features when compared with similar mechanisms that describe biological data (in the selected systems). The study will also identify the deficiencies of biological data models/schemas reflected at the application level. Further direction in CLOSi development can be planned in order to complete it with missing features present in other systems. The extension of CLOSi would be considered on the general level on the contrary to the extension via micro-schemas.

We expect that the outcome of a comparative study would help potential and actual BIS system users to understand the relationships that exist amongst models/schemas concept and implementation functionality. In the longer term, it would provide the basis for interoperability amongst application data files. We are particularly interested in the way other models represent complex data types, recursive relationships (specially related to taxonomic classification and place-names), geometric descriptions, and scientific production originated from the collection items. The system also requires an internal evaluation (within an institutional context) that should be carried out after a complete implementation of a CLOSi-based database.

4.6 Concluding Remarks

We have described the initiative that led to the development of the CLOSi schema, the first concrete output towards biological data management for the Brazilian Amazon. CLOSi was designed based on the structure of the biological collections at INPA as well as from partner institutes in the Amazon region. CLOSi-based implementations have already provided evidence that CLOSi can fulfill the overall needs required by its scientific user community, regarding the abstract representation of biological collections. We have tried to provide a mechanism to describe a general view of biological collections. We are aware that peculiarities present in a given collection cannot always be represented fully. This does not represent a drawback, on the contrary, this would allow the global schema to be extended in such a way that specific details of a collection could still be treated at micro-schema level. To achieve that, a language to describe those particularities and to associate them with a specific collection that can reflect on them in the global

schema representation can be an option. This would provide an enhancement that each schema demands and, at the same time, would provide ways for integration with other systems and data sets.

Since the beginning, the participation of users in the process of identification of the requirements has been fundamental. During the process, we were able to define the extent of interactions that occur with biological data, with the objects in collections, as well as with its characterisation and entire process of management.

With CLOSi available, institutes can benefit a great deal. The major problems described earlier, such as spontaneous development, can be reduced considerably and resources can be diverted to the utilisation of schema representation for database design and implementation. An implemented database system will pave the way for much faster data digitalisation. With these two in place, data exploration and information dissemination can be facilitated through additional functionality or complementary application implementation via database systems in a web environment.

4.6. CONCLUDING REMARKS

Chapter 5

An XML-based Solution for Bio-Metadata *

5.1 Introduction

Biologists at universities, museums, and governmental institutes collect biodiversity data through field surveys. They usually cover a small geographic area over a short period of time (Stockwell, 2000).

Problems concerning data management, analysis, and consequently information dissemination include: data dispersion (data are scattered around institutes without a catalogue), no application of metadata standards (none or poor data set description), limitations on data exchange (data are exchanged amongst researchers at close range of data producers only, with no easy access to the data itself) and lack of data catalogue (the non-existence of catalogue trigger questions such as “Does the data I need exist? Where can I find it? How do I get it?”).

Despite the success of network communication in connecting scientific institutes globally, only a small number of any institute’s environmental information is available on the web. The use of the web to provide scientific information is progressing in phases, from a mere presence, to information and database access, to nodes on exchange, and to information portals. A portal provides access to information on the web and offers a broad array of resources and services, such as e-mail, forums, search engines, and on-

*This chapter is based on Campos dos Santos, J. L. and de By, R. A.(2002). XML-based Metadata Management for Biological Data. In W. Pillmann and K. Tochtermann (eds.): Environmental Communication in the Information Society; EnviroInfo Vienna 2002 (16th International Conference: Informatics for Environmental Protection; Part 1.; September, Vienna, 2002; pp 408-415).

line business (Niemann, 2000). However, the general scientific audience requires a more specific set of resources that can lead to a coherent view of information gathered from disparate sources (White, 1999). By adopting the Enterprise Information Portal (EIP) approach, which focuses on a specific domain, and enables facilities to deal with collaborative and decision processes in organisations, the problem of disparate scientific information can be alleviated (Goldfarb & Prescod, 2002).

Metadata descriptions depend on a formal description of data, which can be of semi-structured form. The best option is the use of the Extended Markup Language (XML), since this language was designed to deliver structured content over the web (Harold & Means, 2001). The strength of XML technology is that it can deal with semi-structured data, which are not suited to be managed by (object) relational database technology (Campos dos Santos *et al.*, 2002b). The development of XML-based metadata solutions that makes biological metadata accessible is urgently needed.

This chapter presents a solution for metadata management, which was implemented for INPA. The solution applies XML technology to manage and distribute biological metadata data for a large-scale audience. It is implemented in the web environment, where users can access metadata and data by querying an XML repository and then searching data for download. We focus on scientists as our prime users who are able to describe metadata. The FGDC metadata standard, which includes the Biological Data Profile, was adopted to be INPA's metadata solution.

Section 5.2 presents important metadata issues covering the scope of the theme, context, and importance, as well as the purposes of clearing-houses, standards and profiles within the FGDC initiative. We present in Section 5.3 the method adopted for mapping the FGDC metadata standard, including the biological profile, to an XML schema representation that was automatically generated from an XML template. Section 5.5 details the implementation of a three tier architecture that provides functionality for client/server components. At the client side, the BioME portal hosts components such as XML file, and schema, an XML Editor and full documentation, to be deployed on request by users at Online Local Nodes (OLNs). Users of an OLN can use the XML Editor to upload the template and describe scientific data sets. Each metadata description corresponds to a data set that can be submitted for storage. The metadata is kept in an XML repository (the XYZFind Server System) and can be accessed via XYZ its Query language (XYZFind Corporation, 2001). The repository maintains a single data representation of all the XML metadata it receives. The metadata can be retrieved by users and can be updated or removed from the repository by the biodiversity metadata editor (BioME) portal manager. Once the repository is indexed, search and query operations are available for global access.

In Section 5.7, we present our concluding remarks indicating some future work. An example of compiled metadata from a crustacean collection is presented in Appendix D.

5.2 Metadata in Brief

5.2.1 Scope

Most scientists have a feeling for data files, however, to a large number of environmental science professionals, the concept of metadata is not so clear. Metadata is data about data. It describes the attributes and contents of a document, work or dataset, and can mitigate data users from having to have complete knowledge of a dataset's existence and attribute characteristics. Standard bibliographic information, summaries, indexing terms, and abstracts are all surrogates for the original material, hence they are metadata. In the last 15 years, the word metadata has been intensively used and has applied to electronic resources, such as datasets, textual information, web pages, graphics, etc. (Milstead & Feldman, 1999).

The scope of questions about metadata covers:

- **Who** collected and **who** distributed the data?
- **What** is the subject, a dataset?
- **When** was the data collected?
- **Why** was the data collected (what is the purpose?)
- **How** was the data collected? **How** should it be used? **How** much does it cost?

Answering such questions will allow users to evaluate whether or not the data is important for them. Figure 5.1 presents in short the understanding of data, metadata and documentation.

5.2.2 Content

Metadata can be categorised as syntactic or semantic. Users need to understand intrinsic factors associated with the digital organisation and electronic structure of data. Such questions related to data are known as syntactic metadata. Semantic metadata are the instructions for users to understand the contents of the data; these must be available for find and use functions (Olsen, 2000). Cornillon (2000) refers to these factors as translational semantic metadata.

According to the FGDC there are seven main sections that describe different aspects of data (see Figure 5.2). They are:

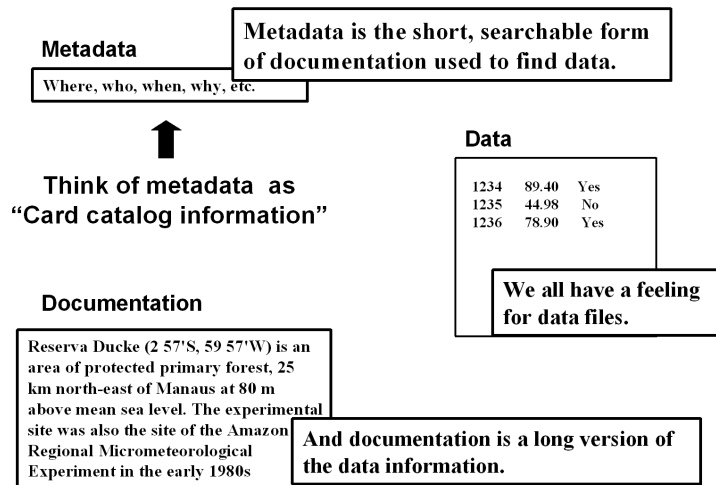


Figure 5.1: A simple understanding about metadata.

1. **Identification** — What is the name of the data set? Who developed the data set? What geographic area does it cover? What topics of information does it include? How up-to-date are the data? Are there any restrictions on accessing or using the data?
2. **Data Quality** — How good are the data? Is information available that indicates the usefulness of the data for a specified purpose? What is the positional and attribute accuracy? Are the data complete? Was the consistency of the data verified? What data were used to create the data set, and what processes?
3. **Spatial Data Organisation** — What spatial data model was used to encode the spatial data? How many spatial objects are there? Are methods other than coordinates, such as street addresses or relative addressing schemas, used to encode locations?
4. **Spatial Reference** — Are coordinate locations encoded using latitude and longitude? Is a map projection or grid system used? What parameters should be used to convert the data to another coordinate system?
5. **Entity and Attribute Information** — What geographic information is available (e.g., roads, rivers, elevation, temperature, etc)?
6. **Distribution** — Who can provide the data? What formats are available? What media are available? Is the data set available online? What is the price to obtain the data?

7. Metadata Reference — When were the metadata compiled? By whom?

Data providers must understand that their data sets will only be of any use if information about the data can be delivered. The non-existence of metadata can make data sets unusable and impossible to be shared. With metadata, we aim to organize and maintain data by alleviating information entropy.¹

5.2.3 Why Metadata?

Metadata is important, but also complex, that is, initially difficult to read, and initially difficult and time-consuming to produce. Metadata can help decision makers, researchers, and managers to find and use data, but they also benefit the primary creator of the data by maintaining the value of the data and assuring their continued use over a span of years. Metadata can also contribute with information to data catalogs and clearinghouses, and provide essential information in case of data sharing.

The importance of metadata is based on the following:

- **Protection of investment** — Data, mainly by effects of staff shift and individual memory loss, can only create means for data reusability and update when documentation of data source and quality control are provided.
- **Allows data understanding** — Data understanding is directly associated with data consistency. This is achieved by using terminology, focusing on key attributes of data. Consistency helps data brokers to establish if data is ready for use, and helps data transfer and understanding by new data providers and brokers. The providers are those users or agents who are responsible for data sets. The brokers are those who use the raw data to extract knowledge from it and describe the metadata.
- **Makes discovery possible** — Metadata can provide information to data catalogs and clearinghouses and allows advanced searching for multiple purposes.
- **Prevents liability** — Metadata provides protection for misuse of data.
- **Prevents mistakes on data processing** — Data updates must be avoided if transformation methods are not recorded. Unrecorded modifications can propagate errors.
- **Careful and responsible data management** — Data providers that make available high quality metadata usually provide data with similar quality.

¹Phenomena that occur during the normal degradation of information content associated with data and metadata over time (Michener *et al.*, 1997).

- **Reduction of data provider dependence** — Data brokers do not need constant assistance from providers to answer questions about the data.
- **Cuts overall costs** — The use of metadata allows development of tools that can play an important part to ease overall load and cost of data and maintenance.

5.2.4 Standards

Metadata and standardisation can be considered time-dependent phenomena, which change when new data and information is emerging from new studies. Therefore, the development is continuous, it requires flexibility and its formalisation are only necessary to make data accessible to an audience that is as wide as possible.

A standard specifies information that helps users to determine what data exist, how fit the data is for their application and purpose and, what is the mechanism to access these data. This can be profitable by saving time and resources, ensuring quality and completeness.

Informally, we create metadata all the time, when we write notes or documents. Also, we all need and use data from different sources. The problem starts when we try to use other people's data, which may not be easy to interpret, or from informal notes or large reports without any standard applied. Standards provide a common set of terms, aiming to reduce discrepancies and misleading descriptions. That means, for every metadata record the terms remain the same. Also, it provides fast and easy location of attributes, since a computer can be programmed to locate interesting data sets. In general standards are federally mandated. For example, the Executive Order No 12906, enforces that the documentation of all federal institutes receiving funding from the American government, must adopt the FGDC CSDGM. Obviously, to understand a metadata standard requires an effort to consolidate skills. However, it is an advantageous investment in the long run.

There are several standards, subsets, supersets of standards available online or documented that to select one constitutes concerns to users. Some of the most mentioned metadata standards include: AGLS, EDNA, ISO, FGDC, JMP, SDDS, and DCES. The cause for the proliferation relies on the application of metadata in different domains, ranging from a simple record to management records. Additionally to that, metadata are used to better describe data sets with essential information on the data life cycle, specification, structures and content.

We have looked into the three commonly used standards: the FGDC CSDGM, International Organisation for Standardisation (ISO), and Dublin

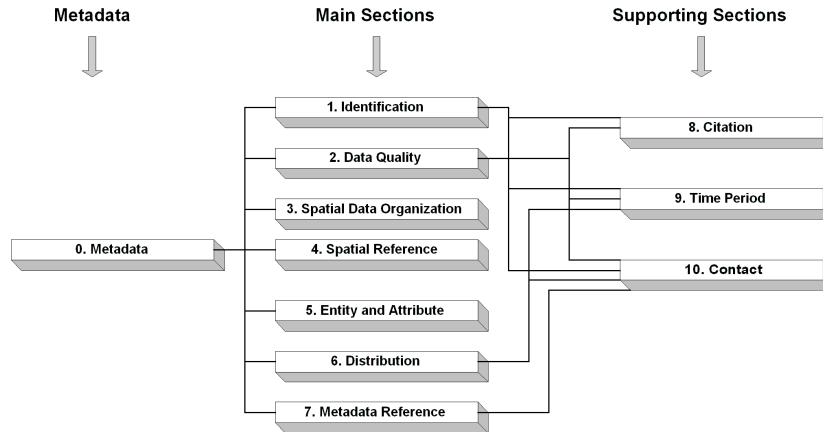


Figure 5.2: Sections of the CDSGM standard (Excerpted from (Federal Geographic Data Committee, 2001)).

Core Element Set (DCES). In the following, we give additional information about these:

- **The FGDC CDSGM** — Established in 1994, has been the most used standard for recording metadata for geospatial data in the USA. Its current status permits recording of spatial and non-spatial data. It presents a generic and flexible form for entering different types of data. For certain types of data, the standard allows the description of profiles as extension. Investments regarding profile production and approval have been made via working groups covering topics such as cadastral data, wetlands, soils, biological data, etc (Gilliland-Swetland *et al.*, 2000). The standard is composed of section definitions, a list of elements, definition of types and values and information on mandatory and repeatable data. The structure of the CDSGM standard, as presented in the Figure 5.2, describes identification, data quality, spatial data organisation, entity and attributes, metadata distribution and references information. The supporting sections provide common method to define citation, time period and contact. The elements utilised in the standard include: compound, data (values and forms for special values) (Federal Geographic Data Committee, 2001).
- **ISO** — Established in 1995—1996, the ISO Committee 211 has been developing the International Metadata Standard. It relies on the participation of the FGDC team, which has made an important contribution to the standard. The member contributors of ISO try to harmonize,

as much as possible, the ISO standard with others already developed. Data providers and data brokers can continue producing metadata standards without having to wait to shift to a completely different standard description in the event of a formalised ISO (Gilliland-Swetland *et al.*, 2000).

- **DCES** — Established in 1995—1996, Dublin Core was developed as a generic metadata standard for use by libraries, archives, governments and for other means of publication of online information. This standard was deliberately limited to fifteen attributes. Because of that, those who are attempting to implement it have raised a number of issues concerning the rules for the context of the fifteen attributes and the rules for structuring and expressing the attributes (Dempsey & Weibel, 1996). In that respect, Dublin Core has been used on a large scale due to its level of simplicity. However, this standard is not satisfactory for our purpose since it is subject to substantial changes, and cannot describe taxonomy classification completely.

5.2.5 The Biological Data Profile

After its approval, the FGDC standard underwent significant enhancements by allowing the metadata producer to profile the base standard by defining a subset to the metadata entities and/or elements to be used by a specific discipline or organisation.

The metadata extended elements include taxonomy, text file structure (default extension 'ASC') and geologic age information as presented in Figure 5.3. The extension were incorporated in the following sections:

1. Section 1 — Identification Information to include Taxonomy profile
2. Section 6 — Distribution Information to include text file structure. The structure provides information about the content and format of an ASC data file.
3. Section 9 — Time Period Information to include Geologic Age.

Figure 5.3 presents the complete extension of the standard indicating the attributes that are mandatory or optional.

The FGDC provided essential elements to research groups at INPA, which, after evaluation of the standard by comparison with ISO and DCES, adopted it for describing biological data and to implement it across its biological collections. The evaluation focused on the properties defined for the profile that can well describe data sets of species classification and determination information (taxonomic information), suitable for biological collections. Regarding biological data dissemination and metadata preparation, during a

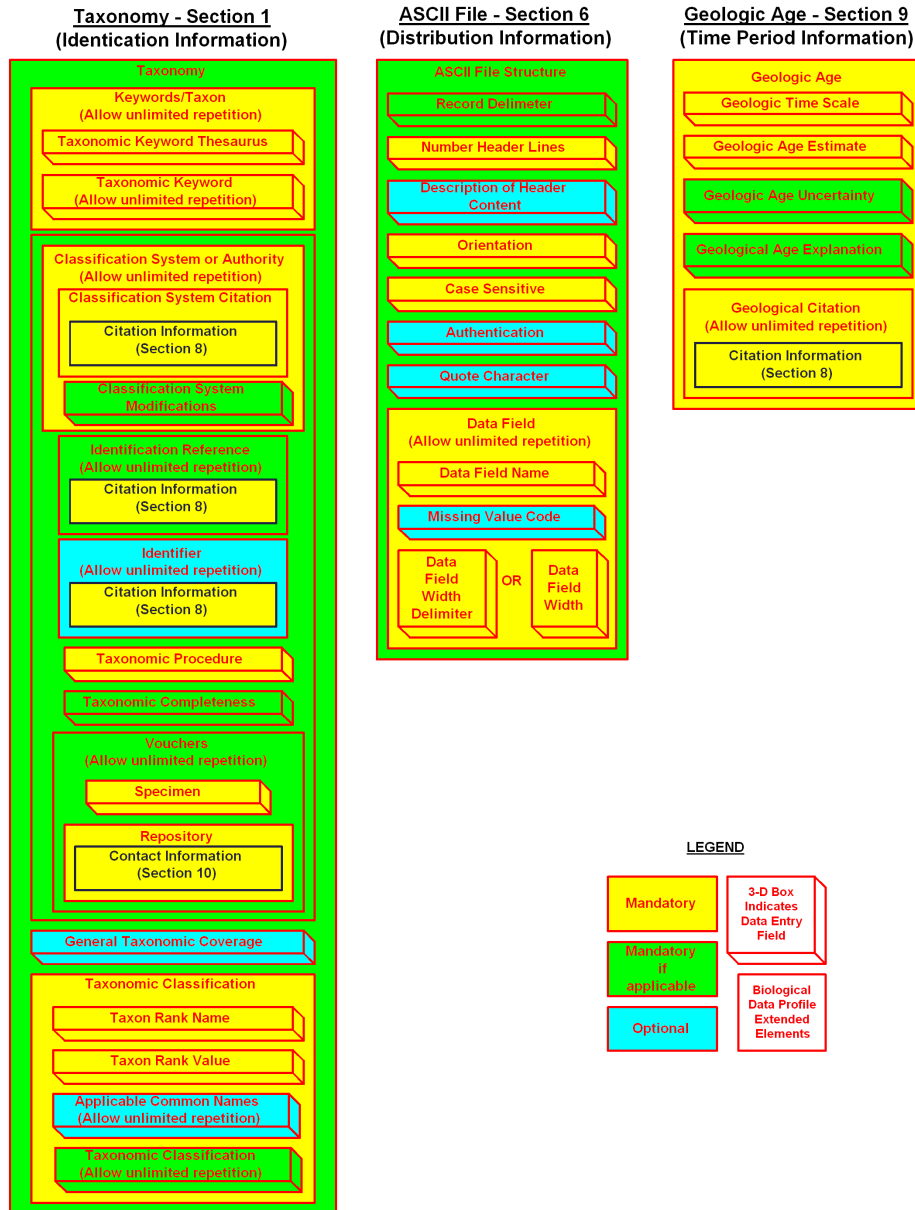


Figure 5.3: Biological data profile extended elements (After Susan Stitt (1998), Center for Biological Informatics, USGS).

scientific workshop held in Belém, Pará, Brazil, the effectiveness and influence of this standard in other research institutes on the Amazon region was discussed, especially those which hold information of a similar nature. The output of the workshop indicates that there is willingness to adopt the standard amongst institutes, especially by the leading institutions like INPA and MPEG (Magalhães *et al.*, 2001).

5.3 From Biological Profile to XML Schema

5.3.1 XML criteria and the mapping process

To provide a solution for researchers to describe their data sets and allow access to them as well as to the data itself, we looked into the XML technology. We consider this to be an extremely useful new technology that can provide good solutions for difficult data problems, such as structuring data in a document which can then be moved to any format on any platform. But this is not enough, and it is worth mentioning that, in some cases, XML is not the recommended solution. The common mistake made by developers is to think that XML should be used for everything they develop. It is advisable to base such decision on user and system requirements rather than on technological fashion only.

For our purpose, we based the choice of XML on four criteria:

1. The need to share biological data and metadata — XML eliminates application dependencies.
2. The need to transfer transient data — XML is suitable for transferring data across the Internet.
3. The need to describe metadata — XML can be used to collect and work with metadata. Since researchers catalogue web content, XML enables easier web searches and queries.
4. The need to receive tailored format output — XML can be used to conduct client-side rendering of web pages by association with Extensible Style Language (XSL). XSL is a language for expressing stylesheets. It comprises XSL Transformations (XSLT, a language used to transform XML documents), the XML Path Language (XPath, an expression language used by XSLT to access or refer to parts of an XML document), and the XSL Formatting Objects (an XML vocabulary for specifying formatting semantics). An XSL stylesheet specifies the presentation of XML documents by describing how an instance of the class is transformed into an XML document that uses the formatting vocabulary (Brundage *et al.*, 2000; Goldfarb & Prescod, 2002). This functionality can reduce the web server side traffic.

Another important reason is that XML is an open format. That is, standards are supported by all major vendors. Large IT companies such as IBM, Microsoft, Oracle, Sun Microsystems, and others have all invested in XML technology and are actively contributing to the standardisation process (Young, 2001).

We used the XML Spy5 system to map all the elements described in the FGDC metadata standard into the W3C (World Wide Web Consortium) XML schema definition (Altova, 2002). The schema is expressed in an XML language, which describes the structure of the content of XML documents. Based on the schema, the system generates an FGDC-based XML output (a well-formed XML FGDC template). Figure 5.4 represents the way we implemented this to produce the FGDC XML Template. Figure 5.5 presents a snapshot of the root level of the BioME (biological metadata) schema representing the elements of the FGDC standard.

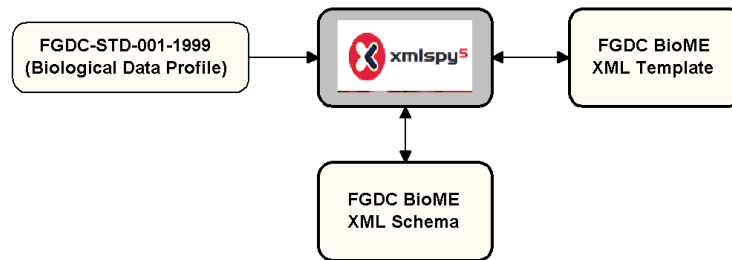


Figure 5.4: Method for mapping FGDC to XML schema and template.

Figure 5.6 presents a segment of the metadata with instance values (in bold) for Citation Information which is part of the Identification Information section.

5.4 Alternatives for Metadata Management

5.4.1 Application deployment and management from a clearinghouse

In certain research domains, the use of a clearinghouses tool kit for managing metadata is an attractive option. A clearinghouse operates by providing mechanisms to institutions to become a node in its infrastructure, linking their data/metadata to a centralised system of servers and search engines with resources to locate and download wanted data and information. Data providers can use any of the nearby nodes to interact with,

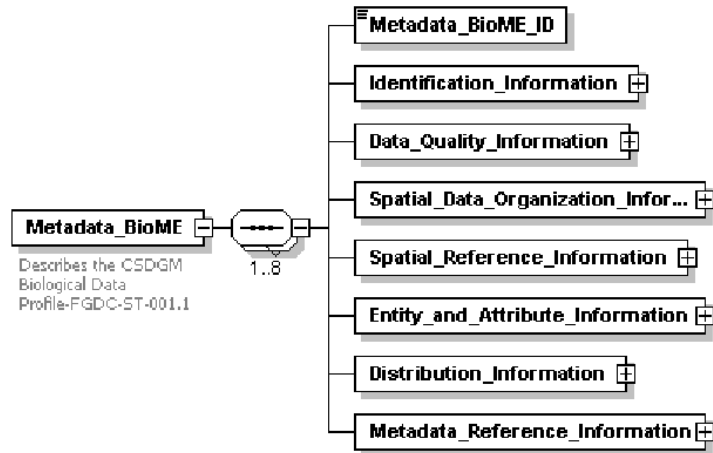


Figure 5.5: FGDC metadata BioME schema - root level.

and at the end, the more data producers participate in metadata and clearinghouse development, the more high quality data can be made available. Such an initiative aggregates a structure of tools and applications to help data providers and data brokers to utilise the web of information (Tsou & Battenfield, 2000). Clearinghouses encompass policies, standards, and procedures under which institutions apply technologies to encourage efficient use, management, and production of data documentation. Users as nodes can download applications, install them at local machines and prepare their own metadata before returning for registration. Users also can count on support and documentation regarding system's operation and the standard that clearinghouse is implementing.

This setup imposes rigidity to nodes (user installations) and specific needs cannot be easily accommodated since they would cascade to a large number of nodes. The same happens to the local computer infrastructure where nodes have to comply with the same software technology adopted by the clearinghouse.

This calls for an evaluation of the complete solutions offered by a clearinghouse and to try to figure out if a complex and large number of variables (500 variables for ISO), a reduction in the number of descriptions (15 variables for DCES) or some specific description is what the institution is aiming at. Additionally, it is important to identify how complex and how costly it would be to comply to the software platform (operating system and program language options) imposed by a clearinghouse.

```

...
- <Identification_Information >
- <Citation_Information >
  <Originator >Sidineia A. Amadio </Originator >
  <Publication_Date >2000-01-10 </Publication_Date >
  <Publication_Time >13:20:00 -05:00 </Publication_Time >
  <Title >CARACTERÍSTICAS POPULACIONAIS DE TRÊS ESPÉCIES DA FAMÍLIA
  ANOSTOMIDAE (OSTEICHTHYES: CHARACIFORMES) DO ECÓTONO CATALÃO
  AMAZÔNIA CENTRAL- AMAZONAS- BRASIL
  </Title >
  <Edition >First </Edition >
  <Geospatial_Data_Presentation_Form >Atlas
  </Geospatial_Data_Presentation_Form >
+ <Series_Information >
+ <Publication_Information >
  <Other_Citation_Details >NA </Other_Citation_Details >
  <Online_Linkage >NA </Online_Linkage >
  <Large_Work_Citation /> NA </Large_Work_Citation >
</Citation_Information >
...

```

Figure 5.6: Segment of a metadata with description for citation information.

5.4.2 Metadata description and management supported by web application

To avoid computer infrastructure similarity amongst nodes there are alternatives where users can use the clearinghouse functionality via a set of web applications. All the resources are available on the web, requiring from the users a reasonable network connection and bandwidth to provide metadata and use the search engine functions.

Researchers indicated they need flexibility to describe their metadata and wanting to do that at a stand-alone basis and when completed to upload and store it at a server. Another inconvenience adopting this alternative is that institutions do not have the same facility regarding network and bandwidth; this is especially the case for Amazonian research institutions. Would it be extremely costly for institutions to comply with these requirements.

Institutions that do not have a clear data policy, which forms the majority in the Brazilian Amazon, view this option as a disadvantage since they do not become a node in that clearinghouse and the data and metadata are stored and managed outside the institution, with no control and no ways to interfere in the clearinghouse plans.

5.5 Implementation: Three-tier Architecture

The reasons for using XML, together with more specific user requirements, have also influenced the choice of a possible architecture for implementation. A three-tier architecture was selected because it is a flexible way of organising distributed client-server systems. By this, we mean that every client is connected to every server. In a three-tier architecture, an intermediate connecting layer is added, that is, tier 2. Figure 5.7 presents the infrastructure was implemented which allows data and metadata management. This will be detailed in the following subsections.

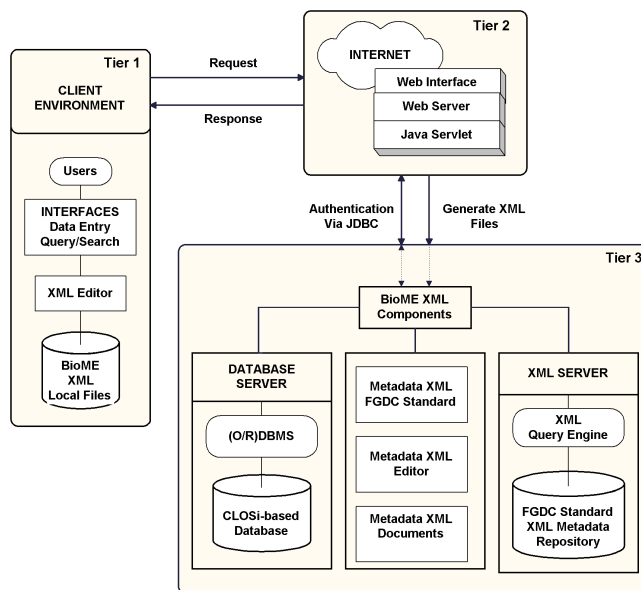


Figure 5.7: The implemented three-tier architecture.

Each of the client applications and servers then communicates with the intermediate layer. This provides a separation of the client applications from the data sources/links and allows them to be maintained more easily. Servers are typically SQL databases but they could equally store other types of databases, XML databases, for instance.

The architecture presents several advantages:

- **Domain separation** — The front-end clients are separated from the back-end data manipulation facilities. This allows details of the data storage mechanisms such as which database is used, record structure

and field names to remain hidden from the client processes. All the front-end client sees is an abstract operation request which involves input and output parameters.

- **Database migration** — Database restructuring, upgrades, migration or any other change can be performed without the need to stop or alter the client applications.
- **Update of front-end applications** — New front-end clients can be added or old ones can be removed without any need to modify the databases or to provide new access mechanisms.
- **Multiple sources Data** — A client user can request data from a number of servers. This is possible due to the split of data operation request into several sub-operations. The sub-operations are performed by different processes at the server-side where the results are combined and sent back as a response to the client.
- **Local caching of data** — Data results from responses can be held at the client side for a specified time period. Thus, common look-up operations can be handled locally without data server access ensuring the data is available faster. This caching is controlled dynamically and in case of any change may result in different data being cached. Client applications have the option of forcing a refresh of any cached data. If such a refresh changes any cached data, the normal data propagation mechanism of the local machine forwards the new data to all interested client applications. Data update operations are typically not cached.
- **Reduce connections in case of central node server (CNS)** — With C client applications and S data servers direct connection requires $C \times S$ connections; use of a CNS-based architecture reduces this to $C + S$ connections.
- **Reduce database loading** — Not only does the database server benefit from fewer connections but any data caching operation results in fewer data operations and therefore less throughput. In addition, this saving is concentrated on those queries that are most commonly performed, thus reducing the potential for conflict on the data. Regarding implementation it eases load balancing. Load balancing refers to the distribution of processes and communications activity equally across a computer network, so that no single device is overloaded. Load balancing is especially important for networks where it is difficult to predict the number of requests that will be issued to a server. Busy web sites typically employ two or more web servers in a load balancing scheme. If one server starts to get inundated, requests are forwarded to another server with additional capacity.

- **Local data** — The central node may also be used for storing local information that does not need to be stored in any of the data servers.
- **Monitoring and auditing** — Since the basic CNS operation is to forward all data changes to interested clients, it is easy to create client applications that have auditing or monitoring functions. This can be done on a system-wide basis or on a cluster-wide basis. Such clients may be used for performance monitoring of operators, tracking exposure in a financial environment or writing a log for close-of-day analysis.

5.5.1 Client Side

The Client Environment (CE) supports two types of users: providers and brokers. Both users operate at Internet and Intranet levels. At Intranet level, networked users can access its local server for data entry, query or search via an XML Editor. For the web server interface application, we adopted the XMLmind XML Editor Standard Edition (XXE for short). XXE is a free software XML editor that supports DTD and XML Schema aware editing commands and a word processor-like view configured using W3C's cascading style sheets (CSS).

For institutional stand-alone users, the BioME XML local files (the biological XML template), can be edited by using an XML editor. For this, we have used the XRay2 system, running on Microsoft Windows 2000 Professional. XRay2 is a free XML editing environment for Windows and provides support for Real-time XML Editing, XML Schema (XSD), XSLT Processing and Schema Validation. Once the metadata has been finalised and validated, it can be stored using the Intranet environment for access.

At the Internet level, the users can access and perform operations available at the BioME portal supported by a web server accessing data and metadata repositories.

5.5.2 Web Server

The selection of a proper server affects performance, reliability, and security but also essential characteristics of implementation, interfacing with legacy technology and some specific user requirements. We adopted the Apache HTTP Server (Coar, 1998) to allow access to the BioME Portal. The reasons for using Apache rely on its dominance, being the most used web server on the Internet. Amongst the main developers, Apache has been the most consistent (Wainright *et al.*, 2002). The main operations at this level include: downloads of data sets specifics, documents (FGDC Standard) and applications (XML Editors like XRay2 and XXE), queries - that can be submitted to a database (CLOSi-based) or the metadata management system,

browser and server specific operations.

5.5.3 BioME XML Components

A portal has been implemented to provide access to the metadata options, which include: metadata catalogue list, full documentation of the FGDC standard, the XML template and an XML editor, as well as to permit on-line editing of metadata. Users can fill up metadata descriptions using the online forms or by installing BioME components. The functions implemented in the web portal are a set of Java classes that use the servlet API for the communication between a browser client and the Java classes available in the web server. An API allows programmers to access the functionality of a pre-built software module through well-defined data structures and subroutine calls. Although programmers often define APIs for private internal codes, network APIs typically are the public entry points to libraries that hide low-level details of computer networking. Traditionally the most popular networking APIs have accompanied socket libraries. Berkeley sockets and Windows Sockets (Winsock) APIs have seen widespread use for many years. More recently, Java network APIs such as servlets have also grown in popularity.

The online entry of metadata is time-consuming, and together with the precarious bandwidth in the institutes where we plan to install, the performance of this process would be affected. An alternative option that we put in place is to deploy the XML template and an XML editor. The editor, once installed in the client environment, uploads the template to allow editing. This guarantees that the final product is a well-formed XML document, fit to be uploaded into the XML server to become searchable.

5.5.4 XML (Database) Server: Metadata Repository

We have tested Tamino (IDC, 2001) and XYZFind Server (XYZFind Corporation, 2001). Tamino is comprised of an engine (called the X-Machine), and an internal XML data store facility, a relational database and SQL engine, a module (called X-Node) for interfacing to external data repositories, an HTTP server, administration tools, and a data map to describe the location of data sources to produce composed-site output from multiple and heterogeneous sources. XYZFind Server is an integrated repository, search, and query server designed for XML. As an XML repository, it stores XML files and maintains a single data representation of all files stored. Files are indexed and can be accessed, updated or deleted from the repository. As an XML query engine, it accepts queries expressed in the XML language (XYZQL). The language provides features that include path-level queries,

5.5. IMPLEMENTATION: THREE-TIER ARCHITECTURE



Figure 5.8: Screenshot of the BioME portal: a way to deploy metadata components.

boolean queries, keyword search, and numeric queries.

We opted for XYZFind due to its simplicity, sufficient features for our needs, easy installation and use of the functions and with no costs involved. The server is operated via a web-based console providing access to status information, a query form with templates, plus full user documentation. Additionally, it offers concurrent read-write operations and Multi-Threaded Server (MTS) application and access control that can be performed by the HTTP basic authentication facility. MTS is a strategic component of server technology that provides greater user scalability for applications supporting numerous clients with concurrent database connections.

Combining the best features of databases and full-text engines, XYZFind delivers robust storage and retrieval access without dependence on schema design, transformation, or DTDs. XYZFind eliminates schema design and data mapping. It achieves this by decomposing XML documents into a single schema data representation. This approach provides persistence, and is comprehensively indexed, that is, at the level of words and numeric values, and all associated path information. XYZFind makes storing XML documents simple and delivers the high performance structured query one expects from a database, and enables seamlessly integrated keyword search (XYZFind Corporation, 2001).

The server accepts four types of request: XYZQL, Zip, Update and GET. XYZQL is the main interface which interacts with the server. The zip format

Table 5.1: HTTP requests of XYZFind

Request Type	HTTP Method	URL Path	Content-type	Body of Request	Server Response
XYZQL	POST	/	text/xml	An <code>< xyz : input ></code> document	An <code>< xyz : output ></code> or <code>< xyz : error ></code> document
Zip	POST	/	application/zip	A zip file containing XML documents to add to the repository	An <code>< xyz : archive ></code> or <code>< xyz : error ></code> document
Update	POST (or PUT)	/docname	text/xml	XML document to add to the repository	An <code>< xyz : update ></code> or <code>< xyz : error ></code> document
GET	GET	/docname	N/A	N/A	The requested XML document or HTTP error 404 (not found)

is a way to add XML documents to the index at once. The Zip request will post a zip file to the server containing only well-formed XML documents. An Update request adds, replaces, or deletes a single XML document. A GET request retrieves one document from the index by URL. Apart from GET request, which uses the GET HTTP method, the others utilize the POST method to request a service. The Update and GET require the URL path to point to the XML file name. The query interface is not sufficient for naive users. Naive users need a much more friendly interface with advanced query capabilities. Table 5.1 shows the properties and requirements of each type of request.

5.6 Discussion of the Implemented Solution

The three implementations, prior to this one, aiming at information access and dissemination, were experienced in the Oiapoque (Marcon, 1999), LBA (LBA Project, 1997) and the monitoring invasive plant species project (Kandeh *et al.*, 2002). The LBA implementation focused on a controlled client/server environment with specific functionality to attend a distributed but small number of users. The Oiapoque project implements remote databases with the same software platform, including the DBMS where these databases are synchronised for replication on the remote sites. The latter of this, implements a stand-alone application with functionality to import/export data and information via a dial-up connection.

This solution goes beyond the needs of those projects mentioned and can present distinctive advantages in functionality against the above mentioned

ones. For instance, the modularisation of client applications and servers allowed to separate and hide the user interface solutions from the database management. This implies that any of the operations performed at the client side do not interfere with the services offered at the server side, and vice-versa. Also, the client environment can hold data for temporary local management and use. For example, exhaustive look-up operations can benefit from this. Local data management will help to reduce the number of connections and the number of data operations on the server side. We did not implement any specific monitoring and audit functions, instead, we used the basic monitoring functions available at the web server and database/repository level (tier 2 and 3) were used. Conclusively, the infrastructure in place is robust and ideal for implementing client applications that can perform audit and monitoring.

The architecture implemented depends on network and internet/intranet capability and the potential user institutions differ in such a capability. Users will have to chose and adapt the application setup to permit most of the operations to run at a local machine; for example, to describe metadata at the client side and when completed to upload it to a server for verification and storage.

We are of the opinion that the use of an XYZFind server provides a simple web-based console, which is not sufficient nor intuitive for a naive users when expressing commands using XYZQL. The development of a more intuitive console interface with advanced query capability is needed. This can be achieved by developing and integrating a client application to the BioMe portal. The portal will need to be dynamic and to be managed by a webmaster to incorporate new requests and updates.

It is important to mention that organisations should be aware that implementing this infrastructure will require expert personnel for development, and maintenance activities at hardware and software level within client and server side.

5.7 Concluding Remarks and Future Work

We have presented an implementation for managing XML-based metadata for biological collection data. The FGDC, a consolidated metadata standard, was adopted and implemented as a template to be deployed on request for editing in stand-alone mode or via the Internet. The solution has been placed into a client/server infrastructure that allows metadata and data dissemination for users on a global scale via the web. The infrastructure comprises a three-tier architecture that accommodates different degrees of implementation robustness and scalability toward an active node network. An example of a biological metadata from INPA's crustacean collection is

presented in Appendix D.

Taking into account the SCP, this work can contribute to data/metadata exchange, particularly across the Amazon region, bringing to light essential information for conservation and sustainable development of the region as well as documenting the existent data sets, a clear sign that information chaos can be controlled. This work is particularly interesting for developing countries since our solution and tools used are either public domain or from free open source, adding low cost as an additional advantage for a robust solution to any particular organisation. We envisage, for later implementation, the need for a complementary module for producing dynamic report outputs at the web interface level. Users are now able to query/search for metadata descriptions and when found, they must be able to produce user-tailored outputs, such as in HTML, PDF formats etc.

5.7.1 An Active Node Approach

Currently, a number of Amazonian research institutes lack clear data policies. This allows either an individual scientist or a scientific team to keep the use of data sets exclusive for a long period. Data policy is an important mechanism within an institute, since it can guide data sharing, citation of data from other researchers, access to restricted data and promote the exchange of quality controlled/quality assured data. Some organisations have advanced on this matter and today have a well-established data and publication policy within their boundaries. These institutes support timely data release for public access and exchange because they advocate the fundamental principle that cooperation and synergy should be a priority in all research activities.

The eco-physiognomy of the world today has put biological data in evidence. As a result, scientific databases on the web can generate a lot of data traffic. One way to dissipate the traffic and provide backup of the database is to establish mirror sites. By storing these mirror sites elsewhere, the intent is to reduce the traffic for regional level. Mirror sites have to provide duplicate copies of the database accessible at several access points. One problem with this approach is how to maintain the same database at other locations. Ideally, every time a central node site is updated, all others must be as well. In the case of a sizeable database undergoing intense updates, to propagate the changes becomes troublesome. Additionally, mirror sites might provide additional regional content.

The important aspects for users are to know if the databases are the same, whether the search/query functions are identical and if there is any advantage to access site A instead of site B. Another problem emerges when adopting different software solutions at different sites. For instance, some

5.7. CONCLUDING REMARKS AND FUTURE WORK

of the mirror sites can have a different database or a more limited version of the database. Given the global scope of the mirror sites, the non-English language search has to be taken into account to deal with the problem of diacritic marks (e.g., tilde and umlaut) that can be solved via character set definitions. We experienced this problem in the Oiapoque project, when implementing mirror sites in Belém - PA, Brazil and Kouru in French Guyana (Marcon, 1999). The solution included three different hardware and software architectures, and the search engine had to provide functionality for Portuguese, French and English. Another experience we had was during the implementation of the metadata harvesting mechanism for the LBA and Nasa LBA Ecology project, which include node solutions. There is a central node located in USA (ORNL-DAAC in Tennessee) interacting with distributed nodes in Brazil (INPE in Sao Paulo and INPA in Manaus).

The solution we implemented here for metadata management can be extended to incorporate an active node approach, where data providers can store data and metadata in specified locations on their own or other web-accessible computer systems and make them searchable. Users can search, query via a web application interface (e.g., BioME), which can store data either in the permanent archive or metadata catalog and CLOSi database, at a local active node. These actions include the ability to support interoperable metadata protocols that allow remote clients to search and retrieve their metadata records. These remote clients can then provide their users with search results from other sources.

This can be in the form of a federation of systems or through harvesting interoperability. In federation form, the user requests are managed through several service providers, agents who harvest metadata from data providers and use the metadata for value-added services as defined in the DLESE's approach described in Lagoze & hunter (2000), where distribute queries to remote metadata providers, obtain the results, and return a response to the requester. In the harvesting form, user requests are managed through a single service provider. The service provider obtains metadata from providers. Service providers import these metadata into their discovery system. Federation requires more effort from data providers but is easier for service providers (no need for parsing and interpreting metadata records). However, it presents problems of scalability. Harvesting requires less effort from data providers, but is more difficult for service providers, that is, the service providers deliver full metadata records; they require an additional process to refine the search to match the requirements and make available in the system (Gilliland-Swetland *et al.*, 2000).

Whatever the architecture proposed, we can implement both forms of interoperability; the harvester seems to be more advantageous, since it will collect metadata from only the files that are signalled as searchable. Each

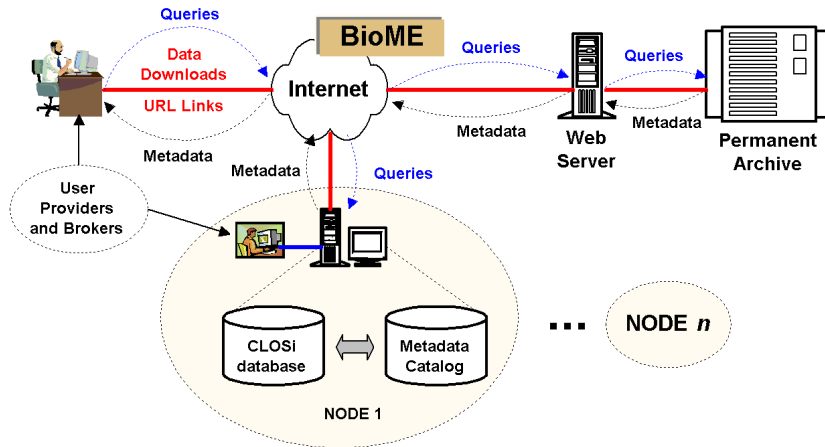


Figure 5.9: Example of the proposed web active node approach.

node harvests metadata and builds a database of the harvest information. Harvested information is shared across the nodes so that users are able to search and review available metadata by accessing the BioME system at any of the active nodes with their web-browser software. Figure 5.9 presents the active nodes approach, an enhancement of the mirror site solution.

5.7. CONCLUDING REMARKS AND FUTURE WORK

Chapter 6

Implementing Biological Data on the Web *

6.1 Introduction

The CLOSi schema has been developed as a proposal to represent the INPA biological collections database as well as being a method for others to follow in managing similar databases. The implementation of CLOSi is intended to reduce the problems of spontaneous development of applications. In CLOSi, clusters are interrelated and all refer to information related to the collection of specimen of organisms. Each of the clusters and their classes were developed to be applicable to a biological collection information system. The CLOSi schema is detailed in Chapter 4 and its complete description can be found in Campos dos Santos & de By (2000).

This chapter focuses on the implementation of the cluster Locality of Biodiversity Data with some development also of the clusters Reference, Agent of Collection and Collecting Event of Collection. The cluster Locality of Biodiversity Data consists of nine classes, containing information pertaining to the locality of a collected organism. The Geo Reference Object class is a superclass with the subclasses point, line and chain each describing a location for a specific collected organism. In addition to the classes and their specific attributes describing location, the CLOSi schema relies on relationships to other clusters to link all pertinent information to a collected organism.

A successful implementation of CLOSi includes the dissemination of

*This chapter is based on Campos dos Santos, J. L., de By, R. A., Magalhães, C., and Apers, P. M. G. (2002). Facilitating interdisciplinary sciences by the integration of CLOSi-based database with bio-metadata. In the International Archives of Photogrammetry and Remote Sensing, Vol. 34 - Part IV, Ottawa, Canada.

data over the web. The web interface developed as part of this prototype is based on Open (free) Source Solutions (OSS). OSS tools allow for projects to be implemented with the advantage of lower costs, vast technical support, fewer restrictions (Perens, 2002). Future projects using the CLOSi schema in developing countries can more feasibly be implemented with OSS tools. Since institutes like INPA, and other similar institutes in developing countries, lack resources, OSS can play an important role in providing a satisfactory solution.

The CLOSi schema aims to be the framework of a biological database within the existing structure of INPA. The role of the database will be to aid in the management and research of biological collections contributing to global conservation and sustainable development. The development of a prototype website interface for such a database is one step in achieving such a comprehensive information system.

It needs to be demonstrated that accessing a biological database founded on the methods of the CLOSi schema can be achieved. This chapter investigates the options for development of a data access oriented web interface as well as implementing a prototype web interface for accessing the INPA data within the Locality of Biodiversity data cluster of CLOSi. The result of this work will also provide recommendations for the implementation of CLOSi in an Amazonian context.

There are a number of relevant questions for which an implementation can provide answers regarding the three-tiers of a client/server architecture, a framework that seems the most natural to be adopted. Questions include:

- What are web server, web database management system and script-side? Why use them?
- What is the standard practice for effective data access web interfaces (design and functionality)?
- What methods of querying are available and which are appropriate for use in the prototype?
- What are the technical requirements for achieving a data access system like the prototype?
- What are the possible constraints to implementation of a data access web interface?
- What are the procedures for development of the prototype and can it be effectively duplicated?

6.2 Tools for Website Development

6.2.1 Open Source Systems

OSS are software programs that are created on the basis of the concept of free and fair distribution, as well as full access to source code. The source codes that are shared support different types of application. This code sharing improves and evolves new applications through a network of support but also leads to a need for code integrity, ensuring reliability. The benefit is that the reliability comes from the users. Over time, any logical error in an OSS program is discovered and repaired by the large number of developers working with the same code. Therefore, it cannot be assumed that expensive software is necessarily better. The conventional software companies develop software and maintain it without the aid of users as a potential quality control check, as is done in OSS. There is a licence that protects the OSS software, so that if anyone improves the software it cannot be redistributed without the source code.

Perens (2002) presents the criteria that define OSS and the requirements for obtaining the OSS licence. The requirements can be summarised as: (a) Free redistribution; (b) Access to source code; (c) Derived works from a large community (allow modifications and redistribution); (d) Integrity of the author's source code (reliability and responsibility); (e) No discrimination against persons or groups; (f) No discrimination against fields of endeavour; (g) Distribution of licence (no additional licences); (h) Licence must not be specific to a product (choice of software design); and (i) The licence must not restrict other software.

Probably the most appealing of these criteria to most users is the 'free redistribution' requirement. Now, institutes that lack financial resources have the ability to develop software applications to suit their purposes. INPA is one of the institutes aiming to benefit from the OSS approach. The use of free software is a suitable alternative for prototyping systems, especially when it is applied in developing countries. INPA's needs for maintaining a database management system for research and management purposes requires the use of the three-tiered architecture. A web server and server-side script are necessary for interaction with the database over the Internet.

6.2.2 The Web Server

A web server is a system that takes a request sent from a web browser in the form of a Universal Resource Locator (URL). After the request is taken, the server finds the file on the local host disk directory and subsequently serves that file back to the client's web browser. The browser making the

request and the server returning the file communicate through the Hypertext Transfer Protocol (HTTP).

The initial function of the web server was to transfer static pages to the browser (i.e., images or text). This has grown to include the transfer of dynamic content, beginning with the Common Gateway Interface (CGI). The concept behind dynamic web content focuses on the ability of the web server to recognise a request from a browser as a request to run a program and then return the program's output as a file result as opposed to simply returning a static file.

There are some complexities in the function of web servers that go far beyond the scope of this work, however, some basic elements include serving content, acceptance of connections, providing security, and running web applications.

Choosing a web server

There is a vast amount of literature available on choosing the right web server (Macehiter, 2000; Brain, 2001; Christy & Katsaros, 2002; Microsystems, 2002). In the following, we briefly discuss the considerations for choosing a web server. They are a good representation of those that frequently occur in the available literature. The items that one must consider include:

- **Compatibility** — The system must be consistent with existing hardware, software, operating system, and needs.
- **Cost** — There are many free packages available, but this should not be the only deciding factor.
- **Security** — Access management via passwords and other mechanisms for authentication and security.
- **Features** — The system should enable setup, logging, web-based applications, content management, indexing and search functions.
- **Ease of Use and Learning** — It should have available (online) tutorials, manuals, help, and an intuitive interface.
- **Upgradability** — The developers commitment of reliable upgrades to solve software implementation problems must exist.
- **Technical support** — The developer must offer training, and on-line support.
- **Review** — Contents of review covering questions, such as What do other users say about the server? What type of application is it recommended for? With what other software does it work well? are extremely important.

We verified five products that match the criteria, they were: Apache (Yerxa, 1999; Aulds, 2000; Laurie & Laurie, 2002; Netcraft Web Server Survey, 2002), Savante (Lamont, 2001), Xitame (Castro, 2001; Hintjens, 2002), GoAhead, and Netscape Enterprise Server (Netcraft Web Server Survey, 2002). From this list, only the latter is a closed source system, the others are OSS. The Apache web server was chosen to support this implementation because it is the most popular web server, holding 60% of the market share with great reviews from its users. It supports nearly universally all platforms with a tradition in Unix, Linux and Windows NT. Although Microsoft IIS and Netscape Enterprise Server have been found to perform better, the recommended server for use with dedicated robust PHP server-side scripts and the RDBMS MySQL is the Apache server. Other attractions with Apache are that it is free, and that it provides full source code and unrestrictive licence.

6.2.3 The Database Management System

Databases are stored collections of data. A relational database stores data in separate tables instead of in one large repository. The data in the tables are related through corresponding attributes, which are called primary and foreign keys. These keys allow the tables to be combined and searched based on a Structured Query Language (SQL) statement. Tables can be joined and stored as new tables or simply related temporarily for the purpose of discovering data within two tables based on certain combinations of values.

The integration of a Relational Database Management System (RDBMS) enhances the functionality of a web server by offering the following additional capabilities:

- **A dynamic data processing engine** — Client requests can be processed concurrently. Data queries will provide the most up-to-date results as they are processed in real-time.
- **Conceptual representation and organisation of data entities and their relationships close to the real world** — Data is stored in tables based on a logical and conceptual structure. In the case of this prototype implementation, the tables will be structured based on the CLOSi schema.
- **A back-end search engine supporting complex queries** — This offers more functionality. SQL features allow more than simple file access for HTML page customisation.

Choosing a DBMS

The most important factor in choosing a DBMS is the model employed to store, manage and retrieve data. For this prototype, we targeted relational technology for practical reasons and also because the main features include the reduction of redundancy, application programming interface (API) support, and tools enabling complex queries. As with the methods for choosing a web server, there are also different reasons for choosing an RDBMS. The following is a summary made up from some of the most important considerations.

- **Cost** — There are a number of DBMSs available under a free licence agreement, but the decision must be based on the analysis of cost and functionality combined.
- **Support from an API and use of programming languages** — It should support server-side scripting in languages that the developer has selected to use.
- **Can stand the ‘ACID test’** — Provide full transactional ACID properties. Regarding transaction process there are four characteristics:
 1. **Atomicity** — None or all operations will complete.
 2. **Consistency** — The database is transformed from one valid state to another valid state.
 3. **Isolation** — The results of a transaction are invisible to other transactions until the transaction is complete.
 4. **Durability** — Once committed (completed), the results of a transaction are permanent and survive future system and media failures.
- **Conformity with platform needs.**
- **Back-up and security features** — Provide users authentication, access trail, back-up and restore regimens.
- **Multi-user support** — Simultaneous access by multiple users. Databases requiring a high volume of multiple accesses at once, must be equipped with transaction and concurrency (locking) control.
- **User friendly tools** — User-friendly tools for direct manipulation by the administrator, good documentation online with provision of paper documentation, sound technical support, etc.

There have been many studies comparing competing (object-) relational DBMS software. Many of them contradict one another and some of them are out of date. (Yeager & McGrath, 1996; Stein, 1996; Ashenfelter, 1998; Morrison *et al.*, 2000). Some basic facts however can be extracted from these studies to make a decision based on the needs for our prototype. Three DBMS

have presented interesting features that we considered as good options to support the prototype implementation: MySQL (Dubois & Widenius, 1999; Adida, 2001; Johnson & Vijayan, 2003), PostgreSQL (Stonebraker & Rowe, 1985; Stinsen, 2001; Geschwinder & Schoening, 2001; Douglas & Douglas, 2003) and Oracle (Kyte, 2001; Lewis, 2001). Despite our aim to use a good OSS solution for the prototype implementation, we included Oracle in the analysis because it is an important proprietary solution and today is considered the leading DBMS in the market. This would provide us with a way to perceive if there are discrepancies in product status and how they differ.

As the best OSS, we have chosen the MySQL because it is fast, reliable, and easy to use. It also has a practical set of features developed in close co-operation with a large user community (around 4 million installations powering websites). The server was originally developed to handle large databases much faster than existing solutions and has been successfully used in highly demanding production environments for several years. Its connectivity, speed, and security make MySQL server highly suited for accessing databases on the Internet. MySQL runs well on several operating system platforms (e.g., Windows NT, Unix, Linux, etc). There are numerous positive reviews for running with Apache and PHP on medium-sized databases. MySQL, however, lacks referential integrity support and nested SQL subqueries. However, there is a promise that soon MySQL will provide these capabilities, and will also become fully ACID compliant. Another constraint is the capability of table level locking only.

On the positive side, MySQL supports replication. For robustness, two (or more) systems can be used to switch to a back-up server in case the master system presents problems. The extra speed is achieved by sending part of the non-updating queries to the replica server.

Despite the fact that MySQL generates a results page about two to three times faster, PostgreSQL has been found to be able to run more concurrent connections. It supports transactions/rollbacks and foreign keys and can do row level locking. The replication mechanism in PostgreSQL is less ideal than in MySQL, and some security features are not as good as in MySQL. On the other hand, PostgreSQL is in a mature stage of development and is ACID test compliant. Amongst its limitations, it does not run as well as MySQL on WindowsNT.

The most recent version of Oracle (Oracle 9i) has focused on becoming an effective platform for Internet and application development, including XML, Enterprise Java Engine, and SQL and PL/SQL improvements. The Oracle database absorbs the WC3 XML data model, and provides new standard access methods for navigating and querying XML. The integration of native XML capability within the database offers a number of benefits, such as management of structured data (tables) and unstructured data (files) or

BLOBs. Such integration has to subject application to different paradigms for managing different kinds of data, valuable repository functionality, and enable data and documents from disparate systems to be accessed and combined into a standard data model. Oracle seems to be a ideal solution as it can provide compatibility for almost all user and system requirements. The main constraint imposed on the kind of application we are involved with is the cost of the software, which reaches \$ 20,000, making this option out of reach for our consideration.

6.2.4 The Server-side Script

There are two methods for serving dynamic information via the web:

1. **The client-side method** — This method transfers code to be executed on the client's web browser. The disadvantages include lack of control over the execution of the code, slow execution (depending on the application and the client's Internet connection), difficulties in controlling the display of the application, and lack of security due to the code being available in text format to the client. Another disadvantage is the lack of database interaction capabilities, as the database will typically be on the server side. After a page is sent to a client, the link between the two will cease to exist and there is no way of communicating with the server until a new request is made. This results in a lack of interaction between the client and the database.

Regarding the code that implements the actions, client-side scripts are mainly useful if the intention is to give the client many controls over the application, such as on-click, mouse-over and data layer manipulation (Williams & Lane, 2002; Welling & Thomson, 2001).

2. **The server-side method** — Also known as the Common Gateway Interface (CGI), the server-side method is based on calling an external program and executing code on the server and returning a static HTML page to the client.

The concept and functionality of server-side scripts can be explained through an analogy. When someone goes to the library and searches the catalogs, s/he is like a user surfing the web. A search is made for a book using keywords or subjects, and a series of addresses is returned, informing the library user where the book is located. Once an address has been selected, the user will go to the librarian and ask for the book. Here, the librarian performs a function similar to the web server. The librarian looks for the requested book (or page) based on a catalog address (or URL address). Once the book is found, the book is simply served to the user. This is the same concept as the web server retrieving a static HTML page based on a URL address.

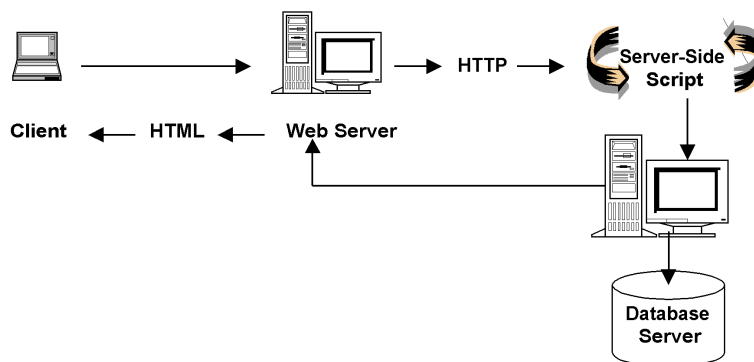


Figure 6.1: Running a server-side script: The three-tier process for dynamic web databases.

The difference between the library analogy to a typical HTML page request and a server-side script process, is that the librarian would be asked to do more than simply return a book. The librarian would have to locate the book and then determine a requested series of questions about that book and then return to the user the answers to those questions. This process is carried out in a computer when the web server recognises the questions (usually in the form of an SQL query) by the extension of the request (i.e., '.php', '.asp', '.jsp', etc.) and then runs the script embedded in the HTML on the server side. The page requested by the client contains the script that could execute many things including interacting with a database or with the file system. After the script is executed, the page is returned to the client as a static HTML page offering no opportunity for the client to see the executable code of the script.

Choosing a server-side script language

After first deciding that the dynamic content will be processed on the server side through a CGI, i.e., ASP or PHP, and not on the client side (i.e., via Java servlets), we determined which language best meets the needs of our prototype. Below, we list some general questions that indicate the limitations and needs that must be considered with respect to development of a biological collections web interface:

- **Simplicity** — Can the script language be learned quickly with little training investment?

- **Availability** — Is the script language OSS and easily accessible for download? Even if the solutions are OSS, some operational costs will be incurred, such as training.
- **Interoperability with other applications** — The database and the web server should be used along with the server-side script language as some combinations are known to have better results due to available functions for particular programs, as well as more support network applications (e.g., Apache web server, MySQL database server, and PHP server-side script language are a well-known combination). Also, the operating system must be considered.
- **Security** — The script language should offer functions for securing the website from clients attempting to perform forbidden actions. Authentication, session tracking and password encryption are examples of some security capabilities that are required.
- **Stability** — The system should be robust and not prone to crashes.
- **Support** — There should be a wide range of tutorials for training, forums for troubleshooting, articles and manuals for available functions. Also, there should be strong recommendations from previous and current users for the chosen server-side script language.

We have looked into three of the most popular solutions of server-side languages: PHP (Hansen *et al.*, 2002; Pushman, 2002; Lerdorf & Tatroe, 2002), ASP (Mitchell & Atkinson, 2000) and Perl (Hansen *et al.*, 2002). As with tools for the web prototype development, we focused on OSS solutions with the necessary features to successfully deliver a prototype quickly and robustly. We selected PHP because of personal experience. It is easy to learn in comparison to other languages such as Perl or Java. Many security features are provided, including sessions and authentication as well as encryption functions.

All three solutions provide means for integration with databases. Perl and ASP have an abstract database access method that makes it easy to change databases. PHP lacks this feature. Although the interfaces are similar, PHP uses another type of function for accessing different databases. When comparing text manipulation, Perl offers the best resources. PHP, on the other hand, also has many of these features. ASP provides several of the same features, but they do not seem to be as well integrated in a non-Windows environment (Hansen *et al.*, 2002). Also, ASP lacks many string handling facilities and is far behind when compared with the integration levels provided by Perl or PHP. All the options presented fail in image manipulation as a native implementation. It can be incorporated however, via plug-in options. Data-driven web pages may be effectively deployed using free OSS.

6.3 Prototype Implementation

The implementation of a CLOSi-based database system following a three-schema architecture as the standard architecture for data modeling, which was first proposed by the ANSI/SPARC committee on databases. The concept is that the design consists of three layers: the internal, external and conceptual - which map to the physical database design, views, and logical database design respectively. The physical database design is invisible or transparent to a user and is concerned with the way the data is actually stored. The external layer or view is a projection of the database — it is concerned with individual user views. The logical database design is the conceptual layer and is the mapping between the external and internal layers. It describes the layout of the database; that is how the tables and columns are actually arranged. The schema architecture is described in Elmasri & Navathe (1994).

A CLOSi description is at the external level (external view of end users). Cardinalities, relationships, classes and attributes with data types are all described within the CLOSi schema. Therefore, the schema contains all the necessary information to form a conceptual design of the biological database. To conceptualise the database, an entity relationship (ER) diagram was created. The ER diagram can be mapped into a relational database structure (tables and attributes with primary and foreign keys identified for referential integrity).

In this particular implementation, the ER schema is extended (EER) to account for multivalued attributes (super/subclass) and specialisation relations or 'ISA' relationships.

Once the conceptual EER has been mapped into relations the internal structure of the database is ready to be created. All the details for creating the database using MySQL as the DBMS and more additional explanations about the physical structure of the database are presented in Bisset (2002).

Web interface heuristics

Bosley and Straub (2002) developed a set of heuristics for web pages focusing on database exploration. The heuristics explained below form a framework for what the INPA database web interface should achieve.

Orient the user to the available body of data

1. Give an overview of the available data: Orientation to the data, highlight scope, restrictions and deliberate omissions.
2. Support situational awareness within the available data: Use text or graphics to propagate the data across levels and tell the users when

they are entering a disjoint partition of the data. Make it easy to return to the initial state.

3. Display and clearly define metadata: Include sufficient metadata and describe technical or unfamiliar terms throughout. Avoid unclear labeling.

Design the interface for interacting with the data

4. Place adequate and clear instructions on the interface: Give explicit instructions for relating with the interactive elements of the interface.
5. Link users to frequently requested analysis: Provide links to frequently requested numbers or data sets. Store common queries and also build an advanced search mechanism.
6. Use simple interaction schemes to accomplish complex query building: Use logical task sequence or natural language to support advanced boolean query syntax.
7. Summarise outcome of complex data specification for review and confirmation: Display a final specification of multiple filters to the user before submitting a request.

Help users anticipate, interpret and evaluate results

8. Offer choices of easy-to-interpret output formats: such as well-known, understandable outputs like tables or simple graphs.
9. Design output formats to facilitate quick and reliable query validation: Output labels should be clear, highly visible and relevant to the queried variables.
10. Help users avoid searching for non-existent or non-available data: warn against or prevent requests for unavailable data. Notify users of null results.

These heuristics should be applied throughout the design process and be refined as the interface is expanded. As the interface is explained below, the involvement of these heuristics will be mentioned.

The interface structure

Flow charts are useful for providing a structure. The flow chart in Figure 6.2 represents the web interfaces within the biological collections website. Users enter through the 'Welcome Page' at the top and may navigate to find each page depending on their access privileges.

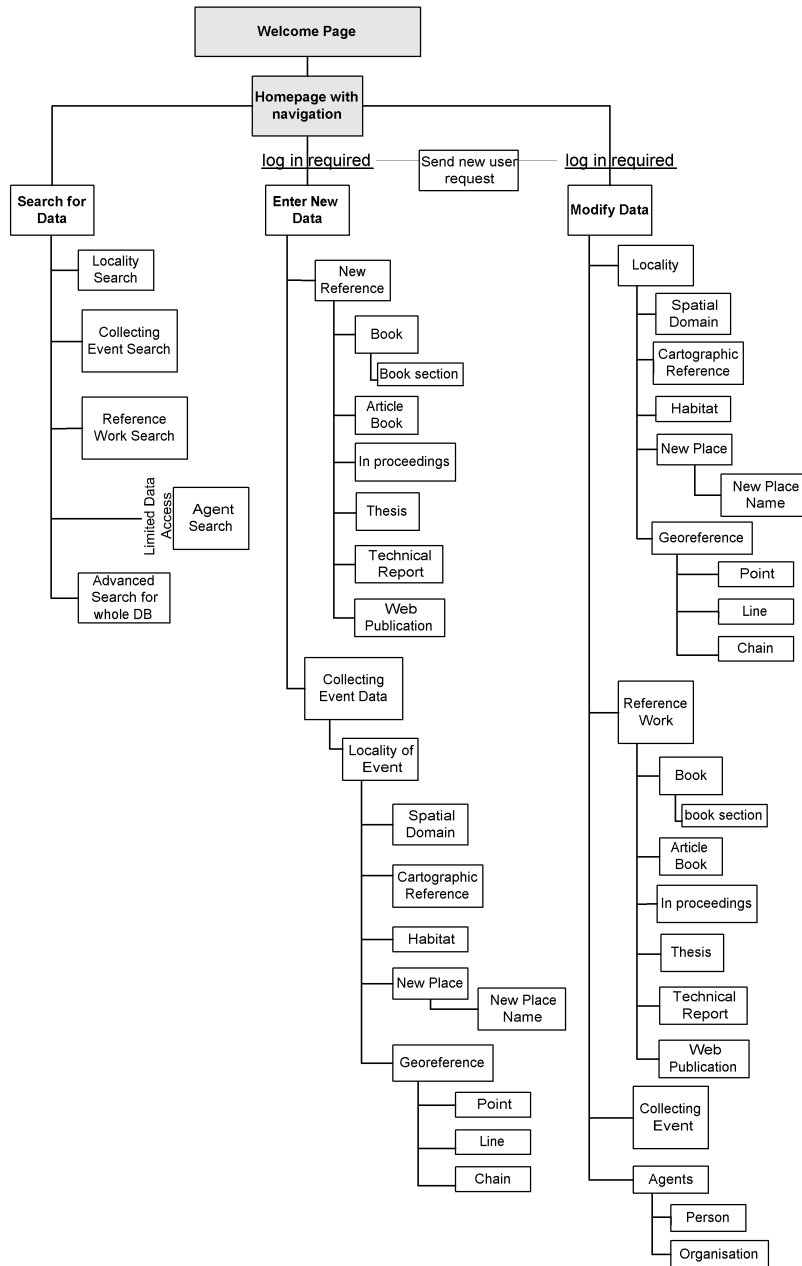


Figure 6.2: Web interface mapping hierarchy.

The mapping of the web interface follows the cluster and class structure of the CLOSi schema. CLOSi provides a well-structured hierarchy, on which all good websites should be built.

The entry page

The first page of the website must be aesthetically pleasing, be fast to load, and give the user a clear and concise description of the website. It should be evident to the user that the page has no further functionality other than to offer situational awareness to avoid wasted time searching for interactive content. The welcome page contains a relevant background image, a meaningful descriptive title, a list of sponsors, and a clearly visible entry portal into the website.

The home page

The home page will provide a more detailed description of the website as well as contact information. The user should be able to easily understand the structure of the website at this point with clear options for proceeding into her/his areas of interest. As a frameset, the home page forms the structure for the entire site, with consistent navigation options within the side-frame. The top-frame has a title and relevant images for carrying a consistent aesthetic style throughout the site. Both frames will reassure the user that they are navigating through the 'CLOSiWEB' biological database with options for navigating. The navigation links within the side-frame are limited, keeping the page clear and easily understandable. The user can quickly see that there are three options for interacting with the website: search, enter data and modify data.

Database search

Following the first heuristic mentioned above, the site offers the user an overview of available data. It is important here to provide an overview of data without confusing the user with too much information. The search page for this prototype will present the user with four areas of search: Locality, Agents, References and Collecting Events. All of these are interrelated clusters and have unique cyclic navigational paths. One may be searching for locality information for a particular event and then wish to find all of the references and agents for that particular collecting event. This means that once a search has been completed, the user should be presented with options for continuing to search related data.

The structure of the search pages can be developed through presenting possible scenarios and ensuring that the page can meet those needs. For

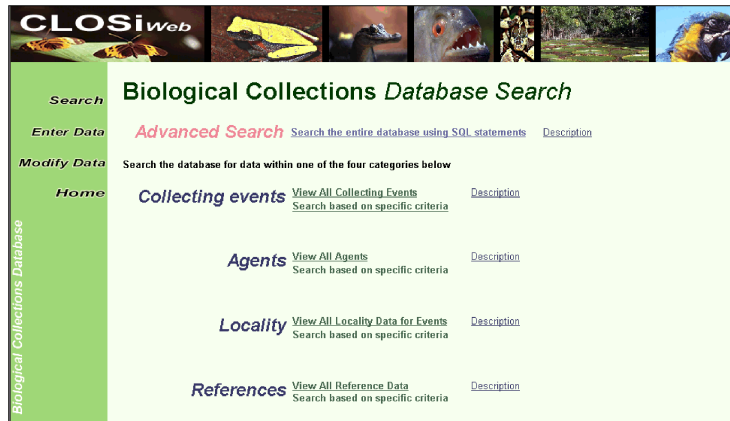


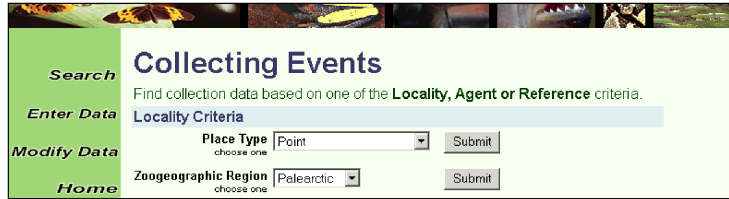
Figure 6.3: Screenshot of the biological collections database search page.

instance, what happens if a user wants to find location information? This is not the same as wanting to find information based on a locality. The user is interested in all locality information within the database, each record of which represents the locality of a collecting event. Therefore, to search for locality information the user is searching for either all of the locality information for all events, or he searches for locality information of a collecting event based on some criteria within any of the tables in the database. Because of this complexity, the search page should offer two additional search options other than simply displaying all locality data. There should be a predefined query search, which is a page made up of popular search options that are likely to help users find what they need, and secondly an advanced search option for entry of SQL statements. All of the four sections should have the same structure with an advanced search being independent of the other two search options within each of the four search categories. This structure is clearly displayed on the first search page seen in Figure 6.3.

The database search steps

Using HTML and PHP scripts to accept a request from a user and then retrieve data from the MySQL database is achieved by the following two steps:

- Create a form within a PHP script including HTML, which contains a variable that will take part in the query and give the form an action that will submit the variable value to the next page, which will be a



The screenshot shows a web interface for 'Collecting Events'. On the left is a vertical navigation menu with links: Search, Enter Data, Modify Data, and Home. The main content area has a title 'Collecting Events' and a subtitle 'Find collection data based on one of the Locality, Agent or Reference criteria.' Below this is a section titled 'Locality Criteria' with two dropdown menus: 'Place Type' (set to 'Point') and 'Zoogeographic Region' (set to 'Palearctic'). Each dropdown menu has a 'Submit' button next to it.

Figure 6.4: Screenshot of specified search page for collecting events.

PHP page containing the SQL statement. An example of a search form is the form for a specified query on collecting event information (see Figure 6.3).

- A PHP page follows and is called by the form action when the submit button is pressed. Within the PHP page there is a set procedure to follow in the script: (a) Define host and user; (b) Connect to the database; (c) Get variable passed from the form; (d) Define an SQL statement using that variable; (e) Loop through the results of the SQL; and (f) Print the results (usually within a table).

In the following, we describe the sequence of events that take place after these two steps:

- User enters a search based on a variable entered in an HTML form.
- The action of submitting the HTML form contains a call to run a PHP page. The Apache server recognises the PHP extension, finds and opens the page, then runs the script.
- The PHP page is retrieved. PHP makes sure it is relating to the right server and user and makes a connection with the MySQL DBMS. The PHP code contains a predefined query that retrieves information from the MySQL database based on the variable that was sent to the server in the HTML form.
- The PHP page sends back the results of the query to Apache as it loops through all the records that match the query statements.
- Apache retrieves the result of the query within a format specified in the PHP code and sends it back to the user's web browser as a static HTML page.
- User sees the result as in the format specified within the PHP code. Usually a table.

Figure 6.5 illustrates the events that produce a predefined query result. Any other search, whether it be an advanced SQL query or a search to re-

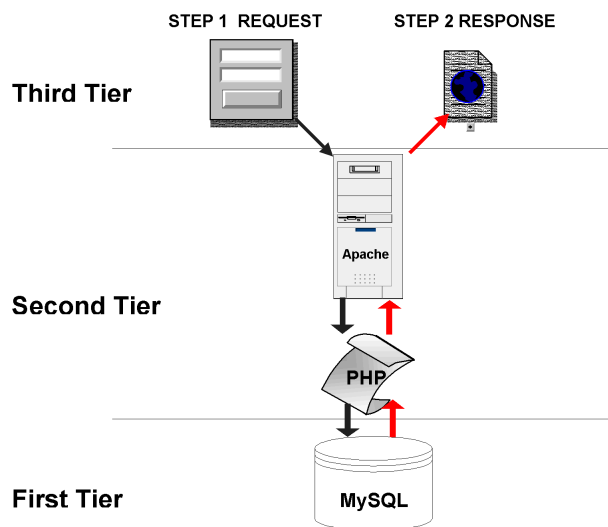


Figure 6.5: The database search sequence.

trieve all information in the table, will be developed in the same sequence but with different queries and submitted variables.

New data entry

The data entry section of the website will only be accessible to users who are authenticated through a login page. The login page allows agents (collection personnel) who collect new data to add them into the database while preventing data entry from unauthorised personnel or outside sources. As with the search pages, the interface heuristics mentioned earlier are important. The first data entry page should present the user with a clear set of data entry options that are available. The user should be able to enter all the required new data in a flowing manner. The CLOSi schema provides a good framework for such an interface mapping flow.

The flow of data entry (represented conceptually in Figure 6.2) begins with the option of entering data into the main tables of each of the clusters within CLOSi. New collecting events or new references are displayed to the user as data entry options. Users are prompted to enter new locality data after the new collecting event data is entered, since the locality information is dependent on the collecting event.

To enter a new agent, the user must fill out an online form and send it to

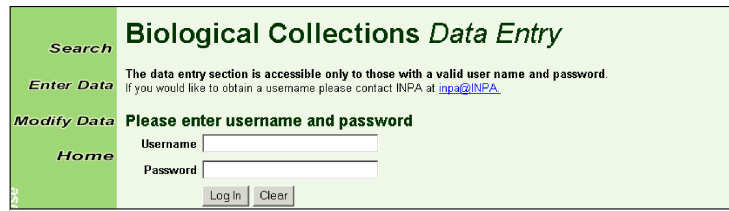


Figure 6.6: Screenshot of the Login page.

the database manager at INPA for authentication. If the user is approved, the appropriate tables will be appended and the new agent will receive a username, user ID and password, to be stored within the MySQL 'user' table for authentication. After the user enters data into the first level, they are presented with options for entering data into the remaining relevant tables.

Example of data entry flow

The login page is displayed for user authentication. The username and password are verified through a query of the 'user' table in the database. When new passwords are entered into the database they are transformed into a set of random numbers and letters that represent that password using PHP's 'MD5' function. This is a security measure to prevent people from seeing the passwords within the database. Figure 6.6 presents the screenshot of the login page.

The user is welcomed to the data entry section after a successful login. At this point, the PHP has started a session. A session ID is created by PHP to authenticate the user as s/he moves through the entry pages. Each data entry page is protected, and will not open unless the current session ID is provided, meaning that the page will only open for a logged in user. A user who is in a session has the opportunity to add a new reference or a new collecting event. A short description is offered for the user to make the choice.

After the user chooses the new collecting event option, s/he is presented with options for data entry (see Figure 6.8). When the data is entered and submitted, the EventID is created automatically as it is an auto-increment variable. The userID is passed from the login page, which is the same as the agent's ID, thereby allowing the database to know who is making the new data entry. The date and time of the new entry are recorded in the database by a PHP 'date' function call that obtains from the computer's internal clock the year, month, day, hour, minutes and seconds. After the new collecting

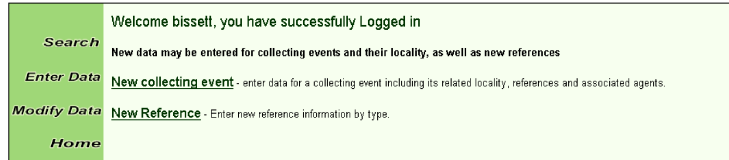


Figure 6.7: Authentication result page.

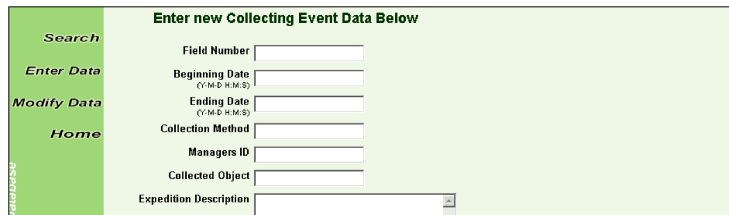


Figure 6.8: Data entry page.

event has been added, the new entry is displayed and more options for entry are offered, with the event and user IDs passed along to the next page.

After entering the collecting event data, the user has an option to go back and enter any reference data, or to proceed and enter locality data. The locality data entry page contains select lists and text areas for the user to enter variables into the appropriate tables via forms. If the collecting event is georeferenced, the type and source are chosen and the user proceeds to enter the specific data for that type (e.g., point, line or chain). A resulting page for the entry is displayed for each of the entry forms on the page with an option to enter more locality data that may be available (see Figure 6.9 and Figure 6.10).

This entry procedure will be the same for all of the tables. There is a



Figure 6.9: Data entry result and proceed prompt.

6.3. PROTOTYPE IMPLEMENTATION

Figure 6.10: Georeferenced object data entry.

The following Data has been entered for the Line

Event ID	Beginning Node Latitude coordinate	Beginning Node Latitude Hemisphere	Beginning Node Longitude Coordinate	Beginning Node Longitude Hemisphere	End Node Latitude coordinate	End Node Latitude Hemisphere	End Node Longitude Coordinate	End Node Longitude Hemisphere	Created	Updated	Updated By (Agent ID)
231	23	N	23	E	23	N	23	E	2002-08-15,02:02:53	2002-08-15,02:02:53	3

[Enter more locality data](#)
[Back to the Homepage](#)

Figure 6.11: Georeferenced object data entry result page.

hierarchy and a flow when entering the data following the CLOSi schema. There are clear navigational options and headings on each page to ensure situational awareness.

The data entry algorithm

The method for developing a data entry interface is composed of five steps:

1. A table must be created for the user names and passwords. The users who are given the authority to enter data will be the agents (persons or institutions). Therefore the userID will be the same as the AgentID. When each of the agents is entered into the table the userID and AgentID must be the same, and there cannot be repetition. When the passwords are entered, they must be entered into the database with the 'md5' function so that they are not visible to outside users.

2. The login form must be created to authenticate users. An HTML form with inputs for username and password is developed. This time the method is posted and the action will take the user to a page where they will begin a session.
3. In this step, a PHP page is developed to perform the task of checking the variables entered as username and password with the MySQL 'user' table. Also the page will begin a session and pass the session ID to the following pages.
4. Once the user has successfully logged in, a session has started and they have selected what data they will enter, a data entry form can be developed. The form will be made up of various input types and select types for the user to enter variables for entry. Many of the attributes will have new data entered automatically, such as the EventID (autoincrement), Created (MySQL date function), Updated (MySQL date function) and UpdatedBy (User ID).
5. Finally, once we have the variables from the entry form, we can pass them and run a query to insert them into the appropriate tables. In some cases, there must be a check that an entry of an EventID has not occurred in another table (there cannot be duplication of the same EventID when there is an 'ISA' relationship for example, i.e., point, line or chain). When the data is entered it is displayed to the user within a table. The user is also presented with the opportunity to return home or to enter further data.

6.4 Conclusions and What Comes Next

This chapter described the three-tiered client/server architecture and also highlighted the tools' main features used to implement such an architecture. The Apache web server, PHP server-side script and MySQL database management system were adopted as tools for this implementation, based on the benefits of being OSS, reliable and well supported. While OSS is free of charge, it should be clear that operational costs, such as training, would still be incurred.

The installation and configuration of the whole environment can be difficult due to the exact requirements for file paths and syntax. Though the tools are complex for non-programmers, the learning curve for web development using the tools is steep.

The development of the interface begins by having a relational database and a relational database begins with a schema identifying external needs. The CLOSi schema provides this structure and allows for an easy transition to the development of an entity-relationship conceptual schema.

Using the standard mapping algorithm described in Elmasri & Navathe (1994), the relational database was then produced from the extended entity-relationship model with cardinalities based on those described within the CLOSi schema. Once the relational database was developed, it was implemented within a physical RDBMS, in our case MySQL.

After developing the database, the interface was built. Using HTML and PHP, an interface that connects to the MySQL database and serves requests through the Apache server can be achieved. The basic construction of the prototype allows search and data entry with some security capabilities with much room for growth in terms of using the full capability of the Apache server and its configuration options.

A successful database-backed web system contains both simple and advanced queries, as well as browsing capabilities. The development of these queries can become rather complex. However, creating a well-designed query interface leads to an easier to understand system and provides effective access to data for the user. The prototype developed for this project has simple search and advanced search capabilities. However, there is much work to be carried out on improving both types of search, partly due to the lack of nested queries in the current version of the adopted MySQL and also due to the need for experience in developing advanced queries within PHP and MySQL.

The simple query interface will be modified over time as the users indicate what is the data that they are most interested in. The advanced query interface must allow the user to interact with and select from the various tables and attributes within the database to form queries based on selected operators. Within the prototype, the advanced search requires the user to know the SQL syntax as well as table and attribute names.

While a web interface can become very dynamic and complex (which PHP allows for), a basic system can also be produced. One of the most important aspects of even the simplest website are the heuristics for database-backed web interfaces. Functionality and interactivity can be utterly confusing if web interfaces do not follow the rules of simplicity, situational awareness and clarity. These heuristics can be more closely adhered to as the project grows and the interface matures with input from users. Within this implementation, it was found that maintaining these heuristics consistently is a challenge. Keeping a log of display settings and functionality for each page as well as an organised interface mapping flow chart proved to be essential.

The search and data entry code follows a sequence that should be understood and duplicated. Further development of algorithms for modification and other important functionality should be produced as the system is extended. Presently, there are no modification capabilities for deletion and

replacement of records.

Extending implementation

This prototype web interface pertains only to a portion of the database schema presented in CLOSi. For the full implementation of a CLOSi-based database there is much work to be done for each of the data entry, search, and data modification elements of the interface to include the remaining clusters, particularly taxonomy and collection management. There are many security features and efficiency improvements that can be made to the existing interface to protect and maintain this important set of data relating to biodiversity. While there may not be many outsiders interested in infiltrating a biological collections database, there still should be precautions taken, as it is valuable data in terms of sustainable development.

With the new version of MySQL, with cascading updates and deletes as well as nested query capabilities, the potential for the system will be improved and it should be modified to take advantage.

Advanced search

More work has to be performed on improving the advanced search capabilities of the website. For example, the user should be offered more capabilities in terms of choosing from existing tables and attributes within the database and then select operators for constructing queries as opposed to writing a plain SQL statement.

Data modification

Data modification will be in the form of deleting a record and entering a new updated one. The essential element of modification is allowing the user to first search for the row to be modified and then to have the ability to select, delete and enter new data into only that record. As with data entries, a login is required. Another table of users should be created, since some agents who may have permission to modify, may not have permission to add new data and vice versa.

The user should be able to search for the record or records through either an advanced query or through browsing the records. Once a selection is made on the records returned, new attribute values may be entered within HTML inputs. The development of this portion of the interface will be part of future work.

6.4. CONCLUSIONS AND WHAT COMES NEXT

Chapter 7

Georeferencing Amazonian Bio-Data *

7.1 Introduction


Biodiversity inventories and monitoring are adding new specimens, and their related information/data to biological collections. Specimens constitutes an estimated number of 2.5 billion items collected worldwide (Agosti *et al.*, 1999). Collections provide some of the most detailed information about the distribution of species in space and time, and are described on labels. Most specimen labels, indicate where the item was located, when it was collected and by whom. These species locations of occurrence can be combined with remotely sensed climate and landscape data for predictive distribution maps (Lawrence *et al.*, 1995).

Meanwhile, the goal is to provide online access to biological data catalogs. The bottleneck in the process of providing access is data entry, more specifically, the entry rate. Agosti *et al.*, (1999) stated that according to experiences in the United States and Canada, label information can be entered into a database at a rate of 10 to 40 specimens per hour. To produce the spatial and temporal data, the specimen labels must be readable. A label identification includes the site of collection, date, name of collector, methods utilised, ecological notes, etc.

Most of the available biological collections and the environmental experiment data are not consistently georeferenced making use of a coordinate

*This chapter is based on Campos dos Santos, J. L., Bissett, M. G. de By, R. A., and Amadio, S. A. (2003). Integrating a CLOSI-based Database with a Retrospective Georeferencing. To appear in the 5th International Symposium on Environmental Software Systems, Sammering, Austria, May 2003.

7.1. INTRODUCTION



Instituto Nacional de Pesquisas da Amazônia

Page No: 287

Expedition No: 028 Locality: *Balbina, Presidente Figueiredo, Amazonas, Brasil*

Hidrological Period: *SECA* Date (dd/mm/yyyy): *28/11/1984* Time (hh:mm:ss): *07:12*

Number:	Order	Family	Genus	Species
<i>24551</i>	<i>Clupeiformes</i>	<i>Pellonidae</i>	<i>Pellona</i>	<i>flavipinnis</i>
<i>24552</i>	<i>Characiformes</i>	<i>Curimatidae</i>	<i>Steindachneria</i>	<i>bimaculata</i>
<i>24553</i>	<i>Characiformes</i>	<i>Curimatidae</i>	<i>Steindachneria</i>	<i>bimaculata</i>
<i>24554</i>	<i>Characiformes</i>	<i>Curimatidae</i>	<i>Psectrogaster</i>	<i>rufiloides</i>
<i>24555</i>	<i>Characiformes</i>	<i>Curimatidae</i>	<i>Psectrogaster</i>	<i>rufiloides</i>

Figure 7.1: Example of INPA's fish expedition note.

system e.g., latitude, longitude or projected coordinates, for instance using the Universal Transverse Mercator — UTM, and consequently the spatial position of the collection site cannot be identified (Huxhold & Levinsohn, 1995; Kerschberg *et al.*, 1996). Instead, the locations are expressed in a free textual format, such as 'Balbina in Presidente Figueiredo, Amazonas state, Brasil' or '5 km North of Rio Negro', as a method of geographic description. Figure 7.1 illustrates an expedition note with data from an event of collection, including its locality description. Over time, place names change and collectors describe locations with different degrees of accuracy. This way of describing locations makes the data collected less useful to GIS tools or spatial analysis software. Therefore, georeferencing the specimens is one of the greatest challenges (cross reference and future analysis of passed present and future collections).

More than a hundred years of sampling activities in the Brazilian Amazon have generated a large volume of non-geospatial data and only recently became the use of global positioning system (GPS) available and affordable. GPS is at the moment widely applied in natural resource data collection for a variety of applications and has proven to be a very suitable tool (Tsui, 2000). Thus, the question to be asked is 'how can one bring this legacy data into a new life format and capacitate it to be further explored with geospatial analysis?' To answer this question, we start by pointing out a motivation fact that geographical data is a critical feature of biodiversity. To integrate GIS databases in a federated environment will provide useful facilities for scientists to georeference their data sets. Georeferencing in short, is the process that assigns coordinates and error estimations to a locality. It is

also called geocoding.

There are three elements or dimensions to geographic information: location, time and attribute. For a georeferencing system, there are three requirements: to be unique, commonly shared, and persistence through time (Aldenderfer & Maschner, 1996).

Clarke (1995) classifies georeference systems in two types:

1. **Nominal** — There are five methods of nominal georeferencing: place-names, postal addresses, postal codes, linear referencing systems, and cadasters. The methods are limited due to lack of uniqueness, not commonly shared and change over time. For example, placenames are limited because they are rarely unique, not commonly shared and change over time. Postal addresses and postal codes are more commonly shared but are limited to buildings.
2. **Metric** — This method, use geospatial coordinates (e.g., latitude and longitude), are more powerful as they have the potential of high spatial accuracy, and permit calculation of distance and area.

The common practice adopted by researchers to generate a distribution map is to compile a site list of the collection referenced for approximate positions, and then, manually plot the distribution map. The main concern with this method is that it is time- and labour-intensive (often requiring interpretations of old and imprecise textual locality descriptions into geospatial values). This is a difficult to maintain method in which the quality of the final product is difficult to guarantee (Cowen, 1997).

The conversion of a vague locality description into a geospatial value produces a false precision, that, in many cases, has proven to be unacceptable due to the need of accurate coordinates (Proctor *et al.*, 2001). There are several sources of uncertainty in locality descriptions. These uncertainties vary in magnitude as well as in their interactions with each other. For this reason, it is important to determine and record uncertainties as a single measurement (a maximum error) of the geographic coordinate determination (Wieczorek, 2002b). The method of deriving geospatial information (coordinates) from textual locality description, together with the association of its respective maximum error distance for those coordinates, is known as retrospective georeferencing or geocoding (Proctor *et al.*, 2001).

In this chapter, we present georeferencing topics regarding methods to describe coordinates, for named places, their offsets and imprecision. Further, we present notions associated with uncertainties in distance and direction. The process of collaborative work available in MaNIS (The Mammal Network Information) gazetteer is also presented together with the benefit it can bring (Wieczorek, 2002b). We report on the integration of a retrospective georeferencing technique using a fish dataset by applying a tool to

a CLOSi-based database. The tool is named CAS, and was developed by California Academy of Science as an extension to ESRI's ArcView software. We comment on the pros and cons of this method and its integration with a CLOSi database. The chapter closes by presenting our conclusions.

7.2 Georeferencing Process

The process of georeferencing can be complex, however, this complexity can be reduced considerably and the consistency of the results can be improved if supported by a comprehensive list of common locality descriptions. These localities present several degrees of uncertainty that vary in magnitude and in the interactions with each other (Wieczorek, 2002b). The process of georeferencing requires the determination and recording of the existing uncertainty.

7.2.1 Expressing Latitude and Longitude

Geographic coordinates can be expressed using different systems (e.g., decimal degrees, degrees/minutes/seconds, degrees/decimal/minutes, or UTM, etc). Conversions can be made, but decimal degrees are the most convenient coordinates, since they permit to describe a locality with two attributes, that is, decimal latitude and longitude. To assign geographic coordinates consists of describing the location by using named places, offsets, or/and vagueness identifications. These are defined and illustrated with examples below:

- **Named Places** — The common location identification consists of only a named place. Gazetteers can provide bounding boxes to define the extents of large areas. For this we should adopt the coordinates of the named place, and the furthest point within the named place as the error distance.

Example: **'Itacoatiara'**

- **Offsets** — An offset is a combination of a distance and direction from a named place. Sometimes the locality identification provide the indication for determining the offset. For instance, 'by road', 'by river', 'by air', 'up valley', etc. In the following we present examples of offsets:

Example: **'5 km S (by river) from Itacoatiara'**

If localities are represented by two orthogonal coordinates, by default, the measures considered are 'by air'.

Example: **'5 km S and 3 km W of Itacoatiara'**

In case the distance provided is along a linear feature such as a road or river, this does not present direction imprecision.

Example: **‘20 km W (by road) from Itacoatiara’**

For rivers, the notions of ‘left’ and ‘right’ bank’ follow the principle of facing toward the source.

Example: **‘left bank of the Rio Negro, 12 km upstream from the Anavilhanas archipelago’**

Locations that are expressed with one linear offset value from a named place and that do not describe the way the measurement was taken, should be analysed individually.

Example: **‘112 km N of Manaus’**

The decision should be based on how the expedition travelled (e.g., ‘by road’ on BR 174, Manaus Boa Vista, RR).

- **Vagueness** — Georeferencing locations that express vagueness should be avoided.

Example: **‘Itacoatiara?’**

It is common to find location described with an offset from a named place without conclusive directions or distances. For such a case, we should adopt the geographic centre of the named place as the geographic coordinates. As the error value, we should adopt the distance to the nearest named place.

Example: **‘near Itacoatiara’**

Offset information can be vague in its direction or in its distance. When the direction information is vague, we should adopt the geographic centre of the named place. The offset distance will be the error distance.

Example: **‘10 km from Itacoatiara’**

Most of localities are recorded without error estimates. For the case of a personal observation (e.g., about) will not affect the identification of a named place or the maximum error.

Example: **‘about 10 km from Itacoatiara’**, the vagueness should be ignored.

One common source of inconsistency in location description comes from the use of elevation information together with the rest of the description.

Example: **‘10 km from Boa Vista, 1400 mt’**.

In this example, there is no place at this location at that elevation.

Another common source of inconsistency occur when the description is inconsistent with the geopolitical area of which it is expected to be part of.

Example: **‘Humaitá, Rondônia state’** (Humaitá is in Amazonas state).

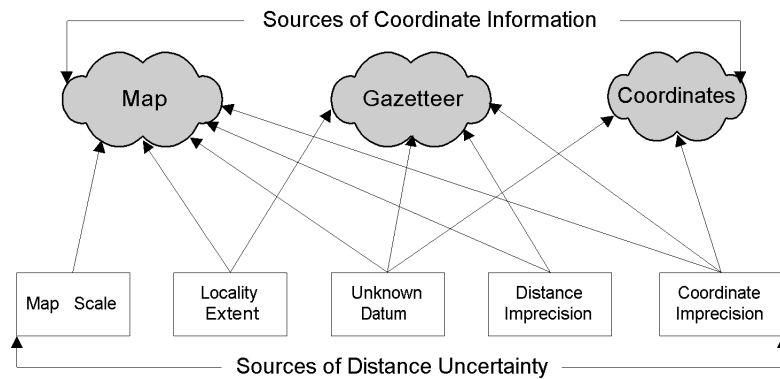


Figure 7.2: Sources of uncertainty.

7.2.2 Dealing with distance and direction uncertainties

In georeferencing, the three basic sources of coordinate information are: maps, gazetteers, and localities recorded with coordinates. As part of the georeferencing process, a verification of the uncertainties in geographic coordinate determinations must be undertaken. Uncertainties can occur from the combinations of the extent of a locality, unknown datum,¹ imprecision in distance, coordinate and direction measurements, and map scale.

The type of uncertainty that may apply for each of the sources of coordinate information are presented in Figure 7.2 and can be understood as follows:

1. **Map** — locality extent, unknown datum, distance imprecision, map scale, and coordinate imprecision. All of these must be all summed up before calculating their combination with direction imprecision.
2. **Gazetteer** — locality extent, unknown datum, distance imprecision, map scale, and coordinate imprecision. They must be summed before their interaction with direction imprecision.
3. **Coordinates** — unknown datum, and coordinate imprecision. They must be summed to get the maximum distance error.

¹A datum is a mathematical model that describes the shape of the ellipsoid. Around the Earth the flattening is not uniform. The shape differs because processes of plate tectonics and there are datums for different parts of the Earth based on different measurements (Snyder, 1982).

Automatic calculations of uncertainty

The CAS tool calculates uncertainties (span measures) from shapes of an area during the process of georeferencing. The MaNIS georeferencing model supports the implementation of the concept of uncertainty, where each source of uncertainty has its specific method of calculating the maximum distance error. A comprehensive description with examples is given in Wieczorek (2002). The methods are described and embedded in the georeferencing error calculator (GEC) tool (Wieczorek, 2002a). The tool is a Java applet that aids in the georeferencing of descriptive localities. The system requires a web browser (Internet Explorer or Netscape Navigator) and runs on Windows, Unix, or PC emulator on the Macintosh operating system.

7.3 Understanding Vagueness and Uncertainties in Distance and Direction

Span measurements can be used to classify locality into more or less vague and help to select localities with appropriate precision (coordinates might not be appropriate for a local map, but are important at national scale).

When not considering the shape of a region, another way to calculate uncertainty depends on to distance, direction or their combination. Distance can be expressed together with the locality description (e.g., 20.6 km). One way to avoid over estimate distance precision is to adopt the distance measurement as integers (e.g., 20 km), otherwise the remainders are treated as fractions - (20.25 becomes $20 \frac{1}{4}$; 20.5 becomes $20 \frac{1}{2}$; 20.6 becomes $20 \frac{6}{10}$; 20.75 becomes $20 \frac{3}{4}$). When distance precision is expressed as integer the distance uncertainty suggested is equal to 1.

Direction is often expressed in locality descriptions (cardinal or intercardinal direction). Such a description is considered incomplete when expressed only 'N' (north). This could be in the direction from 'NW' (northwest) to 'NE' (northeast). The directional uncertainty in this case is 45 degrees between 'NW' or 'NE' (e.g., 20 km N of Manaus).

When a locality description is expressed using two orthogonal directions (e.g., 20 km N and 10 km E of Manaus), the measurement can be made on a map. In this case, the directional uncertainty value can be ignored.

For a locality description expressed with a more specific direction than the cardinal direction 'NE', 'NE' could mean any direction between 'ENE' and 'NNE', which is more precise than when expressing 'N' for instance. The directional uncertainty in this case is 22.5 degrees between 'NNE' and 'ENE' (e.g., 20 km NE of Manaus). Figure 7.3 shows the directional uncertainty of 22.5 degrees.

7.3. UNDERSTANDING VAGUENESS AND UNCERTAINTIES IN DISTANCE AND DIRECTION

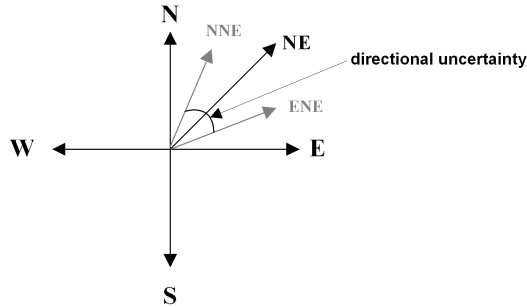


Figure 7.3: Uncertainty in distance precision.

When recording geographic coordinates, it is recommended to keep sufficient precision. The calculated value of coordinate uncertainty is a function of the recorded value as well as of the datum and the coordinate. This is because the degree does not correspond to the same distance elsewhere on the surface of the Earth. The formulas used to calculate coordinate uncertainty were modified after Wieczorek (2002), and are as follows:

$$R = \frac{A}{\sqrt{1 - F^2 \sin^2(\varphi)}}$$

where:

R is the radius of curvature in the prime vertical at the given latitude,

A is the semi-major axis of the reference ellipsoid - the radius at the equator,

F is the first eccentricity of the reference ellipsoid, and

φ is the given latitude.

The geodetic latitude γ can be computed as

$$\gamma = R * \cos(\varphi).$$

First eccentricity is expressed as

$$F = 2 * f - f^2$$

where f is the flattening of the reference ellipsoid.

$$M = \frac{A(1 - F^2)}{(1 - F^2 \sin^2(\varphi))^{2/3}}$$

M is the radius of curvature of the meridian at the given latitude.

$$L1 = \Pi * M * \frac{P}{180}$$

$$L1 = \Pi * \gamma * \frac{P}{180}$$

where:

$L1$ is the Latitude error.

$L2$ is the Longitude error.

P is the latitude longitude precision using the WGS84² reference ellipsoid.

The coordinate uncertainty (U) is:

$$U = \sqrt{(L1^2 + L2^2)}.$$

When more than one direction is used to describe location, uncertainty applies for both cardinal directions and the combination is non-linear. For example: '6 km E and 8 km N of Manaus'. In this case, the description of uncertainty is a bounding box centred on the position 6 km E and 8 km N of Manaus (see Figure 7.4. Each side of the box is 2 km in length (1 km uncertainty in each cardinal direction from the centre). To define the maximum error it is necessary to consider the circle that circumscribes the bounding box.

Another method to calculate uncertainty may involve the shape of the named place. It determines the furthest distance within the named place from the geographic centre of the named place in either of the two cardinal directions mentioned in the locality description. Add this distance to the distance precision and take the square root of 2 times the sum to get the maximum error distance associated with the combination of distance precision and the extent of the named place (see Figure 7.5). Suppose the

²WGS 84 is an Earth global reference frame, including an earth model. It is defined by a set of primary and secondary parameters: the primary parameters define the shape of an earth ellipsoid, its angular velocity, and the earth mass which is included in the ellipsoid reference the secondary parameters define a detailed gravity model of the earth. These additional parameters are needed because WGS 84 is used not only for defining coordinates in surveying, but, for example, also for determining the orbits of GPS navigation satellites.

7.3. UNDERSTANDING VAGUENESS AND UNCERTAINTIES IN DISTANCE AND DIRECTION

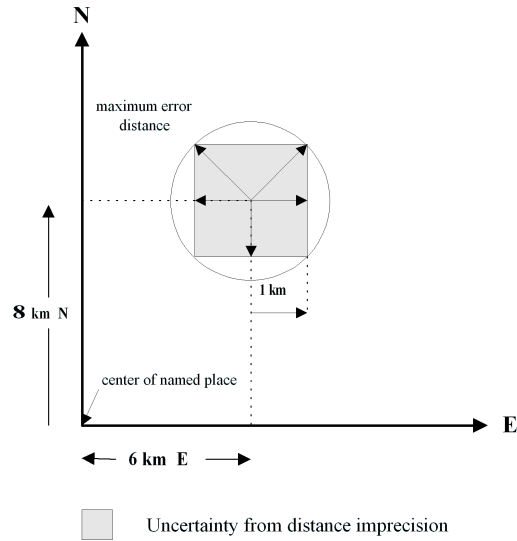


Figure 7.4: Combinations: distances imprecision – After Wieczorek (2002).

furthest extent of the city limits either N or E from the geographic centre is 3 km. There will be a total of 4 km of uncertainty in each of the two directions and the radius of the circumscribing circle would be 4 km times the square root of 2, or a total of 5.657 km.

Distance imprecision in a given direction is linear and additive. The sum of them increases non-linearly the error from directional imprecision. A method is used to account for the correlation between the uncertainties from distance and direction. For example, '10 km NE of Manaus.

Leaving the distance imprecision out, the uncertainty due to direction imprecision is encompassed by an arc (d) from the centre of the locality (Manaus) at coordinate (x, y) at a direction of 45 degrees (θ), extending 22.5 degrees in two directions from that point. The distance (e) from the centre of the arc to the end of the arc (x', y') at directions of 22.5 degree (θ') from the centre of Manaus can be calculated by applying the Pythagorean Theorem:

$$e = \sqrt{(x' - x)^2 + (y' - y)^2}$$

where:

$$x = d * \cos(\theta), \quad y = d * \sin(\theta), \quad x' = d * \cos(\theta'), \quad y' = d * \sin(\theta')$$

The error calculated is 3.90 km. This example can be followed in Figure 7.6.

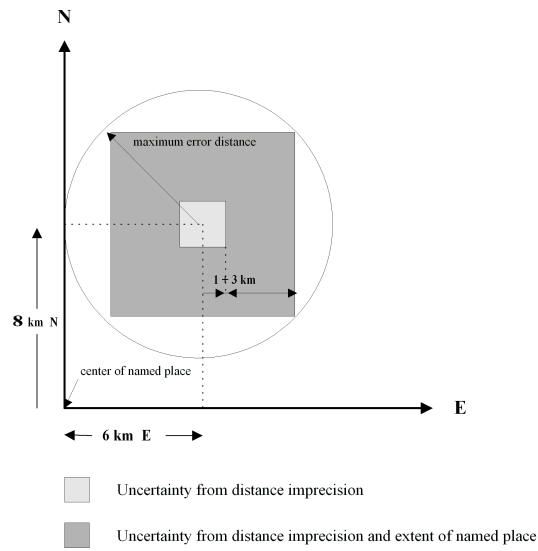


Figure 7.5: Combinations: distance imprecision with named placed – After Wiczorek (2002).

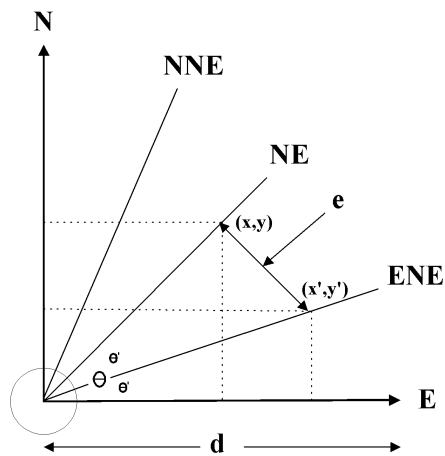


Figure 7.6: Combinations: distance and direction – After Wiczorek (2002).

7.3. UNDERSTANDING VAGUENESS AND UNCERTAINTIES IN DISTANCE AND DIRECTION

7.3.1 Example of distance imprecision

Assuming the distance uncertainties are 3 km (extent of Manaus), 1 km (distance imprecision for '10 km'), 0.079 Unknown datum, coordinates are from the Geographic Names Information System (GNIS³) database, and 0,040 km (gazetteer data) for a sum of 4.119 km.

The error region will be a band twice this width ($2 * 4.119 = 8.238$) km centred on 10 km offset arc spanning 22.5 degree on either side of 45 degrees.

Figure 7.7 details the sum of the distance uncertainties (d'). The error distance uses the Pythagorean Theorem and the parameters are:

$$x = d * \cos(\theta), \quad y = d * \sin(\theta), \quad x' = (d + d')\cos(\theta'), \quad y' = (d + d')\sin(\theta').$$

where: d' is the sum of the distance uncertainties.

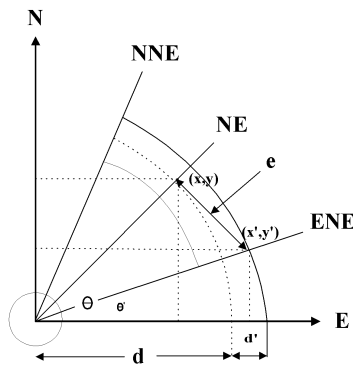


Figure 7.7: Combinations: The sum of distance and direction (d') – Adapted from Wiczorek (2002).

By generalising the geometry, rotating the image that the coordinate (x', y') is on the x axis (see Figure 7.8). The parameters are:

$$x = d * \cos(\alpha), \quad y = d * \sin(\alpha), \quad x' = (d + d') \quad y' = 0.$$

where α is an angle that represents the magnitude of the direction uncertainty. For this example, the distance uncertainty is 4.119 km and the

³GNIS was developed by the USGS in cooperation with the U.S. Board on Geographic Names (BGN), contains information about almost 2 million physical and cultural geographic features in the United States (Information, 2000).

direction uncertainty is 22.5 degrees. The final maximum error distance is 6.210 km.

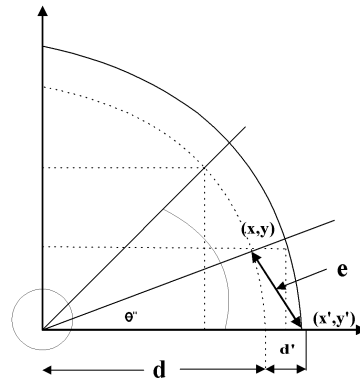


Figure 7.8: Combinations: generalisation distance and direction – Adapted from Wiczorek (2002).

The methods described above were implemented in the GEC tool and made available as utility for collaborative georeferencing works (Wiczorek, 2002a).

7.4 Experiencing Retrospective Georeferencing for Amazonian Data

Within the CLOSi database lies the ability to store georeferenced data. However, much of the legacy data, while not georeferenced, must still be entered into the database.

For these non-georeferenced entries, there is functionality for extracting data, transferring them to an external tool, georeferencing and then returning to the database, for updating the appropriate attributes. For these non-georeferenced entries there exists a functionality for data extraction and transferring to an external GIS environment. Here the retrospective georeferencing can be applied after which the entry can be returned to the database included the updated attributes. George Chaplin developed an ArcView extension to take non-georeferenced data and produce geo-coordinates as well as an accuracy based on ‘allowing the user to search and sort a database of localities, quickly locate a place on a base map, and draw shapes to

represent locality descriptions' (Proctor *et al.*, 2001). We followed the same three-step process recommended by CAS system:

1. **Preparing the specimen database** — The database must be inspected for existing locality variations and its records must be standardised. The locality description can have embedded redundancy. Some redundancy may have valid justification, for example, commonness, many specimens are collected at one site, thus the same place name is associated with each specimen. But, redundancy can also mean trouble: the adoption of different wording used by different collectors, inconsistency in abbreviations, old place names, or typographic errors. In this case, the locality descriptions are used to refer to the same place. We call this phenomena false duplication. The number of records with such redundancy can be high. Tests performed in the herpetological collection in the CAS database indicate that 78% of the localities descriptions were classified as having false duplication (Proctor *et al.*, 2001).

The subsequent inspection attempts to remove redundancy caused by 'false duplication'. This will decrease the amount of georeferencing required as many collecting events occur in the same place which may have been recorded in a different manner in a database (i.e., there might be typing errors, subjective naming, etc.) Despite the normalisation process being time-consuming and laborious, the georeferencing is faster because of the removal of false duplications and searching the database becomes efficient because of standardisation of the database.) The types of inconsistency that increase false duplication of locality records are:

(a) Inconsistent interpretation of original handwritten locality

'Igarapé Tarumã'

'Igarapé do Rio Tarumã'

Handwritten as:

'Ig. R. Taruma., Manaus, AM, BR'

(b) Different wording of the same locality

'Manaus-AM, Universidade Federal do Amazonas'

'Universidade Federal do Amazonas Manaus AM'

(c) Different ordering of the same locality

'INPA, Manaus AM'

'Manaus AM, INPA'

(d) Inconsistency of form — capitalisation, abbreviation, preposition, and punctuation

‘junction/jct/Jct/jct./jctn...etc’
‘highway/hwy/Hwy’
‘Rodovia/Rod./Rod’

‘20 km N of Manaus’
‘20 KM N Manaus’

‘near 3-Hearts’
‘near Three-Hearts’

(e) Typographical errors and misspellings

‘60 86 9.2 N 130 68 24.7 E’
‘60 60 9.2 N 130 68 24.7 E’

‘Flutuante do Carreiro’
‘Flutuante no Carreiro’

‘AM010 na Reserva Ducke’
‘AM010 Reserva Adolpho Ducke’

(f) Inconsistent representation of precision

‘2 vs. 2.0’

(g) Inclusion of micro-habitat as location data

‘5 km N of Reserva da Campina’
‘5 km N of Reserva da Campina, deforestation’

(h) Other typographical inconsistency

‘5 km N of Reserva da Campina (on BR174)’
‘5 km N of Reserva da Campina [on BR174]’

- 2. Georeferencing** — The process of georeferencing involves a dataset (legacy data) and the preparation of base layer data (base maps). These are used to find geographical areas of a collecting event, after which locations are digitised as point, lines or polygons with of course coordinates and accuracy calculated. The accuracy value will determine the precision of location of the collecting event. Digital base maps to be used must be georeferenced and usable in ArcView. It should depict one’s geographical area of interest with enough detail and text to visually identify features and places. Despite the Digital Chart of the World (DCW) or other international providers, biological collectors

7.4. EXPERIENCING RETROSPECTIVE GEOREFERENCING FOR AMAZONIAN DATA

need a much better base maps data. To depict this it is recommended one refer to local national organisations for acquiring base mapas or to adopt a specific georeferencing system. Usually base maps can be available online and can be downloaded, converted, if necessary, and saved.

To access the dataset that will be georeferenced, a logon prompt requires the user identification. This information together with a session timestamp (logon data and time) will be recorded and will record in the database and metadata.

Functions for interactive geocoding are made available for associated base maps to data or zoom steps to find locality on the base maps. When a selection is made, a list of candidates for geocoding becomes available for the best matching place name that will automatically zoom to the position on the base map.

The point selected is identified in the base map. If the location description expresses an offset, a circle is draw with the radius representing this offset. If there is a direction in the description, a second point identifies the direction. This locality can be re-shaped using options and tools for drawing shapes (point, line, circle, and polygon), using a base map for reference, to represent that locality.

- 3. Integrating the new data** — Once the locations are georeferenced, the tables within the database, from which the data was originally extracted, are updated with the new geo-coordinates and span information. The attributes in the tables include: Place index as primary key of the normalised locality, logname, the name of user, X_coord and Y_coord, the latitude and longitude of the centroid of the finalised shape (in decimal degrees), and span, the length of the longest distance across the finalised shape (in meters). Span measures express vagueness of a locality – the more vague the description, the larger the span. Also, any additional geographical information that may not have been within the extracted data will be instantiated, such as town, state, or elevation, etc.

The system also creates an update table, in tab-delimited format, to store any edits suggested by the user. This table can be inspected and used as a batch update on the original collection database.

We used the database that contains 'Event of Collection' for fish species that took place in the region of Presidente Figueiredo, Amazonas, Brasil, during expeditions between November 1984 and November 1985. To accomplish the CAS steps in the context of INPA's biological collection database, four tables were created: EVENT (details the place of the event and the collector), TAXA (describes the taxa and the place of the occurrence), SDO

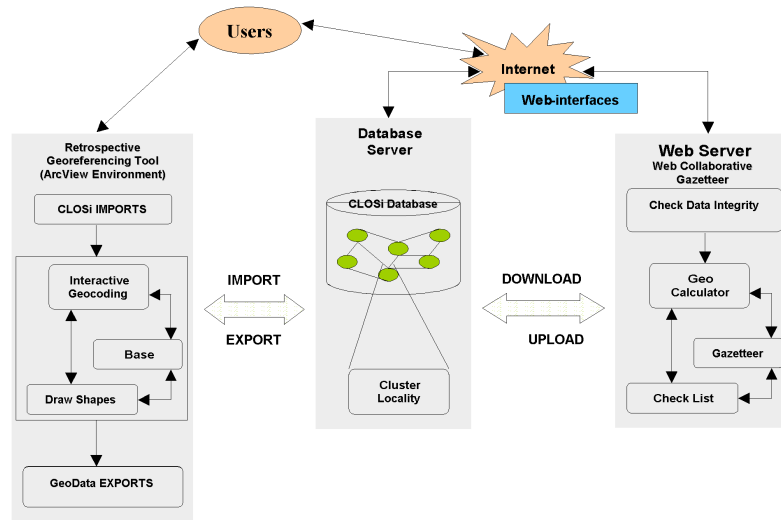


Figure 7.9: Georeferencing process and collaborative gazetteer within the CLOSi web context.

(information about the spatial data objects) and SDOSTATUS (maintains the status of each place regarding georeferencing). These tables are necessary to conduct the georeferencing in the manner required by the CAS tool. Data available within the CLOSi-based database was used to populate some of the attributes of these required tables, and those attributes with many null attributes (the Spatial Data Object — SDO table, for instance) is to be populated when the georeferenced data is transferred to the database.

Figure 7.9 represents the process of integration of our CLOSi-database with the CAS application and a collaborative gazetteer. The georeferencing process within the biological database and web interface context is described as follows:

1. When inputting new data into the database, there are many types of locality data to enter. Once all the locality data are entered, the user is presented with the option to georeference the data that activates PHP code (see Figure 7.10). This operation will extract all relevant data from the biological collections database into the CAS environment to prepare for georeferencing. This is done based on a query, which retrieves any data within the database that is not within the georeferenced object class, and then inserts into the SDO table specified fields, neces-

7.4. EXPERIENCING RETROSPECTIVE GEOREFERENCING FOR AMAZONIAN DATA

sary for georeferencing. Other tables that will be extracted related to the georeferenced object include a TAXA table, an EVENT table and a table to track status. The SDO table will have a PlaceIndex attribute that will hold a unique identifier for all of the events to become georeferenced through a link to the newly georeferenced place. The data redundancy problems must be solved prior to data entry.



Figure 7.10: CLOSi web interface with georeferencing tool.

2. Once the PHP code has extracted all of the necessary data from the biological database, it then needs to be converted to the MySQL. Figure 7.11 shows a snapshot of the data exported to the CAS application.
3. The SDO table is opened within the ArcView environment with the CAS extension on, as presented in Figure 6. The shapefiles available in the system for the region specified within the SDO table are utilized to digitize a location (see Figure 7.12). Using the extension a selection is made from one of the records in the extracted SDO table. The place name available from the SDO table is used to find an area within the available shapefile layers and create a point line or polygon for the events locations based on any information available within the SDO table. An 'x' and 'y' coordinate as well as accuracy are retrieved for the digitised locations and automatically entered into the SDO table.
4. The SDOout table is then uploaded into the MySQL database structure with the data now having geo-information. At this point any event within the legacy data that occurred in that place and at the same collecting event will be georeferenced. It is not necessary to georeference all the events if the place they are located is the same as one that already has been georeferenced.
5. After localities have being georeferenced, export the complete set of data from the CLOSi-database to a tab-delimited text file. This file

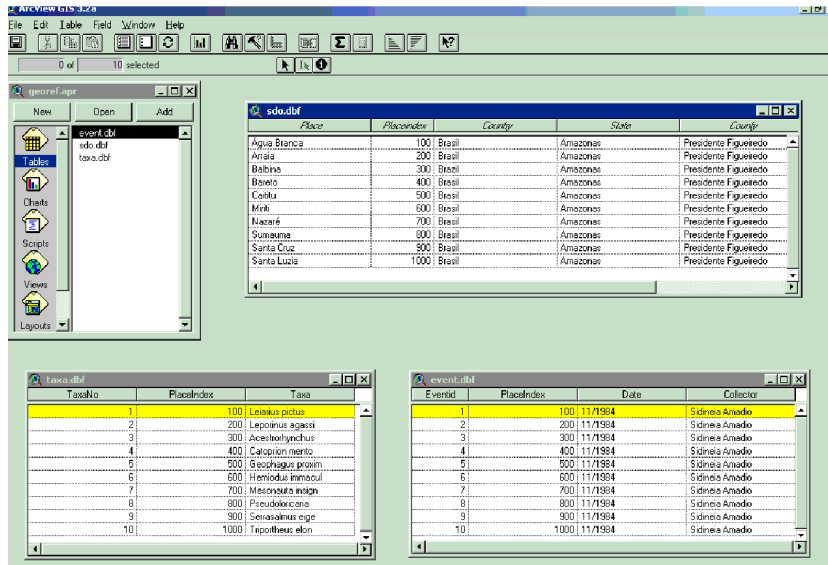


Figure 7.11: CLOSi data exported to the CAS system.

should be uploaded by anonymous FTP. The file will be checked for data integrity and loaded into the MaNIS.

7.5 Potential for Collaborative Gazetteer

The United Nations system definitions use names specifically as part of their gazetteer definition. The standard ISO 211 states that a gazetteer is a directory of instances of a class or classes of real world phenomena, containing some information about their locality (Hill *et al.*, 1999). The concepts embedded within the MaNIS gazetteer focus on the build-up of locality data retrieved from databases of collaborating institutions. In such an approach, collaborators need to follow a consistent methodology to get access to a georeferenced area of interest or to contribute by submitting locality information from areas that have not been georeferenced before (see also Figure 7.9).

Their solution is supported by web technology and client/server applications. A user can access an online checklist of localities already georeferenced or assigned for georeferencing. From this, the users access names of geographical locations, who was responsible for georeferencing, when it was started, how many records comprises the location georeferenced (the record

7.5. POTENTIAL FOR COLLABORATIVE GAZETTEER

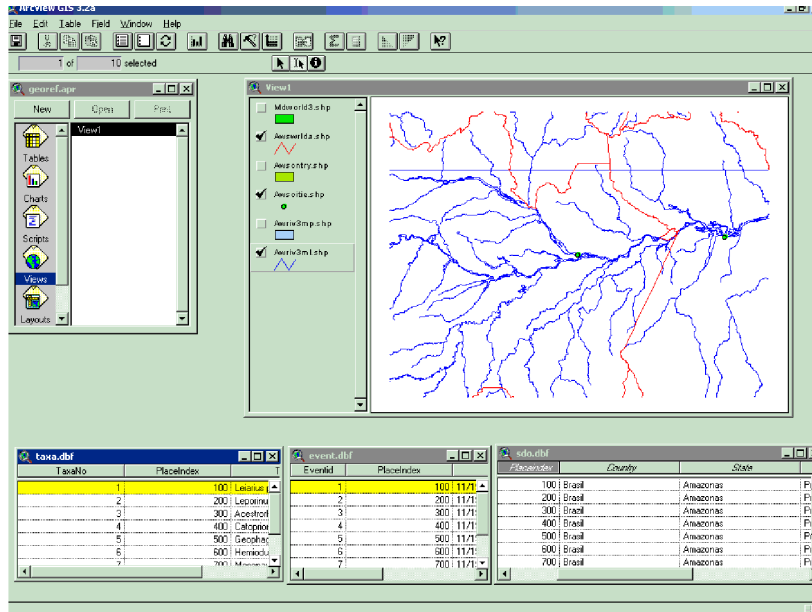


Figure 7.12: View based on Amazonian shape files.

count is the number of localities that have the given value in a respective geographical area), and the status (complete or in progress) can be derived.

Contributors can commit themselves to georeference a geographic area for the MaNIS list, by communicating to the gazetteer manager. The framework for collaborative georeferencing includes: import of locality information from the MaNIS database, georeference new localities, export newly georeferenced localities, and upload to the MaNIS infrastructure.

The process for downloading localities begins with a search using an online form to build and submit queries. This allows to narrow down the list of localities one is interested in. Once a locality is selected, it can be downloaded and imported to any spreadsheet or database system at the client side. An Access database template is also provided for downloading, in which data can be imported. For any other type of destination for the imports, a tab-delimited text file should be adopted and the column order and structure of the original file should be preserved.

The georeferencing process requires the use of a calculator (the GEC tool) to determine the maximum error distance for each locality as the coordinates are determined. The complete description of data fields are presented in Wieczorek (2002).

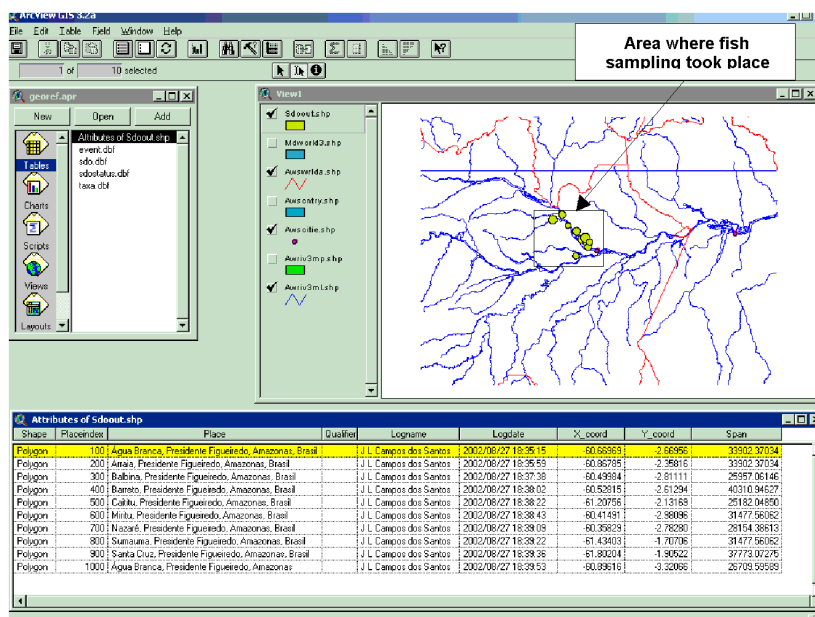


Figure 7.13: Localities georeferenced in Presidente Figueiredo, Amazonas, Brazil displaying the area where fish were catch.

7.6 Appraisal of the Pros and Cons

The process of retrospective georeferencing is relatively new and there is no standardised way to execute it. The entire process is laborious, requiring interpretations of old, imprecise locality descriptions into Cartesian coordinates. The mapping of a vague locality description into a single location creates the problem of false precision, sometimes to a scientifically unacceptable degree. The good thing is that such a process only needs to be done once and there are sufficient functions for modification, import/export and batch updates to the original data.

Tools: proprietary vs. open source software

One disadvantage of the CAS tool is the use of terms and concepts with the assumption that the users do know desktop ArcView GIS software. In the same line, base maps to be used during the georeferencing process may need to be converted to ArcView format. Further, ArcView GIS is a proprietary software. However, many argue that open source software (OSS) is more

appropriate for the technology realities in developing countries, our main target community. A good option to avoid proprietary solutions is to extend freeware similar software like Spring. Spring is a mature system running under UNIX, Linux and Windows (Cámara *et al.*, 1996).

7.6.1 Improvements from previous methods

This method of retrospective georeferencing has provided a step forward to help to speed the process of georeferencing. Some noticed improvements include:

- The integration of place-name look-ups with a variety of spatial measuring tools to help the replacement of localities.
- The way for unexperienced users to represent localities as two dimensional shapes (circles, buffered lines, or polygons).
- The functionality allows quantification of the precision of localities for non-well refined direction/distance.
- The capture of metadata to document the georeferencing process.
- The increase of the overall speed of the georeferencing process.

After a georeferencing session, information about the session is recorded, including the date, time, user's logon identification, together with additional information about the tool project file, i.e., filename, projection, datum, and reference themes. These recorded information is transformed in a XML document and stored in a separated table in the CAS environment. We consider this feature useful interesting, because a CLOSi database is integrated to a bio-metadata management. We can enrich the bio-metadata by exporting the table to our XML repository, which is associated to the CLOSi database.

7.6.2 Foreseen future

We expect advancements in the following topics:

- Optimisation of the quantity and quality of data for spatial analysis.
- Maximisation of precision, thus will empower means for answering research questions. Consequently, it will enable determination of critical habitat requirements for species.
- The capabilities to retrieve maps and analyse collections data are growing, creating a large demand for georeferenced data.
- Analytical tools represents a more specialised class of users for georeferenced collection data.
- Development of distributed query systems for collections database.

7.7 Conclusions

The CLOSi-based biological database web interface implementation utilises a three-tiered client/server architecture (see Chapter 6) integrated with a retrospective georeferencing ArcView extension. The CAS tool used for retrospective georeferencing was found to require a PHP converter to allow for the extraction of data from MySQL into CAS environment. The integration with the georeference tool was found to be a successful and valuable addition to the web system that provides great benefit to researchers of biological collections that need geospatial results. The advantages of this process can include: increasing speed in georeferencing, maximising consistency between users, allowing the incorporation of interpretation standards established by researchers/curators and quantifying textual locality vagueness.

Further work envisaged includes testing the integration of CLOSi-based database with the MaNIS collaborative gazetteer. Amazonian datasets georeferenced by retrospective process can integrate the MaNIS catalogue as well as that other geopolitical regions already available in MaNIS Checklist, can help INPA to accelerate the necessary and laborious process of reviving legacy data.

7.7. CONCLUSIONS

Chapter 8

Automatic Reconciliation of Taxonomic Belief Differences

8.1 Introduction

Classifying animal species by observation started in the ancient times with the work of Aristotle (384 B.C. – 322 B.C.). Aristotle's most successful scientific writings were those on biology; as a result of observe actions he classified living things into hierarchy. Taxonomy is the science of classifying organisms into species and logical groups of species. Linnaeus (1707 – 1778) created a system for naming ranking, and classifying organisms. The system has been extended and is still used widely (Linnaeus, 1758; Blunt & Stearn, 2002).

Regarding to its importance, a taxonomic reference system is at the centre of any biological scientific activity. Systematics is the research field within biology that attempts to unfold the (evolutionary) tree of life, i.e., the biological taxonomy. Biologists do not know that tree, and are in the process of slowly uncovering its secrets. Consequently, the current beliefs differ between them, and in their communications they go to great length in properly identifying the taxa of their discourse. When such is human communication, differences in believe are usually fairly quickly identified and resolved.

In this chapter, we look at the same problem when two automated systems, each with its own 'taxonomic belief' attempt to agree on the taxa in their communication. We propose a negotiation protocol that help systems

do this. We abstract away from certain complexities of taxonomic practice, such as the presence of synonyms and misspellings, and assume error- and redundancy-free taxonomic data sets on both ends, which may (and will) however differ in structure and completeness.

8.2 Systematics in brief

Systematics is the biological research field that attempts to unravel the (evolutionary) tree of life. It deals with questions such as which living organisms exist or have existed, what is their evolutionary history, and how are they related to each other. All these questions can be viewed as questions of classification.

For our purpose, we may define the term *taxon* as a maximal group of organisms that are classified in the same way, given a set of criteria. The criteria applied depend on the purpose of the classification; they can be morphological or otherwise phenotypical, chemical, specifically genotypical, or ecological.

Biological classification takes many different forms. In ecology, ecosystems are important units in the classification. They are formed from living organisms that over time have found a way of communal coexistence.

Throughout this work, we will tacitly assume that the purpose of building a biological taxonomy is to unravel the actual tree of life, seen as a collection of separate taxa and the full extent of their ancestral relationships. We do not know this tree, but man's understanding of it is improving. Especially in recent years, with genetic techniques having become more commonly applied, a number of important, and sometimes surprising, finds have been reported (Gaston & May, 1992).

8.2.1 Biological classification

In biological classification, four fundamental structuring and naming systems can be identified:

Taxonomic reference system — The TRS is the taxonomic theory and paradigm adopted. It determines the taxonomic principles, definitions of units, in- or exclusion of higher-order rank taxa, et cetera. The question as to which TRS best serves which purpose is both a fundamental and a controversial one within systematics (Taxonomy, 2002).

As we will discuss below in Section 8.2.4, the last 20 years have witnessed another domain of systematics activity, namely phylogenetic taxonomy.

Nomenclature — This lays down the rules that govern the naming of identified taxa. As such, it depends on the TRS adopted, as the latter determines the types of taxonomic unit that must be named.

Biological nomenclature of living and fossil organisms is regulated by a number of international codes (ICs). They have been developed under the auspices of either the International Union of Biological Sciences (IUBS) or the International Union of Microbiological Societies (IUMS). Five ICs are currently in operation:

- the IC of Zoological Nomenclature,
- the IC of Botanical Nomenclature,
- the IC of Nomenclature for Cultivated Plants,
- the IC of Nomenclature of Bacteria, and
- the IC of Virus Classification and Nomenclature.

These concerted efforts have resulted into the 1996 Draft BioCode (Greuter *et al.*, 1996).

The above are all codes based on the principles of Linnaean nomenclature. The phylogenetic school has also produced a draft code, the PhyloCode (Cantino & de Queiroz, 2000).

Actual taxonomy — The actual taxonomy defines the taxa identified, using the primitives of the TRS adopted, for instance, assigning a rank to each taxon. It also identifies the classification relationships thought to exist between taxa, adhering to TRS rules, and in so doing defines a structure on the identified taxa. (This is usually a tree or a forest.)

Actual naming scheme — The actual naming schema provides names for all identified taxa. The nomenclature adopted, again, prescribes what are (and are not) allowed names. All of this relates to the scientific naming scheme. Usually, an informal vernacular naming scheme is also provided, but the rules for this are largely unwritten.

The first two issues in the above list are *prescriptive* in nature; the last two are *descriptive*. The choice of TRS and nomenclature, in database terminology, fixes the schema of the taxonomic data set; the choice of one may not be entirely independent of that of the other. For a specific implementation of a bio-information system, one should make a choice in prescriptive systems (choose the laws to adhere), and then describe the taxonomy and naming according to these. The actual taxonomy and naming scheme are the actual data sets.

The Linnaean rank order system is a prominent example of a TRS. The associated nomenclature uses binomials such as *Empidonax flavescens* to name a species, and trinomials to name a species' races (or subspecies). *Empidonax* is the genus name; the genus comprises a group of rather similar species. There are more ranks above the genus level: subfamily, family,

order, et cetera. This is not to say that amongst scientists who subscribe to the Linnaean rank order system as their TRS there is full agreement about which are the species, genera, families and so on. They agree to use the system, but may well disagree on the actual taxonomy, i.e., the recognition of species and/or their relationships. In the partial actual taxonomy exemplified below, the Yellowish Flycatcher is mentioned as a taxon at species rank, and the name associated to it in the applied actual naming scheme is *Empidonax flavescens* (Lawrence, 1865).

For instance, the work by (Sibley *et al.*, 1988), which adopts a version of the Linnaean system, classifies the Yellowish Flycatcher from the root of the avian taxonomy as follows:

class *Aves*
infraclass *Neoaves*
parvclass *Passerae*
superorder *Passerimorphae*
order *Passeriformes*
suborder *Tyranni*
infraorder *Tyrannides*
parvorder *Tyrannida*
family *Tyrannidae*
subfamily *Tyranninae*
genus *Empidonax*
species *flavescens*

Ranks like infraclass and suborder form the kernel of the taxonomic reference system adopted, here the Linnaean rank order system. Two ranks are not present in the above example: tribe and subspecies. In the case of *Empidonax flavescens*, the term tribe is not used as no tribes are currently recognized within the subfamily *Tyranninae*; the species' subspecies (*flavescens*, *imperturbatus*, *salvini*) have simply been left out of the example.

In the above example, the adopted nomenclature prescribes various rules. Simple examples of these are: standardized endings of names per rank (like *~nae* for subfamilies), and gender agreement between generic and specific name (*Empidonax* being a masculine noun requiring *flavescens* to have a masculine ending (David & Gosselin, 2002a; David & Gosselin, 2002b)). The listing for *E. flavescens* requires eleven taxa, assigns them a rank according to the TRS, and provides ten taxonomic relationships between them, in this case forming a direct tree traversal from the avian root to a single species.

Scientists involved in systematics do not work on the complete tree; typically, they are specialists working on understanding just a little part of

it, for instance, the taxonomic relationships between some or all (bird) species of (the New World's) tyrant flycatchers (*Tyrannidae*). But even in these much more restricted domains, unravelling such parts of the tree is not easy, and different research efforts sometimes lead to mutually inconsistent results. In other words, there is room for different interpretations, different beliefs.

Figure 8.1 serves to illustrate a small reconstructed part of the taxonomic tree. For instance, it shows the relative position of the Yellowish Flycatcher, compared to a number of other flycatcher species, as inferred by (Cicero & Johnson, 2002). The authors studied the genetic differences between these different species to test an interesting biogeographic hypothesis.

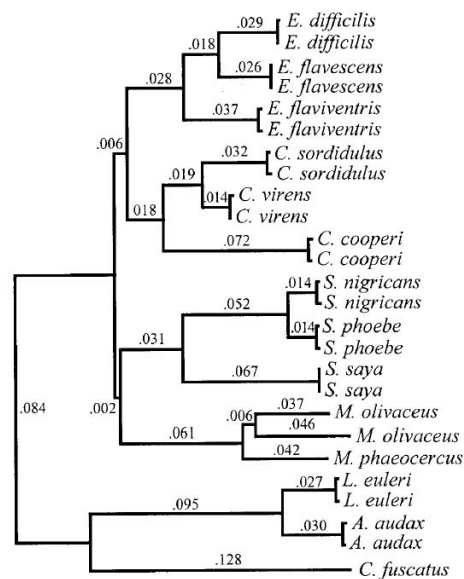


Figure 8.1: Example of a taxonomic subtree indicating evolutionary history of some new world flycatchers . Values provided on the branches can be interpreted as a measure of genetic distance. Internal nodes denote the most recent common ancestor of all taxa in the subtree rooted by that node. These internal nodes represent inferred taxa: an ancient ancestor's genetic make-up, determined from the genetic material of its present-day descendants. With permission from Cicero & Johnson, 2002

8.2.2 Some basic terminology

Taxonomy is the study of biologically classifying live forms into defined taxa. A *taxon* is a named group of organisms. It may denote a maximal population of organisms, maximal relative to a set of shared characteristics, as is typical in the case of a species or subspecies. It can also denote a higher order grouping of taxa — in phenological work called a *clade* — where the grouping is often based on the assumption or belief of sharing a set of common characters or sharing an ancestral history.

An equally important term is the *taxon definition*: a description of the meaning of the taxon name. Related to it is the *taxon diagnosis*: (the description of) a method of recognising or identifying the members of the taxon.

Especially in the Linnaean system, which we discuss slightly more extensively below, it is not uncommon that taxa are rearranged in the actual taxonomy applied, i.e., they change position in the taxonomic tree upon new scientific finds. A taxon, on the basis of research carried out might for instance move up or down a rank in the actual taxonomy. This may lead to merges or divisions of nodes, known as *lumps* and *splits*, respectively.

Another issue very much related with the Linnaean system is the existence of synonymous and homonymous names. Multiple names (*synonyms*) may have been proposed for the same population, for instance, or the same name (a *homonym*) may have been used for different populations. The ICs mentioned above provide the rather intricate laws to deal with such cases. Their understanding, and consequently their implementation, is far from trivial, and has led over the centuries to many violations of the codes, with subsequent corrections; see (David & Gosselin, 2002a; David & Gosselin, 2002b) for examples.

8.2.3 The Linnaean TRS

Historically, the most widely used TRS is the Linnaean rank order system. Since the publication of Linnaeus' seminal work (Linnaeus, 1758), this reference system has been the *de facto* standard of biological taxonomy.

It is instructive to compare the illustrated class-species path for Yellowish Flycatcher with the partial flycatcher tree illustrated in Figure 8.1. The latter tree consists only of branches well inside the tree for the subfamily of *Tyranninae*! By itself, it was not a complete subtree, as many more species of *Empidonax*, for instance, are known. But the flycatcher tree also showed nodes that we can not associate with one of the TRS's ranks. For instance, *E. flavescens* and *E. difficilis* are in the same small subtree, the root of which has no rank in the Linnaean TRS. That node merely reflects the fact that these two species had a common ancestor relatively recently, in existence at least after their shared ancestor with *E. flaviventris*.

In general, nodes within the tree represent dual notions: the node for ‘family of *Tyrannidae*’ stands for the *most recent common ancestor* of all tyrant flycatcher species (the historic, evolutionary perspective), but more informally it also represents the extant collection of tyrant flycatcher species (the contemporary perspective).

This leads us to the following intermediate conclusions:

- Taxonomists attempt to classify biological taxa in a tree that reflects (their belief of) evolutionary history.
- To this end, they adopt a TRS so that they can associate some (but not all) nodes in their tree with ranks. The TRS adopted defines a total order of ranks.
- Not all nodes in the tree need to be associated with a rank.

8.2.4 The Phylogenetic TRS

The use of the Linnaean TRS dates back to the year 1758, though the beginning of modern evolutionary theory is associated with Darwin’s *Origin of Species* (Darwin, 1859).

Advocates of phylogenetic taxonomy, or cladistics, point out that by its very design, the Linnaean TRS is not based on evolutionary principles, but takes more a phenetical approach, grouping organisms on overall similarity. They argue that this shows in the procedures of the associated nomenclatures (de Queiroz & Gauthier, 1990; de Queiroz, 1988). For instance, a taxon definition may have been done in terms of characters and their states, and not in terms of evolutionary primitives. As a consequence, changing insights of taxon characteristics may lead to changes in rank of such a taxon, leading to instability of the taxon name.

In phylogenetic systematics, the balance between taxon definition and taxon diagnosis is much more towards emphasizing a common history of ancestry within the definitions. This common ancestry is of course measured through shared characters (or traits). Two terms are essential in this domain. An *apomorphy* is a derived character state, i.e., a trait that is an evolutionary novelty. A *synapomorphy* is an apomorphy found in more than one taxon, i.e., a shared character state that evolved in their most recent common ancestor. It is through the analysis of synapomorphies that taxa are grouped in phylogenetics.

On top of differences in interpretation, taxonomists are not even agreed on the fundamental notions that they work with. The important term ‘species’ has different definitions, depending on the taxonomic school of thought. There are too many definitions to mention, and this is certainly not the place to provide an exhaustive overview. It can be said that in general the phylogenetic school does not support the recognition of (Linnaean) ranks, such

as family, order, or parvclass, above the species level. There are species, representing (current) populations, and there are clades, representing a latest common ancestor and all its descendants. The situation is more complicated, and doubts have been raised as to the validity of even the species rank (Pleijel & Rouse, 2000).

8.3 Integration of taxonomic data sets

8.3.1 Why do taxonomic beliefs differ?

The above discussion exemplifies the various reasons for differences in taxonomic belief between experts, and consequently their supporting administrative systems:

- The adopted TRS and/or nomenclature may differ;
- With identical TRS and nomenclature, still the actual data may differ, because:
 - Local expertise may have been imposed on the taxonomic data set.
 - Remote expertise may not have been accepted, or may have been judged irrelevant or unimportant.
 - The administrative system may suffer from backlog in taxonomic updates.

Taxonomic experts typically restrict their domain of work: there can be geographic restriction, a taxonomic restriction, but in practice both these restrictions apply.

There are thus good reasons why taxonomic beliefs differ, and the purpose of any administrative system should be not to eradicate such differences, but to accommodate them, allowing communication over taxa by taking away possible causes for misunderstanding. In the curatorial domain of application, the problems at least are aggravated, as the collections have been built up over many years, and specimen labels are written only once, at the time of inclusion in the collection. It is clearly infeasible to rewrite labels with changes in taxonomic belief.

8.3.2 Components of the problem of communicating over taxonomy

It is in the above context that we want to look at the problem of data integration between biological collection data sets. Biological collections consist of collected specimens, each of which is labelled. The label, amongst others, provides the name of the species, collecting date and place, and name of the

collector. The administrative system will take care of such data, and will be the agent in (automated) inter-collection communication.

One would generally expect the data of one biological collection to be based on one consistent taxonomy for the life forms that are present in the collection. But problems with this assumption can arise in various scenarios. First, the collection is built up over many years, under changing taxonomic beliefs of its owner(s). Secondly, with the volume of specimens involved, it is only natural that misspellings and misidentifications occur. Both of these cases cannot be expected to be entirely and satisfactorily handled by an automated system, but what such a system can do is accommodate the administration of such changes and errors. An important subsystem for taxonomic data sets, certainly for the Linnaean TRS as we discussed above, deals with keeping track of synonyms and homonyms (Embury *et al.*, 1999). We will assume here such a subsystem to be available.

Another important problem arises when such taxonomic data is looked at, studied and interpreted by systems or scientists with another ‘taxonomic belief’. Generally, these problems are resolved on an *ad hoc* basis: the scientist in question knows who has set up the collection, knows that person’s taxonomic assumptions, and adapts to this ‘belief’ for the time being. The general problem is recognised as such, but has never been solved generically by taxonomists. This ‘taxonomic chaos’ is an accepted fact of life.

When trying to achieve interoperability between biological collection data sets, we want to enable the exchange of data about specimens, and overcome these differences in taxonomic interpretation. They are, after all, a potential source of misunderstanding, as they should be the basis of deeper systematic understanding. The need for taxonomic explicitness, universality, and stability has been emphasised by many, see for instance (de Queiroz, 1988; de Queiroz & Gauthier, 1990; Lee, 1996).

We may conclude that systems that interoperate about biological species data must be able to identify the ‘taxonomic belief’ under which a local data set has been built. They must also be able to explicitly identify what that belief consists of. This may include identifying the populations that are included, and the populations that are not included within a species. There should be some negotiation protocol that allows systems to resolve such taxonomic differences, and provide an explanation about it to their users.

8.3.3 Problem definition

The problem that we want to address is the following. An information service (system) *R* (for requester) intends to request information from another service *P* (for provider) related to biological taxa. To this end, it needs to identify one or more taxa, and ensure some level of guarantee of semantic

agreement, i.e., the two systems try to ensure that they agree on the taxa under discussion.

We can formally define the ‘semantics of a taxon’ as the content of the taxon under the taxonomic belief B_s of the system s at hand. Given the assumption that a taxonomic belief is represented as a tree, or at least a forest, by the *content of a taxon* t , we mean the complete set of leaves of the subtree rooted at t , so the collection of taxa falling under it. Observe that this is a notion relative to the taxonomic tree (belief) with which one works. Since it is system R that requests, the communication between R and P should resolve the semantics of taxon t under R ’s belief, which one could denote as $B_R(t)$.

For the time being, we make two simplifying assumptions. These are:

- the communication between R and P concerns just a single taxon t , not a set of taxa;
- both systems have a means of resolving issues of synonymy and homonymy;

The first assumption is not a strong one: if one knows how to resolve the semantics of a single taxon, the resolution of a set of taxa is straightforward. There will be issues of optimisation involved when resolving a complete set, as there will typically be differences in taxon contents between systems, and such differences could be covered by contents of other taxa in the same request set of taxa. In other words, resolving the semantics of a set of taxa can be achieved by resolving one taxon at the time, but this is likely to be far from optimal, and more intelligent techniques could be developed. One such idea that we will not elaborate on, for instance, is to take a taxonomic generalisation approach by determining the taxonomic least upper bound of the set of taxa, and using its semantic resolution as a basis.

So, the problem at hand is a semantic agreement problem: how does system R indicate to system S which biological units it is filing its request about?

8.4 Framework for Automatic Negotiation

8.4.1 Requirements

Several issues need to be taken care of in the implementation of this protocol. Important considerations derive from the fact that neither actual taxonomy may be complete, even relative to the high level query at hand, and similarly that their shapes may differ. As a consequence, a leaf node in one taxonomy may not be a leaf node in the other, given that it exists there. Looking at this problem, we foresee a need for a negotiation set-up that

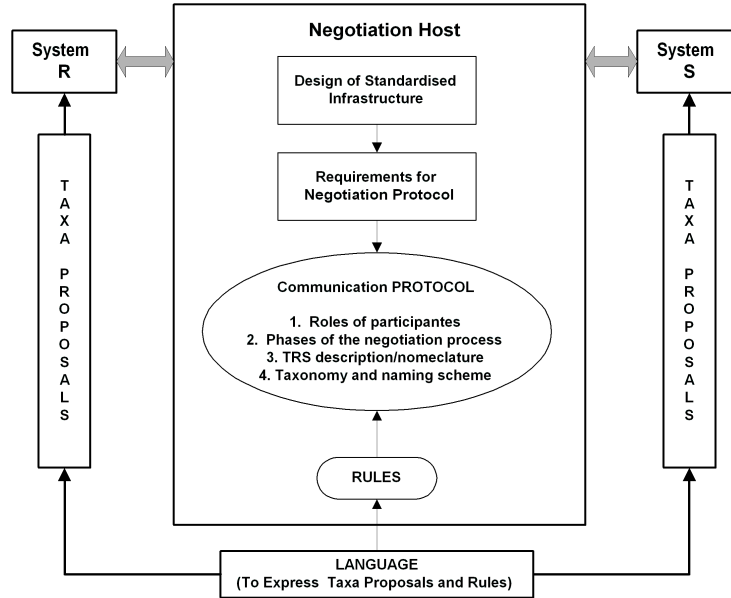


Figure 8.2: Framework to negotiation participants.

enables independent taxonomic representation to interact using a form of negotiation. This set-up would cover aspects including defining a protocol for negotiation (definition of participants, roles and phases of negotiation); defining language(s) for negotiation rules and to express negotiation proposals.

Although many definitions of negotiation exist, we use the definition described by Elfatraty (2002), who defines negotiation as the communication between two or more agents (i.e., taxonomy belief) who are initially ready to accept a range of different outcomes for the purpose of achieving an agreement on a topic over which there is an initial disagreement.

In negotiation, we must distinguish between protocol and strategy. Negotiation protocol determines the flow of messages between the participants, dictating who can say what, when and it acts as the rules by which the negotiation participants must abide by if they are to interact. The protocol should be public and open. A strategy is the way in which a given participant acts within those rules and makes an effort to get the best outcome of the negotiation (when and what to concede, and when he should not concede). Therefore, a strategy of each participant is necessarily private (Bartolini *et al.*, 2003).

The design of a standardised set-up is needed to allow independent inter-

actions over time to reach agreement. A framework for negotiation between two or more taxonomic classifications should have the following characteristics: participants must be free to have their own infrastructure, providing prerequisite functions to support for the automation of the negotiation process; the infrastructure should enforce the standardisation of basic interaction rules; the protocol must guarantee that no party has access to extra information or to be able to forge false information.

We identify the following requirements for our negotiation protocol:

1. A sufficient formalism that automates taxonomy belief interaction.
2. Support negotiation about simple and complex taxonomy descriptions.
3. Accommodate a third party participant to arbitrate in the negotiation.

8.4.2 A communication protocol

To solve the issue, we must understand the potential differences between the (data of) the systems. They can be summarised as follows:

- different TRS and/or nomenclature,
- different actual taxonomy and naming scheme.

A taxon may exist in one but not in the other system; it may be in both but be differently ranked; its taxonomic content may differ. As indicated above, we assume the naming differences are taken care of.

Figure 8.3 illustrates a typical protocol unfolding of the taxonomic disambiguation process that we have in mind. This process consists of the following steps:

1. $R \rightarrow P$ Full data request, depending on metadata availability, this could for instance be a high level query on the data set of P .
2. $R \leftarrow P$ Acknowledgement (or negative acknowledgement), indicating that P is (not) willing to accept the request, and in affirmative case is prompting for additional taxonomic information, regarding taxa (one taxon at the time) mentioned in full data request.
3. $R \rightarrow P$ Initial request, providing relevant data on taxon t like its name, the TRS and Nomenclature in use, as well as contextual information on the position of t in the actual taxonomy. The latter could entail some parent taxon, for instance.
4. $R \leftarrow P$ Initial reply, which will include a proposed definition of the content of t , from perspective of P , indicating also the TRS and Nomenclature applied at P .
5. $R \rightarrow P$ Optional additional request, indicating the content that R finds missing in P 's initial reply. This constitutes a request for more taxa.

6. $R \leftarrow P$ Additional reply, providing added answers (leaf taxa) to the initial reply given earlier.
7. $R \rightarrow P$ Full taxonomic request, upon selection by R of taxonomic content elements earlier proposed by P .
8. $R \leftarrow P$ Full data reply by P taking into account the high level query earlier submitted by R .

We have in the above assumed that there is no a priori knowledge on either system with respect to the other. Clearly, the protocol can be simplified if R knows P 's TRS and/or Nomenclature. Such knowledge can also be used intelligently for the definition of the optional additional request.

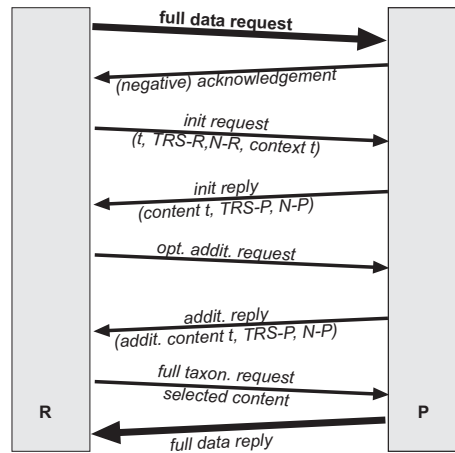


Figure 8.3: Protocol for taxonomic disambiguation between two autonomous taxonomic information services R and P .

8.4.3 Rules of negotiation

The taxonomic protocol is defined independently from the negotiation rules supported by the negotiation host. The rules for negotiation will be used during the phases of the negotiation process. They should be able to deal with well-formedness of taxon proposals. Rules should regulate the alternation of participants in submitting proposals via systems (e.g., systems R and S). Also, there must be rules to dictate the security and visibility aspects of taxon proposals in any negotiation. Rules for supporting methods to ensure knowledge proofs to avoid revealing important information during the negotiation.

8.4.4 To express rules and negotiation proposals

The format to express negotiation proposals has to be standardised. A language to attend the basic requirements must take into account the following:

- Support for ontology and namespace.
- High degree of expressiveness.
- Ability to express fully bound specifications.
- Ability to express constraints over all possible values.
- Loose support of types and inheritance.
- Support for complex queries.
- Support for complex matching.

A language we envisage that can fulfill the requirements for expressing rules and negotiation proposals is DAML+OIL (DAML-Org, 2003). The XML allows information to be more accurately described using tags. The use of XML to provide metadata markup, makes the meaning of the words unambiguous. However, XML has a limited capability to describe the relationships (schemas or ontologies) with respect to objects. The use of ontologies provides a powerful way to describe objects and their relationships to other objects (Guarino & Welty, 2001).

The DAML language is being developed as an extension to XML and the Resource Description Framework (RDF). The latest release of the language (DAML+OIL) provides a rich set of constructs with which to create ontologies and to markup information so that it is machine readable and understandable. The Web Ontology Language OWL is another option that can be considered. OWL is a semantic markup language for publishing and sharing ontologies on the web. OWL is developed as a vocabulary extension of RDF (the Resource Description Framework) and is derived from the DAML+OIL Web Ontology Language. This document contains a structured informal description of the full set of OWL language constructs and is meant to serve as a reference for OWL users who want to construct OWL ontologies.

In Table 8.1, we present a summarised language feature scope covering XML, RDF, DAML+OIL, and OWL. A brief discussion about these features is presented in (DAML-Org, 2003).

The languages DAML+OIL and OWL appears to be better suited to deliver the definition of taxa proposals, rules and the communication protocol.

Table 8.1: Language feature comparison (source DAML-Org, 2003).

Features	XML DTD	XML Schema	RDF(S)	DAML+OIL	OWL
bounded lists	x	x		x	x
cardinality				x	x
class expression				x	x
data types		x		x	x
defined classes				x	x
enumerations	x	x		x	x
equivalence				x	x
extensibility				x	x
formal semantics				x	x
inheritance			x	x	x
inference				x	x
local restrictions				x	x
qualified constraints				x	
reification			x	x	x

8.5 Conclusions

In this chapter, we described fundamental problems present in biological classification. Further, we presented concepts, terminology of taxonomy rank systems and discussed the problem of data integration between biological collection data sets, when trying to achieve interoperability between biological collection data sets. We highlighted the reasons for differences in taxonomic beliefs as well as the problems of communicating over taxonomy. Such a problem was defined in the context of a proposed framework for automatic negotiation. The requirements and a protocol sequence have been described and indication for a possible implementation was given, which covers the rules of negotiation, taxon proposals and language features required for a possible implementation. Regarding a candidate languages, it must have distinctive features, since it must specify rules, taxon proposals and the negotiation host protocol. One can consider potential candidates languages like XML, DAML, OWL. We can envisage a new language with specific features for fulfilling the requirements of such a protocol demand, Taxonomy Markup Language type.

The implementation of the proposed protocol would enable the exchange of data about specimens, and overcome these differences in taxonomic interpretation. Such differences are a potential source of misunderstanding, as they are the basis a lack of deeper systematic understanding. Systems that interoperate about biological species data must be able to identify the ‘taxonomic belief’ under which a local data set has been built. They must also be able to explicitly identify what that belief consists of. This may include

8.5. CONCLUSIONS

identifying the populations that are included, and the populations that are not included within a species. We believe that a negotiation protocol would allow systems to resolve differences, and provide an explanation about it to their users.

Chapter 9

Conclusions and Future Research

This chapter provides a summarised description of the work presented in this thesis. The contributions of the work presented, together with some considerations are discussed. Further, we comment on topics we left out, and discuss the overall scenario of a computer infrastructure to host biological data, metadata management, georeferencing processes applied to Amazonian biological datasets, and a proposed automatic framework for a negotiation protocol for reconciling taxonomic belief differences. In the end, we indicate from our work, directions for further research.

9.1 Summary

In Chapter 1, the scenario of biological data was presented. The chapter described, as the background, the global outcry for the need of high quality information about biodiversity. It emphasises the problems of biological data and how BIS can help researchers to share and disseminate their scientific findings. Important scientific experiments in the Brazilian Amazon were introduced. With that, we aimed to show the amplitude and importance of scientific data produced that need management and the efforts embedded in those experiments. The chapter indicated some issues investigated in this research, pointing to observations related to problems in environmental information systems, ranging from data production to data management. Other issues indicated for investigation included: data representation and integration, computer architecture for information distribution on the web, and georeferencing biological legacy data. Further, it was presented the

motivations and objectives, concentrated in addressing some critical deficiencies in biodiversity information systems.

Chapter 2 presented an overview of biological collections and related activities in the main institutes in the Amazon region. It identified the main problem that can interfere with data management in the context of museum data. The chapter highlighted efforts and drawbacks that can help the development of better BISs. More detailed observations made at INPA described the approach to deal with collection data and the information technology that would help to deliver an integrated system.

Chapter 3 described the functional and system requirements for a comprehensive representation and implementation of a BIS. The functions presented the existing interrelated activities in research, conservation, and education. It covered collection elements and their management (creation, maintenance, transaction processing and information retrieval and reporting). The system requirements provided information for a specification of a system architecture that can respond to the functional needs. It covered processing, data types, volume and usage needs. The chapter also presented the requirements for security (system and data) as well as the length of flexibility during maintenance related to some architectural constraints.

Chapter 4 presented a schematic representation of a biological collection, named CLOSi, for use in database design and implementation. The schema represents functional groups that are specified as object classes and relationships, describing information about collection management, events of collection, localities, species taxonomy, users of collections and publications related to species in a collection. A syntactic definition and attribute constraints related to the schema representation were provided in Appendix A and B.

Chapter 5 described a solution for biological metadata management. The solution was based on XML and client/server technology, which can describe and distribute biological metadata on the web. The chapter presented metadata issues, including context, importance, aspects of a clearinghouse, and standards. The described implementation comprised of a mapping from the FGDC standard to an XML schematic representation. A three-tier architecture (client and server side, and web) supports an XML repository, which is integrated with a web portal for interactions. Further, an extension of this approach has been proposed, which is to enlarge the user capacity (providers and brokers) for accommodating a distributed metadata repository (the nodes approach).

Chapter 6 focused on the prototype implementation of a CLOSi-based database, and a web interface. The options available for the implementations were investigated and we elaborated on free open source products for client, server side and for the web. Details of the implemented prototype

were presented and relevance of the achievements and future steps was also discussed.

Chapter 7 presented a retrospective georeferencing method integrated with a CLOSi-based database. In the chapter, georeferencing issues, like expressing latitude and longitude, distance and direction uncertainties that are involved in the calculation of maximum error distance were discussed. The georeferencing process used a legacy fish dataset implemented with CLOSi. We also discussed the positive and negative aspects in the process and alternative ways to calculate span for more detailed locality descriptions. The process integration, the CAS tool with CLOSi database can be extended to collaborative gazetteer initiatives.

Chapter 8 presented a proposal for an automatic negotiation protocol to deal with taxonomic belief differences. Fundamental problems present in biological classification and the basic terminology and taxonomic reference system (Linnaean and phylogenetics) were described. The problems around integration of taxonomic data sets were detailed as components of the problem of communicating over taxonomy. Following, we presented a framework, comprised of requirements, the protocol, its rules and negotiation strategy and the features that a computer language must have to implement it.

9.2 Achievements

This thesis has been written in the context of computer infrastructure functionality, database and metadata developers, users, providers and brokers who are involved with biological collections needs and their major problems. The main contribution of the thesis covers:

- The problem of ad hoc system development and how to reduce it, so that resources can be diverted using a CLOSi database design and implementation.
- Support for faster data digitising.
- The framework for data exploration and information dissemination, are essential for conservation and for sustainable development of natural biological resources.
- Data and metadata exchange, particularity across institutes in the Amazon region.
- Effective aids in documenting existing data sets.
- Solutions suitable for developing countries, since the system developed and tools used are either public domain or free open source, adding low cost for robust solution.

- The integration with the georeferencing application was found to be a successful and valuable addition to the web system, providing great benefit to researchers of biological collections that need geospatial analysis results. The advantages of this process include: increasing speed in georeferencing, maximising consistency between users, allowing the incorporation of interpretation standards established by researchers, specially curators, and quantifying textual locality vagueness.

9.3 Topics not addressed

The research presented in this thesis had the ambition to also define a system architecture for a biodiversity analysis tool, with special emphasis on the spatial (database) functionality required, together with GIS capabilities. Further, we have thought to define a schema model to support the proposed system's architecture, with emphasis on spatial aggregates and generalisation functions and to study the usefulness of the implemented system in a chosen Amazonian biodiversity context. Obviously, our ultimate goal was not achieved, and we needed to scale down for some important problems not visible at the time we started.

9.4 Future research

There are some issues that need to be investigated, and some of the aspects dealt with in this research still need further study and development. The main areas for future work directly relevant to the research are the following:

1. Study of CLOSi effectiveness — CLOSi has been partially implemented using INPA insect and fish collections. Unfortunately, we were not able to compare its representation effectiveness against similar models/schemas developed to describe biological collections. The systems available that could provide essential information about the conceptual models embedded on those system include: BASIS, Biolink, BIÓTICA, BIOTA, KE EMu, MUSE, SAMPADA, SPECIFY and TAXIS. The study will help BIS designers and users to perceive the relationships amongst model/schema concepts, and functionality, the basis for interoperability amongst software solutions.
2. Interoperability in a multi-taxonomic system — Interoperability in taxonomic systems represents a motivation and will provide us an opportunity for the implementation and evaluation of the framework for automatic negotiation, as presented in Chapter 8. In the frameworks

the language to be applied must have distinctive features and are the heart of the protocol specification, rules and taxa proposal. Amongst the candidates languages to be considered are, XML, DAML, OWL. Alternatively, a new language with specific features for fulfilling the requirements of such a protocol, we can envisage a Taxonomy Markup Language.

3. Integration of analytical tools — Analytical tools that can use data stored in a database, as well as generate and store derived data. We believe that there are two primary types of analytical tool that can be used: modeling and simulation tools, and GIS analytical functions. Regarding the visual analysis of stored data, there is also a need for more powerful visualising tools normally not available in GIS.

To integrate these analytical tools with the database environment we envisage two options: (1) to implement the needed functions as extensions to the primitive functions of some DBMS, and (2) to implement interfaces between the analytical tools and the DBMS to provide means to import/export data.

Implementing extensions to DBMS functionality may require much effort but, in some cases, there may be applications of such functions available to the specific DBMS. These extensions would allow users to store, access, manage, and manipulate spatial data in the same database as the rest of the traditional data.

Concerning the second option, it is more flexible than the first since it does not deprive scientific teams of using their favourite analytical tools. This is particularly important to the use of modeling tools which, in many cases, are systems previously developed that work well for the desired goals.

9.4. FUTURE RESEARCH

References

- Adida, Ben. 2001 (July). *Why Not MySQL*. <http://openacs.org/philosophy/why-not-mysql.html>, accessed on 2002/07/22.
- Aelfataty, Ahmed. 2002. *A Framework For Negotiating Software Services*. Ph.D. thesis, UMIST.
- Agosti, Donat, Moskovits, Debra, & Wang, Yeqiao. 1999 (June). *Biodiversity indicators and modeling understanding the distribution and conditions of biodiversity*. Proceeding from a Symposium on Conservation Biology and Nasa. Washington, DC.
- Alcamo, Joseph. 2002. Global Change Meets Global Policy: A New Impulse for Global Environmental Informatics. *Pages 22–28 of: Pillmann, W., & Tochtermann, K. (eds), Environmental Communication in the Information Society*. International Society for Environmental Protection, Vienna, Austria.
- Aldenderfer, Mark S., & Maschner, Herbert D. G. 1996. *Anthropology, Space, and Geographic Information Systems (Spatial Information Series)*. Oxford University Press.
- Allehaibi, M. 1998 (July). *Choosing a Web Server*. <http://slistwo.lis.fsu.edu/planning/server/choose.htm>, accessed on 2002/07/22.
- Allkin, R. 1997. Data management within collaborative projects. *Pages 5–24 of: J. Dransfield, M. J. E. Coode, D. A. Simpson (ed), Plant Diversity in Malaysia*. Royal Botanic Garden Kew, London, England, UK.
- Allkin, R. 1998. *Collection Information Management at INPA*. Tech. rept. 01. Instituto Nacional de Pesquisas da Amazônia, Manaus, Amazonas, Brasil.
- Altova. 2002. *XML Spy 4.4 User and Reference Manual*. Vervanté. Mountain View, CA.
- Ashenfelter, Paul. 1998. *Choosing a Database for Your Web Site*. John Wiley & Sons. Hoboken, NJ.

REFERENCES

- Association of Systematic Collections. 1997 (June). *Association of Systematic Collections: The ASC Reference Model*.
- Aulds, Charles. 2000. *Linux Apache Web Server Administration*. Linux Library. Sybex Inc. Alameda, CA.
- Bartolini, Claudio, Preist, Chris, & Kuno, Harumi. 2003 (February). *Requirements for Automated Negotiation*. <http://www.w3.org/2001/03/WSWS-popa/paper19>, accessed on 2003/02/27.
- Batini, Carlo, Lanzerini, Maurizio, & Navathe, Shamkant B. 1986. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, **18**(4), 323–364.
- BCDAM-MMA. 1998 (Março). *Sistemas de Bases Compartilhadas sobre a Amazônia - Conceção e Funcionamento*. Tech. rept. Ministério do Meio Ambiente.
- Benchimol, Samuel. 1992. *Amazônia: A Guerra na Floresta*. Civilização Brasileira.
- Benchimol, Samuel. 1996. *A Amazônia e o Terceiro Milênio: Antivisão: O Brasil no Terceiro Milênio - O Livro da Profecia*. Senado Federal - CEGRAFI.
- Bisby, F. A. 2000. The quite revolution: Biodiversity Informatics and the Internet. *Science*, **5488**(289), 2309–2312.
- Bissett, Michael. 2002 (August). *An Implementation of a CLOSi-based Database for Managing Biological Collections on the Web*. Professional Master Thesis, International Institute for Geo-Information Science and Earth Observation.
- Blum, Stanley D., Stein, Barbara R., & Beach, James H. 1995. *Museum of Vertebrates Zoology: Requirements Analysis - Functional Requirements and System Requirements*. University of California at Berkeley.
- Blunt, Wilfrid, & Stearn, William T. 2002. *Linnaeus: The Compleat Naturalist*. Princeton University Press.
- Bortoni, Larissa, & de Moura, Ronaldo. 2002. *Mapa da corrupção no governo FHC*. Fundação Perseu Abramo.
- Bosley, J. J., & Straub, K. 2000 (July). *Data exploration interfaces: Meaningful web database mining*. http://www-3.ibm.com/ibm/easy/eou_ext.nsf/Publish/2773, accessed on 2002/08/08.
- Bourret, R. 1999 (September). *XML and Databases*. Technical University of Darmstadt.
- Brackett, M. H. 1997. *The Data Warehouse Challenge: Timing Data Chaos*. Wily Computer Publishing.

- Brain, Marshall. 2001 (December). *How Web Servers Work*. Tech. rept. Computer How Stuff Works. <http://www.redlinesnetworks.com>, accessed on 2001/12/20.
- Brigadão, Clóvis. 1996. *Inteligência e Marketing: O Caso Sivam*. Editora Record.
- Brown, J. H., & Roughgarden, J. 1990. Ecology for a changing earth. *Bulletin of the Ecological Society of America*, **1**(71), 173–188.
- Brundage, M., Dengler, P., Gabriel, J., Hoskinson, A., Kay, M., Maxwell, T., Ochoa, M., Papa, J., & Vanmane, M. 2000. *Professional XML Databases*. Wrox Press Inc. Chicago, Illinois.
- Câmara, G., Freitas, U. M., Souza, R. C. M., & Garrido, J. 1996. SPRING: Integrating Remote Sensing and GIS by Object-Oriented Data Modelling. *Computers and Graphics*, **15**(6).
- Campos dos Santos, J. L. 2000 (January). *Biodiversity Information Systems in an Open Analytical Database Architecture*. Tech. rept. International Institute for Aerospace Survey and Earth Sciences.
- Campos dos Santos, J. L., & de By, R. A. 2000 (March). *A Clustered Object Schema for INPA's Biodiversity Data Collections*. Tech. rept. International Institute for Aerospace Survey and Earth Sciences.
- Campos dos Santos, J. L., de By, R. A., & Magalhães, C. 2000. A Case Study of INPA'S Bio-DB and an Approach to Provide an Open Analytical Database Environment. *Pages 155–163 of: International Archives of Photogrammetry and Remote Sensing. ISPRS 2000, Amsterdam, The Netherlands*, vol. XXXIII.
- Campos dos Santos, J. L., de By, R. A., Apers, P. M. G., & Magalhães, C. 2002a. Clustered Object Schemas for INPA's Biological Collections Data. *Pages 38–44 of: Nagib Callaos, Luis Hernandez-Encinas, Fahri Yetim (ed), Volume I - Information Systems Development I of the SCI 2006 - 6th World Multiconference on Systemics, Cybernetics and Informatics - 2002, Orlando, Florida (USA)*, vol. 1. International Institute of Informatics and Systemics.
- Campos dos Santos, J. L., de By, R. A., Magalhães, C., & Apers, P. M. G. 2002b. Facilitating Interdisciplinary Sciences by the Integration of CLOSi-based Database with Bio-Metadata. *In: International Archives of Photogrammetry and Remote Sensing. ISPRS 2002, Ottawa, Canada*, vol. 34. Canadian Institute of Geomatics.
- Cantino, Philip D., & de Queiroz, Kevin. 2000. *PhyloCode: A Phylogenetic Code of Biological Nomenclature*. URL www.ohiou.edu/phylocode/.
- Castro, Elizabeth. 2001. *Perl and CGI for the World Wide Web: Visual QuickStart Guide*. 2nd edn. Peachpit Press.

REFERENCES

- Chen, I-Min A., & Markowitz, Victor M. 1995. An Overview of Data of the Object-Protocol Model OPM and OPM Data Management Tools. *Information System*, **20**(5).
- Chen, I-Min A., & Markowitz, Victor M. 1996 (June). *The Object-Protocol Model*. Technical Report LBNL 32738. Lawrence Berkeley National Laboratory, Information and Computing Sciences Division, 1 Cyclotron Road, Berkeley, CA 94720.
- Christy, Peter, & Katsaros, John. 2002 (June). *Web I/O Servers: Rethinking Web Servers Architecture*. White Paper. Redline Networks - Net-Edge Research Roup, 334 State Street, Suit 201, Los Altos, California. <http://www.redlinesnetworks.com>, accessed on 2002/12/12.
- Cicero, Carla, & Johnson, Ned K. 2002. Phylogeny and character evolution in the *Empidonax* group of Tyrant Flycatchers (Aves: Tyrannidae): A test of W. E. Lanyon's hypothesis using mtDNA sequences. *Molecular Phylogenetics and Evolution*, **22**(2), 289–302.
- Clarke, Keith C. 1995. *Analytical and Computer Cartography*. Prentice Hall Series in Geographic Information Science. Prentice-Hall.
- Cleary, A., Kohn, S., Smith, S. G., & Smolinski, B. 1999. *Language Interoperability Mechanisms for High Performance Scientific Applications*.
- Coar, Ken A. L. 1998. *Apache Server For Dummies*. John Wiley & Sons. Hoboken, NJ.
- Cornillon, Peter. 2000. *An Organizational Structure for Metadata*. White Paper.
- Cowen, D. J. 1997. *Discrete Georeferencing*. NCGIA Core Curriculum in GIScience. <http://www.ncgia.ucsb.edu/giscc/units/u016/u016.html>.
- Crowder, David, & Crowder, Rhonda. 2000. *Building a Web Site for Dummies*. John Wiley and Sons. Hoboken, NJ.
- da Fonseca, C. R. V., Salem, J. I., & Weigel, P. 2002. *Bioacervos em Instituições da Amazônia*. Tech. rept. BIOAMAZÔNIA, Manaus, Amazonas, Brasil.
- da Silva, Cylon Goncalves, & de Melo, Lucia Carvalho Pinto. 2001. *Ciência, tecnologia e inovação: desafios para a sociedade brasileira - O Livro Verde*. Ministério da Ciência e Tecnologia e Academia Brasileira de Ciência.
- DAML-Org. 2003 (March). *Language Feature Comparison*. <http://www.daml.org/language/features.html>, accessed on 2003/03/29.
- Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

- David, Normand, & Gosselin, Michel. 2002a. Gender agreement of avian species names. *Bulletin British Ornithologists Club*, **122**(1), 14–49.
- David, Normand, & Gosselin, Michel. 2002b. The grammatical gender of avian genera. *Bulletin British Ornithologists Club*, **122**(1), 257–282.
- de Queiroz, Kevin. 1988. Systematics and the Darwinian revolution. *Philosophy of Science*, **55**, 238–259.
- de Queiroz, Kevin, & Gauthier, Jacques. 1990. Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. *Systematic Zoology*, **39**(4), 307–322.
- Deitel, Harvey M., Deitel, Paul J., & Nieto, T.R. 2001. *Internet and World Wide Web How to Program*. Prentice Hall.
- Delbaere, Ben C. W. 1998. *Facts & Figures on Europe's Biodiversity: state and trends 1998-1999*. Tech. rept. European Centre for Nature Conservation, Tilburg.
- Dempsey, L., & Weibel, S. L. 1996 (July-August). *The Warwick Metadata workshop: a framework for the deployment of resource description*. D-Lib Magazine.
- Douglas, Korry, & Douglas, Susan. 2003. *PostgreSQL*. 1st edn. New Riders Publisher. Indianapolis, IN.
- Dubois, P., & Widenius, M. 1999. *MYSQL*. New Riders Publisher. Indianapolis, IN.
- Edwards, J. L., & Nielsen, E. S. 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, **5488**(289), 2312–2314.
- Elmasri, R., & Navathe, S. B. 1994. *Fundamentals of Database Systems*. 2nd edn. World Student Series. Benjamin/Cummings Publishing Company, Inc. Redwood City, California.
- Embury, Suzanne M., Jones, Andrew C., Sutherland, Iain, Gray, W. A., White, Richard J., Robinson, John S., Bisby, Frank A., & Brandt, Sue M. 1999. Conflict Detection for Integration of Taxonomic Data Sources. *Pages 204–213 of: Statistical and Scientific Database Management*.
- Fairchild, Alea. 1996. *Interoperability for Enterprise Information Systems*. Computer Technology Research Corporation.
- Federal Geographic Data Committee. 2001 (March). *Biological Data Profile Workbook*. Tech. rept. FGDC.
- Fedra, K. 1994. *GIS and Environmental Modelling*. Oxford Press.
- Feldmann, R. M., & Manning, R. B. 1992. Crisis in systematic biology in the “age of biodiversity”. *Journal of Paleontology*, **1**(66), 157–158.

REFERENCES

- Fonseca, O. J. M., & Ferreira, E. J. G. 1998. *Guia de referência dos pesquisadores do INPA*. Série Documentos. no. 2:37 p.
- Gangeni, Aldo. 2002 (November). *Using Foundational Ontologies for Conceptual Domain Analysis and Terminology Merging*. Workshop on Ontology for Biology, Heidelberg, Germany.
- Gaston, K. J., & May, R. M. 1992. Taxonomy of taxonomists. *Nature*, 281–282.
- Geschwinder, Ewald, & Schoening, Hans-Juergen. 2001. *Postgresql: Developer's Handbook*. 1st edn. Sams Publishing. Indianapolis, IN.
- Gilliland-Swetland, Anne, Bacca, Murtha, & Gill, Tony. 2000. *Introduction to Metadata: Pathways to Digital Information*. Getty Information Institute.
- Goldfarb, C. F., & Prescod, P. 2002. *XML Handbook*. 4th edn. Prentice-Hall PTR, New Jersey.
- Goycochea, A. P. 1998. *Report de la Consultória para diagnosticar a situação atual do gerenciamento das coleções botânicas do INPA e recomendar procedimentos para sua modernização*. Tech. rept. Instituto Nacional de Pesquisas da Amazônia, Manaus, Amazonas, Brasil.
- Greuter, W., Hawksworth, D. L., McNeill, J., Mayo, M. A., Minelli, A., Sneath, P. H. A., Tindall, B. J., Trehane, P., & Tubbs, P. 1996. Draft BioCode: the prospective international rules for the scientific names of organisms. *Taxon*, **45**, 349–372.
- Guarino, N., & Welty, C. 2001. *Ontological Analysis of Taxonomic Relationships*. Proceedings of the 19th International Conference on Conceptual Modeling.
- Gunther, Oliver. 1998. *Environmental Information Systems*. Springer Verlag, Berlin.
- Hammer, M., & McLeod, D. 1981. Database Description with SDM: A Semantic Database Model. *ACM Transactions on Database Systems*, **6**(3), 351–386.
- Hansen, Michael Schacht, Dorup, Jens, Ribe, Lars Riisgaard, & Larsen, Kristoffer W. 2002 (December). *A Comparison of technologies for database-driven websites for higher education*. <http://www.intermed.dk/datadriven>, accessed on 2002/12/22.
- Harold, E. R., & Means, W. S. 2001. *XML in a nutshell: A Desktop Quick Reference*. Natshell Handbook.
- Hatley, Derek J., Hruschka, Peter, & Pirbhai, Imtiaz A. 2000. *Process for System Architecture and Requirements Engineering*. Dorset House.

- Hill, L. L., Frew, J., & Zheng, Q. 1999 (January). *Geographic names: The implementation of a gazetteer in a georeferenced digital library*. D-Lib. <http://www.dlib.org/dlib/january99/hill/01hill.html>.
- Hintjens, Pieter. 2002 (December). *iMatix Editorial*. <http://www.imatix.com>, accessed on 2002/12/22.
- Hopkins, M. 1999. *Projeto Flora da Reserva Ducke*.
- Huxhold, William E., & Levinsohn, Allan G. 1995. *Managing Geographic Information System Project (Spatial Information System (Cloth))*. Oxford University Press.
- IDC. 2001. *Vendor Profile: Software AG and Tamino*. IDC.
- Information, USGS National Mapping. 2000 (November). *GNIS database Online Data Base*. http://geonames.usgs.gov/gnis_users_guide_toc.html, accessed on 2002/01/05.
- International Conservation Organizations Consortium. 1998. *Better Data for Better Decisions*.
- Jarke, M., & Vassiliou, Y. 1985. A framework for choosing a database query language. *ACM Computing Surveys*, **17**(3), 313–340.
- Johnson, Maryfran, & Vijayan, Jaikumar. 2003. NcNealy: 'Let Us Run the Machines'. *Computerword*, February.
- Johnson, Norman F. 1997. *The Ohio State University Insect Collection - OSUC Database Structure*.
- Kandeh, J. M. K., Campos dos Santos, J. L., & Kumar, L. 2002 (September). Automatic Mapping and Monitoring of Invasive Alien Plant Species, The South African Experience. *In: Proceeding of the 18th International CODATA Conference — Frontier of Scientific and Technical Data*. CODATA, Montreal, Canada.
- Karaiva, P., & Anderson, M. 1988. *Spatial aspects of species interactions: the wedding of models and experiments*. Springer Verlag, New York. Community Ecology.
- Kerschberg, L., Michaels, G., Bressler, J., Muir, P., Piselli, T., & Ramnathan, R. 1996 (November). *A GIS Application to The Development of Biodiversity Information Systems*. Project NSF Form 1358 - CS1810/INFT864.
- Kirchner, T. B. 1994. *Data management and simulation modelling*. Taylor and Francis, London, UK. Chap. Environmental information management and analyses: ecosystem to global scales, pages 357–375.
- Knuth, Donald. 1964. Backus normal form vs. Backus Naur form. *Comm. ACM*, **7**(12), 735–736.

REFERENCES

- Kobler, Ben, Berbert, John, Caulk, Parris, & Hariharan, P.C. 1995. *Architecture and Design of Storage and Data Management for the NASA Earth Observing System Data and Information System*. IEEE Symposium on Mass Storage Systems.
- Koschel, Arne, Kramer, Ralf, Nikolai, Ralf, Hagg, Wilhelm, & Wiesel, Joachim. 1996 (January). *A Federation Architecture for an Environmental Information System incorporating GIS, the World-Wide Web, and CORBA*. In Third International Conference/Workshop Integrating GIS and Environmental Modeling, Santa Fe, New Mexico.
- Kyte, Thomas. 2001. *Expert One on One Oracle*. Wrox Press Inc. Chicago, Illinois.
- Lagoze, Carl, & Hunter, Jane. 2000. An Event-Aware Model for Metadata Interoperability. *ECDL*, September.
- Lamont, Michael. 2001 (February). *Savant Documentation*. <http://www.nwc.com/1020/1020f1.html>, accessed on 2002/011/17.
- Lane, M. A. 1996. Roles of natural history collections. *The Missouri Botanical Garden*, 1(83), 536–545.
- Laurie, Ben, & Laurie, Peter. 2002. *Apache: The Definitive Guide*. 3rd edn. O'Reilly & Associates, Inc.
- Lawrence, William, Saatchi, Sasan, DeFries, Ruth, Dietz, James, Rice, Richard, Dietz, Lou Ann, de Araújo, M. Siqueira, & Alger, Keith. 1995. Utilization of SAR and optical remote sensing data for habitat conservation in the tropical forest of Brazil. *International Geoscience and Remote Sensing Symposium (IGARSS) - IEEE*, 2, 1480–1482. Piscataway, NJ.
- LBA Project. 1997 (October). *LBA-DIS Recommendations Report*. Tech. rept. NASA-INPE CPTEC. College Park, Maryland.
- Lee, Michael S. Y. 1996. Stability in meaning and content of taxon names: an evaluation of crown-clade definitions. *Proceedings Royal Society of London, Series Biological Sciences*, 263, 1103–1109.
- Lerdorf, Rasmus, & Tatro, Kevin. 2002. *Programming PHP*. O'Reilly & Associates.
- Levi, S. A. 1992. The problem of pattern and scale in ecology. *Ecology*, 73, 1943–1967.
- Lewis, Johnathan. 2001. *Practical Oracle 8i: Building Efficient Databases*. Addison-Wesley Pub Co.
- Linnaeus, Caroleus. 1758. *Systema Naturae Per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis*. Editio decima, reformata edn. Laurentii Salvii.

- Macehiter, Neil. 2000 (November). *Web Server Performance Scalability*. White Paper. Zeus Technology, Inc. <http://www.zeus.com>.
- Maciaszek, Leszek A. 2001. *Requirements Analysis and System Design: Developing Information System with UML*. 1st edn. Addison-Wesley Publishing.
- Magalhães, C., Campos dos Santos, J. L., & Salem, J. I. 2001. Automação de Coleções Biológicas e Informações sobre a Biodiversidade da Amazônia. *Parcerias Estratégicas - Centro de Estudos Estratégicos - MCT*, 294–312.
- Mannino, Michael V. 2000. *Database Application Development & Design*. McGraw-Hill/Irwin.
- Marcon, Eric. 1999. *Oiapoque Project: Multi-lingual gateway towards brazilian and french grey literature concerning Amazonia*. Tech. rept. INPA and Silvolab-Guyane.
- Martinelli, Pedro. 2001. *Amazônia - O Povo das Águas*. Terra Virgem Editora.
- Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S.G. 1997. Nongeospatial metadata for the ecological sciences. *The Ecological Applications*, 7(1), 330–342.
- Microsystems, Sun. 2002 (September). *A Introduction to appliance Servers*. White Paper. Sun Microsystems. <http://www.itpapers.com>.
- Milstead, J., & Feldman, S. 1999 (January). *Metadata: Cataloging by any other name*. Online Inc Magazine.
- Ministério de Ciência e Tecnologia. 2002 (January). *PROBEM DA AMAZÔNIA: Desenvolvimento de Pesquisas sobre o Uso Sustentável dos Recursos Naturais da Região Amazônica*. Relatório MCT 2001.
- Mitchell, Scott, & Atkinson, James. 2000. *Sams Teach yourself Active Server Pages 3.0 in 21 Days*. Sams Publishing. Indianapolis, IN.
- Morrison, Mike, Morrison, Joline, & Morrison, Morrison. 2000. *Database-Driven Web Sites*. Course Technology.
- Netcraft Web Server Survey. 2002 (July). *Web Server Software Usage on Internet Connected Computers*. <http://www.netcraft.com/survey>, accessed on 30-07-2002.
- Newsome, M., Pancake, C., & Hanus, J. 1995. *HyperSQL: Web-Based Query Interface for Biological Databases*.
- Niemann, B. L. 2000. An integrated repository and registry for environmental information: An XML portal for public access and data exchange. *INFOTERRA 2000: Global Conference on Access to Environmental Information*, 1(September).

REFERENCES

- Nobre, C.A., Dolman, A.J., Gash, J.H.C., Hutjes, R.W.A., Jacob, D.J., Janetos, A.C., Kabat, P., Keller, M., Marengo, J.A., McNeal, R.J., J. Melillo, P.J. Sellers, Wickland, D.E., & Wofsy, S.C. 1996. *LBA Science Planning Group*. Tech. rept. SC-DLO, Wageningen, Netherland. 44p.
- Oak Ridge National Laboratory. 1999a (March). *BOREAS Follow-On Project / Campaign Document*. ORNL-campaign2802.
- Oak Ridge National Laboratory. 1999b (March). *FIFE Follow-On Project / Campaign Document*.
- Oliveira, N. M. 1999. *Projeto e Implementação do Modelo de Banco de Dados Entomológico do INPA*. Tech. rept. Universidade do Amazonas, Departamento de Ciência da Computação.
- Olsen, Lola M. 2000. *Global Environmental Databases - Present Situations; Future Directions*. 1st edn. Vol. 1. International Society for Photogrammetry and Remote Sensing. Chap. Discovering and Using Global Databases, pages 220–233.
- Pandolfo, Clara Martins. 1991. *Amazônia brasileira: ocupação, desenvolvimento e perspectivas atuais e futuras*. Editora Cejup.
- Perens, B. 2002 (July). *The Open Source Definition*. <http://www.opensource.org/docs/definition.html>, accessed on 2002/07/22.
- Piatetsky-Shapiro, Gregory, & Frawley, Willian J. 1992. *Knowledge Discovery in Databases*. MIT Press.
- Pleijel, F., & Rouse, G. W. 2000. Least-inclusive taxonomic unit: a new taxonomic concept for biology. *Proceedings Royal Society of London, Series Biological Sciences*, **267**(1443), 627–630.
- PNE Project. 2001a (July). *Managing biodiversity data within PNE*.
- PNE Project. 2001b (July). *Plantas do Nordeste*.
- PNE Project. 2001c (July). *PNE's Subprogramme for Information, Dissemination and Training (SIDT)*.
- Primack, R. 1993. *Essentials of Conservation Biology*. Sinauer Associates Inc., MA. 564pp.
- Proctor, E. J., Blum, S. D., & Chaplin, G. 2001 (August). *A Software Tool for Retrospectively Georeferencing Specimen Localities using ArcView*. California Academy of Sciences.
- Pushman, J. 2002 (August). *Why PHP?* http://www.webdevelopersjournal.com/articles/why_php.html, accessed on 2002/07/22.
- Secretaria de Assuntos Estratégicos. 1993. *SIPAM - Sistema de Proteção da Amazônia: estudo de viabilidade técnica, econômica e financeira*.

- Sibley, C. G., Ahlquist, J. E., & Jr., B. L. Monroe. 1988. A classification of the living birds of the world, based on DNA-DNA hybridization studies. *Auk*, 409–423.
- Silva, Marina. 2001 (Dezembro). *Biodiversidade: Oportunidade e Dilema*. Carta de São Luiz do Maranhão.
- Snyder, John P. 1982. *Map Projections Used by the U.S. Geological Survey*. 2nd edn. United States Government Printing Office. Washington, D.C.
- Sonderegger, J., Petry, P., Campos dos Santos, J. L., & Alves, N.F. 1998. An Entomological Collections Database Model for INPA. *Pages 421–434 of: Ling, Tok Wang, Ram, Sudha, & Lee, Mong-Li (eds), Conceptual Modeling - ER '98, 17th International Conference on Conceptual Modeling, Singapore, November 16-19, 1998, Proceedings*. Lecture Notes in Computer Science, vol. 1507. Springer Verlag.
- Staab, Steffen. 2002 (November). *The Semantic Web - New Way to Integrate and Present Information*. Workshop on Ontology for Biology, Heidelberg, Germany.
- Stein, Lincon D. 1996. *How to Set-Up and Maintain a Web Site*. 2nd edn. Addison-Wesley Pub Co.
- Stinsen, Barry. 2001. *PostgreSQL*. New Riders Publisher. Indianapolis, IN.
- Stockwell, D. R. B. 2000. Understanding Biodiversity Through Environmental Informatics. *NPACI and SDSC En Vision*, 16(3).
- Stonebraker, Michael, & Rowe, Lawrence A. 1985. The design of the POSTGRES. *Pages 374–387 of: Proceedings of 1985 SIGMOD International Conference on Management of Data*. ACM/SIGMOD.
- Stonebreaker, Michael, Moore, Dorothy, & Brown, Paul. 1999. *Object Relational DBMS: Tracking The Next Great Wave*. 2nd edn. Morgan Kaufmann Publisher.
- Sventek, Joseph, & Coulson, Geoffrey. 2000. *Middleware 2000: IFIP/ACM International Conference on Distributed System Platforms*. Lecture Notes in Computer Science, vol. 1, no. 1795. New York, NY: Springer Verlag.
- Taxonomy. 2002 (August). *Taxonomy and Classification*. Sultan Qaboos University - College of Agriculture. <http://squ.edu.om/agr/OnlineCourses/Biol12020/Taxonomy>, accessed on 19/08/2002.
- Terborgh, John. 1992. *Diversity and The Tropical Rain Forest*. W H Freeman & Co.;
- The United Nations Program of Action from Rio. 1992 (January). *Agenda 21*. Report.

REFERENCES

- Tsou, M. H., & Battenfield, B. P. 2000 (October). *Agent-based mechanism for distributing geographic information services on the Internet*. GIScience 2000: The First International Conference on Geographic Information Science.
- Tsui, James Bao-Yen. 2000. *Fundamentals of Global Positioning System Receivers: A Software Approach*. 1st edn. Wiley-Interscience.
- Wainright, Peter, Ahmad, Afrasiab, & Lint, Michael. 2002. *Professional Apache 2.0*. 1st edn. Wrox Press Inc. Chicago, Illinois.
- Welling, L., & Thomson, L. 2001. *PHP and MySQL Web Development*. Sams Publishing. Indianapolis, IN.
- Wheeler, Q. D., & Cracraft, J. 1997. Taxonomic preparedness: are we ready to meet the Biodiversity Challenge? *Understanding and Protecting Our Biological Resources*, 435–446.
- White, C. 1999 (April). *Using information portals in the enterprise*. DM Review Magazine.
- Wieczorek, J. 2002a (January). *Manual of the MaNIS Georeferencing Error Calculator*. University of California, Berkeley.
- Wieczorek, J. 2002b (June). *The Mammal Networked Information System - Georeferencing Guidelines*. University of California, Berkeley.
- Wijnstekers, Willem. 1992. *The Evolution of CITES: a reference to the Convention on International Trade in Endangered Species of Wild Fauna and Flora*. 6th edn. Lousanne: CITES Secretariat.
- Williams, Hugh E., & Lane, David. 2002. *Web Database Applications with PHP & MySQL*. O'Reilly & Associates.
- Wilson, E. O., & Peters, F. M. 1988. *Biodiversity*. National Academy Press. Washington, D.C.
- World Conservation Monitoring Centre. 1999 (June). *Feasibility Study for a Harmonised Information Management Infrastructure for Biodiversity-related Treaties*. Report.
- XYZFind Corporation. 2001. *XYZFind Server User's Guide - Version 2*. XYZFind Corporation.
- Yeager, Nancy J., & McGrath, Robert E. 1996. *Web Server Technology: The Advanced Guide for World Wide Web Information Providers*. Morgan Kaufmann.
- Yerxa, G. 1999 (July). *The Best Bets for Web Development, Servers and Peripherals*. <http://www.nwc.com/1020/1020f1.html>, accessed on 2002/07/22.
- Young, Michael J. 2001. *XML Step by Step*. Microsoft Press.

Author's Bibliography

Papers published related to this study:

1. Sonderegger, J. Petry, P., **Campos dos Santos, J. L.** and Alves, N. F. (1998). *An entomological collection database for INPA*. In: Ling, T.W., Ram, S and Lee, M.L. (eds.) Proceedings of the 17th International Conference on Conceptual Modeling — ER'98, Singapore p.421-434.
2. **Campos dos Santos, J. L.**, de By, R. A. and Magalhães, C. (2000). *A case study of INPA's Bio-DB and an approach to provide an open analytical database environment*. In International Archives of Photogrammetry and Remote Sensing, 33(B4): 155-163.
3. Magalhães, C., **Campos dos Santos, J. L.** and Salem, J. I. (2001). *Biological Collection Automation and Information about the Amazonian Biodiversity*. In the Parcerias Estratégicas, Brasília, 12: 294-312. URL:<http://www.mct.gov.br/CEE/revista/Parcerias12/16celio.pdf>, 2001/17/12 — (In Portuguese).
4. Souza, E. N. and **Campos dos Santos, J. L.** (2001). *NetR-X: A tool to control the access to the INPA's Local Area Network*. In the Proceedings of the VI ENAPI — Annual meeting of research and scientific initiation of the UNOESTE (Universidade do Oeste Paulista, Presidente Prudente/São Paulo, Brazil — (In Portuguese).
5. Roriz Filho, H. S. and **Campos dos Santos, J. L.** (2001). *Validating the spatial aspects of a biological collection data model*. In the Proceedings of the VI ENAPI — Annual meeting of research and scientific initiation of the UNOESTE (Universidade do Oeste Paulista, Presidente Prudente, São Paulo, Brasil — (In Portuguese).

6. Abinader Junior, F. and **Campos dos Santos, J. L.** (2001): *Georeferencing and the implementation of a INPA's fish collection database*. In the Proceedings of the 1st Conference of Informatics in Amazonian, Manaus, Amazonas, Brasil, April 2001, pp 295-302. — (In Portuguese).
7. Cohen, D. and **Campos dos Santos, J. L.** (2001). *Management of URL access in an integrated network environment*. In the Proceedings of the 1st Conference of Informatics in Amazonian, Manaus, Amazonas Brasil, April 2001, pp 303-314. — (In Portuguese).
8. Ribeiro, L. E. and **Campos dos Santos, J. L.** (2001). *Study of implementation and performance of Lynx Servers in the Inpanetwork environment*. In the Proceedings of the 1st Conference of Informatics in Amazonian, Manaus, Amazonas, Brasil, April 2001, pp 315-326. — (In Portuguese).
9. Rios, R. M. C., and **Campos dos Santos, J. L.**, 2002 (March). Backbone da INPANetwork. Tech. rept. MCT/INPA/GTI. — (In Portuguese).
10. **Campos dos Santos, J. L.**, de By, R. A., Apers, P. M. G. and Magalhães, C. 2002. *Facilitating interdisciplinary sciences by the integration of CLOSi-based database with bio-metadata*. In the International Archives of Photogrammetry and Remote Sensing, Vol. 34 — Part IV, Ottawa, Canada, July 2002.
11. **Campos dos Santos, J. L.**, de By, R. A., Apers, P. M. G. and Magalhães, C. (2002). *Clustered object schemas for INPA's biological collections data*. In: Callaos, N., Hernandez-Encinas, L. and Yetim, F. (eds), ISBN: 980-07-8150-1. In the Proceedings Volume I (Information Systems Development I) of the SCI 2002 (6th World Multiconference on Systemics, Cybernetics and Informatics), Orlando, Florida — USA, July 2002, pp 38-44.
12. **Campos dos Santos, J. L.** and de By, R. A. (2002). *XML-based Metadata Management for Biological Data*. In W.Pillmann and K. Tochtermann (eds.): Environmental Communication in the Information Society; EnviroInfo Vienna 2002 (16th International Conference: Informatics for Environmental Protection); Part 1.; September, Vienna, Austria, 2002; pp 408-415.
13. Kandeh, J. M. K., **Campos dos Santos, J. L.** and Kumar L. (2002). *Mapping Invasive Alien Plant Species in South African*. Poster presented at the Environmental Communication in the Information Society; EnviroInfo Vienna 2002 (16th International Conference: Informatics for Environmental Protection); September, Vienna, Austria, 2002.

14. Kandeh, J. M. K., **Campos dos Santos, J. L.** and Kumar L. (2002). *Automatic Mapping and Monitoring of Invasive Alien Plant Species, The South African Experience*. In the Proceeding of the 18th International CODATA Conference — 'Frontiers of Scientific and Technical Data', Montreal, Canada, September-October, 2002.
15. **Campos dos Santos, J. L.** , Bissett, M. G., de By, R. A. and Amadio, S. (2003). *Integrating a CLOSI-based Database with a Retrospective Georeferencing*. In the 5th International Symposium on Environmental Software Systems, Sammering, Austria, May, 2003.

Author's Bibliography

Abbreviations

ABI	Association for Biodiversity Information
AES	Advanced Encryption Standard
AGLS	Australian Government Locator Service
ASC	Association of Systematic Collections
ASP	Active Server Pages
ATM	Asynchronous Transfer Mode
BCDAM	Banco de Dados Compartilhado da Amazônia
BDT	Base de Dados Tropicais
BINbr	Rede de Informação em Biodiversidade
BIN21	Biodiversity Information Network
BioME	Biodiversity Metadata Editor
BIONTE	Biomass and Nutrients
BIOTA	Programa de Pesquisas em Conservação Sustentável da Biodiversidade
BIS	Biodiversity Information Systems
BLOBs	Binary Large Objects
BOREAS	Boreal Ecosystem-Atmosphere Study
CAS	California Academy of Science
CBD	Convention on Biological Diversity
CCD	Convention to Combat Desertification
CCG	Centro de Coordenação Geral
CE	Client Environment
CGI	Common Gateway Interface
CITES	Conventions on International Trade in Endangered Species of Wild Fauna and Flora
CLOSi	Clustered Object Schema for INPA's Biodiversity Data Collections
CMS	Convention on Migratory Species of Wild Animals

Abbreviations

CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CNS	Central Node server
CONABIO	Comisión Nacional de Biodiversidad
COTS	Commercial-Off-The-Shelf
CPAA	Centro de Pesquisas Agroflorestal da Amazônia Ocidental
CPATU	Centro de Pesquisas Agroflorestal da Amazônia Oriental
CRV	Centro Regional de Vigilância
CSDGM	Content Standards for Digital Geospatial Metadata
CSS	Cascading Style Sheets
CSSs	Closed Source Systems
DAA	Digest Access Authentication
DAML	DARPA Agent Markup Language
DBMS	Database Management System
DCES	Dublin Core Element Set
DCW	Digital Chart of the World
DFID	Department for International Development
DIP	Digital Image Processing
EDMI	Entomological Database Model
EdNA	Education Network Australia
EIP	Enterprise Information Portal
EIS	Environmental Information Systems
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FCCC	Framework Convention on Climate Change
FGDC	Federal Geographic Data Committee
Gb	Gigabyte
GBIF	Global Information Facility
GEC	Georeferencing Error Calculator
GIF	Graphics Interchange Format
GIS	Geographical Information Systems
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
IABIN	Inter-American Biodiversity Information Network
IBGE	Instituto Brasileiro de Geografia e Estatística
IEPA	Instituto de Pesquisas Científicas e Tecnológicas do Estado do Amapá
INBio	Instituto Nacional de Biodiversidad
INPA	Instituto Nacional de Pesquisas da Amazônia
INPE	Instituto Nacional de Pesquisas Espaciais

IP	Internet Protocol
IPGRI	International Plant Genetic Resources Institute
ISLSCP	International Satellite Land Surface Climatology Project
ISO	International Organisation for Standardisation
IT	Information Technology
ITC	International Institute for Geo-Information Science and Earth Observation
JMP	Japan Metadata Profile
Kb	Kilobyte
LAN	Local Area Network
LBA	Large-Scale Biosphere-Atmosphere Experiment in Amazonia
MB	Megabyte
MCT	Ministério da Ciência e Tecnologia
MMA	Ministério do Meio Ambiente
MPEG	Museu Paraense Emílio Goeldi
MTS	Multi-Threaded Server application
MVZ	Museum of Vertebrate Zoology
NBII	National Biological Information Infrastructure
NEODAT	Inter-Institutional Database of Fish Biodiversity in the Neotropics
NGOs	Non-governmental organisations
NPACI	National Partnership for Advanced Computational Infrastructure
NSCA	Natural Science Collections Alliance
NSDI	National Spatial Data Infrastructure
ODBC	Open Data Base Connectivity
OLN	Online Local Node
OPM	Object Protocol Model
OSS	Open Source Software
OSUIC	Ohio State University Insect Collection
OWL	Ontology Web Language
PDA	Personal Digital Assistants
Perl	Practical Extraction and Report Language
PHP	Personal Home Page
PI	Primary Investigator
PNE	Plantas do Nordeste
PPA	Plano Plurianual
PPG-7	Programa Piloto para a Proteção das Florestas Tropicais do Brasil
PPP	Point-to-Point Protocol

Abbreviations

Ramsar	Convention on Wetlands of International Importance
RDBMS	Relational Database Management System
RDF	Resource Description Language
SCP	Scientific Collections Program
SDDS	Special Data Dissemination Standard
SDM	Semantic Data Model
SDO	Spatial Data Object
SDSC	San Diego Super Computer Center
SHIFT	Studies of Human Impact on Floodplains and Forests in the Tropics
SIVAM	Sistema de Vigilância da Amazônia
SQL	Structured Query Language
SLIP	Serial Line Internet Protocol
SSL	Secured Socket Layer
TCP	Transmission Control Protocol
TTA	Three Tier Architecture
UFAM	Universidade Federal do Amazonas
URL	Universal Resource Locator
UTP	Unshield Twisted Pair
WCNH	World Cultural and Natural Heritage
WDCM	World Data Centre for Microorganisms
XML	Extended Markup Language
XSD	XML Schema Definition
XSL	Extensible Style Language
XXE	XML Editor Standard Edition

Appendix A: The Syntactic Definition for CLOSi Schemas

BNF Grammar for CLOSi

BNF or Backus-Naur Form, after Donald Knuth's suggestion, is one of the most commonly used metasyntactic notations for specifying the syntax of programming languages and command sets (Knuth, 1964). Here, we present the BNF defined for clusters, object classes and relationships, Is.a object classes, attributes, attributes constraints and control value classes of CLOSi.

```
<CLOSi_Schema> ::= <cluster>
                  | <cluster> <CLOSi_schema>;

<cluster> ::= <cluster_name>
              <object_classes>;

<cluster_name> ::= <text>;

<object_classes> ::= <object_class>
                    | <object_class> <object_classes>;

<object_class> ::= <object_class_name>
                  <class_relationship>
                  <object_class_description>
                  <object_class_attributes>;

<object_class_name> ::= 'CLASS' <string>;

<class_relationship> ::= <Is_a>
                       | <null>;
```

Appendix A: The Syntactic Definition for CLOSi Schemas

```
<Is_a> ::= 'Is_a Relationship Type to' <super_class_name>;

<super_class_name> ::= <string>;

<object_class_description> ::= 'CLASS DESCRIPTION' ':' <text>;

<object_class_attributes> ::= <object_attribute>
                             <attribute_description>
                             | <object_attribute> <attribute_description>
                             <object_class_attributes>;

<object_attribute> ::= <standard_attribute>
                       | <composite_attribute>
                       | <derivation_attribute>
                       | <geo_attribute>;

<standard_attribute> ::= 'ATTRIBUTE'
                        <attribute_name> ':'
                        <single or set-of or list-of>
                        <cardinality> '-' <value_class>;

<composite_attribute> ::= 'COMPOSITE ATTRIBUTE'
                        <attribute_name> ':'
                        <single or set-of or list-of>
                        <cardinality> '-' <value_class>;

<derivation_attribute> ::= 'ATTRIBUTE'
                        <attribute_name> '(!)'
                        <derivation_class_from>
                        <derivation_class_to>;

<geo_attribute> ::= 'ATTRIBUTE'
                  <attribute_name> '(Latitude,Longitude):'
                  <single or set-of or list-of>
                  | <Coordinates>
                  | <Distance>;

<Coordinates> := <geographic>
                | <rectangular>
                | <nodes>;

<geographic> ::= <cardinality> <value_class> ,
                <cardinality> <value_class>
                | <cardinality> <value_class> ,
                <cardinality> <value_class> ,
                <cardinality> <value_class> ,
                <cardinality> <value_class>
```

```

    <hemisphere>;

<rectangular> ::= <cardinality> <value_class>
    | <cardinality> <value_class>
    <cardinality> <value_class>
    <hemisphere>;

<nodes> ::= <cardinality> <value_class> <hemisphere>
    | <cardinality> <value_class>
    <cardinality> <value_class>
    <hemisphere>;

<Distance> ::= <cardinality> <value_class> ,
    <cardinality> <value_class>;

<hemisphere> ::= <value_class>;

<attribute_name> ::= <string>;

<derivation_class_from> ::= 'DERIVATION' ':' '!'<string>;

<derivation_class_to> ::= '[' <string> ']';

<single or set-of or list-of> ::= set-of
    | list-of
    | <null>;

<value_class> ::= '['<$> <controlled_value_class> '<$>'
    | <code_type>
    | <object_class>
    | '[' <object_classes> '+' ']';

<attribute_description> ::= 'DESCRIPTION' ':' <text>;

<cardinality> ::= [ <min> , <max> ]
    | [ <min> , ]
    | <null>;

<min> ::= <integer_value>;

<max> ::= <integer_value>;

<controlled_value_class> ::= <controlled_value_class_name>
    { <enum_constant> }
    <class_standard> <code_type>
    <class_description>
    <declared_domain>

```

Appendix A: The Syntactic Definition for CLOSi Schemas

```

| <controlled_value_class_name>
  { <ranges> }
  <class_standard> <code_type>
  <class_description>
  <declared_domain>;

<controlled_value_class_name> ::= <class_name>;

<enum_constant> ::= <constant_code_desc>
  | (<enum_constant> ,
    <constant_code_desc>);

<constant_code_desc> ::= <constant> [<code>]
  [<value_description>];

<value_description> ::= " <text> ";

<code> ::= (alphanumeric characters or special symbols : + -
  * / \ ! = . */);

<ranges> ::= <range_desc>
  | <ranges> , <range_desc>;

<range_desc> ::= <range>
  | <range> [, <value_description>];

<range> ::= <number>
  | (- <number> )
  | <number> - <number>
  | ( - <number> ) - <number>
  | ( - <number> ) - ( - <number> );

<code_type> ::= <null>
  | <code_data_type>
  | CODE_TYPE ':' <code_data_type>;

<code_data_type> ::= CHAR '(' <integer> ')
  | VARCHAR '(' <integer> ')
  | INTEGER
  | SMALLINT
  | TINYINT
  | REAL
  | FLOAT
  | FLOAT '(' <integer> ')
  | DECIMAL
  | DECIMAL '(' <integer> ')

```

Appendix A: The Syntactic Definition for CLOSi Schemas

```
| DECIMAL '(' <integer> ',' <integer> ')'  
| NUMERIC  
| NUMERIC '(' <integer> ')'  
| NUMERIC '(' <integer> ',' <integer> ')'  
| DATETIME  
| TIMESTAMP  
| TEXT  
| IMAGE  
| CURRENCY;
```


Appendix B: CLOSi

Controlled Value Classes

Control Value Classes are optional lists of values that are associated with attributes of classes in a cluster described in a CLOSi schema. Each Control Value Class is described by the following descriptors: the class name, a list of values available to the class, the code type, the name of the cluster and the object class in which the controlled value class was associated, the standard value defined for the class and the description of the class. The CLOSi schema is detailed in Chapter 4.

TypeOfObject

```
CONTROLLED VALUE CLASS TypeOfObject
{"Lot", Lot},
("Specimen", Specimen),
("Part of Specimen", Part of Specimen),
("Product of Specimen", Product of Specimen),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster CollectionManagement in
Class Collection_Object.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the object types.
```

CountModifier

```
CONTROLLED VALUE CLASS CountModifier
{"=", "indicates count is exactly the given number"},
{"+", "+", "indicates count is at least the given number"},
{"ca.", ca, "circa; indicates the count is approximate" },
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(3)
DESCRIBED_IN: Cluster CollectionManagement in Class Lot.
STANDARD: Unknown
DESCRIPTION: Controlled vocabulary for CountModifier.
```

Sex

```
CONTROLLED VALUE CLASS Sex
{"Masculine", M},
{"Feminine", F},
{"Not Examined", Ne},
{"Indeterminate", Ind},
{"Hermaphrodite", H}
CODE_TYPE: VARCHAR(3)
DESCRIBED_IN: Cluster CollectionManagement in Class Specimen.
STANDARD: Not Examined
DESCRIPTION: Control vocabulary for sex types.
```

NoAvailability

```
CONTROLLED VALUE NoAvailability
{"Loaned", Loaned},
{"Inspection", Inspection},
{"Research", Research},
{"Display", Display},
{"Exchange", Exchange},
{"Gift", Gift},
{"Under Repair", Under Repair},
{"Destroyed", Destroyed},
{"Unknown", Unknown}
CODE_TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster CollectionManagement.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the attribute
NoAvailabilityDue.
```

PreparationMethod

```
CONTROLLED VALUE PreparationMethod
{"Alcohol", Alcohol },
{"Dry", Dry},
{"Humid", Humid},
{"Mounted Slides", Mounted Slides},
{"Layered", Layered},
{"Pinned", Pinned},
{"Glued", Glued},
{"Fluid", Fluid},
{"Fluid and Skull", Fluid and Skull},
{"Parts in Alcohol", Parts in Alcohol},
{"Parts Pinned", Parts Pinned},
{"Parts Glued", Parts Glued},
{"Indeterminate", Indeterminate}}
```


CODE_TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster CollectionManagement in
Class Object.Situation.
STANDARD: Indeterminate
DESCRIPTION: Control vocabulary for the attributes preparation.

TypeOfPlace

CONTROLLED VALUE TypeOfPlace
{("Point", Point),
("Hydrographic basin", Hydrographic basin),
("Sub Hydrographic basin", Sub Hydrographic basin),
("Tributary", Tributary),
("Community", Community),
("Region", Region),
("District", District),
("Municipality", Municipality),
("State", State),
("Country", Country),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(10)
DESCRIBED_IN: Cluster Locality in Class Named.Place.
STANDARD: Unknown
DESCRIPTION: Describes what kind of locality the name defines.

HabitatType

CONTROLLED VALUE HabitatType
{("River", River),
("Mud of River Banks", Mud of River Banks),
("Sand of River Banks", Sand of River Banks),
("Lake", Lake),
("Mouth Bay", Mouth Bay),
("Beach", Beach),
("Flooded Area", Flooded Area),
("Flooded Forest", Flooded Forest),
("Primary Forest", Primary Forest),
("Secondary Forest", Secondary Forest),
(" Deforested Area", Deforested Area),
("Canopy", Canopy),
("Wood Debris", Wood Debris),
("Aquatic Vegetation", Aquatic Vegetation),
("Marginal Vegetation", Marginal Vegetation),
("Rock Bottom", Rock Bottom),
("Loose Rocks", Loose Rocks),
("Falls", Falls),

```
("Rapids", Rapids),
("Creeks", Creeks),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Locality in Class Habitat.
STANDARD: Unknown
DESCRIPTION: Defines a type for the habitat.
```

GeoRefObjectType

```
CONTROLLED VALUE GeoRefObjectType
{"Point", Point },
("Bounding Box", Bounding Box),
("Line Segment", Line Segment),
("Chain", Chain),
("Polygon", Polygon),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(12)
DESCRIBED_IN: Cluster Locality in Class Geo-Reference-Object.
STANDARD: Unknown
DESCRIPTION: Controlled vocabulary for indicating the
subclass of a GeoReferencedObject.
```

LocalityValueSource

```
CONTROLLED VALUE LocalityValueSource
{"Map", Map },
("GPS", GPS ),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(03)
DESCRIBED_IN: Cluster Locality in Class Geo-Reference-Object.
STANDARD: Unknown
DESCRIPTION: Controlled vocabulary for indicating the source
of the longitude and latitude values.
```

TypeOfAgent

```
CONTROLLED VALUE TypeOfAgent
{"Person", Person),
("Organisation", Organisation),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(12)
DESCRIBED_IN: Cluster Agent-Of-Collection in Class Agent.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the type of an agent.
```

PhoneType

```
CONTROLLED VALUE PhoneType
{"Commercial", Commercial},
("Fax", Fax),
("Lab", Lab),
("Residential", Residential),
("Pager", Pager),
("Cellular", Cellular),
("Telex", Telex),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Agent.Of.Collection in Class Agent.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the type of phone access.
```

RefType

```
CONTROLLED VALUE RefType
{"Journal", Journal},
("Article", Article),
("Book", Book),
("Book Chapter", Book Chapter),
("PhD Thesis", PhD Thesis),
("MSc Thesis", MSc Thesis),
("Technical Report", Technical Report),
("In Proceedings", In Proceedings),
("Miscellaneous Publication", Miscellaneous Publication),
("Unpublished Material", Unpublished Material),
("Web Publication", Web Publication),
("Museum Notes", Museum Notes),
("Research Project Report", Research Project Report),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(25)
DESCRIBED_IN: Cluster Reference in Class Reference.Work.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for reference work or types
of publication.
```

PublicStatus

```
CONTROLLED VALUE PublicStatus
{"Manuscript", Manuscript},
("Publication", Publication),
("In Press", In Press),
("Submitted", Submitted),
("Unknown", Unknown)}
```

CODE.TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Reference in Class Reference.Work.
STANDARD: Publication
DESCRIPTION: Control vocabulary for status of publication
of reference work.

StudyType

CONTROLLED VALUE StudyType
{("Field Notes", Field Notes),
("PhD Thesis", PhD Thesis),
("MSc Dissertation", MSc Dissertation),
("Monography", Monography),
("Museum Note", Museum Note),
("Research Project", Research Project),
("Unknown", Unknown)}
CODE.TYPE: VARCHAR(12)
DESCRIBED_IN: Cluster Reference in Class Scientific.Study.
STANDARD: Unknown
DESCRIPTION: Describes the type of the study.

TitleType

CONTROLLED VALUE TitleType
{("Mr", Mr),
("Mrs", Mrs),
("Ms", Ms),
("Sir", Sir),
("Ir", Ir),
("Eng", Eng),
("BSc", BSc),
("MD", MD),
("Dr", Dr),
("Prof", Prof),
("Unknown", Unknown)}
CODE.TYPE: VARCHAR(5)
DESCRIBED_IN: Cluster Agent.Of.Collection in Class Person.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for name of suffixes,
which distinguish members of families with the same name.

FieldOfActivity

CONTROLLED VALUE FieldOfActivity
{("Research Institute", Research Institute),
("University", University),

```
("Museum", Museum),
("Non-Governmental Organisation", Non-Governmental Organisation),
("Federal Government", Federal Government),
("State Government", State Government),
("Industry", Industry),
("Private Enterprise", Private Enterprise),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(25)
DESCRIBED_IN: Cluster Agent in Class Organisation.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the field of activity
of an Organisation.
```

TaxonRankName

```
CONTROLLED VALUE TaxonRankName
{"Kingdom", Kingdom},
("Division", Division),
("Phylum", Phylum),
("Subphylum", Subphylum),
("Superclass", Superclass),
("Class", Class),
("Subclass", Subclass),
("Infraclass", Infraclass),
("Superorder", Superorder),
("Order", Order),
("Suborder", Suborder),
("Infraorder", Infraorder),
("Superfamily", Superfamily),
("Family", Family),
("Subfamily", Subfamily),
("Tribe", Tribe),
("Subtribe", Subtribe),
("Genus", Genus),
("Sub Genus", Sub Genus),
("Species", Species),
("Sub Species", Sub Species),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Taxonomy in Class Classification.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the taxon rank value.
```

TypeOfZoogeographicRegion

CONTROLLED VALUE TypeOfZoogeographicRegion
{("Palearctic", Palearctic),
("Nearctic", Nearctic),
("Holarctic", Holarctic),
("Neotropical", Neotropical),
("Ethiopian", Ethiopian),
("Oriental", Oriental),
("Australian", Australian),
("Oceanic", Oceanic),
("Unknown", Unknown)}
CODE.TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Locality.Of_BiodiversityData in Class.
Named.Place
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the Zoogeographic regions.

RelationTypes

CONTROLLED VALUE RelationTypes
{("Competition", Competition),
("Predation", Predation),
("Parasitism", Parasitism),
("Mutualism", Mutualism),
("Detritivory", Detritivory),
("Unknown", Unknown)}
CODE.TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Taxonomy in Class Taxon_Relation.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the RelationTypes.

DegreeOfConfidenceInDetermination

CONTROLLED VALUE DegreeOfConfidenceInDetermination
{("Expert in the group", Expert in the group),
("Scientist not expert in the group", Scientist not expert
in the group),
("PhD Student", PhD Student),
("MSc Student", MSc Student),
("Undergraduate", Undergraduate),
("Trainee", Trainee),
("Parataxonomist", Parataxonomist),
("Technical Assistant", Technical Assistant),
("Unknown", Unknown)}
CODE.TYPE: VARCHAR(20)
DESCRIBED_IN: Cluster Taxonomy in Class Determination.

STANDARD: Unknown
DESCRIPTION: Control vocabulary for the DegreeOfConfidenceInDetermination.

TransportationInformation

CONTROLLED VALUE TransportationInformation
{("First Class", First Class),
("Air Mail", Air Mail),
("Parcel Post", Parcel Post),
("Insured", Insured),
("Registered", Registered),
("Special Handling", Special Handling),
("Library Rate", Library Rate),
("Freight", Freight),
("Air Freight", Air Freight), ("Express", Express),
("Prepaid", Prepaid),
("UPS", UPS),
("FedEx", FedEx),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(25)
DESCRIBED.IN: Cluster CollectionManagement in Class Loan.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the TransportationInformation.

TypeOfLoanCategory

CONTROLLED VALUE TypeOfLoanCategory
{("As a gift", As a gift),
("As an exchange", As an exchange),
("As a loan at your request", As a loan at your request),
("For examination", For examination),
("As a return of material borrowed by us", As a return of material
borrowed by us)
("As a return of material sent to us for identification", As a return
of material sent to us for identification)
("As a loan to us", As a loan to us),
("For identification", For identification),
("Unknown", Unknown)}
CODE_TYPE: VARCHAR(50)
DESCRIBED.IN: Cluster CollectionManagement in Class Loan.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the TypeOfLoanCategory.

TypeOfThesis

```
CONTROLLED VALUE TypeOfThesis
{"PhD", PhD},
{"MSc", MSc},
{"Unknown", Unknown}
CODE.TYPE: VARCHAR(5)
DESCRIBED_IN: Cluster Reference in Class Thesis.
STANDARD: Unknown
DESCRIPTION: Control vocabulary for the TypeOfThesis.
```


Appendix C: Examples of CLOSi Instantiated Values

To illustrate CLOSi components, we present values that were instantiated to object classes and attributes. The instances in the examples use information from INPA's Vertebrate collection. For a comprehensive descriptions of CLOSi components refer to Chapter 4 of this thesis or (Campos dos Santos, 2000).

Cluster Collection Management

Object Class: Biological_Collection

ATTRIBUTE CollectionID: INPA-Invertebrates-001
ATTRIBUTE Curator: C. Magalhaes
ATTRIBUTE OrganisationAcronym: INPA
ATTRIBUTE TaxonomicGroup: Crustacea

Object Class: Collection_Object

ATTRIBUTE AccessionNumber: 85-012
ATTRIBUTE ObjectID: Catalogue-Book-150
ATTRIBUTE ReceivedAs: Gift
ATTRIBUTE ObjectType: Lot
ATTRIBUTE PlaceInCollection: Invertebrates Collection

Object Class: Object_Situation

ATTRIBUTE Cupboard: INPA-Invertebrates-001-001-XA1
ATTRIBUTE Drawer: INPA-Invertebrates-001-XA1-001
ATTRIBUTE Cataloger: L. Rapp
ATTRIBUTE RegisDate: 20-03-1985 - 12:00:00
ATTRIBUTE Preparation: Alcohol 70%
ATTRIBUTE Preparator: C. Magalhaes

Object Class: Lot

ATTRIBUTE LotDescription: MZUSP6550
ATTRIBUTE FemaleCount: 1
ATTRIBUTE MaleCount: 1
ATTRIBUTE TotalCount: 2

Object Class: Loan

ATTRIBUTE InvoiceNumber: MZUSP-Inv-356-2001
ATTRIBUTE InvoiceDate: 20-03-1985 - 17:00:00
ATTRIBUTE PackedBy: C. Magalhaes
ATTRIBUTE MaterialCategory: Loan at your request
ATTRIBUTE Loaner: G. Rodriguez
ATTRIBUTE Controller: L. Rapp
ATTRIBUTE ShipToAddress: Centro de Ecologia-IVIC,
Apartado 21827, Caracas, Venezuela
ATTRIBUTE LoanPeriod: 12 months

Cluster Collecting Event of Collection

Object Class Collecting Event

ATTRIBUTE FieldNumber: EPA 73-07245
ATTRIBUTE BeginDate: 20-07-1973
ATTRIBUTE EndDate: 28-07-1973
ATTRIBUTE PlaceOfEvent: Rio Tapajos - Monte Cristo,
Para, Brazil
ATTRIBUTE Created: 20-02-2002
ATTRIBUTE Update: 20-02-2002
ATTRIBUTE UpdatedBy: C. Magalhaes

Cluster Locality of Biodiversity Data

Object Class Locality

ATTRIBUTE PlaceNames: Monte Cristo

Object Class Named Places

ATTRIBUTE Name: Monte Cristo
ATTRIBUTE PlaceType: Community
ATTRIBUTE Region: Rio Tapajos
ATTRIBUTE ZoogeographicRegion:
Neotropical

Object Class Habitat

ATTRIBUTE Type: Mud of River Banks

Object Class Cartographic Reference

ATTRIBUTE Name: SB-21 Carta do Brasil ao Milionesimo
ATTRIBUTE Created: 20-02-2002
ATTRIBUTE Updated: 20-02-2002

Cluster Taxonomy

Object Class Taxon Name

ATTRIBUTE Repository: INPA-Invertebrates Collection
ATTRIBUTE TaxAuthors: C. Magalhaes, M. Turkey
ATTRIBUTE ParentTaxon: Pseudothelphusidae
ATTRIBUTE TaxonOrigRef: C. Magalhaes, M. Turkey, 1986
ATTRIBUTE References (!): C. Magalhaes, M. Turkey
ATTRIBUTE Classific: Arthropoda, Crustacea, Malacostraca,
Decapoda, Brachyura,
Pseudothelphusidae, Brasiliiothelphusa,
Brasiliiothelphusa tapajoense

Object Class Classification

ATTRIBUTE Name: Arthropoda
ATTRIBUTE TypeTaxonRankName: Phylum
ATTRIBUTE Name: Crustacea

--> New entry for Object Class Classification

ATTRIBUTE TypeTaxonRankName: Subphylum
ATTRIBUTE Name: Malacostraca

--> New entry for Object Class Classification

ATTRIBUTE TypeTaxonRankName: Class
ATTRIBUTE Name: Decapoda

--> New entry for Object Class Classification

ATTRIBUTE TypeTaxonRankName: Order
ATTRIBUTE Name: Pseudothelphusidae
ATTRIBUTE Name: Brasiliiothelphusa

--> New entry for Object Class Classification

ATTRIBUTE TypeTaxonRankName: Genus
ATTRIBUTE Name: tapajoense

--> New entry for Object Class Classification

ATTRIBUTE TypeTaxonRankName: Species
ATTRIBUTE ClassificationOrigRef: T. Bowman, L. G. Abele, 1982

--> New entry for Object Class Classification

ATTRIBUTE Author: T. Bowman, L. G. Abele
ATTRIBUTE UpdatedBy: C. Magalhaes

Object Class Determination

ATTRIBUTE CollObject: Brasiliothelphusa tapajoense
ATTRIBUTE Taxon: Brasiliothelphusa tapajoense
ATTRIBUTE Determiner (Specialist): C. Magalhaes
ATTRIBUTE Date: 15-07-1985
ATTRIBUTE Created: 20-02-2002
ATTRIBUTE Update: 20-02-2002
ATTRIBUTE UpdatedBy: C. Magalhaes

Cluster Agent of Collection

Object Class Agent

ATTRIBUTE: Person

Object Class Person

ATTRIBUTE Title: Dr
ATTRIBUTE FirstName: M.
ATTRIBUTE MidName:
ATTRIBUTE FamName: Turkey
ATTRIBUTE Position: Curator

--> New entry for Object Class Person

ATTRIBUTE Title: Dr
ATTRIBUTE FirstName: C.
ATTRIBUTE MidName: U.
ATTRIBUTE FamName: Magalhaes
ATTRIBUTE PersonAcronym: Celiomag
ATTRIBUTE Position: Senior Researcher

Cluster Reference

Object Class Reference_Work

ATTRIBUTE Author: C. Magalhaes, M. Turkey
ATTRIBUTE Title: Brasiliothelphusa, a new Brazilian
freshwater-crab genus (Crustacea: Decapoda:
Pseudothelphusidae)
ATTRIBUTE Year: 1986
ATTRIBUTE Keyword: Crustacea, Decapoda, Freshwater-crab
ATTRIBUTE Author: C. Magalhaes
ATTRIBUTE Title: Taxonomic review of Brazilian freshwater
crab from the family Pseudothelphusidae
(Crustacea, Decapoda)
ATTRIBUTE Year: 1986
ATTRIBUTE Keyword: Crustacea, Decapoda, Freshwater-crab

ATTRIBUTE Author: T. Bowman; L. G. Abele
ATTRIBUTE Title: Classification of the recent Crustacea. P. 1-27.
 In Abele, L. G. (ed.), Systematics, the
 Fossil Record, and Biogeography. New
 York, Academic Press. 319 pp.
 (The Biology of Crustacea, v.1).
ATTRIBUTE Year: 1982

Object Class Book

ATTRIBUTE Publisher: L. G. Abele (ed.), Academic Press, New York
ATTRIBUTE Volume: The Biology of Crustacea - 1
ATTRIBUTE Series: Systematics, the Fossil Record, and
 Biogeography

Object Class In book

ATTRIBUTE Chapter: 1
ATTRIBUTE Pages: 1-27

Object Class Article

ATTRIBUTE Journal: Senckenbergiana biologica
ATTRIBUTE Volume: 66
ATTRIBUTE Number: 4/6
ATTRIBUTE Pages: 371-376
ATTRIBUTE Created: 20-02-2002
ATTRIBUTE Update: 20-02-2002
ATTRIBUTE UpdatedBy: C.
Magalhaes

--> New entry for Object Class Article

ATTRIBUTE Journal: Amazoniana
ATTRIBUTE Volume: 9
ATTRIBUTE Number: 4
ATTRIBUTE Pages: 609-636
ATTRIBUTE Created: 20-02-2002
ATTRIBUTE Update: 20-02-2002
ATTRIBUTE UpdatedBy: C. Magalhaes

Appendix D: Metadata of a Crustacean Collection

This metadata description is the output of the implemented system, described in Chapter 5. The description was originated from INPA's crustacean collection. This example is provided only to illustrate metadata that are compliant with the FGDC standard (elements mandatory, applicable and mandatory if applicable). Note that variables which can be instantiated are those followed by the character ':', otherwise, the variables are of type compound. For details about the FGDC standard and the bio-metadata implementation, see Chapter 5.

IDENTIFICATION INFORMATION:

Citation

Originator: Amazoniana

Publication date: 1988

Title: A catalogue of type specimens of Crustacea
in the Invertebrate Collection of the
Instituto Nacional de Pesquisas da Amazonia,
Manaus, Brazil, up to January, 1988.- Amazoniana,
10(3): 267-282. [1988]

Publication information:

Publication place: Kiel, Germany

Publisher: Max-Planck-Institut fur Limnologie,
AG Tropenokologie

Description

Abstract: Data and pertinent information on the
type of specimens of 45 species of
crustaceans deposited in the Crustacean
Section of the Systematic Invertebrate
Collection of the Instituto Nacional de
Pesquisas da Amazonia (Manaus, AM) are
presented. The names of museums and

Appendix D: Metadata of a Crustacean Collection

institutions where the paratypes of some of these species were deposited are indicated, along with their respective registration numbers, when possible.

Purpose: Listing the type species of crustaceans deposited in the Crustacea Collection of INPA until January, 1988.

Type period of content:

Multiple Date/Time

Calendar date: 197610

Single Date/Time

Calendar date: 198801

Currentness reference: source type specimens deposited in the collection

Status

Progress: Complete

Maintenance and Update Frequency: Irregular

Spatial Domain:

Keywords

Theme

Keyword Thesaurus: None

Theme Keyword: Crustacea

Theme Keyword: INPA

Theme Keyword: Type collection

Theme Keyword: Catalogue

Place

Place Keyword Thesaurus: None

Place Keyword: Amazon region

Place Keyword: Brazil

Place Keyword: Colombia

Place Keyword: Amazonas

Place Keyword: Para

Place Keyword: Mato Grosso

Place Keyword: Departamento de Valle

Access Constraints: Partial Use Constraints: Scientists, governmental agencies, persons intending research activities

Point of Contact

Contact Organisation Primary

Appendix D: Metadata of a Crustacean Collection

Contact Organisation: INPA
Contact Position: Curator of the Invertebrate
Collection
Contact Address:
Address Type: postal and electronic mailing
and physical address
Address: Av. Andre Araujo, 2936
City: Manaus
State: Amazonas
Postal Code: 69060-001
Country: Brazil
Contact Voice Telephone: (+55 92) 643-3334

Native Data Set Environment: Data set uses Sistema
de Gerenciamento de Colecao - SGC, v. 3.1, running
under DOS 3.x

TAXONOMY INFORMATION

Keywords/Taxon

Taxonomic Keyword Thesaurus: None
Taxonomic Keywords: Collection
Taxonomic Keywords: Multiple species
Taxonomic Keywords: Invertebrates
Taxonomic Keywords: Crustacea
Taxonomic Keywords: Branchiopoda
Taxonomic Keywords: Maxillopoda
Taxonomic Keywords: Malacostraca

Taxonomic System

Classification System/Authority
Classification System Citation: According to
Bowman & Abele (1982)
Classification System Modification: None
Identification Reference:
Identifier:
Taxonomic Procedure: Morphological analysis, comparisons
with previously identified material,
use of available keys
Taxonomic Complexness: High degree of completeness; specimens
identified and described by specialists

Voucher

Specimen: zoological specimens
Repository: INPA, Invertebrate Collection, Curator of

Appendix D: Metadata of a Crustacean Collection

Non-Insect Invertebrates

General Taxonomic Coverage: all type specimens of Crustacea deposited in the Invertebrate Collection of INPA until January, 1988

Taxonomic Classification

Taxon Rank Name: Kingdom

Taxon Rank Value: Animalia

Taxonomic Classification

Taxon Rank Name: Phylum

Taxon Rank Value: Arthropoda

Taxonomic Classification

Taxon Rank Name: Sub-Phylum

Taxon Rank Value: Crustacea

Access Constraints: None Use Constraints: Users are recommended to cite Invertebrate Collection of INPA as the repository of the type specimens

Point of Contact

Contact Organisation Primary

Contact Organisation: INPA

Contact Position: Curator of the Invertebrate Collection

Contact Address:

Address Type: postal and electronic mailing and physical address

Address: Av. Andre Araujo, 2936

City: Manaus

State: Amazonas

Postal Code: 69060-001

Country: Brazil

Contact Voice Telephone: (+55 92) 643-3334

Browse Graphic:

Data Set Credit:

Security Information:

Native Data Set Environment

Cross Reference: Crustacean Collection of Museu Paraense

Appendix D: Metadata of a Crustacean Collection

Emilio Goeldi (Belm, PA - Brazil), data set organised in the same SGC platform. This is a complementary data set of Amazonian crustaceans.

Analytical Tool

Analytical Tool Description: None

DATA QUALITY INFORMATION

Variables to be described are as follows:

Attribute Accuracy

Attribute Accuracy Report:

Logical Consistency Report:

Completeness Report:

Positional Accuracy

Horizontal Positional Accuracy

Horizontal Positional Accuracy Report:

Lineage

Source Information

Source Citation

Originator:

Publication Date:

Title:

Geospatial Data Presentation Form:

Source Scale Denominator:

Type of Source Media:

Source Time Period of Content

Single Date/Time

Calendar Date:

Source Currentness Reference:

Source Citation Abbreviation:

Source Contribution:

SPATIAL DATA ORGANISATION INFORMATION

Variable to be described is as follows:

Direct Spatial Reference Method:

SPATIAL REFERENCE INFORMATION

Variables to be described are as follows:

Horizontal Coordinate System Definition

Appendix D: Metadata of a Crustacean Collection

Planar

Grid Coordinate System

Grid Coordinate System Name:

Universal Transverse Mercator

UTM Zone Number:

Transverse Mercator

Scale Factor at Central Meridian:

Longitude of Central Meridian:

Latitude of Projection Origin:

False Easting:

False Northing

Planar Coordinate Information

Planar Coordinate Encoding Method:

Coordinate Representation

Abscissa Resolution:

Ordinate Resolution:

Planar Distance Units:

Geodetic Model

Horizontal Datum Name:

Ellipsoid Name:

Semi-major Axis:

Denominator of Flattening Ratio:

ENTITY AND ATTRIBUTE INFORMATION

Entity and attributes to be described are as follows:

Detailed Description

Entity Type

Entity Type Label:

Entity Type Definition:

Attribute

Attribute Label:

Attribute Domain Values

Codeset Name:

DISTRIBUTION INFORMATION

Distributor

Contact Organisation Primary

Contact Organisation: INPA

Contact Position: Curator of the Invertebrate
Collection

Appendix D: Metadata of a Crustacean Collection

Contact Address:

Address Type: postal and electronic mailing
and physical address
Address: Av. Andre Araujo, 2936
City: Manaus
State: Amazonas
Postal Code: 69060-001
Country: Brazil
Contact Voice Telephone: (+55 92) 643-3334

Distribution Liability: none

Standard Order Process

Digital Form
Digital Transfer Information
Format Name: MDB

Digital Transfer Option

Online Option

Computer Contact Information

Network Address

Network Resource Name: ftp://inpa.gov.br/colecoes

Access Instructions: Anyone with access to the Internet may connect to INPA's server via anonymous ftp and download available INPA crustacean digital data. To access ftp to INPA's server, login as anonymous, enter your e-mail address at the password prompt, change to the colecoes directory for a sampling of digital data files. Use the ftp 'get' command to transfer readme file for further instructions.

Offline Option

Offline Media: 3-1/2 inch floppy disk

Recording Capacity

Recording Density: 1.44

Recording Density Units: megabytes

Recording Format: tar

Recording Format: MS-DOS

Appendix D: Metadata of a Crustacean Collection

METADATA REFERENCE INFORMATION

Metadata Date: 200212

Metadata Review Date:

Metadata Future Review Date:

Metadata Standard Name: Crustacea Type Collection - INPA

Metadata Time Convention: Manaus, 13 Dec 2002

Metadata Access Constraints: None

Metadata Use Constraints: None

Metadata Security Information

Metadata Security Classification System:

Metadata Security Classification: Free

Metadata Security Handling: None

Metadata Extensions:

Online Linkage: <http://www.inpa.gov.br/colecoes>

Profile Name: FGDC-STD-001-1998

List of ITC Ph.D. Thesis

1. **Akinyede**, 1990, Highway cost modelling and route selection using a geotechnical information system
2. **Pan He Ping**, 1990, 90-9003757-8, Spatial structure theory in machine vision and applications to structural and textural analysis of remotely sensed images
3. **Bocco Verdinelli, G.**, 1990, Gully erosion analysis using remote sensing and geographic information systems: a case study in Central Mexico
4. **Sharifi, M.**, 1991, Composite sampling optimization for DTM in the context of GIS
5. **Drummond, J.**, 1991, Determining and processing quality parameters in geographic information systems
6. **Groten, S.**, 1991, Satellite monitoring of agro-ecosystems in the Sahel
7. **Sharifi, A.**, 1991, 90-6164-074-1, Development of an appropriate resource information system to support agricultural management at farm enterprise level
8. **Zee, D. van der**, 1991, 90-6164-075-X, Recreation studied from above: Air photo interpretation as input into land evaluation for recreation
9. **Mannaerts, C.**, 1991, 90-6164-085-7, Assessment of the transferability of laboratory rainfall-runoff and rainfall - soil loss relationships to field and catchment scales: a study in the Cape Verde Islands
10. **Ze Shen Wang**, 1991, 90-393-0333-9, An expert system for cartographic symbol design
11. **Zhou Yunxian**, 1991, 90-6164-081-4, Application of Radon transforms to the processing of airborne geophysical data
12. **Zuviria, M. de**, 1992, 90-6164-077-6, Mapping agro-topoclimates by integrating topographic, meteorological and land ecological data in a geographic information system: a case study of the Lom Sak area, North Central Thailand
13. **Westen, C. van**, 1993, 90-6164-078-4, Application of Geographic Information Systems to landslide hazard zonation
14. **Shi Wenzhong**, 1994, 90-6164-099-7, Modelling positional and thematic uncertainties in integration of remote sensing and geographic information systems
15. **Javelosa, R.**, 1994, 90-6164-086-5, Active Quaternary environments in the Philippine mobile belt
16. **Lo King-Chang**, 1994, 90-9006526-1, High Quality Automatic DEM, Digital Elevation Model Generation from Multiple Imagery
17. **Wokabi, S.**, 1994, 90-6164-102-0, Quantified land evaluation for maize yield gap analysis at three sites on the eastern slope of Mt. Kenya
18. **Rodriguez, O.**, 1995, Land Use conflicts and planning strategies in urban fringes: a case study of Western Caracas, Venezuela
19. **Meer, F. van der**, 1995, 90-5485-385-9, Imaging spectrometry & the Ronda peridotites
20. **Kufoniyi, O.**, 1995, 90-6164-105-5, Spatial coincidence: automated database updating and data consistency in vector GIS
21. **Zambezi, P.**, 1995, Geochemistry of the Nkombwa Hill carbonatite complex of Isoka District, north-east Zambia, with special emphasis on economic minerals

22. **Woldai, T.**, 1995, The application of remote sensing to the study of the geology and structure of the Carboniferous in the Calañas area, pyrite belt, SW Spain
23. **Verweij, P.**, 1995, 90-6164-109-8, Spatial and temporal modelling of vegetation patterns: burning and grazing in the Paramo of Los Nevados National Park, Colombia
24. **Pohl, C.**, 1996, 90-6164-121-7, Geometric Aspects of Multisensor Image Fusion for Topographic Map Updating in the Humid Tropics
25. **Jiang Bin**, 1996, 90-6266-128-9, Fuzzy overlay analysis and visualization in GIS
26. **Metternicht, G.**, 1996, 90-6164-118-7, Detecting and monitoring land degradation features and processes in the Cochabamba Valleys, Bolivia. A synergistic approach
27. **Hoanh Chu Thai**, 1996, 90-6164-120-9, Development of a Computerized Aid to Integrated Land Use Planning (CAILUP) at regional level in irrigated areas: a case study for the Quan Lo Phung Hiep region in the Mekong Delta, Vietnam
28. **Roshannejad, A.**, 1996, 90-9009284-6, The management of spatio-temporal data in a national geographic information system
29. **Terlien, M.**, 1996, 90-6164-115-2, Modelling Spatial and Temporal Variations in Rainfall-Triggered Landslides: the integration of hydrologic models, slope stability models and GIS for the hazard zonation of rainfall-triggered landslides with examples from Manizales, Colombia
30. **Mahavir, J.**, 1996, 90-6164-117-9, Modelling settlement patterns for metropolitan regions: inputs from remote sensing
31. **Al-Amir, S.**, 1996, 90-6164-116-0, Modern spatial planning practice as supported by the multi-applicable tools of remote sensing and GIS: the Syrian case
32. **Pilouk, M.**, 1996, 90-6164-122-5, Integrated modelling for 3D GIS
33. **Duan Zengshan**, 1996, 90-6164-123-3, Optimization modelling of a river-aquifer system with technical interventions: a case study for the Huangshui river and the coastal aquifer, Shandong, China
34. **Man, W.H. de**, 1996, 90-9009-775-9, Surveys: informatie als norm: een verkenning van de institutionalisering van dorps - surveys in Thailand en op de Filipijnen
35. **Vekerdy, Z.**, 1996, 90-6164-119-5, GIS-based hydrological modelling of alluvial regions: using the example of the Kisaföld, Hungary
36. **Pereira, Luisa**, 1996, 90-407-1385-5, A Robust and Adaptive Matching Procedure for Automatic Modelling of Terrain Relief
37. **Fandino Lozano, M.**, 1996, 90-6164-129-2, A Framework of Ecological Evaluation oriented at the Establishment and Management of Protected Areas: a case study of the Santuario de Iguaque, Colombia
38. **Toxopeus, B.**, 1996, 90-6164-126-8, ISM: an Interactive Spatial and temporal Modelling system as a tool in ecosystem management: with two case studies : Cibodas biosphere reserve, West Java Indonesia: Amboseli biosphere reserve, Kajiado district, Central Southern Kenya
39. **Wang Yiman**, 1997, 90-6164-131-4, Satellite SAR imagery for topographic mapping of tidal flat areas in the Dutch Wadden Sea
40. **Asun Saldana-Lopez**, 1997, 90-6164-133-0, Complexity of soils and Soilscape patterns on the southern slopes of the Ayllon Range, central Spain: a GIS assisted modelling approach
41. **Ceccarelli, T.**, 1997, 90-6164-135-7, Towards a planning support system for communal areas in the Zambezi valley, Zimbabwe; a multi-criteria evaluation linking farm household analysis, land evaluation and geographic information systems
42. **Peng Wanning**, 1997, 90-6164-134-9, Automated generalization in GIS
43. **Lawas, C.**, 1997, 90-6164-137-3, The Resource Users' Knowledge, the neglected input in Land resource management: the case of the Kankanaey farmers in Benguet, Philippines
44. **Bijker, W.**, 1997, 90-6164-139-X, Radar for rain forest: A monitoring system for land cover Change in the Colombian Amazon
45. **Farshad, A.**, 1997, 90-6164-142-X, Analysis of integrated land and water management practices within different agricultural systems under semi-arid conditions of Iran and evaluation of their sustainability
46. **Orlic, B.**, 1997, 90-6164-140-3, Predicting subsurface conditions for geotechnical modelling
47. **Bishr, Y.**, 1997, 90-6164-141-1, Semantic Aspects of Interoperable GIS
48. **Zhang Xiangmin**, 1998, 90-6164-144-6, Coal fires in Northwest China: detection, monitoring and prediction using remote sensing data
49. **Gens, R.**, 1998, 90-6164-155-1, Quality assessment of SAR interferometric data

50. **Turkstra, J.**, 1998, 90-6164-147-0, Urban development and geographical information: spatial and temporal patterns of urban development and land values using integrated geo-data, Villaviciencia, Colombia
51. **Cassells, C.**, 1998, Thermal modelling of underground coal fires in northern China
52. **Naseri, M.**, 1998, 90-6164-195-0, Characterization of Salt-affected Soils for Modelling Sustainable Land Management in Semi-arid Environment: a case study in the Gorgan Region, Northeast, Iran
53. **Gorte, B.G.H.**, 1998, 90-6164-157-8, Probabilistic Segmentation of Remotely Sensed Images
54. **Tenalem Ayenew**, 1998, 90-6164-158-6, The hydrological system of the lake district basin, central main Ethiopian rift
55. **Wang Donggen**, 1998, 90-6864-551-7, Conjoint approaches to developing activity-based models
56. **Bastidas de Calderon, M.**, 1998, 90-6164-193-4, Environmental fragility and vulnerability of Amazonian landscapes and ecosystems in the middle Orinoco river basin, Venezuela
57. **Moameni, A.**, 1999, Soil quality changes under long-term wheat cultivation in the Marvdasht plain, South-Central Iran
58. **Groenigen, J.W. van**, 1999, 90-6164-156-X, Constrained optimisation of spatial sampling: a geostatistical approach
59. **Cheng Tao**, 1999, 90-6164-164-0, A process-oriented data model for fuzzy spatial objects
60. **Wolski, Piotr**, 1999, 90-6164-165-9, Application of reservoir modelling to hydrotopes identified by remote sensing
61. **Acharya, B.**, 1999, 90-6164-168-3, Forest biodiversity assessment: A spatial analysis of tree species diversity in Nepal
62. **Akbar Abkar, Ali**, 1999, 90-6164-169-1, Likelihood-based segmentation and classification of remotely sensed images
63. **Yanuariadi, T.**, 1999, 90-5808-082-X, Sustainable Land Allocation: GIS-based decision support for industrial forest plantation development in Indonesia
64. **Abu Bakr, Mohamed**, 1999, 90-6164-170-5, An Integrated Agro-Economic and Agro-Ecological Framework for Land Use Planning and Policy Analysis
65. **Eleveld, M.**, 1999, 90-6461-166-7, Exploring coastal morphodynamics of Ameland (The Netherlands) with remote sensing monitoring techniques and dynamic modelling in GIS
66. **Yang Hong**, 1999, 90-6164-172-1, Imaging Spectrometry for Hydrocarbon Microseepage
67. **Mainam, Félix**, 1999, 90-6164-179-9, Modelling soil erodibility in the semiarid zone of Cameroon
68. **Bakr, Mahmoud**, 2000, 90-6164-176-4, A Stochastic Inverse-Management Approach to Groundwater Quality
69. **Zlatanova, Z.**, 2000, 90-6164-178-0, 3D GIS for Urban Development
70. **Ottichilo, Wilber K.**, 2000, 90-5808-197-4, Wildlife Dynamics: An Analysis of Change in the Masai Mara Ecosystem
71. **Kaymakci, Nuri**, 2000, 90-6164-181-0, Tectono-stratigraphical Evolution of the Cankori Basin (Central Anatolia, Turkey)
72. **Gonzalez, Rhodora**, 2000, 90-5808-246-6, Platforms and Terraces: Bridging participation and GIS in joint-learning for watershed management with the Ifugaos of the Philippines
73. **Schetselaar, Ernst**, 2000, 90-6164-180-2, Integrated analyses of granite-gneiss terrain from field and multisource remotely sensed data. A case study from the Canadian Shield
74. **Mesgari, Saadi**, 2000, 90-3651-511-4, Topological Cell-Tuple Structure for Three-Dimensional Spatial Data
75. **Bie, Cees A.J.M. de**, 2000, 90-5808-253-9, Comparative Performance Analysis of Agro-Ecosystems
76. **Khaemba, Wilson M.**, 2000, 90-5808-280-6, Spatial Statistics for Natural Resource Management
77. **Shrestha, Dhruva**, 2000, 90-6164-189-6, Aspects of erosion and sedimentation in the Nepalese Himalaya: highland-lowland relations
78. **Asadi Haroni, Hooshang**, 2000, 90-6164-185-3, The Zarshuran Gold Deposit Model Applied in a Mineral Exploration GIS in Iran
79. **Raza, Ale**, 2001, 90-3651-540-8, Object-oriented Temporal GIS for Urban Applications

80. **Farah, Hussein**, 2001, 90-5808-331-4, Estimation of regional evaporation under different weather conditions from satellite and meteorological data. A case study in the Naivasha Basin, Kenya
81. **Zheng, Ding**, 2001, 90-6164-190-X, A Neuro-Fuzzy Approach to Linguistic Knowledge Acquisition and Assessment in Spatial Decision Making
82. **Sahu, B. K.**, 2001, Aeromagnetics of continental areas flanking the Indian Ocean; with implications for geological correlation and Gondwana reassembly
83. **Alfestawi, Y.**, 2001, 90-6164-198-5, The structural, paleogeographical and hydrocarbon systems analysis of the Ghadamis and Murzuq Basins, West Libya, with emphasis on their relation to the intervening Al Qarqaf Arch
84. **Liu, Xuehua**, 2001, 90-5808-496-5, Mapping and Modelling the Habitat of Giant Pandas in Foping Nature Reserve, China
85. **Oindo, Boniface Oluoch**, 2001, 90-5808-495-7, Spatial Patterns of Species Diversity in Kenya
86. **Carranza, Emmanuel John**, 2002, 90-6164-203-5, Geologically-Constrained Mineral Potential Mapping
87. **Rugege, Dennis**, 2002, 90-5808-584-8, Regional Analysis of Maize-Based Land Use Systems for Early Warning Applications
88. **Liu, Yaolin**, 2002, 90-5808-648-8, Categorical Database Generalization in GIS
89. **Ogao, Patrick Job**, 2002, 90-6164-206-X, Exploratory Visualization Of Temporal Geospatial Data Using Animation
90. **Abadi, Abdulbasit M.**, 2002, 906164-205-1, Tectonics of the Sirt Basin - Interferences from tectonic subsidence analysis, stress inversion and gravity modeling
91. **Geneletti, Davide**, 2002, ISSN 0169-4839, Ecological evaluation for environmental impact assessment
92. **Sedogo, Laurent G.**, 2002, ISBN 90-5808-751-4, Integration of Participatory Local and Regional Planning for Resources Management Using Remote Sensing and GIS
93. **Montoya, Ana Lorena**, 2002, ISBN 90-6164-2086, Urban Disaster Management: A Case Study of Earthquake Risk Assessment in Cartago, Costa Rica
94. **Ahmad, Mobin-ud-Din**, 2002, ISBN 90-5808-761-1, Estimation of net groundwater use in irrigated river basins using geo-information techniques: A case study in Rechna Doab, Pakistan
95. **Said, Mohammed Yahya**, 2003, ISBN 90-5808-794-8, Multiscale perspective of species richness in East Africa
96. **Schmidt, Karin S.**, 2003, ISBN 90-5808-830-8, Hyperspectral Remote Sensing of Vegetation Species Distribution in a Saltmarsh
97. **Binnqüist, Rosaura C. L.**, 2003, ISBN 9036519004, The Endurance of Mexican Amate Paper: Exploring Additional Dimensions to the Sustainable Development Concept
98. **Zhengdong, Huang**, 2003, ISBN 90-6164-211-6, Data Integration for Urban Transport Planning
99. **Jianquan, Cheng**, 2003, ISBN 90-6164-212-4, Modelling Spatial and Temporal Urban Growth

Samenvatting

De rijkdom aan genetisch materiaal en aan soorten in verschillende ecosystemen is een belangrijke vorm van biodiversiteit. In een poging een antwoord te vinden op belangrijke vragen omtrent hun rol daarin, zijn miljoenen organismen in het verleden verzameld in het tropisch regenwoud en zijn wateren, om opgenomen te worden in natuurhistorische collecties. Grote hoeveelheden gegevens werden verzameld tijdens talloze onafhankelijke en ongerelateerde onderzoeken in het hele Braziliaanse Amazonegebied over de afgelopen eeuw. Individuele onderzoekers hebben grote moeite om deze gegevens en informatiebronnen geheel te overzien, mede omdat hun onderzoeksvragen gesteld worden in een multidisciplinaire context en afhankelijk zijn van goed gedocumenteerde gegevens. Omdat niet al deze onderzoeksinspanningen even goed op elkaar zijn afgestemd, is er sprake van een serieus probleem op het vlak van gegevensredundantie, -inconsistentie en -interpretatie, en dit leidt tot hoge kosten van gegevensverwerking en niet-optimale wetenschappelijke resultaten.

Informatietechnologie is belangrijk gereedschap voor het beheer van natuurhistorische informatie. Daartoe dient aan een aantal vereisten te worden voldaan: het bestaan van een geschikt gegevensmodel, een juist beheer van gegevens en metagegevens, de beschikbaarheid van methodes om oudere gegevensbronnen weer tot leven te roepen en te integreren, onder andere door de toevoeging van geografische informatie en analyse-mogelijkheden.

Dit proefschrift verschaft een overzicht van natuurhistorische collecties, hun complexiteit, en hun informatiebeheersactiviteiten voor de belangrijkste instituten in het Amazonegebied. De gebruikte functie- en systeemanalyse is gebaseerd op interviews, informatie-analyse, data flow-analyse en evaluatie van systeembeschrijvingen, met de hulp van wetenschappers als gebruikersgroep en curatoren als informatiebeheerders en -verzorgers.

Een conceptueel databaseschema, CLOSi ('Clustered Object Schema for INPA's biodiversity data collections'), werd ontwikkeld om daarmee de ontwikkeling van een database voor natuurhistorisch collectiebeheer te faciliteren en te stimuleren.

Normaliter beschouwen onderzoekers hun brongegevens als onbewerkt, en bestaan deze meestal uit tabellen van numerieke of gecodeerde gegevens betreffende (veld)waarnemingen. Het nut van zulke gegevens kan alleen bepaald worden wanneer ze geassocieerd zijn met een theoretisch of conceptueel interpretatiemodel. Dit vereist begrip van de vigerende variabelen en hun types en eenheden, mogelijke vertekening in gemeten waarden, begrip van de waarnemingsmethoden, en enkele additionele feiten die zich slechts in de metagegevens laten uitdrukken. Het is een feit dat informatie verloren kan gaan door degradatie van de brongegevens of door afwezigheid van metagegevens. Deze laatste staan de gebruikers toe brongegevens makkelijker te lokaliseren en begrijpen over de tijd heen. De combinatie van gegevens met metagegevens binnen een conceptueel raamwerk heeft een positief effect op de informatieproductie. Om die reden werd een op XML gebaseerde oplossing voor het beheer van biologische profielen als metagegevens ontwikkeld op het web. De hier voorgestelde oplossing gebruikt de standaard voor metagegevens van FGDC, waarin een biologisch gegevensprofiel opgenomen is, en die hier werd gerepresenteerd en opgeslagen als een XML schema in een web-gebaseerd opslagsysteem.

Het proefschrift beschrijft ook het ontwerp en de implementatie van een gegevensbank, een web-interface voor natuurhistorische collecties en de integratie met een retrospectief georeferentie-gereedschap, die als eigenschap heeft dat ermee bijdragen geleverd kunnen worden aan gemeenschappelijke geografische gazetteers (indices van plaatsnamen). De gegevensbankontwikkeling werd gebaseerd op het CLOSi schema, maar ook op de vereisten van een web-interface, ondersteund door een drielaags systeemarchitectuur. Deze bestaat uit een web-interface voor gebruikers, een applicatie-server, en een database management systeem.

Deze architectuur werd afgeleid middels onderzoek naar behoeften van gebruikers van gegevens over natuurhistorische collecties. De voorgestelde oplossingen kunnen instituten vergelijkbaar met INPA enorm van dienst zijn. Het probleem van ad hoc systeemontwerp kan op deze manier fors gereduceerd worden zodat middelen vrijkomen om meer te doen aan het gebruik (niet de ontwikkeling) van het ontwikkelde systeem. Een zodanig geïmplementeerd gegevensbanksysteem staat vervolgens een snellere digitalisering van biologische gegevens toe. Als deze twee activiteiten netjes zijn opgezet, kan de exploratie en verspreiding van gegevens worden toegelaten middels additionele functionaliteit of applicatie-ontwikkeling.

Voor het beheer van metagegevens in XML werd een client-server ontwerp opgezet dat gebruikers wereldwijd toestaat de (meta)gegevens te bespelen. Dit werk is bijzonder zinvol voor gebruik in minder ontwikkelde landen omdat het ontwikkelde systeem en de gebruikte gereedschappen of 'public domain' of 'free open source' zijn, hetgeen betekent dat instituten een robuuste oplossing kunnen verkrijgen tegen betrekkelijk geringe kosten.

De integratie van een georeferentie-toepassing in de web-omgeving bleek een succes- en waardevolle, met name voor onderzoekers van natuurhistorische collecties met een behoefte aan resultaten van ruimtelijke analyse. De voordelen omvatten onder andere: de toegenomen snelheid van het georeferentieproces, het maximaliseren van consistentie tussen gebruikers, het mogelijk maken van de toepassing van interpretatiestandaarden, zoals bijvoorbeeld opgesteld door curatoren, en het kwantificeren van vaagheid in lokatie-omschrijvingen.

Het ondersteund worden door een taxonomisch referentiesysteem is uiterst belangrijk voor iedere biologisch-wetenschappelijke activiteit. Taxonomie is het onderzoeksveld binnen de biologie dat tracht de evolutionaire afstammingsboom (de biologische taxonomie) te ontrafelen. Biologen kennen deze afstammingsboom niet echt, maar langzamerhand worden meer en meer onzekerheden daaromtrent uit de weg geruimd. Verschillen in interpretatie blijven vooralsnog echter bestaan, en systemen die communiceren over taxonomische eenheden zoals soorten, hebben dus een communicatieprobleem. Een onderhandelingsprotocol wordt voorgesteld in dit proefschrift dat tracht daar een oplossing voor te bieden. Wel werden daartoe enkele vereenvoudigende aannames gepleegd ten opzichte van de taxonomische realiteit, zoals de afwezigheid van synoniemen en spelfouten, fout- en redundantievrije taxonomische gegevensverzamelingen in beide systemen, die overigens (uiteraard) wel onderling kunnen verschillen.

Op deze wijze zouden we een zinvolle oplossing kunnen bieden voor de communicatie over en de uitwisseling van biologische informatie voor een breed publiek van biologen, met het doel hun onderzoek te vergemakkelijken, om zo aan de toenemende noodzaak van het vinden van antwoorden op belangrijke vragen omtrent onze biosfeer te voldoen.

Sumário

Especies e diversidade genética nos diferentes ecossistemas representam importante componentes da biodiversidade. Para entender como os ecossistemas funcionam, milhões de organismos de florestas tropicais e de suas águas foram coletados e continuam a ser depositadas em coleções biológicas. Um grande conjunto de dados de toda a região da Amazônia Brasileira foi obtido e compilado, desde o século passado, por meio de diversos estudos não relacionados e independentes. Pesquisadores, de forma individual são incapazes de compreender totalmente esses dados e informações, uma vez que as questões mais recentes podem depender de um contexto multidisciplinar e de dados bem documentados. Como algumas atividades estão relacionadas mas outras não, a maioria das soluções adotadas para a destinação desses dados nesse momento, são caracterizadas por redundância, inconsistência de dados, interpretação de lacunas, levando a um alto custo de trabalho, processamento e infra-estrutura, e consequentemente, apresentando resultados científicos insipientes.

A tecnologia da computação tem sido um recurso fundamental aplicado para gerenciamento da bio-informação. Para o uso adequado dessa tecnologia, existe um número de exigência: um modelo de informação, gerenciamento formal de dados e metadados, assim como métodos para integrar e restaurar dados antigos, entre outros, pela inclusão de informações geográficas e da habilidade para análises.

Esta tese apresenta uma visão geral das coleções biológicas dos principais institutos da região Amazônica, suas complexidades e atividades de gerenciamento relacionadas a bio-informação. A análise funcional e de sistemas adotada foi o resultado da interação por meio de entrevistas, análise de requisitos de informação, fluxo de dados e avaliação de descrições, contando com a participação de pesquisadores como usuários, e curadores como gerente de informações e provedores de dados.

Um esquema conceitual de banco de dados, CLOSi (Clustered Object Schema for INPA's biodiversity data collections), foi concebido para facilitar e estimular o desenvolvimento de banco de dados de coleções biológicas.

Normalmente, pesquisadores se referem aos seus dados como dados brutos, os quais são estruturados em linhas e colunas ou contendo observações coletadas na forma numérica ou codificada. A utilidade de tal dados pode

ser levantada apenas quando os mesmos são associados a um modelo de interpretação teórica ou conceitual. Isso requer o entendimento dos tipos de variáveis, das unidades adotadas, das potenciais implicações nas medidas, métodos de coletas e um número de fatos que não são representados nos dados brutos, e sim nos metadados. De fato, a informação pode ser perdida devido a degradação dos dados brutos ou pela ausência de metadados. Metadados proporcionam aos usuários de dados a capacidade de localizar e entender os dados através do tempo. A combinação de dados e metadados dentro de um cenário conceitual propicia a produção de informação. Uma solução baseada em XML foi implementada para o gerenciamento de perfis de metadados biológico via da web. A solução apresentada utiliza o padrão de metadados do FGDC, o qual incorpora a descrição biológica, e a qual é representada como um esquema XML e armazenado no repositório XML based on the web.

Ainda, o projeto e implementação de um banco de dados, uma interface web para coleções biológicas e sua integração a ferramenta para georeferenciamento retrospectivo, com a vantagem de contribuir para a iniciativa de um gazetteer colaborativo é detalhado. O desenvolvimento de banco de dados é baseado no esquema CLOSi assim como na construção de uma interface web para acessar o banco de dados, apoiado pela arquitetura de um sistema de três categorias. A arquitetura consiste de uma interface cliente na web, uma aplicação servidora e um sistema de gerenciador de banco de dados.

Essa estrutura foi produzida como resultado de pesquisa aprofundada das necessidades dos que usam dados de coleções biológicas. As soluções disponibilizadas podem beneficiar tremendamente institutos similares ao INPA. O problema de desenvolvimento de sistemas *ad hoc* pode ser reduzidos consideravelmente e recursos podem ser direcionados para a utilização no projeto de banco de dados apresentado e implementado. Um sistema de banco de dados implementado pavimentará o caminho para uma digitalização mais rápida de dados biológicos. Com esses dois recursos disponíveis, a exploração de dados e disseminação de informação pode ser permitida por meio de funcionalidades adicionais ou implementação de aplicativos complementares.

Para gerenciar metadados baseado em XML, foi utilizada uma configuração cliente-servidor que permite a disseminação de metadados e dados para usuários em uma escala global via web. Este trabalho é particularmente adequado para países menos desenvolvidos pois o sistema desenvolvido e ferramentas utilizadas são de domínio público ou código fonte aberto e livre, assegurando baixo custo para uma solução robusta que pode ser utilizada em qualquer ambiente de instituições particulares.

A integração com aplicações de georeferenciamento foi reconhecido ser um sucesso e valiosa contribuição para o sistema web, proporcionando grande benefício para pesquisadores de coleções biológicas que precisam de resultados de análise geoespacial. As vantagens desse processo incluem: aumento de velocidade no georeferenciamento, maximizando consistência en-

tre usuários, permitindo a incorporação de interpretações padrões estabelecidas por pesquisadores, especialmente curadores, e quantificando incerteza de localidade textualmente descrita.

Um sistema de referência taxônomica é extremamente importante para qualquer atividade científica em biologia. Sistemática é um campo de pesquisa da biologia que tenta desvendar (evolucionariamente) a árvore da vida, isto é, a taxonomia biológica. Biólogos não conhecem a árvore, e estão em processo de descobrir seus segredos gradualmente, mas com visões diferentes de suas crenças. Para tratar desse problema, nós propomos um cenário para um protocolo de negociação que auxiliaria sistemas automáticos. Nós nos abstraímos, entretanto, de certas complexidades da prática taxônomica, tal como a presença de sinônimos e erro de escrita, e assumimos erros e conjunto de informações taxônomicas livres de redundância nos dois pontos, o qual pode (e poderá), entretanto, diferenciar em estrutura e complexidade

Desta forma poderíamos estar proporcionando uma solução útil para a troca de informações biológicas dentro da ampla comunidade de pesquisadores em biologia, objetivando facilitar a investigação de nossos recursos biológicos, satisfazendo nossas necessidades crescentes de oferecer respostas acerca da importância dos mesmos para nossa biosfera.

Resumen

Las especies y la diversidad genética en diferentes ecosistemas son componentes importantes de la biodiversidad. Para dar respuesta a preguntas relacionadas con su funcionamiento, millones de organismos son recolectados del bosque lluvioso tropical y sus acuíferos, los cuales continúan siendo depositados en colecciones biológicas. Grandes recopilaciones de datos fueron recolectados y compilados durante muchos estudios independientes a lo largo de la región amazónica brasileña durante el último siglo. Los investigadores independientes no han sido capaces de comprender estos datos y grupos de información dado que las preguntas en boga pueden depender de contextos multi-disciplinarios y en datos bien documentados. Dado que algunas actividades institucionales están relacionadas, pero otras no están del todo relacionadas, la mayoría de las soluciones adoptadas en este momento en cuanto a ordenanzas de datos sufren de redundancia, inconsistencia de datos y brechas de interpretación, causando altos costos laborales y de procesamiento e infraestructura, los cuales producen resultados científicos menos óptimos.

La tecnología computarizada ha sido un recurso fundamental aplicado al manejo de la bio-información. Para un exitoso uso de esta tecnología, existen un número de requisitos: un modelo de información apropiado, manejo de datos formales y de metadata, así como de métodos para integrar y revivir datos legados, entre otros, al agregar información geográfica y la capacidad de análisis.

Esta tesis proporciona un vistazo a colecciones biológicas, su complejidad y actividades relacionadas al manejo de bio-información en los principales institutos de la región amazónica. El análisis de sistema y función adoptado fue el resultado de la interacción con entrevistados, análisis de requerimientos de información, flujos de datos y evaluación de descripciones, con participación de los investigadores en calidad de usuarios, curadores en calidad de administradores de información y proveedores de datos.

Un esquema conceptual de la base de datos, CLOSi (esquema de objeto agrupado para colecciones de datos de bio-diversidad del INPA) fue concebido para facilitar y estimular el desarrollo de bases de datos de colecciones biológicas.

Usualmente los investigadores se refieren a sus datos como datos crudos, los cuales son estructurados en filas y columnas de observaciones de muestreos ya sean numéricos o codificados. La utilidad de esos datos puede ser evaluada únicamente cuando están asociados a un modelo ya sea teórico o de interpretación conceptual. Esto requiere de entendimiento sobre los tipos de variables, las unidades adoptadas, los prejuicios potenciales en las medidas, métodos de muestreo y un número de hechos que no están representados en los datos crudos sino más bien en la metadata. De hecho, la información puede perderse por degradación de los datos crudos o por la inexistencia de metadata. La metadata permite que el usuario de datos localice y comprenda los datos a través del tiempo. Los datos y la metadata combinados con un marco conceptual mejora la producción de información. Se implementó una solución por vía de la web para el manejo de perfiles biológicos de metadata. La solución presentada utiliza el estándar de metadata FGDC, el cual incorpora el perfil de datos biológicos y el cual es representado como un esquema y almacenaje XML en un depósito basado en la web.

Además, se detalla el diseño y la implementación de la base de datos, una interface web para colecciones biológicas y su integración con una herramienta de geo-referenciación retrospectiva, con la ventaja de que contribuye a iniciativas colaborativas de tipo gazetteer. El desarrollo de la base de datos se centra en el esquema CLOSi así como en la construcción de la interface web para el acceso a la base de datos, la cual está apoyada por una arquitectura de un sistema de tres hileras. La arquitectura consiste de una interface de cliente-usuario en la web, un servidor para aplicaciones y un sistema de manejo de bases de datos.

Esta estructura fue producida a partir de una investigación exhaustiva sobre las necesidades de los usuarios de datos de colecciones biológicas. Las soluciones presentadas pueden beneficiar tremendamente a institutos similares al INPA. El problema del desarrollo de sistemas ad hoc puede ser reducido considerablemente y los recursos pueden ser canalizados hacia la utilización del diseño e implementación de la base de datos presentada. La implementación de un sistema de base de datos permite una digitalización más rápida de datos biológicos. Con ambos en existencia, la exploración de los datos y la diseminación de la información puede permitirse a través de funcionalidad adicional o implementación de una aplicación complementaria.

Para manejar metadata basada en XML, se utilizó un sistema de cliente-servidor para permitir la diseminación de los datos y de la metadata a usuarios a escala global a través de la web. Este trabajo es particularmente adecuado en países en desarrollo ya que el sistema desarrollado y las herramientas utilizadas son de dominio público o de fuentes libres, lo cual asegura un bajo costo para una solución robusta para cualquier ambiente institucional particular.

Se encontró que la integración con la aplicación de georeferenciación es exitosa y es una adición valiosa al sistema web, la cual provee gran beneficio a los investigadores de colecciones biológicas que necesitan resultados de

análisis geo-espacial. Las ventajas de este proceso incluyen: mayor rapidez en la geo-referenciación, maximización de la consistencia entre usuarios, lo cual permite la incorporación de estándares de interpretación establecidos por los investigadores, especialmente los curadores, y la cuantificación de la vaguedad de localidad textualmente descrita.

Un sistema de referencia taxonómico es extremadamente importante para una actividad científica en el campo de la biología. La sistemática es el campo de investigación dentro de la biología que trata de revelar el árbol (evolucionario) de vida, la taxonomía biológica. Los biólogos no conocen ese árbol, y se encuentran en un lento proceso del descubrimiento de sus secretos, el cual difiere de sus creencias. Para encarar este problema, nosotros proponemos un marco para un protocolo de negociación que pueda ayudar a sistemas como este. Sin embargo, nos abstraímos de ciertas complejidades de la práctica taxonómica, tales como la presencia de sinónimos y errores ortográficos, y asume grupos de datos taxonómicos libres de error y redundancia en ambos extremos, los cuales no obstante pueden (y van a) diferir en estructura y complejidad.

De esta forma, nosotros podríamos proveer una solución útil para el intercambio y la comunicación de información biológica hacia una audiencia amplia de investigadores biológicos, satisfaciendo nuestra necesidad creciente de respuestas sobre la importancia de nuestra biosfera.

Curriculum vitae



José Laurindo Campos dos Santos was born in Manaus, Amazon in Brazil. From 1975 until 1978 he attended Electronics studies at the Federal Technical School of the Amazon State, in Manaus. In 1980 he started university at the Amazonian Institute of Technology in the area of Civil Engineering and Technology, which was concluded in 1984. He attended a diploma course in System Analysis at the Federal University of Amazonas in 1985. He obtained his Master of Science degree in Computer Science in 1988 from the Federal University of Paraíba, in Campina Grande, Brazil, in the area of Database and Artificial Intelligence.

Since 1988, he has been involved in research and education in the field of database technology at the National Institute for the Amazon Research and at the Federal University of Amazonas in Manaus, Amazonas. At INPA, his various activities include research, teaching, supervision of students and development of computer solutions applied to environmental research. Also, he has participated in several national and international research projects, such Oiapoque Project (Silvolab – French Guyana, INPA – Brazil), Large Scale Biosphere Atmosphere Experiment in Amazônia – Data Information System (INPE – Brazil, NASA – USA), Base de Dados Compartilhada da Amazônia (MCT, MMA – Brazil), INPA Network Design and Implementation (PPG7, INPA – Brazil).

In 1999 he joined the International Institute for Aerospace Survey and Earth Sciences (ITC) and started his research studies in Spatial Information Theory and Applied Computer Science, jointly with the University of Twente, Enschede, The Netherlands. His research interests are conceptual database design and tools, theory of database, scientific data and metadata integration and environmental informatics.

The research reported in this book was partially supported by the International Institute for Geo-information Science and Earth Observation (ITC), The University of Twente (UT), The Netherlands and the Instituto Nacional de Pesquisas da Amazônia (INPA), Brazil.

José Laurindo Campos dos Santos

Instituto Nacional de Pesquisas da Amazônia
Avenida André Araújo, 2936 - Petrópolis
69.083-000 - Manaus - Amazonas, Brasil
Telephone: +55 (0)92 643 3032
Facsimile: +55 (0)92 643 3111

Email: *lcampos@inpa.gov.br*
INPA Home Page: *http://www.inpa.gov.br*

and

International Institute for Geo-Information Science
and Earth Observation
P.O. Box 6, 7500 AA Enschede, The Netherlands
Telephone: +31 (0)53 487 4444
Facsimile: +31 (0)53 487 4335

Email: *santos@itc.nl*
ITC Home Page: *http://www.itc.nl*