UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Extensions to Metadata Languages for Environmental Primary Data

Permalink
https://escholarship.org/uc/item/8q09s2nn

Author
IBANEZ, LUIS M.

Publication Date
2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Extensions to Metadata Languages for Environmental Primary Data: the Forest Cloud

By

Luis Maria Ibanez de Garayo

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Environmental Science , Policy, and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Greg Biging, Chair
Professor Ray Larson
Professor Dennis Baldocchi

Fall 2011

para mis grandes amores

for their support and comprehension

Barbara ,Maite, eta Robertorentzat, nere ondoan egoteagatik eta emandako indar eta maitasun guztiagatik.

Maite zaituztet

Abstract

During the last decade environmental scientists and managers have found in the Internet a new communication venue that has improved their productivity by allowing them to share data and knowledge more fluently. Since the invention of the eXtensible Markup Language (XML), XML has brought the attention of many researchers to improve the communications among people working with data-centric documents since XML was supposed to be the correct approach to standardize data. During this decade, many papers have been published under the pretenses of XML being the new and paradigmatic standard to share data even though no study has proved that the XML languages have been used by researchers or managers working with data-centric documents during this period of time. This thesis by researching all possible spaces proves that, on the contrary, that after more than a decade from its invention, XML is still not used by the vast scientific community that works with data-centric documents who are still using data archives with legacy formats. Therefore, if data standardization is difficult to attain, and facilitating sharing data a goal to reach, the other clear venue to follow to achieve the goal is to use metadata information. However, metadata languages such as the Ecological Modeling Language (EML) and others have no intrinsic features to complete and directly describe the information conveyed in many important types of data-centric documents used by environmentalists. By carefully studying the nature of data-centric archives and the process of metadata creation, this thesis shows that any data archive can be easily described using an "a posteriori" approach where the lexical descriptors of the physical data from a data-centric file are developed by inspection of the file instead of by following the specifications of the format of the file. In addition, following the principles of the Linked Open Data project, the lexical tree is mapped into a simple logic model with semantic annotations from controlled vocabularies which can be easily serialized for data exchange or data syndicalization. With the metadata extensions researched in this thesis, metadata languages such as EML can be improved by increasing its expression power. Environmental scientists and researchers can us this to exchange data-centric documents, and multidisciplinary projects can easily syndicate data from different authors in different formats in a data-centric cloud.

Preface

Ten years ago while learning about renewable forest resources and having known first hand and read about unscrupulous forest operations, I had a dream of a new forest regime where the information about forest operations could travel through the Internet free of any impedance with contrasted and true information.

With this thesis, I have tried to pave the road for a better Internet. However this thesis could no have been materialized without the support of my wife who for eighteen long years has supported me and our family every single day of the trip.

My thanks also go to all people that with their writings and tools have enlighten my ideas and work. Special thanks to my professors for their infinite patience and my apologies to them for my lack of communication skills. I thank Prof. Larson for teaching me the power of databases and his willingness to listen my ideas every time I stepped into his office, to Prof. Baldocchi for showing me his enthusiasm, the multidisciplinary nature of environmental sciences and the ubiquitous need of sharing primary datasets for the good of the people and their environment, to Prof. Tim Berners-Lee and Prof. Goodchild for having the time to read my spam-like email without knowing me and support my ideas and given me interesting feedback, and last but not least, to Prof. Biging for believing in me and taking me under his wing to the very very very last day. To all of you I thank again from the bottom of my heart. Gracias!

Contents

1 The XML is Dead **1**

1.1 Introduction 1

1.2 Networking 2

1.2.1 The Paleozoic-network Era: 1880's to 1900's 2

1.2.2 The Mesozoic-network Era: 1900's to 1950's 3

1.2.3 The Tertiary-Network Period: 1950's to 1980's 3

1.2.4 The Quaternary-Network Period: 1980's to 1990's . . . 3

1.2.5 The Contemporary Period: 1990's to 2000s 4

1.2.6 The Semantic Web Period: since 2000 5

1.2.6.1 The Web Portal 5

1.2.6.2 The GRID 7

1.2.6.3 The Cloud 10

1.3 Environmental Collaboration 11

1.4 Environmental Data Exchange 15

1.5 XML Spaces 19

1.5.1 Bibliography 19

1.5.2 Public Web Space 21

1.5.3 Private Spaces 23

1.6 Conclusions 24

2 Long Live Metadata **25**

2.1 Introduction 25

2.2 Environmental Archives 26

2.3 Environmental Primary Data 29

2.4 Metadata about Data-centric Documents. 31

2.4.1 Archival Metadata 33

2.4.2 Ecological Metadata Language (EML) 33
2.4.3 Earth Science Markup Language (ESML) 34
2.4.4 Data Format Description Language (DFDL) 34
2.4.5 PADS 34
2.4.6 Extensible Scientific Interchange Language (XSIL) . . . 35
2.4.7 Binary Format Description (BFD) 35
2.5 Metadata Comes after Data 35
2.6 Proof of Concept for 'A Posteriori' Metadata to Describe Primary Data 36
2.7 Hierarchical Engine for Metadata Processing: HEMP 41
2.7.1 Hierarchical Scanner: Lexical Description 41
2.7.1.1 Pointers 42
2.7.1.2 Parallel Vs Sequential Scanning 42
2.7.1.3 Stream Concatenation 43
2.7.1.4 Lexical Description Examples of Common Data Files 43
2.7.2 My Scientific Object Notation (MySON): Logical Structure 48
2.7.2.1 My Scientific Object Notation: MySON . . . 48
2.7.2.2 Serialization of MySON Objects 51
2.7.3 Controlled Vocabularies: Semantic Mapping 53
2.7.4 HEMP: Processing description 56
2.7.4.1 Containers 56
2.7.4.2 Lexical Functions 57
2.7.4.3 Arithmetic Functions 58
2.7.4.4 Logical Functions 58
2.7.4.5 Engine Description 59
2.7.4.6 Example: Metadata File for a Compressed Shapefile with Tree Data 60
2.8 Conclusions 69

3 The Forest Cloud: Link between Local and Global Regimes 75

3.1 Introduction . . . 75

3.2 Local forest governance . . . 76

3.3 Global Forest Governance . . . 77

3.4 International Governance . . . 77

3.5 The Tragedy of the Forest . . . 78

3.6 Tree Identification . . . 79

3.6.1 Where Is It From? . . . 79

3.6.2 Which Specie Is It? . . . 80

3.6.3 Which One Is It? . . . 81

3.7 The Forest Cloud . . . 81

3.7.1 Cloud Concept . . . 82

3.7.1.1 Software on Demand . . . 82

3.7.1.2 Computational Power . . . 82

3.7.1.3 Illimitable Data Store . . . 82

3.7.1.4 Layer of Common Communications . . . 83

3.7.2 Forest Cloud Philosophy . . . 84

3.7.3 Forest Cloud Infrastructure . . . 85

3.7.3.1 Metadata Templates . . . 85

3.7.3.2 Data Store . . . 99

3.7.3.3 The Forest Cloud Services . . . 104

3.8 The Future of The Forest Cloud . . . 105

List of Figures

1 Authors per Article From Different Disciplines during Last Decade 14

2 Authors per Article across Decades in Ecology 15

3 Normalized View of the Number of Scientific Publications sorted by year and Keywords 20

4 Number of XML Documents Retrieve by Google Search Engine by Language 22

5 Estimated Lifespan of XML Languages 23

6 XML vs Legacy Formats 24

7 Logic Equivalence between a Table and a List of Triplets . . . 30

8 Physical, Logical, and Semantical View of a Data Stream . . . 31

9 Injective Mapping between the Physical Stream and Logical Object 37

10 Synoptic Tiff file Representation with non-Contiguous Data . 45

11 Synoptic Tiff file Representation with Contiguous Data 46

12 Railroad Diagram of MySON Object 50

13 Railroad Diagram for the Self-BAr Value object of MySON . . 53

14 Semantic Layering 54

15 Tree Location Map in Corner Heights at Ithaca, NY 61

16 Street Map of Cornel Heights at Ithaca, NY 62

17 Logical Description of a Compressed Shapefile with Tree Data 68

18 Tim-Berner-Lee's Layered Graph of the Semantic Web 72

19 Modified Layered Graph of the Semantic Web for Data-centric Documents 73

20 Walls in the Cloud created by private competitive communication protocols 84

21 General Physical Description for Containers of an Excel Dataset 87

22 Physical Description for Containers of an Excel Dataset with regular structure 88
23 Physical Description for Containers of an Excel Dataset with a mixed regular structures 89
24 The Forest Cloud partial view of trees from research project in central California 91
25 General Physical Description for Containers of a Shapefile with point data 92
26 General Physical Description for Containers of a DBase file . . 92
27 The Forest Cloud partial view of trees from Cornel University 94
28 The Forest Cloud partial view of singular trees from Cadiz province, Spain 97
29 The Forest Cloud partial view of urban trees from Philladephia 98
30 Data Store Entity-Relationship diagram 104

1 The XML is Dead

1.1 Introduction

Since almost the beginning of telecommunications, environmental data such as weather has been encoded and transported by wired and wireless media. During the last decade, access to Internet has modified research patterns among scientists and managers working with natural resources. Along with a fast improvement on data gathering and analysis, the epistemological grounds of empirical sciences have been moved in new directions. Scientists working with natural resources, nowadays, collaborate with other scientists via interdisciplinary, transdisciplinary or multidisciplinary studies more often than they did in the past. By working collaboratively together, often times they have to share and integrate data-centric documents, communicate them fluidly, and in many instances learn new methods to accomplish the goals of the project [130]. Open access to environmental data has been positively cited in numerous articles under the pretenses that data can be reused by other researchers to accelerate scientific discoveries and move science at a faster pace [237].

Researchers and managers working with natural resources work with many data formats many of which are discipline specific and a few common ones such as the ubiquitous spreadsheet and comma-separated-value files. Even disciplinary collaborative teams have found that data integration becomes a spinous and problematic matter since integration of data-centric documents with different formats might require expertise in information systems which sits beyond their reach. The problems generated by data integration in projects regarding natural resources have been documented profusely during the last decade, and the number of attempts to provide technological solutions for data integration becomes a regular research exercise [69]. Multidisciplinary scientific collaboration by sharing data needs common understanding of data, and for a decade the prevailing research to attain data integration was to develop standardize data formats using the XML toolbox and a lot of middleware [229]. Therefore, a large number of XML standards were developed by the members of different groups working with natural resources that brought an overpopulation of XML standards with overlapping functionality and disparate semantics creating a Babelian chaos that did not benefit the primary goal of researchers and managers working in multidisci-

plinary projects [139].

Nonetheless XML has been used to encode popular metadata languages used to publish, archive and syndicate electronic resources[151][123]. The development of metadata standards to describe data-centric documents is becoming a interesting area of research [114] since it represents an alternative to data standardization not only for the communities that used the Internet to work collaboratively working with natural resources, but for a larger community with different goals such as data mining and discovery of semi-structured documents [133].

1.2 Networking

Environmentalists by definition have been collecting and using primary data since the first day that humankind started to study the laws that regulate nature, however networking and technological advances during the last century have changed the working methodologies of many scientists and managers.

1.2.1 The Paleozoic-network Era: 1880's to 1900's

After the telegraph and Morse's code was invented and commercialized, governmental agencies from most advanced countries started collecting, archiving and publishing environmental data. For instance, the Weather National Service from the United States of America started collecting weather in 1865 [231]. This office, during the last part of the nineteenth century, collected data to forecast weather, survey storms, and monitor river streams. This data was used with a vision of public interest to help society to manage natural resources. Early in twentieth century, scientists started to use numerical models to forecast weather and floods. For instance, Vilhelm and Jacob Bjerknes pioneered modern Meteorology and Climatology from calculations based upon natural laws [105]. On the same manner, hydrologists were able to improve our knowledge about the hydrological cycle already established by researchers during previous centuries such as Dalton, Halle, or Darcy. The availability of rain data on-the-fly helped to develop advisory protocols, and seed the field with primary data to develop improved systems to help societies to manage disastrous situations [80].

1.2.2 The Mesozoic-network Era: 1900's to 1950's

During the first half of the twentieth century teleprinters using Baudot's code were used to communicate weather data from point to point. Primary weather data were collected in weather stations all over the world, manually transcribed and electronically transmitted to many clients . These clients were involved in human activities such as aviation and agriculture that benefit from using this data [179].

1.2.3 The Tertiary-Network Period: 1950's to 1980's

The transistor invention was an inflection point that marked the beginning of an exponential curve in computational power and opened the gates to the creation of the Internet. In 1969 the UNIX system was developed with networking capacity that allowed the deployment of different networking architectures and applications. At the same time, several scientific groups created computer networks to satisfy their collaborative needs. During the next decade the ARPANET grew rapidly to link computers from many universities around United states to support the needs of a growing scientific community [202]. During that decade using the UNIX environment, researchers started using common networking protocols to transfer text and data between computers such as electronic mails and data files, and to operate remotely. At the same time, digital spatial information started to be created, collected, archived and distributed to clients using these networks and other venues. For instance, remote sensing data was collected from early earth observing satellites launched by different countries, photogrammetric imagery was scanned into digital files, and the United States Geological Survey (USGS) developed the first digital line format maps to storage spatial data [164]. Parallel to this technological advances in spatial information, scientists started to developed new methods and analysis pioneering the rich field of geographical information systems [156].

1.2.4 The Quaternary-Network Period: 1980's to 1990's

During the next decade, the Internet double its size every year [203] and several organizations from the university spheres developed new important scientific networks: NASA, NFS, DOE, and CERN are good examples. At

the same time computer availability and computational power followed an explosive phenomena benefiting scientific research. At the same pace, the number of scientific data-centric documents grew creating the first dilemmas about data-format disparities, data standardization, and disciplinary data aggregation under public data banks [102]. The GenBank in 1982 was one the first public funded project with open access to store and annotate collections of publicly available nucleotide sequences and their protein translations [230]. Since then, the creation of new public data repositories with primary data useful for scientists has been an ever growing phenomenon that have lasted until today[21].

1.2.5 The Contemporary Period: 1990's to 2000s

During the last decade of the last century, computer power followed the same exponential path, but the Internet grew by an even faster pace [124] because it came out from the scientific closet, and reached the civil society as a whole. During the first luster of the decade, the standardization of the HyperText Markup Language (HTML) along with the HyperText Transfer Protocol (HTTP), pushed to the limits from a larger commercial interest, brought the Internet to the large society who embraced it immediately. This unparalleled networking phenomenon brought the environmental communities into even a closer contact than ever before allowing multidisciplinary work by easing communication strategies [204]. By the end of the decade, a number of initiatives from scientific communities emerged impetuously to cover data standardization, metadata annotations, and syndication of information. The Word Wide Web Consortium (W3C) patronized the creation of the eXtensible Markup Language (XML) as a formal markup language that could be tailored through specific schemas to define specific XML languages to encode data-centric documents. This new paradigmatic language was portrayed in many instances as the new and first lingua franca for scientific data-centric documents [27]. In 1995 a group of people interested in finding and sharing Internet resources developed the Dublin Core Metadata Initiative (DCMI) to create simple standards which will play a fundamental role in the Internet space where millions of new documents where populating the ever growing population of nodes in the Internet. The DCMI developed a controlled structures with enough terms and semantics to build vocabularies to describe Internet resources with an archival perspective, and with

the vision that new generations from different disciplines can aggregate more semantics and terms into the metadata. Therefore, this metadata language was quickly embraced by other initiatives from a number of diverse disciplines to annotate their resources [194]. By the end of this decade, private companies patronized the development of the Rich Site Summary as a very simple and straightforward language encoded in XML format to syndicate Internet resources that could be properly broadcast to the appropriated recipients [153]. This simple language coupled to a simple syndication protocol was rapidly accepted by the Internet community since it had a clear purpose, a clean interface to implement, and a succinct standard language to convey the information [219].

This important period saw how private and public agencies dealing with environmental data started to publish their repositories, and environmental data becomes abundant on the Internet; how academics projects started sharing their information with the public through public accessible websites; how multidisciplinary research projects were on the rise, and their members created private intranets to keep communications among peers fluid; and how information technologists created 'ad-hoc' solutions to integrate data from disparate sources and different formats on data repositories [46]. By the end of the decade, several disciplinary data languages encoded in XML had been published with the goal of becoming the standard formats to share and exchange data.

1.2.6 The Semantic Web Period: since 2000

During the last decade technological advances, computer power, and networks communications kept an exponential growth pace in number of users [47]. Scientific data and inventories about natural resources grew also at faster pace since primary data acquisition became cheaper and reliable, and the public was becoming more conscious about the environmental issues that our planet was facing. Even in this short period of time, scientists and researchers working with natural resources have gone through three very distinctive Web applications with the same basic goal to support collaborative projects [38].

1.2.6.1 The Web Portal Hundreds of Web Portals with databases with environmental and biological primary data claimed their Web space, but

these databases grew as independent islands with disparate communication protocols and languages to encode data-centric documents which made collaboration harder than it needed to be since data discovery became something but trivial, and data integration from different sources a problem tractable only by specialists in information systems [112].

Public administration agencies offered disciplinary data-centric documents through their public Web Portals, and multidisciplinary scientific projects aggregated data information in file repositories reachable by the Internet [111]. Private companies and non-governmental agencies collected and provided environmental information through their servers. However, from the data management there has been a clear evolution from simple file repositories to semantic integrated databases [10]. Nonetheless, this evolution has not happened across the full spectrum of disciplines involved in environmental research and management. For instance, bioinformatics has became a multidisciplinary discipline in itself with a dedicated community of researchers pushing the limits of data integration among interdisciplinary groups publishing their work in well-recognized scientific journals. Geographers are another community that has pushed the envelope of Web Portals through novel ideas and research about new information models to integrate data-centric documents, and protocols to retrieve and serve the spatial data.

Syndication of scientific information has become a regular practice by publishing news using the RDF Site Summary (RSS) language. This language with very simple structure and semantics has been popular for Web portal administrators to aggregate and broadcast scientific news, and has been studied to extend its semantics to syndicate environmental data-centric documents [123]. Among several variants , the most common scientific syndication protocol is managed by a central computer that tracks known sites designed by researchers and managers where they post RSS documents which point to information that wants to be published, and collects the feeds into a database which is routinely surveyed by news aggregators which have been implemented in Web browsers since the beginning of the decade. This very simple voluntary broadcasting method has become a well-recognized standard protocol among Internet users in general and environmental scientists and managers in particular [113].

Besides data management, Web portals have become the scientific melting pot for multidisciplinary communion of ideas, and the public visible showcase of the projects. It must be difficult, nowadays, to find an environmental

project without its proper Internet space and Web portal that tells the world who they are, what their goals are, and how they work. Modern Web portals provide their members with tools and services to ease their work by allowing easy communications among their members since as many researchers have concluded fluid communications are a cornerstone for multidisciplinary research [**?**]. Software engineers and companies made it possible to develop and tailor Web portals with ease that supported many types of synchronous and asynchronous communication protocols allowing researchers and managers to fit the expecting demands [243]. However, the lack of common standards to integrate multidisciplinary data-centric documents forced the designers of Web portals to provide 'ad-hoc' solutions [161]. Therefore, integration of data-centric documents across Web Portals remained untamed.

1.2.6.2 The GRID After the advent of parallel computing during the 1960's and three decades of research, parallel computing has become a well-known and established paradigm that has helped to improve the computational capabilities of our systems, and, therefore, accelerate the unveiling of scientific knowledge since impossible methods from the past became doable, and some scientific frontiers started to close under the crunching computing power of the GRID. During this decade many grids have been created by the scientific community with different purposes and philosophies, but with the common goal of improving scientific inquiry.

Volunteering computing has been one of the most common GRID computing system developed during the last decade in which computational power has been attained by volunteering the computational capacity of hundreds of computers. For instance, the Berkeley Open Infrastructure for Network Computing (BOINC)[17] was the first GRID to used donated time from networked computers to be established just at the beginning of this decade. The BOINC has been hosting a number of research projects which benefit from the computational capacity of networked computers whose computational capacity has been donated to the GRID. Right after BOINC, others volunteer GRIDs have been developed at other academic institutions such as the Folding@HOME launched at Stanford University[29] or XtremWeb developed by a group based at the University of Paris[94].

On the other hand, multi-institutional groups have joined efforts to create extremely fast computing networks to host projects with intensive calculation needs by networking large supercomputers. Projects like the TeraGRID

in the USA, DEISA in Europe and NAREGI in Japan represent parallel efforts from different scientific communities seeking to provide computing power to scientific projects that require intensive computations as it was also the case for the projects hosted by volunteer computing. A data GRID is often called for with a project whose main focus is data integration instead of fast computation. The Biomedical informatics Research Network teams have developed its own data GRID to share data among people involved in the medical field through many campuses and hospitals in English-speaking countries. This data GRID offer them particular rules to share data, manage metadata, and create collaborative projects. RoadNet is another example of data GRID project under the TeraGRID umbrella whose mission is to integrate real-time data from a network of different type of sensors that collect real-time data.

Under the obvious growth in the number of grids, and the multidisciplinary and collaborative view of modern sciences, the Globus Alliance, a community of organizations and individuals is developing fundamental technologies behind the GRIDs, it has created the Open GRID Services Architecture (OGSA) to document "best practices" [18], guidelines, and standards for open GRID technologies. This Alliance has hosted a conglomerate of working groups developing a common basis to pursue a functional open GRID based in shared practices regarding the different aspects of the open GRID . Within OGSA, the data format definition language working group whose goal is pivotal for data integration in a network of disparate resources, has produced an incomplete but complex language whose description capabilities remain to be tested with real data-centric documents [3]; therefore, data integration across open GRIDs is still an open problem as it was for Web portals and will be for next projects. To resolve the data integration issue, OGSA members have developed a simple Open GRID Services Architecture Data Access and Integration (OGSA-DAI) that allows each project to develop ad-hoc solutions on each particular basis; therefore, favoring the integration of data resources using a black-box approach which is a pragmatic but non-ideal solution.

Different communities from different fields working with natural resources have establish their grids to facilitate and encourage collaborative partnerships among their community members. For instance, the biological community has created the BioGRID to house and distribute collections of protein and genetic interactions from major model organism species. The Biomedi-

cal Informatics Coordinating Center (BIRN-CC) has developed, implemented and supported the grid-infrastructure necessary to achieve distributed collaborations and data sharing among the test bed participants. The RoadNet wants to enhance the capacity to monitor and respond to changes in our environment by developing both the wireless networks and the integrated, seamless, and transparent information management system that delivers seismic, oceanographic, hydrological, ecological, and physical data to a variety of end users in real-time. The Long Term Ecological Research group and the National Ecological Research Council have consolidated two grids with many similar and equivalent goals such as to integrate ecological data-centric documents from different sources in different formats under the same own information system; however both systems don't have any formal mechanism to exchange or coordinated efforts which by definition contradicts the spirit of both systems. The TeraGRID is a gateway to share data, computing power, software applications with scientists from very different disciplines: astronomy, chemistry, earthquake mitigation, geophysics, global atmospheric research, neuroscience, molecular biology, cognitive science, physics, and seismology. However the TeraGRID system does not establish any protocol for an interdisciplinary usage of the GRID [30], and the same problem of data integration happens as it happened since the beginning of the Internet .

Each GRID will ask new members to install its particular client application and modules into your system in order to be able to participate [170]; therefore, for truly multidisciplinary GRID research across particular GRID boundaries, researchers will have to install as many clients as grids you might access to retrieve the needed data or functionality; besides these data sets will be unreadable for you unless you have the correct software to read them; therefore it is sound to infer that you might need as many software applications as GRID clients to be able to read the data into your computers besides the software tools to manipulate the data-centric documents. Ergo, data integration is not improved at all, since there is only file integration. These disparate GRID solutions do not lower the impedance against the flow of data through a distributed network of environmental actors which come truly short of the main goals among researchers and managers who works collaboratively.

Because of this manifest lack of generalizable technologies to integrate disparate data, and the need to merge and manipulate data from different sources to satisfy the multidisciplinary projects, the open GRID standards

and particular grids projects are favoring the idea of GRID as a service whose main goal is to ease collaboration by providing scalable computing services to clients with simple interfaces while using a big black box commonly called the Cloud [20].

1.2.6.3 The Cloud The Cloud is the third paradigmatic networked architecture that has became popular during the last luster of this decade as an extension of the GRID concept where computing becomes a service coming from a big black-box called the Cloud that offers services such as data storage, publishing, fast computing, and software tools. The Cloud has no concrete technical definition, but shares the values of the GRID as a network of distributed computers which offer scalable services to clients. The Cloud could offer its scalable infrastructure as a service to store from a couple of files to large databases, or to accommodate computational requirements from one processor to thousands of them. On the other hand, clients can also use the Cloud as a platform to build Internet applications from a website with few webpages to a Web Portal with millions of documents, tools and services [201]. Ultimately, the Cloud has been used to serve software applications on demand to clients allowing software manufacturers providing their services to exert better control over the usage of the software while reducing maintenance costs.

Private enterprises are developing their own clouds to serve their commercial interests and the community given the capabilities of the network, computer hardware, and software tools. Since the capabilities of the networks and computers are expected to keep exploding with the use of better knowledge [16], Cloud computing will become a main utility among environmental communities even though during the last couple of years Cloud computing has been used exclusively by large private for-profit corporations.

Scientific groups are starting to develope clouds to serve large-base domain-specific communities to address their peculiar needs. For instance, the members of the interdisciplinary medical community are implementing clouds [218]where it is advised that researchers will require of fundamental changes in their 'modus operandi' [213]. Members from the environmental community are also moving in the Cloud direction implementing this type of cyberstructure for different disciplines such as forestry [132], oceanography [55], and or soils science[178].

From the Web Portals to the Cloud, a common and general goal among people working with natural resources collaboratively has been to create a distributed information system where data can be easily discovered, retrieved and manipulated. For instance, researchers monitoring climate change would like to access data from different sources such as social , dendrochronological, geographical and weather sources, and to integrate these data into its logical framework for analysis. Different members of the scientific community have other particular goals which are not shared among all members such as fast and reliable data transfer of large data-centric documents, high computational needs for modeling and analysis, artificial intelligence for automatic reasoners, or real-time integration of disparate resources. However, they share a common goal of attaining seamless data integration across different projects which unfortunately remains inalterable and intractable. One pervasive problem across all these initiatives which collect and present data through the Internet is that data integration from different sources is solved using a black-box approach which by definition defies generalizability and does not apply well to the open multidisciplinary distributed data-centric system which is required for so many researchers and managers working with natural resources.

Managing the Earth planet which has a limited amount of natural resources, a rich biodiversity in peril, and a conscious human specie that can change the fate of the planet are the driving forces of a sole scientific movement; therefore, all scientific activity and findings are interlinked as concepts in the intellectual space, and the primary data which supported these concepts should be open linked in the Web space [74]. This is the driving motto for the members of the W3C who are patronizing the Data Link project as an open Cloud where all data is linked, and accessible, and where generalizable solutions are researched to avoid black-box solutions [35].

1.3 Environmental Collaboration

Environmental research and management have had a history of bright personal ideas and successful collaboration among peers with basic common goals such as better understanding of the interaction among the planet and the living beings to provide humans with the knowledge to sustain and maintain the capacity of the planet for future generations.

Collaboration is not a singularity from environmental research, is the cornerstone of classical fields such medicine or emergent ones such bionanotechnology. Members of the medical field have pioneered collaborative work at research and administration levels [215]. Modern medicine has aggregated a number of disciplinary researchers recognizing the holistic nature of human medicine and the absolute need for a fast pace of medical findings[82]. This fast medical pace would be impossible without the collaboration of a multidisciplinary group of scientists such as engineers, mathematicians, physicists or computer savvies. Bionanotechnology, for instance, is a new important collaborative field that requires expertise from different fields to attain results at a record pace [195]. Collaboration is justified because the sum of efforts can not only expedite the research process, which is nowadays a critical aspect, but provide solutions that cannot result without it [237].

Collaboration has been ontologically classified under different symbols based on certain aspects such as the discipline or epistemic values of collaborators. Multidisciplinary, interdisciplinary, cross-disciplinary or trans-disciplinary are few of the symbols used to defined collaboration [238].

Departments of interdisciplinary studies are now common in numerous universities around the world where researchers from different disciplines aggregate together to seek solutions to complex problems which cross the boundaries of several disciplines [119]. For instance, to evaluate the impacts that human activities have on ecosystem is a common research topic for multidisciplinary teams. Marine ecosystems are a good example to see collaborative efforts at different levels in order to understand, manage and save marine biodiversity [63]. Atmospheric scientists have developed a FLUXNET network to exchange and ease collaboration being data exchange a cornerstone of the project[24] .Forest management is another classical field where researchers, managers and people from different backgrounds and knowledge have joined efforts to defend the forests of the planet since most of the people understand the importance of trees for all living beings in this planet[226].

The analysis of bibliographical records has been used in the past to characterize scientific activities. The number of authors per article has been used as a proxy for measuring collaboration [158]. Numerous papers have shown trends that collaborative scientific work is becoming more prevalent. This thesis also confirms that the same patterns happen in environmental sciences. As can be seen from Figure 1, the signature of collaboration in environmental studies is clear. During the last ten years more than half of

the references from environmental articles analyzed were signed for three or more authors. This statistic contrasts significantly with the results from other non-related disciplines as distant as Sociology and Mathematics which show different publication patterns during the last decade where single-authored papers are the most frequent type of work. On the other hand, looking at the trend of number of authors per publication in the ecological field in Figure 2 , it can be seen that the number of single-authored papers has decrease significantly while the multi-authored papers have increased during the last decades showing distinctive trends. Both analysis clearly define the nature of environmental research in terms of collaboration.

Collaborative environmental work is not a field of roses where efforts are counted by successful projects; different scholars have analyzed the barriers to environmental collaboration and a number or recommendations have been proposed . Besides the ones which refer to personal and social human behavior, data sharing is a common direction advised to bring better collaboration among environmentalists since their work is based on primary data [129].

Collaboration by data sharing could have different communicative pathways and goals. Most collaborative efforts are limited to few people who have overlapping interest and complementary knowledge. These co-authors communicate with each other using private venues. These authors often times live and work in distant places and use the Internet to keep in touch and share data and ideas. The significance of collaboration between authors from different countries goes beyond the pragmatical aspects of collaboration and into the philosophical aspects of a worldwide environmental effort.

The other common pathway to data sharing as it has been already mentioned in the previous section is by creating public databases with environmental data where group of scientists come together on a disciplinary or multidisciplinary basis to share data. This data is deposited in these data banks by its authors to be used by a third party. The data is then publicized and made available through public gateways for the environmental community.

Figure 1: Authors per Article From Different Disciplines during Last Decade

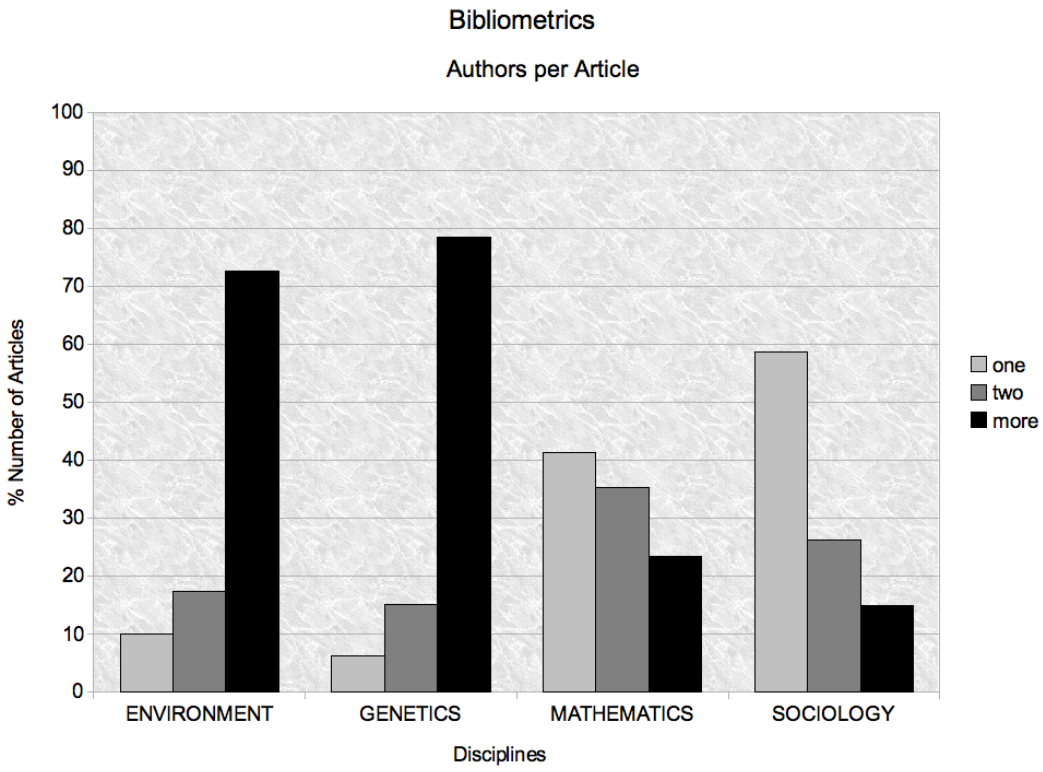
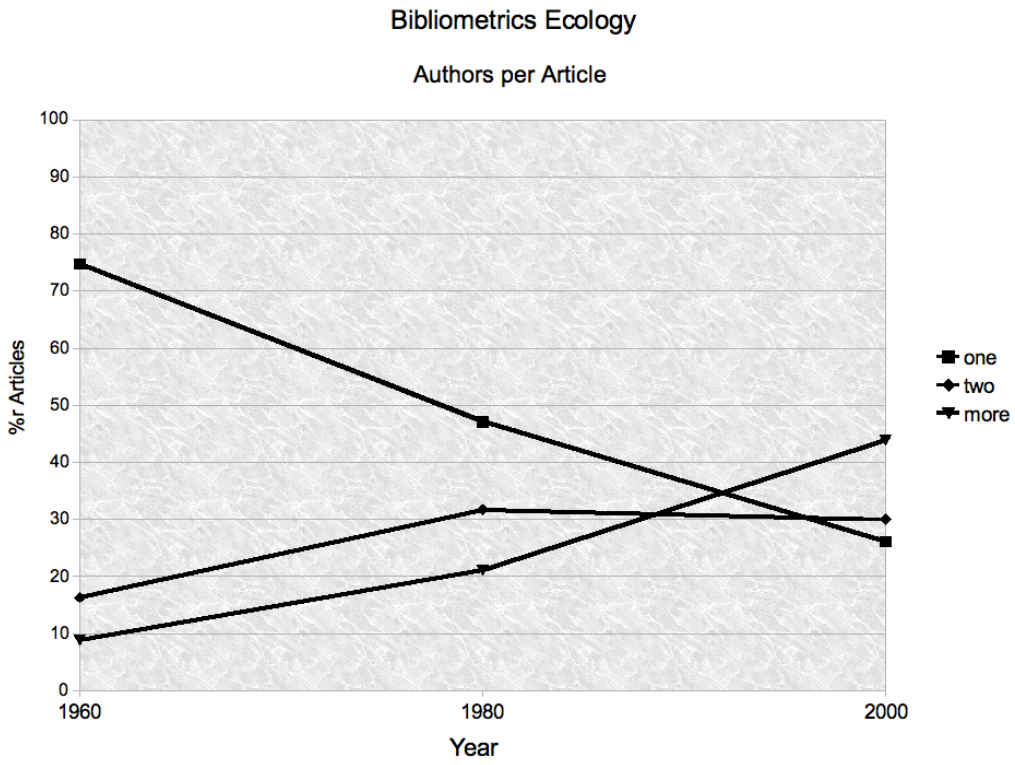


Figure 2: Authors per Article across Decades in Ecology



1.4 Environmental Data Exchange

Since telecommunication started with the commercialization of the electrical telegraph, and the invention of the Morse code in the 1830's , society has significantly changed its 'modus operandi' for collecting and sharing data. During the 1860's, telecommunications were an important factor for the arrival of the capitalistic era from the previous mercantilistic society. Until this decade, capital and labor were physically conjoined factors [53] , but since the commercialization of telecommunications, both factors could be physically distant, i.e., the natural resources in the fields of Texas, and the capital in the banks of New York. The modern ability of exploitation, often

times spoliation, of natural resources has been linked to this new capitalistic system were telecommunications have been a cornerstone tool to support it.

Therefore, the financial sector was one of the first groups that took advantage of the capabilities of data telecommunications. During the early twentieth century, the financial sector develop the Baudot's code and teletypewriter systems to support its capitalistic activities. Both, Morse and Baudot's codes can be regarded as the first coded languages to transfer data using electromagnetic devices. However, they were human-to-human systems which could not be automatized. To solve this shortcomings, standardization groups were formed to develop standards for Electronic Data Interchange (EDI). These groups have developed more than three thousand standards to address data exchange between financial partners to enhance business processes, reduce costs and expand the reach of their organizations.

While the financial sector has benefited from telecommunications and actively developed it to satisfy its needs, the environmental sector has lagged for decades. Scientists and managers started to exchange data electronically right after the Internet was first developed and the first four nodes were established in four different universities in 1968. This year marked a point of inflection for the world in general and for scientists and managers working with natural resources. Since then, the capacity of information exchange on the Internet has grown following Moore's law doubling its size every lustrum [203].

During the first three decades of the Internet, data among environmental researchers and managers has been exchanged using a myriad of formats. These formats have a formal structure defined in their respective specifications and have plagued the modus operandi of environmental researchers and managers. These large number of data formats have inflicted a complex impedance to the ability to share data in a multidisciplinary field where collaboration has grown in importance. Many of this formats are disciplinary and they are linked to specific software tools while other few formats perform as general containers of environmental data used in almost all disciplines. While the former files represent niche communities, the later ones represent the body of environmentalists as a whole. Some of the general data containers have become 'de facto' standards to exchange data such as spreadsheets and DBF files. While spreadsheets and tabular files are the most popular data containers to exchange data, a number of formats developed specifically to resolve data integration have not attained their goals such as CDF, netCDF, or HDF.

During the last decade and given the exponential growth of the Web, the need for data exchange has been exacerbated and the search and research for solutions have exploded. The Extensible Markup Language surged as the new paradigmatic format to encode data-centric documents and thousands of languages under its umbrella were originated. During its first lustrum of existence, XML attracted a number of researchers and managers to develop and write about it. Thousand of research papers have been written to illustrate the capabilities of the new language in the toolbox of scientists and managers. However, very few revisions have been written, and no one has brought any quantitative data to support the success and penetration of this paradigmatic language among researchers and managers.

Geographers have always been providing basic primary data for others by creating and distributing spatial information. During the last four decades, geographical information systems have been developed all over the world. This wealth of information whose importance for environmental analysis and management is beyond any doubt, sits in many computers, under diverse formats, and in most cases hidden from public access making its retrieval a costly enterprise. To ameliorate this situation, several groups created the Open Geospatial Consortium (OGC), for developing standards to represent spatial data. From these efforts, a vast set of standards and recommendations have been developed. The Geographic Markup Language (GML) has been the fundamental recommendation created by the members of the OGC to house and distribute spatial data. However, the usage of the Geographic Markup Language (GML), an XML language, is minimal after more than 10 years of constant reformulation of the standard to fit the increasing commercial aspects of geographical services [143].

Earth scientists who monitor and study the Earth's atmosphere come from different backgrounds working together to understand and defend this singular planet in this particularly difficult age where the only blue planet is under risk of profound changes due to the human activities. The functional relationships between the objects of study by earth scientists, the array of the different disciplines involved monitoring Earth, and the multitude of different methodologies to study the subjects make Earth scientists a body with special needs to work together, and share primary data. Earth scientists use a plethora of data formats . National, supranational and global organizations have created websites to collect and publish primary data under discrete efforts incompatibles among each other at the data handling level. Unlike

geographers, Earth scientists have not created any global strategic alliance to normalize data communications and systematize its access. Nonetheless, the bounteous XML-thinkers have been prolific in creating markup languages to encode for the different disciplines but not following any common guidelines. Therefore, there are many singular ad-hoc efforts which remediate the collaboration needs of specific groups. For instance, there are Internet portals to compile, archive and publish weather, paleontology, atmospheric, soil, water, or satellite data using very different protocols which makes data integration a Babelian problem.

Life scientists conformed another very diverse group of researchers and managers who try to unravel the secrets of life. As with the previous group, there is not a comprehensive alliance to integrate their data-centric documents, and there is a proliferation of ad-hoc efforts to create Internet gateways for sharing data and knowledge. As in many other scientific fields, comparative analysis is an epistemic foundation for life scientists that requires primary data from different sources to validate knowledge. Microbiologists, taxonomists, botanists,zoologists, ecologists, and geneticists, have created numerous portals to publish their data. For example, ecologists have created a global Long Term Ecological Research (LTER) group to satisfy the needs of its members to share data, reuse it and open wider the door of collaboration. This group was the first group to understand that metadata can be used to describe the physical structure of data-centric documents in order to achieve the goal of data integration. They developed the Ecological Metadata Language which has structures to define the physical data in tabular data-centric documents.

Social scientists work tirelessly to understand how the environment influences culture and vice versa. Among all environmental-related disciplines, they have done the least to integrate and publish primary data. They use legacy datcentric documents to store and share their data. In addition, most of these social scientists do not know about XML at all.

Environmental professionals such as foresters, and agriculturists have not formed any coalition to integrate their data for the most part. Their data often times is not only divided by profession but also by administrative boundaries which atomize the data in almost uncountable places. For instance, it is materially impossible to know how many trees are in the world, and where they are. Most data about trees has been usually deposited in private spaces using different legacy formats since the first tree inventory was completed back in the nineteenth century. Even taking into account the explosion of

XML languages during the last decade, only few were created to support the activities of these professionals working with natural resources.

1.5 XML Spaces

After XML was standardized and became a recommended language by the W3C in 1998, many researchers and managers have worked to promote XML as the perfect tool with many usages such as to standardized data communications among peers [141], to federate databases [118], to syndicate data-centric documents [114], or just to save data sets in the hard drive. Therefore, XML has been presented to the society as the swiss-knife for data management. However, no work has been presented to study the penetration of the xml technology in general; ergo, the reach of XML among environmental scientists and managers is unknown. Moreover, given the fact, that collaboration is a requirement for many projects, and sharing data a basic component of this collaboration, it is long overdue to ascertain the penetration of XML technology among the members of the scientific community.

1.5.1 Bibliography

The bibliographical space is a good mirror to analyze the activities of the scientific society to look for patterns of behavior. By analyzing a large corpus of scientific references (see appendix Bibliometrics) and, looking at the trend of the XML curve in Figure 3, it is evident that since its beginning XML had the attention of many scientists working in environmental sciences. The peak of the curve matches the era of scientific Web Portals which need technological solutions to aggregate and serve data.

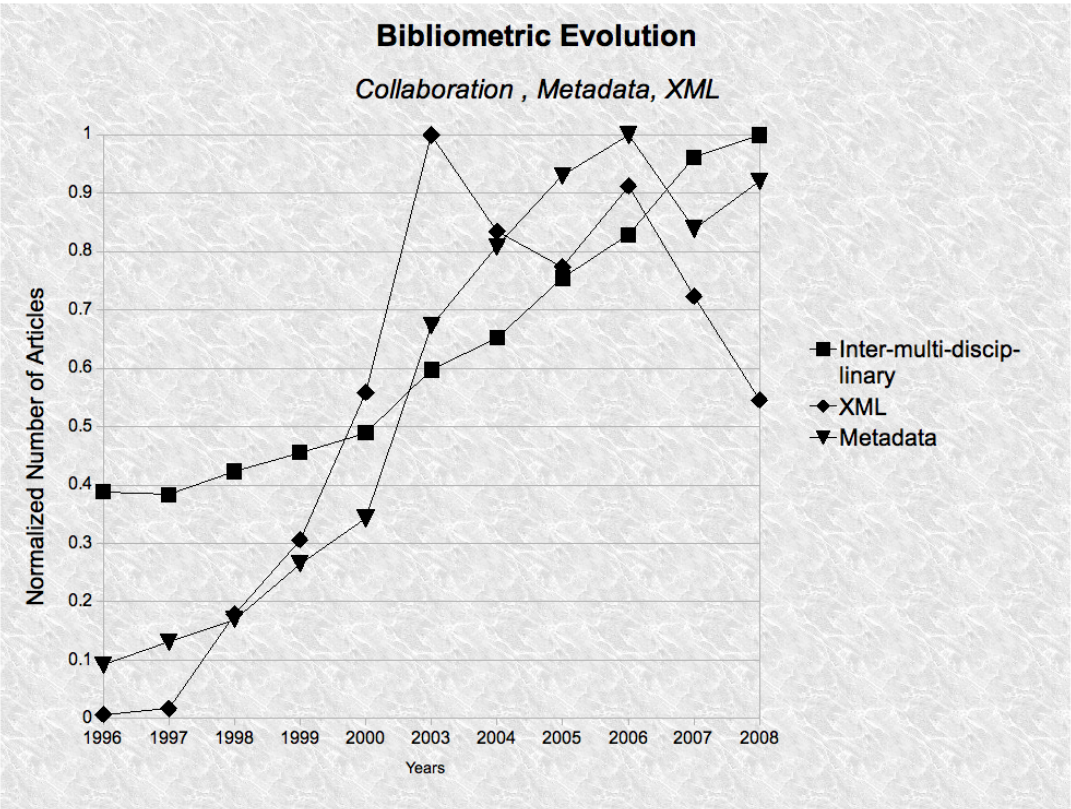
Scientists and managers at the time often claimed that there was a Babelian chaos regarding the formats used by scientists, and XML was the perfect technology in the toolbox to solve this data chaos which was encroaching the possible growth of scientific Web portals. By standardizing data using domain-specific XML, the claim that XML was the key technology to integrate data-driven projects was prevalent across many scientific papers.

The number of XML languages grew considerably during this first period of the XML as can be seen in Figure 5 where many competing languages were

developed by scientists from different projects. This XML confusion also created the a similar semantic chaos [139].

After reaching the maximum peak in number of publications in the year 2003, there is a clear negative trend for the number of publications about XML that could have several causes. One of these causes, is that XML has not penetrated inside the environmental community as it was predicted in those numerous papers.

Figure 3: Normalized View of the Number of Scientific Publications sorted by year and Keywords



1.5.2 Public Web Space

The Web is another space that can reflects the patterns of a community by analyzing its contents. There should be a positive correlation between the number of soft-documents of a determined kind in the world and the number of documents of the same kind in the Web space. Therefore, by looking at the number of XML documents in the Web space, the penetration of the XML technology can be assessed.

Using the programmable search interface from Google that can be programmed to retrieve documents by a given type and having a determined keyword in the text of the document (see Appendix XML documents in the Web Space), the number of XML documents per language can be measured. The results of this analysis that appear in Figure 4 shows that metadata languages which use XML as a language are significantly more popular than the data-centric documents. The lack of data-centric documents in the Web space using XML languages as an encoding format is striking given the fact that XML has been a popular research topic.

Another parallel reading from the results of this analysis is that the the popularity of the XML depends of the purpose, and not in the XML itself. Metadata providers have been very comfortable producing metadata documents using XML; on the contrary, data providers either have provided few documents or have few documents to provide.

The lack of existence of XML documents used to store scientific data in the Web space may be because these documents can be stored in private spaces beyond the reach of the Web.

Figure 4: Number of XML Documents Retrieve by Google Search Engine by Language

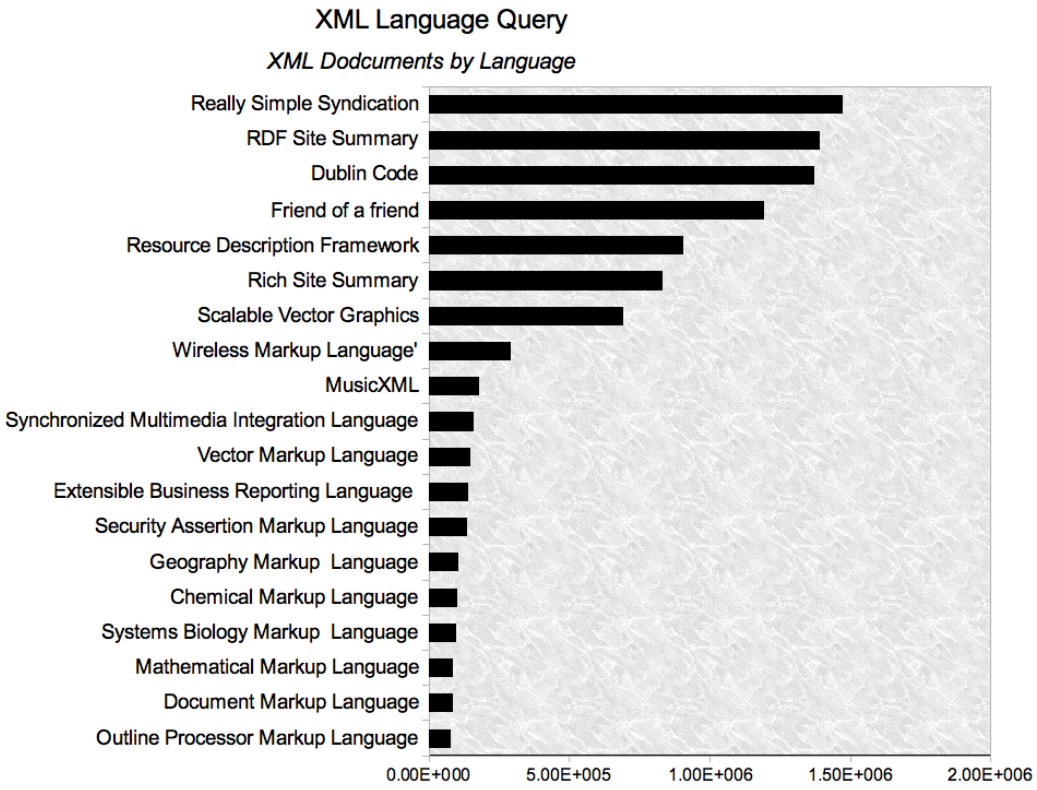
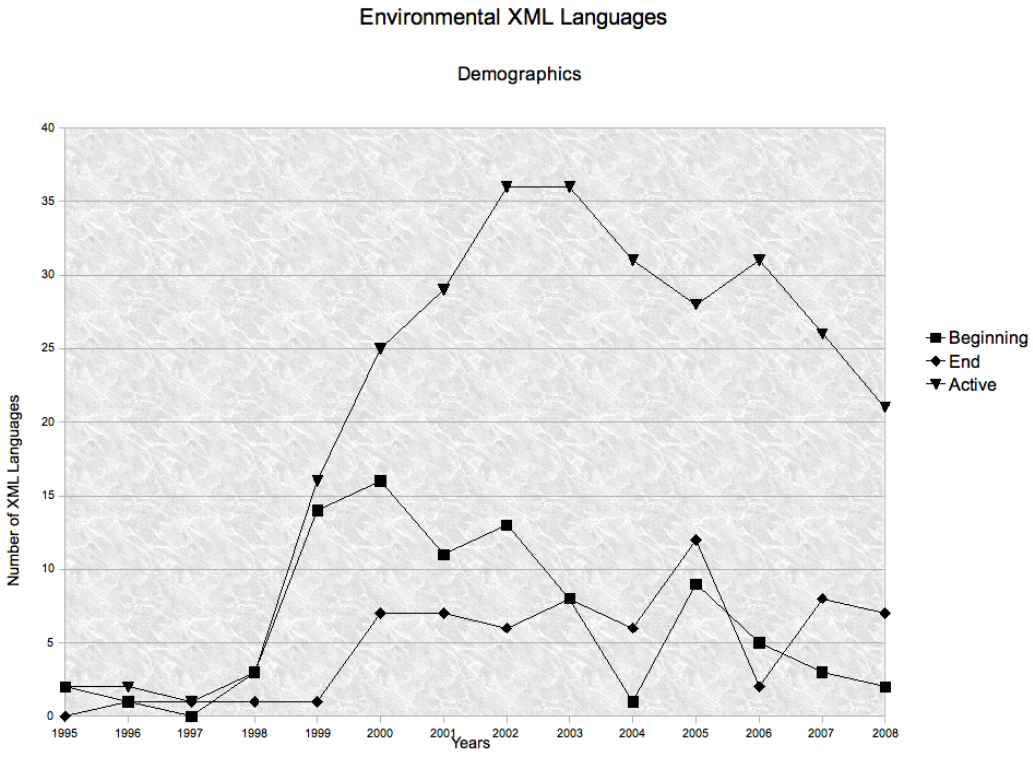


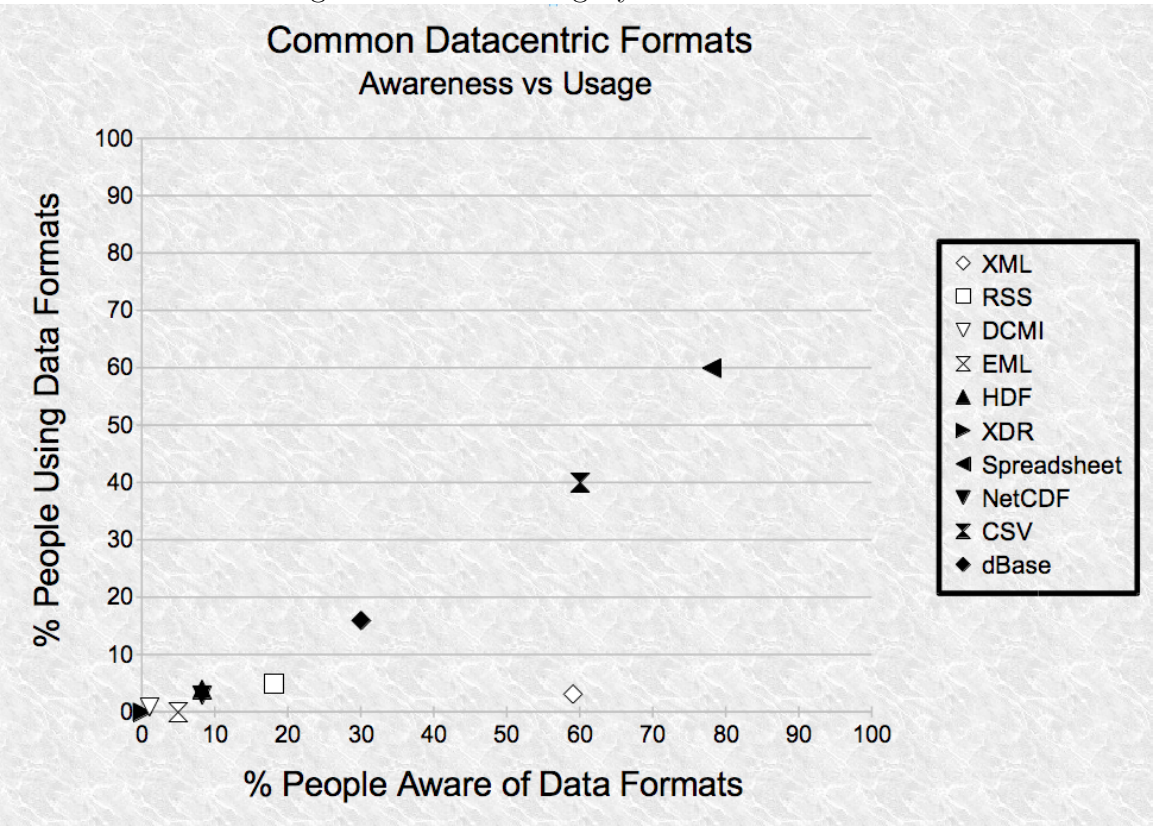
Figure 5: Estimated Lifespan of XML Languages



1.5.3 Private Spaces

Since scientists are reluctant to openly share their data in Web spaces, it is plausible that many data-centric XML documents are stored beyond the reach of the Web. Therefore, a voluntary survey was distributed across empirical disciplines across at all UC campuses to measure the usage of data-centric formats. The results of the survey (see survey in appendix n) as can be seen in Figure 6 indicate that the XML format for data-centric languages is well known and not used. From the results of the survey, it can be inferred that the legacy formats are still preferred, and XML while well known is not used because there must be logistical and methodological barriers that impede its penetration.

Figure 6: XML vs Legacy Formats



1.6 Conclusions

The reality is that XML technology as a data exchange mechanism has not been embraced by the environmentalist community and professionals working with natural resources. This simple, but important statement can be grounded as truth by looking at the Internet space and computer habits of environmentalists. Using the search interfaces of two important Internet companies to look for the different XML languages that have been developed to exchange environmental data residing in the public Internet space, similar results are found. This corroborates the robustness of the analysis and shows the lack of environmental data encoded using the XML languages. This lack of XML data indicates that the penetration of XML technology on the environmental community is extremely low. By surveying the researchers

and managers working with natural resources, the lack of usage of XML is unquestionable even though the XML technology is known by them. Therefore, there are barriers that disallows its usage when collaboration among researchers and manager working with natural resources is on the rise.

Big science is now e-science, and e-science implies collaboration which implies among others issues the sharing of data. XML was the mechanism that was recommended by the W3C and recommended by many scientist and managers as the tool to standardize data-centric documents. Even though XML is a known language among scientists and researchers, only few use it. Instead, they keep using tabular data-centric legacy formats such as CSV, DBF or spreadsheets. After more than a decade of environmental data-centric XML, it is clear that XML has not achieved its goal of widespread use and other alternatives need to be pursued.

2 Long Live Metadata

2.1 Introduction

The multidisciplinary environmental community has grown, and collaborative efforts have multiplied to discover and manage the rules of nature, therefore, sharing data have become a pivotal aspect on environmental research and management [190]. After more than ten years of extensible markup languages, and thousands of publications defining markup languages and web-based frameworks to share or exchange data, the usage of these auto-proclaimed standard environmental markup languages is very minimal among environmentalists who still use legacy data files and tools. Therefore, metadata languages which describe environmental data documents are becoming more important since it is the other clear alternative for improving collaboration through automatic data exchange mechanisms. In order to use metadata languages for data exchange, these languages have to be able to completely define the environmental archives and their contents. During the last decades some metadata languages have been developed to characterize data-centric archives. These metadata languages have been designed with different goals, but mainly they have been designed by librarians with archival purposes to be able to retrieve the datasets [101]. However, the optimal metadata language should have enough expressive power to completely define the data

from any data-centric archive while using a simple language which should be processed with ease in a timely manner; at the same time, this ideal metadata language must be alive having a large support group of researchers and managers who must provide oversight and development of it to facilitate its usage.

2.2 Environmental Archives

Environmental archives are where environmental researchers and managers store their precious environmental data using software tools following determined specifications. Nowadays and because of the inertia of the Semantic Web[1]several authors wants to use its infrastructure to maximize the reach of the data; moreover some of them are trying to transform the scientific epistemology from a document-centric to a data-centric science [81]. Therefore, if scientists and managers want to share the data from their archives besides needing to find solutions to the reluctancy of many of them to share their data [64], we have to understand the nature of the archives from different points of view to offer general solutions that can be applied openly.

Environmental data archives for the purpose of this thesis can be classified based on their encoding, syntax, logic structure and semantics.

Typically environmental data archives have been classified based on their encoding as binary- or text-based archives.

Binary-based archives are those whose values are stored in a stream following common binary encoders. The main advantage of these type of archives is their high data density and closeness to the binary format that the computers uses in their control process unit. These are two very likable characteristic since environmental sciences and management are working with large datasets and complicated algorithms; therefore, binary archives can be store more efficiently, transferred quicker and processed faster.

Text-based archives, in the other hand, are those whose values are stored in a stream using standard character encoders. The main advantage of these

[1]"The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by World Wide Web Consortium with participation from a large number of researchers and industrial partners."

environmental data archives is therefore that they can be easily inspected using an ubiquitous text application; therefore, they are easy to read and edit. In the other hand, these archives when they store numeric data can increase the size of the archive considerably. Because working with text-based files is straight-forward, many scientific disciplines have designed specific formats to collect their specific data; therefore, the proliferation of file formats is very understandable.

Mix-based archives using both type of encoders to store the data on a stream are also commonly used in environmental science and management, and they offer a mixture of advantages and handicaps.

Environmental data archives can also be classified based on their syntactical structure into two-way class system: context-free and non-context free archives.

Context-free archives are those archives whose stream can be defined using a context-free language, and parsed with a deterministic pushdown automata [12]. The physical structure of the stream of these archives has an important hierarchical structure that can be used to parse the stream into a logical hierarchical structure to represent the data in the memory of the computer. The physical structure can be defined using the Extended Backus-Naur Form (EBNF) notation since it is a common notation used to define context-free languages.

Non-context free archives languages are those which cannot be defined using a context-free language because the encoding rules have contextual values which make them not amenable for being processed or recognized with a pushdown automata. For instance, variable-length records in binary files cannot be defined with a context-free grammar following the common methodology to scan and parse the data stream since the length of the records are not known at the time the specification of the format of the file is designed. This common methodology follows the classical "a priori" approach to syntactically analyze streams where the grammar and the automata are designed before the existence of the stream that belongs to the language defined by the grammar [12]. However, a context free language grammar can be designed "a posteriori" to analyze a given binary file with any sort of variable-length records, i.e. by inspecting the physical structure of the records on the file, a context-free grammar or parser can be designed to extract the information as needed. Therefore, even if the specification of the format of one file does not

belong to the realm of the context-free languages, the physical data of the file can be extracted with a deterministic pushdown automata by inspecting the offset and lengths of the physical data structures inside the file and designing a contex-free language accordingly.

The specifications of data-centric archives can also be classified based on the their continuity of the records as contiguous and non-contiguous specifications. While contiguous specifications require that the records have to be written in a orderly sequence to the file, non-contiguous specifications use pointing structures allowing the records be written in any position in the file. Moreover, it is common for some specifications to have physical data structures whose members are placed in non determined order; instead the specification uses physical tags to identify the identity and semantics of the members [5] [7].

At the logical level, there are two classical models for representing the data: hierarchical and relational. Many of the small environmental archives can be regarded as having an implicit hierarchical logical model associated with them, and few data formats are of relational nature. In addition, is well know that a relational logical structure can be mapped into a hierarchical structure without loss of information [197]. Therefore, from a conceptual analysis, a hierarchical model should be able to represent all data-centric documents used by environmental scientists and managers.

Taking into account the semantics of the environmental archives, these archives can be classified as grounded, non-grounded and semi-grounded files. In logic parlance, a ground term is an expression with no variables where all terms are known and constant, and it can be used to conceptualize the grounding regarding the semantics of environmental archives. Many specification of archives have been developed by domain-researchers to satisfy their own requirements for storage of concrete pieces of information, and therefore, the meaning of the data from these archives is grounded, e.g., the decadal specification is a classical archive where we can know the semantics of every piece of data since the specification of its format specifies not only the physical data structures, but also the semantics of these structures. On the other hand, non-grounded archives are those files where the semantics of the data cannot be known before the existence of the file, e.g., a DBF file is a classical tabular archive used to store records of information with a constant number of fields whose semantics can only be known after the archive is written. In between these two extremes, we can find archives where only some pieces of

information from the specifications are grounded. e.g., a Shapefile is a classical archive used to storage geographical information where the semantics of the geographical information are grounded in the specification, and the semantics of the non-geographical attributes are not defined in the specification and will be defined by the geographer after creating the archive.

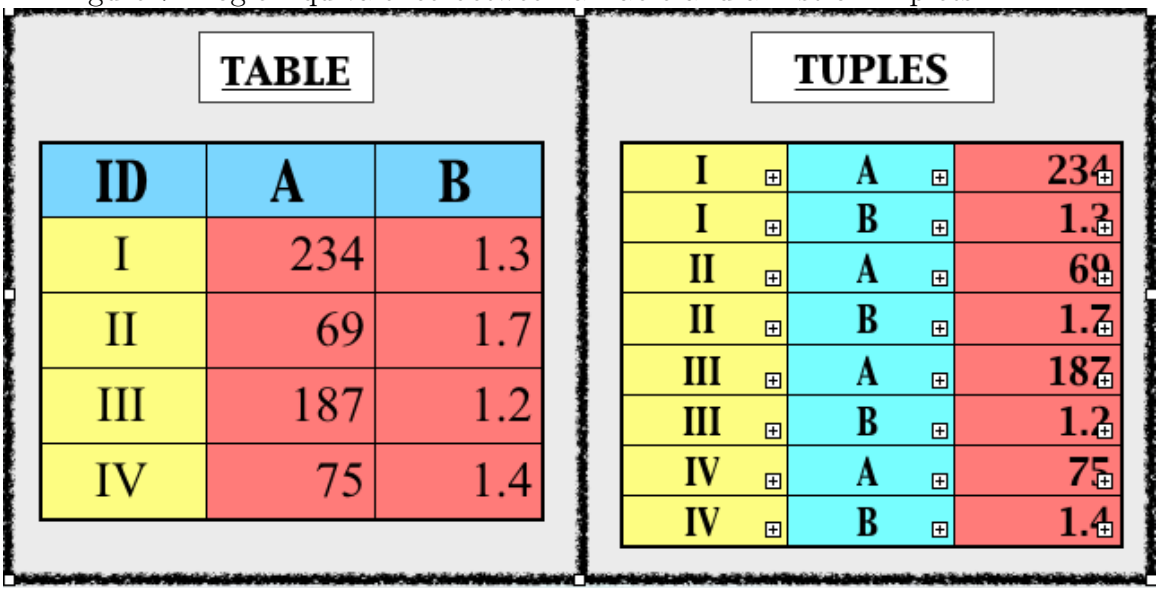
2.3 Environmental Primary Data

Environmental primary data is of fundamental importance for environmental scientists and managers to prove theories, model environmental variables, or provide statistics to governments and people . Recently, it has been shown how the scientific community benefits from sharing primary data [181]. This emphasis in data sharing has nurtured numerous projects to accomplish this general goal, which has not been resolved given the epistemological differences among providers of primary data [232].

By looking at the physical level of the environmental files with primary data, the classical problem of sharing data among scientists and managers by having a multitude of different formats appears inevitably. This problem gets exarcebated in the environmental community given the numer of different disciplines that come together to work in environmental projects [174]. Therefore, searching for a solution for this problem is more than ever a peremptory necessity.

At the logical level, environmental primary data often times comes in tables , series of data, or singular values and the logical structure of primary data can be reduced to a very simple model where where any observation recorded can be translated into a triplet where the first member of the triplet represents a reference to the subject under observation, the second member of the triplet represents the feature being observed, and the last member of the triplet represents the value being recorded. Therefore, any environmental file with primary data can be logically represented as a set of triplets as we can see in Figure (7) where a classical tabular file has been transformed in a non-ordered list of triplets. This simple logical model is general enough to record common primary data. For instance, an image will be a single tuple being the value the array of pixels, a decadal file will be a list of tuples having as values the single arrays, and a polygon map will be a list whose values will be a list of two dimensional irregular arrays.

Figure 7: Logic Equivalence between a Table and a List of Triplets

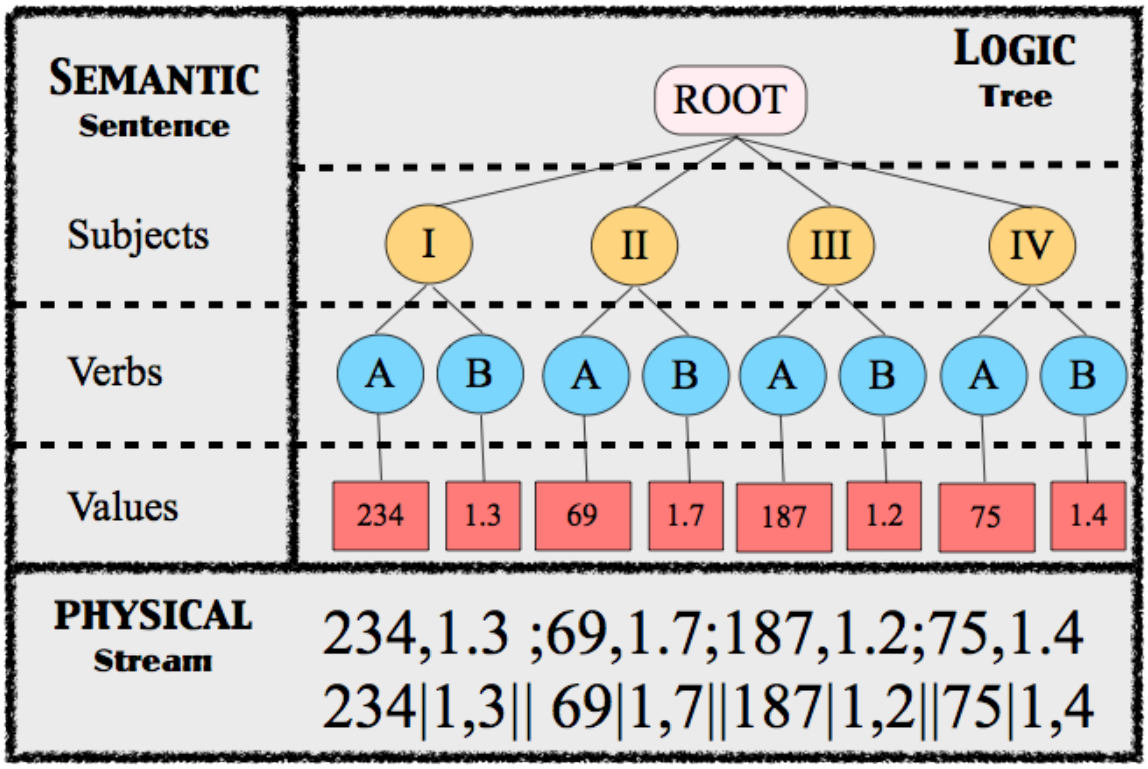


At the semantic level, each of the members of the triplets belongs to disjoint conceptual universes. The first members of the triplets belong to the universe of references to things, the second members of the triplets belong to the realm of verbs, and the third members of the triplets belong to the universe of the metrology. The semantic description of these files can be simplified even further since the subjects of the file often belongs to the same kind; therefore, just one annotation should be enough to describe all the subjects. In addition, primary data cannot be logically reified because of its ontological essence: it is the first time being written. Therefore, common reification conceptual problems are not in the semantic scope of this semantic model as have been shown in other more complex semantic models [217]. Furthermore, a very human and common semantic classification regards triplets as a sentence. The first member of the triplet is the subject, the second member is the verb and the third member is the object [35]. However, given the ontological essence of the primary data, it is more appropriated to refer to the members of the triplet as subject , verb, and value.

Summarizing, any data-centric document with primary data has a very common and defined structures at the logical and semantical levels. It is at the physical level where the differences become a source of problems since the stream of data is often encoded differently as illustrated in Figure (8).

Therefore, the challenge to describe data-centric files becomes more arduous at the lexical level since there are more varieties of structures while the logic and semantic levels are less complicated, and based in very well known and common standard structures: trees and simple sentences.

Figure 8: Physical, Logical, and Semantical View of a Data Stream



2.4 Metadata about Data-centric Documents.

Metadata is defined as being data about data [34]and can be used to define the information about a given dataset that helps to better interpret the information contained in it. Semiotically, the bare numbers and identifiers inside a dataset are naked signs that need extra information for being understood:the metadata.

Data-centric documents have been also described using metadata languages from two different perspectives: archival and sharing purposes.

In the past, metadata has been used commonly to collect bibliographical information about data-centric documents with archival purposes to facilitate their retrieval. These metadata languages have been developed by the bibliographical community, and have the capacity to collect only bibliographical information such as authorship, creation-date, and keywords. Therefore, the data itself inside the data-centric document remains undefined when using this type of metadata languages.

Recently and with the advent of the Web, sharing data between researchers and managers has become a prevalent activity that has pushed the development of metadata languages which have facilities to describe not only the bibliographical information, but also information about the data inside the archive. This has been an important step for the environmental community; e.g., the EML has been one of the cornerstones of the Long Term Ecological Research project [212].

Syntactically, metadata files have been defined using text-files to facilitate their reading and editing. During the last decade the metadata information has been collected in files with XML format . For instance, the geographers have been a very active discipline developing metadata schemas since the beginning of the Web using tex-based schemas [61], and nowadays the geospatial metadata is coded in XML-schemas [13].

Since metadata about data-centric documents can be classified as primary data about data, the same logical model with triplets described to represent environmental data can be used to render metadata in the memory of our computers [168]. This has been the case for the W3C promoting the use of RDF to represent metadata information. The use of RDF-metadata has been very popular to syndicate information via automatic systems [151], in medicinal applications [73], in online news [123]and many other fields.

Because of the essence of metadata, the semantic level is the most important aspect of metadata languages. As it was the case for environmental archives, three disjoint conceptual spaces can clearly be defined, but the ontological space for the object member of the metadata sentences is quite larger and more diverse. The members of the W3C and people from the research community have developed formal ontologies using controlled vocabularies and simple logics to formalize this complex semantic space[36].

Syndication of information has been the main incentive for the growth of metadata languages such as RSS, RDF or ATOM. However, the syndication

of data-centric documents has not been as prevalent as the syndication of text-based documents such as news, papers, and blogs even though from inside the scientific community there have been voices to use these popular syndication models [114].

Syndication of data and formalization of semantics is pushing metadata from the card catalogs of the libraries to the semantic web of data using the mechanisms of the Linked Data project. This open project is creating a new open Web for sharing based on independent, non-overlapping, and related projects that offer data, metadata, controlled vocabularies, and logic services [35]. Given the fact that environmental scientists and managers have terabytes of data in legacy formats that can populate the open web data, the quest for extensions to metadata languages for improving the description of data-centric documents is highly needed [147].

2.4.1 Archival Metadata

Under this category we can aggregate a number of metadata languages that have been used to archive data-oriented archives. None of this languages have the ability to directly describe the data of a data-oriented document. They have been designed following the spirit from metadata languages used to classify and catalogue bibliographical resources. For instance, the community working with geographical information systems has developed several standard metadata languages such as the Content Standard for Digital Geospatial Metadata (CSDGM) , or the ISO 19115 which do not offer information to read and understand the data from the geographical files. For instance, the CSDGM will weakly describe the format of the archive by referencing it with a popular name such as "Shapefile". Therefore, any person or robot will have to know the specifications of this file to read its contents since the metadata offered by the CSDGM does not provide information for reading the data from the archive stored as a Shapefile.

2.4.2 Ecological Metadata Language (EML)

The EML was conceived for the members of the Biocomplexity Network as a main tool for the Long Term Ecological Research project to facilitate collaboration among researchers and perform comparative analysis. This metadata

language is well alive and used by the researchers from the biocomplexity network, and is open to further collaborations. However, this metadata language has a very limited capacity to describe data-centric documents extensionally; therefore, most of the files used by environmental researchers must be defined intensionally by naming the name of the format for the file. For instance, a comma-separated-value-file will be defined explicitly but a tiff-image will be defined implicitly because EML has not enough expression power to describe files with complex formats. Therefore, this metadata language needs extensions to explicitly describe data-oriented document in order to be properly called a metadata language for environmental data-oriented documents.

2.4.3 Earth Science Markup Language (ESML)

The metadata ESML was created with similar intentions to the EML, however, since 2006 the language has had no support and little documentation is available. This language will have more expressiveness than the EML based on the scarce available documentation , however, it will not be able to describe a number of common formats used by environmentalists such as tiff-images and decadal files.

2.4.4 Data Format Description Language (DFDL)

DFDL is actively developed by the Open Grid Forum with the goal to parse data files into a Document Object Mode. DFDL is already being used today in commercial systems such as IBM's WebSphere Business Integrator Message Broker and Microsoft's BizTalk flat file. As the previous metadata languages, DFDL lacks not only semantics to encode important files used by environmentalists such as tiff-images and decadal archives, but also to attach any semantic information which is of pivotal importance for data exchange in environmental projects.

2.4.5 PADS

PADS is a domain domain-specific language for processing ad-hoc data. PADS is an active language developed collaboratively by researchers at Princeton University, AT&T and Google. Unfortunately, it often takes substantial

time and expertise to write a PADS description since PADS is a complex low level data language only amenable for computer scientists. The researchers developing PADS have created a tool to automatically infer PAD description from data files which alleviates the burden of writing PAD description files from scratch[88]. PADS, using a recursive descent algorithm can encode and describe any continuous structure, array, or trees structure. However it will not be able to encode important environmental formats such tiff-images or zip archives which contain non-sequential data and used pointing structures.

2.4.6 Extensible Scientific Interchange Language (XSIL)

The XSIL is a metadata language which can encode text and binary files, however the encoding capacity of the language is limited; several common environmental archives used by researchers and managers cannot be decode with this metadata language. XSIL was developed at CalTech, and lacks support and documentation.

2.4.7 Binary Format Description (BFD)

The BFD is a superset of the XSIL by providing more control structures. The BFD was developed by researchers working at the Pacific Northwest National Laboratory from the US Department of Energy [177]. The BFD has not been developed fully and lacks publications and support. The BFD cannot encode common environmental data files such as tiff-images or decadal archives.

2.5 Metadata Comes after Data

As a matter of fact, metadata should come after data because if metadata is data about data, metadata cannot exist before the data is laid out in a file. This tautological information, and perhaps because of its obviousness, has not been used in any metadata research to date[2]. Nonetheless, the usage of this information can relax the problem of creating metadata descriptors and its complexity to more basic levels.

[2]By searching the Web using common search engines the expression "metadata comes after data" has not being used in any document with its temporal meaning.

An evident example of relaxing the metadata description requirements is the case of tiff-images. The specification of tiff-images are rather involved since the image data does not need to be contiguous and could be divided into several non-sequential strides. However, in most of the tiff files, the image data is laid out as a sequential array of pixel-data; therefore parsing the pixel-data of a tiff-image file can be achieved using a trivial regular expression which can be further processed to extract pixel level information. This easy parsing process contrasts with the complex parsing requirements described in the tiff specifications.

Most of computer scientists who have been trained in formal grammars, languages and parsing structured documents think of the parsing process following an "a priori" schema in which they create a machine able to recognize any structured document that belongs to the set defined by a formal grammar prior to existence of the data files. Therefore, we can see this "a priori" metadata descriptor information as a classical decoder which serves as a template to process all archives of the same format since they share the same formal grammar rules.

However, in metadata terms, only semantically grounded archives can be described before the data file is defined and most data files have to be described after the data file is written. Therefore, a classical "a priori" parsing schema mimicking the rules of the specification of the format of the file has a limited interest for just describing environmental datasets with simple physical structures with ground semantics such as decadal files.

Moreover, all data-centric files can be described "a posteriori" regardless of their complexity; this "a posteriori" approach also permits us to define different views of the datasets which is a good characteristic since data providers can offer different views of the same dataset to different people based on their requirements while a classical "a priori" decoder will describe the full dataset.

2.6 Proof of Concept for 'A Posteriori' Metadata to Describe Primary Data

Primary data in environmental projects correspond to values that have been observed on subjects to describe their features; thereafer, this information is recorded onto magnetic physical media for preservation and ulterior usage. It has been a common practice to store the data in files with common formats,

i.e., formal grammars that define particular languages. However, there is not a common grammar for datacentric documents, and many grammars have been used to define the different file formats that are commonly used today.

Since there is not a common language to 'a priori' describe the data from all data files, but we have tools to explore "a posteriori" the structure of the information of these common data files, it is possible to describe 'a posteriori' the data inside the file using a simple language and processed with a machine that will map the data items inside the physical file to the slots of the logical object.

The data items inside the stream lay one after another clearly defined by its initial position and length, and at the same time, the logical data lay in a transparent indexed list of triplets. Therefore, as can be seen in figure(9), the injective mapping between the pysical stream and the logical object is well-defined and unique, i.e., for every item in physical the stream a number of fixed relations are mapped into the logical object. By 'a posteriori' analyzing the physical files, the set of relations that define the mapping can be expresed in an extended context-free language that can be used to generate the machine that can generate the logical object from the physical file.

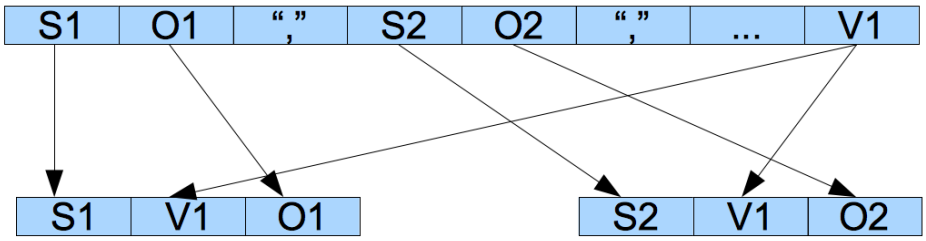


Figure 9: Injective Mapping between the Physical Stream and Logical Object

Every data item inside a physical stream can be positioned using a pair that defines its offset and length. Many times, the data items of interest lay down sequentially one after another without stream-space between items; therefore, only the length of the item is the required information to correctly position the item inside the stream. Therefore, there are two basic systems to reference the position of the data items:

- Absolute positioning of the container where the offset and length have to be defined,and

- Relative positioning of the container where ony the length of the item has to be defined.

As seen in the previous sections in figure(8), the data items inside a physical file can be organized using a containment relationship. For example, the physical values of the cell of one row are inside a physical container in the stream, and the physical container that defines the table in the stream contains an orderly sequence of containers where each one contains one row of values from the table. Therefore, this hierarchical relationship of containment can be describe using the productions of a deterministic context-free grammar. However, a simple extension has to be made to a context-free grammar to handlle the injective mapping to create a deterministic push-down transducer that can effectively translate the physical stream into a logical object. This mapping has to be declared at the end of a production rule of the context-free grammar that produces a single value as a fourplet where the first element of the fourplet gives the indexing value of the list for the other three values that sequentially rerpresent the subject, verb and object of a logical sentence as shown in algorithm(1).

The right-hand side of the rules of these container grammars, are absolute or relative containers. Absolute containers, are delcared between braquets and used the positional information to defined the stream of the container while relative containers will extract the stream of the container given the end-position of the previous container and its length.

The containers in the left-hand side of the rules contain the stream that the right-hand side containers are going to parsed. Unlike regular context-free grammars, it is possible to have several containers on the left-hand side of the rule whose streams are going to be concatenated before being processed by the containers of the right-hand side of the rule. This last requirement, even though is not conceptually necessary, simplifies the declaration of the container grammar because the fomal specifications of many file foramts allow to break the stream of data into non-sequential streams inside the archive[2].

Algorithm 1 Container Grammar Definition

$$Container\ Grammar = (C\ , R\ , F\ , S)\ where :$$

$$C\ is\ a\ set\ of\ cotainers$$

$$R\ is\ a\ set\ of\ productions\ rules$$

$$F\ is\ an\ injective\ function$$

$$S\ is\ the\ stream\ of\ the\ data\ file$$

$$R = \{R_1, R_2, \ldots, R_n\}$$

$$\forall\ R\ \sum |C| \Rightarrow \sum C \Rrightarrow F(ID)[SUBJECT\ , VERB\ , VALUE]$$

$$\forall\ absolute\ container\ C_i \equiv (offset, length, ending, parent)\ where :$$

$$ending = offset + length$$

$$parent = parent\ container$$

$$\forall\ relative\ container\ \ < C_i > \equiv (offset, length, ending, parent)\ where :$$

$$offset = C_{i-1}[ending] + 1$$

$$ending = offset + length$$

$$parent = parent\ container$$

$$\forall\ container\ C\ |C| = C[parent][C[offset] \leftrightarrow C[ending]$$

For instance, the data presented in figure(9) can be defined using the container grammar from algorithm(2) where the hierarchical processing of the containers finalize with the declaration of the injective between the physical and logical spaces. As it can be seen in algorithm(3), the algoritm to define the mapping is not unique and different set of production rules can be used to define the injective operation having the same logical object as a result; ergo, the rules from algoritm(3) will also parse the file into the same logical object.

Therefore, regardless of the complexity of the specifications of any physical file, if there is an ad-hoc tool to explore the physical structure of the information on the file, a simple 'a posteriori' description of the data from the file can be generated by 'a posteriori' processing the given file at hand, and develop the metadata needed to parse its information. Moreover, given the regularities of some files, the 'a posteriori' description of files can be concise an susccint since few production rules are needed to describe them.

Algorithm 2 Example of Container Grammar Using Relative Positioning

$$S \Rightarrow DATA + VERB$$

$$|DATA| \Rightarrow \sum_{ID=1}^{ID=n} DATUM$$

$$|DATUM| \Rightarrow SUBJECT{+}VALUE \Rrightarrow \digamma(ID)[SUBJECT, VERB, VALUE]$$

Algorithm 3 Example of Container Grammar Using Absolute Positioning

$$S \Rightarrow < VERB > + < DATA >$$

$$|DATA| \Rightarrow \sum_{ID=1}^{ID=n} < DATUM >$$

$$|DATUM| \Rightarrow < VALUE > + < SUBJECT > \Rrightarrow F(ID)[SUBJECT, VERB, VALUE]$$

2.7 Hierarchical Engine for Metadata Processing: HEMP

Environmental files with primary data haven been written following different rules, and unfortunately, it will be impossible to develop a classical push-down automata that could process all the files because some of these files have structures that cannot be expressed following the rules of context-free grammars. Therefore, an extended pragmatic approach to describe the physical structure of the primary data from the different environmental files is an obvious necessity. However, due to the regularity of the physical data of archives with primary data, and the hierarchical nature of the logical structure of the environmental files, the problem of describing a file structure and its primary data can be resolved using an extended deterministic hierarchical scanner made of regular expressions which can transform a stream of bytes from the environmental archive into a logical object annotated with semantic values to fully qualify the primary data from the environmental archive.

2.7.1 Hierarchical Scanner: Lexical Description

To represent the lexical structure of data-centric files, the standard metasyntax notations like the Extended Backus-Naur form (EBNF) used to describe context-free grammars do not have enough facilities to describe the lexical structure of some important file formats commonly used by environmentalists since this metasyntax cannot represent for instance, pointing structures, backtracking and concatenation of streams. Therefore the expressive power

of these types of metasyntaxes have to be extended to reflect and operate in the the lexical structure of these files for being able to synoptically represent the lexical manipulation of the stream of the archive.

2.7.1.1 Pointers Pointers and pointer notation are common concepts from compiled programming languages where one structure in memory has information that points to another data structure and provided with the information needed to read into memory this last structure when the pointer is dereferenced. In some common data formats a conceptually similar structure is used to represent the logic of the lexical operations. For instance, it is common to see in binary files that before reading the data structure of interest as a contiguous block of bytes, first, a pointing structure whose position and structure is known has to be read because it contains the offset and length of the block of interest. Therefore, one structure is pointing to the position of the other. Conceptually a non-terminal symbol from the metasyntax should have a double structure: one that represents a block of data and a second structure that could represent a block being pointed. Therefore, and following the custom notation of pointers in languages such as C or C++ , the presence or lack of an asterisk prefixing a non-terminal symbol represents the semantics of the pointer structure in the lexical descriptions in this thesis.

2.7.1.2 Parallel Vs Sequential Scanning It is common among environmental archives that physical data structures follow a sequential order predefined in the specification of the format for these environmental archives. This feature is a necessary condition for having an amenable definition of the grammar following the rules of a context free grammar. Otherwise, if the specification of the format does not enforce a sequential order for the data structures, a rule from a context-free grammar that would have n productions in its right side, it would have to have n-factorial rules to be able to handle any of the possible sequential orders for the productions which would complicate tremendously the declaration of the grammar and the parsing of any stream belonging to the language defined by this type of grammar.

To relax the definition of our scanner, a parallel processing of determined rules where there is not a predefined sequence of the productions is desirable. This means that the stream of the left non-terminal symbol will have to be scanned more than one time as needed by moving the index of the scanner to

the beginning of the stream. This feature departs from the classical scanning techniques where the index of the scanner moves only forward after a match has been encountered [91]. In this case the reading head of the scanner has to be moved back to the beginning of the stream any time a match has been encountered.

2.7.1.3 Stream Concatenation Having a single non-terminal symbol in the left hand side of a production rule is a basic characteristic of context-free grammars due to the fact that there is a physical hierarchy of streams where the stream at the root is splinted successively at the different levels of the hierarchical tree into smaller streams until the last level where the terminal stream cannot further being divided. For some non-contiguous files such as tiff files and pdf files, the physical structure of the data could be chopped into several blocks with no specific order among blocks and to generate the contiguous physical block of data for further processing the different blocks have to be concatenated following a determined order. Therefore, meta-syntaxes for data-centric documents have to have the facilities to represent concatenation rules. Therefore, more than one non-terminal symbol could appear in the right left side of the production rule. As it is well known, context-free languages only allow one non-terminal symbol in the left side of its production rules [12].

2.7.1.4 Lexical Description Examples of Common Data Files

2.7.1.4.1 DBF database: Lexical Description "A Posteriori" DBF archives have been used to store tabular data of any kind and have been a very popular database format for many applications. A DBF archive has a header with metadata and the rest contains the data in a physical table where each row represents a record, each column represents a feature that has been measured and the intersection of a record and a feature represents the registered value

In addition, a DBF file is non-grounded and the information of the file cannot be described until the file has been created, and is a perfect candidate format for being described "a posteriori". In the example in Listing (1), since the length and position of the DATA, the length of RECORD, and the length and relative position of CYTY, STATE, AND POPULATION are known before

the description of the file the lexical description of the file can be attained using the regular facilities of the EBNF.

Listing 1: DBF Lexical Description "A Posteriori"

```
ROOT      :=  TRASH DATA
DATA      :=  RECORD+
RECORD    :=  CITY  STATE TRASH POPULATION  TRASH
```

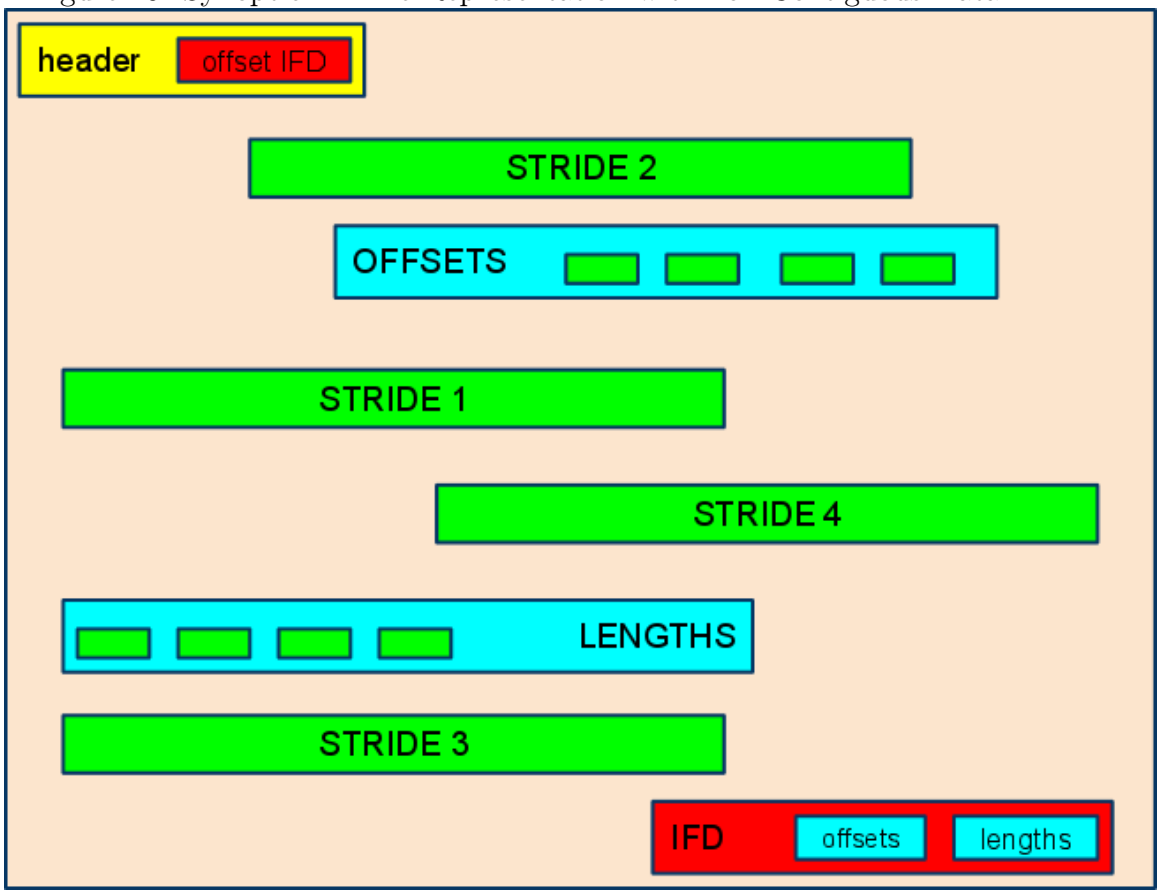
2.7.1.4.2 Tiff Image: Lexical Descriptions Tiff-files following the Tiff-specification have been used to archive image information, and it is a very popular format to exchange imagery among environmental researchers and managers. The complete specification goes beyond the description capabilities of context-free languages.

2.7.1.4.3 Tiff "A Priori" The physical layout of image data in Tiff-file does not follow a hierarchical structure since its specification uses pointer structures to address this lack of physical contiguity as can be seen in Figure (11). Therefore we have to extend the expressivity of our metasyntax language by using pointers and stream concatenation as illustrated in Listing (2) to get the physical data of the IMAGE. This lexical description will work for all files.

Listing 2: Tiff-image Lexical Description "a priori"

```
ROOT              :=    HEADER  IFDOFFSET        IFD
*IFDD             :=    IFS
IFS               :=    WIDHT | LENGTH | ROW | BYTE
*(*ROW *BYTE)+    :=    IMAGE
IMAGE             :=    ROW+
ROW               :=    PIXEL+
PIXEL             :=    R  G  B
```

Figure 10: Synoptic Tiff file Representation with non-Contiguous Data



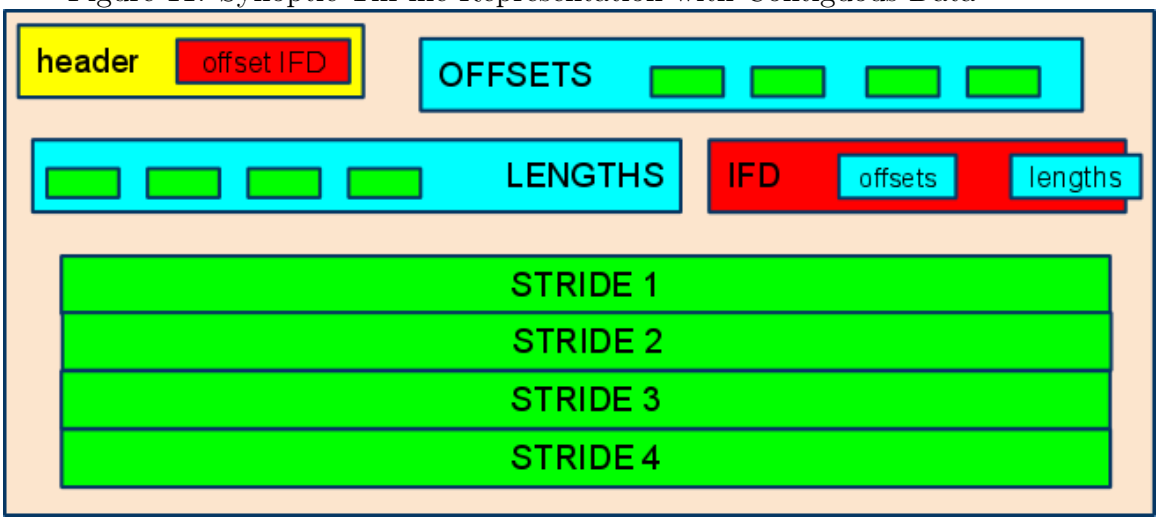
2.7.1.4.4 Tiff "A Posteriori" Tiff-images almost always have the physical data for the image in a single contiguous block (see Figure (11)) and by describing the image "a posteriori" the lexical description of the physical data becomes quite simple and can be represented using the EBNF notation as can be seen in Listing (3). In this case, the offset and length of IMAGE, and the length of the ROW will be known before developing the lexical description of the file. Therefore, this lexical descriptor will only work for just the tiff-image described.

Listing 3: Tiff-Image: Lexical Description "a posteriori"

```
Contiguous Block of Data
```

```
ROOT        :=  IMAGE
IMAGE       :=  ROW+
ROW         :=  PIXEL+
PIXEL       :=  R  G  B
```

Figure 11: Synoptic Tiff file Representation with Contiguous Data



If the physical data of the image in a tiff file is chopped into different blocks, also called strides, the "a posteriori" approach also facilitates the process of lexical description of the physical data. As shown in Listing (4), in this case the concatenation of the blocks to generate the physical data of the image is required. The offset and length of the strides can be read before to write the lexical description of the image.

Listing 4: Tiff-Image: Lexical Description "A Posteriori"

```
Non-contiguous Block of Data

ROOT          :=  STRIDE1  STRIDE2 STRIDE 3
STRIDE1 + STRIDE 2 + STRIDE 3    :=        IMAGE
IMAGE         :=  ROW+
ROW           :=  PIXEL+
PIXEL         :=  R  G  B
```

2.7.1.4.5 Decadal File: Lexical Description "A Priori" The decadal file is a contiguous, and semantically grounded file used to archive dendrochronological series. Decadal files are a clear example of simple text-based files used by members of an environmental discipline to store their primary data. Because of their characteristics, the lexical description form Listing (5)can be used to describe the physical data of all decadal files.

Listing 5: Lexical Description Decadal File

```
ROOT        :=    SERIE+
SERIE       :=    FIRSTLINE  LINE+
FIRSTLINE   :=    CODE  FIRSTYEAR  DECADE
LINE        :=    TRASH  TRASH  DECADE
DECADE      :=    GROWTH+
```

2.7.1.4.6 ESRI 2D Polygon Shapefile: Lexical Description An ESRI shapefile is an archive that contains three files:

- Main file: it is the file that contains a sequence of records as geographical shapes of the same kind (2D-Point, 3D-Point, 2d-polzyong, etc)
- DBF file: is the file that contains the feature attributes for each record in the previous file
- Index file is the file that contain information to accelerate the reading of the records in the main file.

The main file of the shapefile is a contiguous grounded file since its semantic information is attached in the description of the file and the records are sequentially stored in a single block; therefore, it is an obvious candidate for a template that can describe the file "a priori". However, as it has been previously mentioned, a DBF file only can be described "a posteriori".

The Shapefiles have been the most popular archive to work with geospatial data to represent the features of objects that occupy any space over the surface of the earth such forest areas, urban areas, etc. The physical data of the main file can be described using a EBNF notation as shown in Listing (6). This lexical description can describe the physical data of all files of this kind.

Listing 6: 2D-Polygon Shape Lexical Description "A Priori"

```
ROOT          :=  HEADER BODY
BODY          :=  (RECORD_HEAD  RECORD)+
RECORD_HEAD   :=  RECORD_ID
RECORD        :=  SHAPEFILE  BOX  NUMRINGS  RING+
RING          :=  POINT+
POINT         :=  X          Y
```

2.7.2 My Scientific Object Notation (MySON): Logical Structure

There are a number of standard models to represent documents logically, and the Document Object Model (DOM) is the most popular of them since it has been developed to store XML documents over the Internet; however, the DOM is a general hierarchical model that can store every document that has being encoded following the XML specifications since the DOM and the XML have the same origins. It is, therefore, a general model that will need a schema document to define any formal data structure.

Along with the XML, the RDF model has become an important model used on the modern Web. The RDF model is a network model that is oriented to linked data using triplets, where an element from a given domain and range values are linked through properties. The basic triplets of this logical model resemble the reviewed logical structures of primary data, since the domain class can represent the subject of the triplets for primary data, the range class can represent the values, and the properties can represent the verbs.

MySON represents a specialized subset of the RDF document object model to represent primary data that should contain all the information needed to understand its data. Therefore, the object can be exchange among researchers or syndicated in the Web of open data since its contents are completely defined using the principles of the Linked Data project .

Moreover, archives with primary data have a peculiar logical and semantic structures, and MySON objects should be tailored to fix their logical structure that also implies the semantic structure of the triplets: [subject, verb, value] (see Figure (8)).

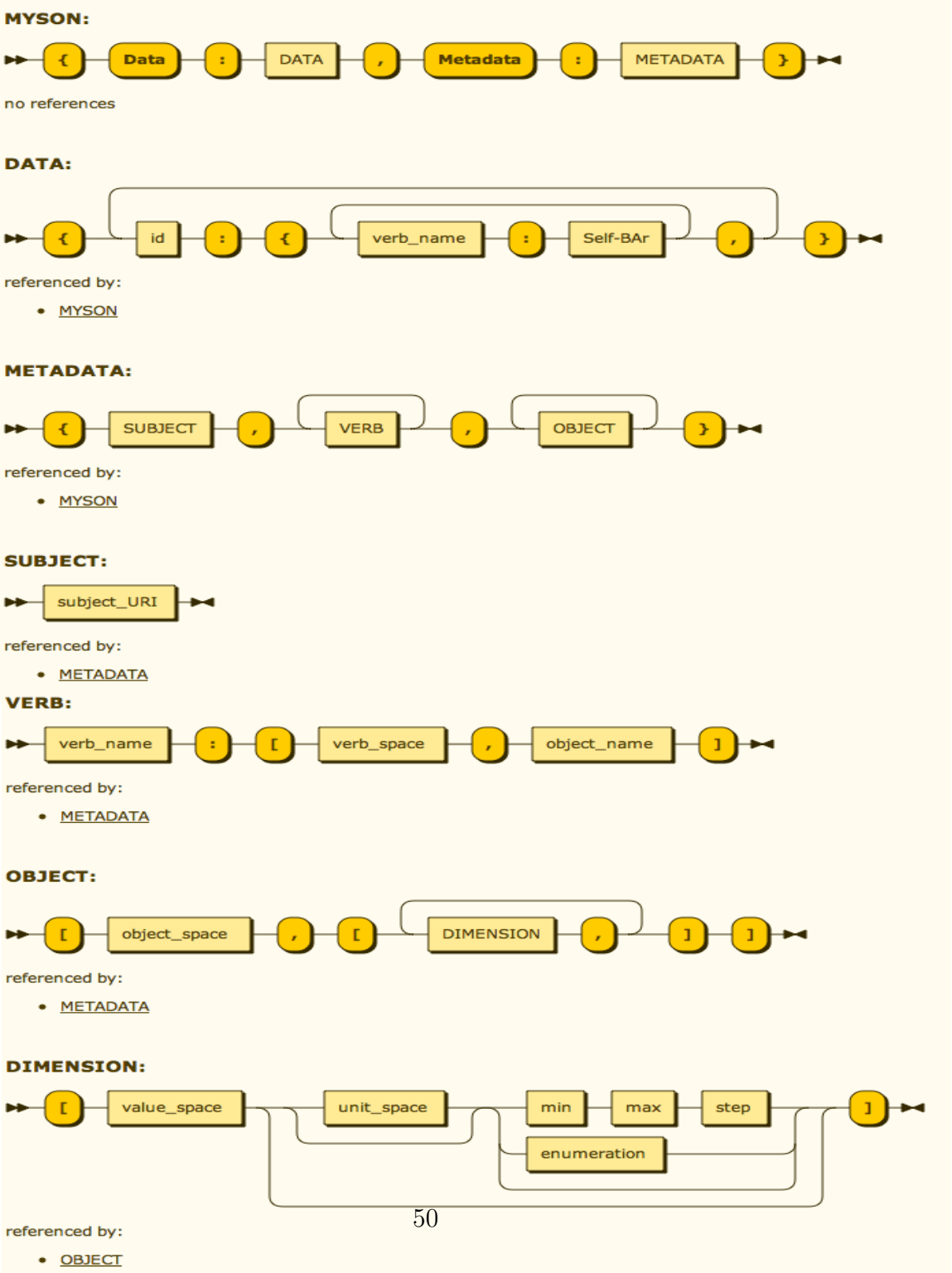
2.7.2.1 My Scientific Object Notation: MySON MySON objects can be defined as a hierarchical tree with four levels of height corresponding

to the document and three members of the triplets. These tree is modeled using dictionaries, which is a common data structure in modern computer languages, as can be seen in Listing (7), and Figure (12). The first level of the tree represents the data-centric document with primary data, the second level of the tree represents the references to the actual things or subjects of observation, the third level of the tree represents the properties that have being observed, and the forth level of the tree represents the values that have been registered.

Listing 7: Lexical Description of MySON object

```
MYSON     ::= "{" "Data" ":" DATA "," "Metadata" ":" METADATA "}"
DATA      ::= "{"( id ":" "{"(verb_name ":" Self-BAr)+ "," )+"}"
METADATA  ::= "{" SUBJECT "," VERB+ "," OBJECT+ "}"
SUBJECT   ::= subject_URI
VERB      ::= verb_name ":" "[" verb_space "," object_name "]"
OBJECT    ::= "[" object_space "," "["(DIMENSION ",")+ "]" "]"
DIMENSION ::= "[" value_space (unit_space?((min max step)|
                                          enumeration)?)?"]"
```

Figure 12: Railroad Diagram of MySON Object



2.7.2.2 Serialization of MySON Objects Serialization of logical objects has been a common task, and common routine to either save the state of a software application on the hard drive or to exchange objects among people. For instance, most of the modern programming languages have facilities to dump and load logical objects from the hard drive, and lately private Internet platforms such as Google [45] of FaceBook [220] have developed methods to serialize objects than can be exchanged between users in the network to send back and forth messages with objects between their applications.

The facilities that modern languages offer to serialize data structures are language-dependent; for instance, the Python programming language offers two methods to dump logical objects into a stream and load the stream back into a logical object [159]; however, theses methods do not apply when using other programming languages since the serialized stream is only compatible with facilities of one specific programming language, i.e., other computer language does not have routines to read objects that have been serialized with Python routines.

The serializing alternatives offered by Google, called Protocol Buffers, and FaceBook, called Thrift, to bring logical objects into streams have not only clear incompatibilities between services based on both technologies since neither of them can exchange a message serialized with the services of the other company , but also a complicated and verbose interface definition language that resemble the facilities offered by low-level programming languages such as C and C++.

On the other hand, JSON is a recent very popular notation, and with ample support since it uses lexical structures commonly used in modern languages to define logical objects; therefore, it makes sense since the mapping between logical and lexical objects is easy and very well understood [67]. On the other side, JSON lacks a mechanism to describe classes or types like in the aforementioned services from Google and FaceBook.

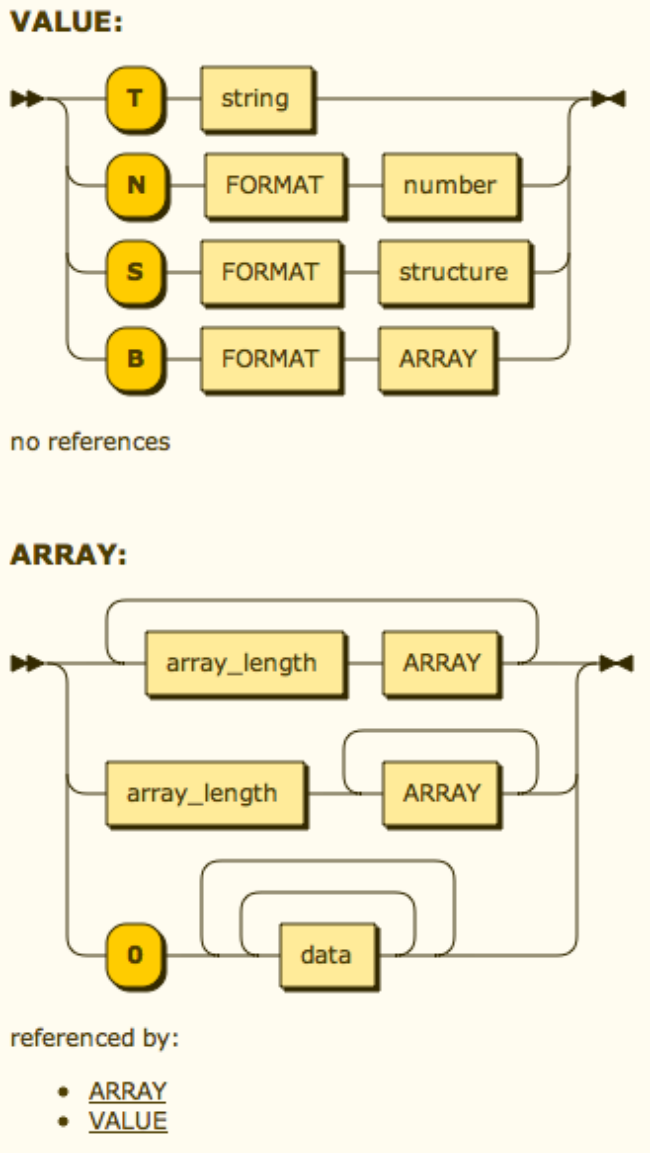
Since the logical structure of MySON objects is fixed and defined by the triplet structure, there is not a necessity for a metalanguage to define its structure like in Protocol Buffers and Thrift ; however, the array structure of the value member of the triplet needs a definition since it can and should vary from triplet to triplet. Therefore, the MySON object uses a self-defined binary array to serialize the values based on the grammar shown on Listing(8) and Figure(13). The Self-describing Binary Array (Self-BAr) structure can

be used to serialize all common values used in data-centric documents such as single values, arrays, multidimensional irregular arrays and multidimensional regular arrays. Therefore, there is not a need for a metadata language to define the structure of MySON objects because they are self-defining objects. For any level of the array structure the array can be regular or irregular; if the array is regular (all the members of the array have the same length) only one array_length value is needed to define the members of the array; if the array is irregular one array_length per member will be needed to define the members of the array; and if the array is a terminal element, the header has the information to decode it.

Listing 8: Lexical description of the Value object for MySON

```
VALUE ::= ('T' string) |
          ('N' FORMAT number) |
          ('S' FORMAT structure) |
          ( 'B' FORMAT ARRAY)
ARRAY ::=( (array_length  ARRAY)+|array_length ARRAY+) | '0' (data+)+
```

Figure 13: Railroad Diagram for the Self-BAr Value object of MySON



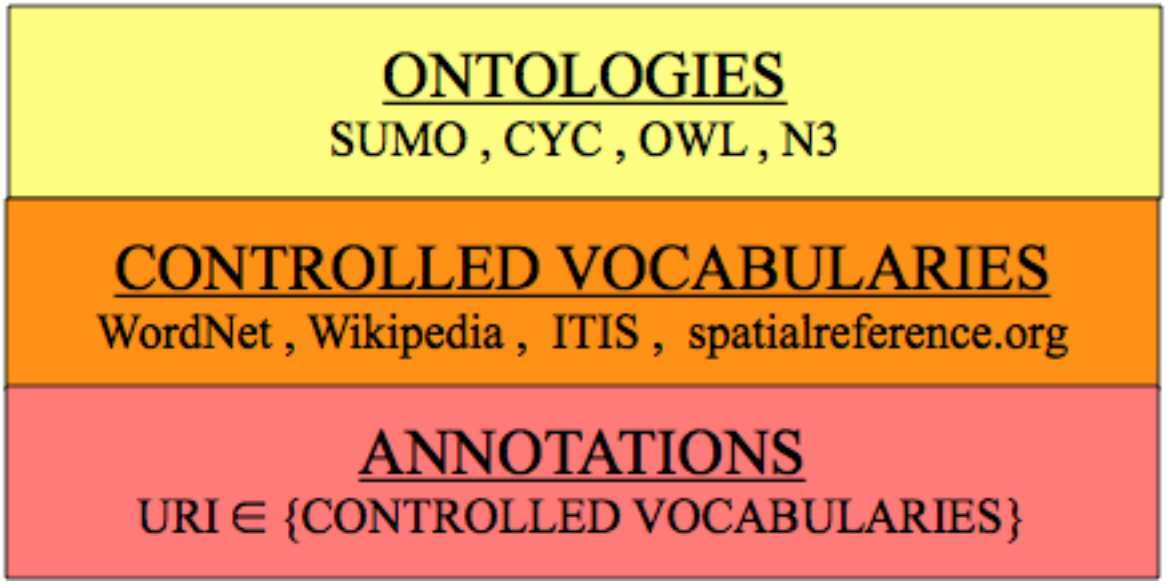
2.7.3 Controlled Vocabularies: Semantic Mapping

Semantic mapping is of vital importance since it is the information that linked our data with the rest of the Web as seen in Figure figure 14. It can be conceptualized as three layers with distinctive characteristics where for

every pair of sequent layers the layer above is functional related with the layer just below. The bottom layer is the space of URIs that link the data from the data-centric documents with terms from controlled vocabularies, i.e., the three terms from a logical triplet representing an observation are linked to terms in controlled vocabularies. Similarly, the ontologies terms from the controlled vocabularies might be linked to one or many ontologies.

The annotations from bottom layer have to be created by the data producers and linked to their data. Therefore, the basic and simple task for a data provider should be to annotate their data with terms from standard and common dictionaries. This should not be the complex, time-consuming, and cost-expensive task that many authors have reported as one of the main causes for not creating metadata. On the contrary, this semantic task is as easy as connecting the two terms with a single URI.

Figure 14: Semantic Layering



The controlled vocabularies are the middle man between the data producers and the ontologists. However not every dictionary on the Internet can be regarded as a controlled vocabulary since the dictionary has to have persistence, need to be peer-review, and have a RESTfull architecture [200]. Several Controlled vocabularies can be reached on the internet which have

had persistence, community support and their terms can be accessed using URIs. Some of these vocabularies are of general usage such as WordNet and Wikipedia containing human semantic descriptions for nouns verbs and objects. Other vocabularies, like ITIS, only offer domain-specific terminology; e.g., ITIS is an integrated standard taxonomy for living organisms, and all its terms (groups) are reachable using URIs. Therefore, these terms can be easily used as values in sentences where they are required. Another important example of a domain-specific vocabulary to georeference spatial systems is the website spatialreference.org which contain a large collection of reference systems reachable through their URIs.

The lack of persistence of resources on the internet is a common malady that has been defined as link-rot problem, and has plagued every corner of the Internet. It is common on the Internet to find URIs that point to nowhere. Therefore controlled vocabularies used for metadata definitions have to have a recognized and ample social recognition.

Being Peer-reviewed is a certification of quality that is a necessary requirement for any publication. However peer-reviewing is also evolving along with the Internet, and there are now flexible models to certify the quality of documents published in the Internet. Therefore, quality of articles and definitions posted in collaborative space can be computed without the formality of the classical approach.

A RESTfull website has a friendly interface to integrate data in the web of open data under the Linked Data project [74]. This architecture basically offers a clear interface for accessing resources in the Internet via URI [200].

The top layer of the semantic cake is occupied by the formal ontologies which provide reasoning power to the semantic web. They are used in expert systems and are of little interest for data providers since they are not in direct contact with this layer. SUMO and CYC are upper-ontologies commonly used in expert systems since they offer advanced reasoning capabilities; on the other side, N3 and OWL are languages with a very limited subset of logical functions; therefore, the reasoning power of the former ontologies is greater than the reasoning power of the second ones later ones. However, N3 and OWL are ontologies that have been designed and promoted by the members of the W3C, their structure are oriented to handle the needs of the Semantic Web.

Good semantic accuracy is a fundamentally important aspect of semantic

annotation of data. As in any conceptual mapping, the conceptual disparity between the concept linked to the data, and the concept in the controlled vocabularies should be minimum [106]; e.g., if the subjects of the sentences of the data-centric document with primary data are white horses, and the concept of white horse is linked to a the horse entry in a controlled vocabulary, because the dictionary that has been used only has the general concept of a horse, the conceptual disparity is going to be large since most of the horses are not white. Therefore, there is a large conceptual disparity between the annotated concept and the concept in the dictionary. It is the responsibility of the researcher who provides metadata for data-centric documents to avoid large conceptual disparities by attaching the information; e.g., if there is not white horse entry in the controlled vocabulary, the metadata has to have information to add to every subject into the data-centric document (horse) a verb (fur_color) and a value(white).

2.7.4 HEMP: Processing description

The HEMP is a complex engine that hierarchically transforms a file-stream based in the information supplied by the metadata file. The engine has a unique namespace with a determined and constant number of containers defined in the metadata file. The engine processes one container at a time and creates the new containers based on the information inside the container; each time a container is created, this container is processed by the engine. Therefore, the scanner generates a scanning tree following a deep-first in-order path. During the lexical processing of the stream, the containers also have information to construct the logical tree and annotate with semantic information the data. The engine ends the transformation when it finishes the first container and has no more information to be processed. Another characteristic of this engine is that it is deterministic, and therefore, the processing time of a file would be proportional to its size [12].

2.7.4.1 Containers For this engine, the container structure is the basic logical element that has enough information to define the behavior of the scanner, to construct the logical object, and link the semantic annotations. These containers are a five-value structure, and each container has a stream, a pointer, a value, an annotation, and a machine.

The stream is the physical data that the container represents; the pointer is a pointing structure with two members that define the offset and length of the stream which is being pointed having the stream of the root-container as the basic stream, the value is either a single numerical or string variable whose value can be changed by the engine, the annotation is the semantical information need to be transferred to the logical object, and the machine is the alma mater of the container since is the logic unit that operates creating new containers of a given type, modifying variables, creating and populating the logical object, and attaching semantic annotations. All this functionality can be grouped in lexical, logical, and arithmetic functions.

2.7.4.2 Lexical Functions Lexical functions are functions that operate at the stream level of the containers. The main lexical functions are functions which are based on regular expressions which can create new containers of a given type. There are also transforming functions that can operate and modify the stream of a container. The following are the lexical functions used by the engine:

2.7.4.2.1 WORD This lexical function recognizes a string based on a regular expression, and can generate one new container of a given type.

2.7.4.2.2 WORDS This lexical function splits the stream of the container based on a character, and can generate a list of containers of the same type.

2.7.4.2.3 RECORD This lexical function recognizes a a string of a given length, and can generates one new container of a given type.

2.7.4.2.4 RECORDS This lexical function splits the stream of the container based on a given length and can generate a list of containers of the same type.

2.7.4.2.5 CONCATENATE This lexical function creates a new container of a given type by concatenating the streams of the containers provided as arguments.

2.7.4.2.6 ZIP This lexical function creates a new container of a given type by deflating the stream of a container provided as an argument.

2.7.4.2.7 POINTER This lexical function creates a new container with the stream pointed by the container.

2.7.4.2.8 TOBINARY This lexical function transforms the value of the container into a binary stream.

2.7.4.3 Arithmetic Functions Arithmetic functions perform regular arithmetic operations on the values of the containers

2.7.4.3.1 SUM This function modifies the value of its container by adding its value to the value of the container passed as an argument.

2.7.4.3.2 MULTIPLICATION This function modifies the value of its container by multiplying its value by the value of the container passed as an argument.

2.7.4.3.3 EQUAL This function modifies the value of its container by assigning its value to the value of the container passed as an argument.

2.7.4.4 Logical Functions The logical functions operate to create and populate the logical object model. The following are the logical functions.

2.7.4.4.1 DOM This function creates a new empty document object model.

2.7.4.4.2 OBJECT This function adds and object to the document object model.

2.7.4.4.3 PROPERTY This function adds a semantic property to an object.

2.7.4.4.4 LIST This function creates an empty list.

2.7.4.4.5 APPEND This function appends an object or a value to a list.

2.7.4.4.6 STRUCTURE This function creates an empty structure.

2.7.4.4.7 MEMBER This function assigns a value or an object to a structure.

2.7.4.4.8 VALUE This function assigns a value to an object.

2.7.4.5 Engine Description The HEMP has an engine that processes the containers as soon as the containers are produced by the lexical functions mimicking a pre-order depth-first transversal of the lexical tree after reading the data-centric file and the metadata file. The metadata file which describes the file has the format described in Listing (9) and the grammar generates streams that belong to the JSON[3] language which is a common format used in the Web 2.0[4].

Listing 9: Metadata File Lexical Structure

```
METAFILE   ::= "metadata" "=" "{" SUBJECT "," VERBS "," VALUES "," AUTOMATAS "}"
SUBJECT    ::= "subject" "=" "'" subject_URI "'"
VERBS      ::= "verbs" "=" "{" ( verb_name ":" verb_URI ",")+  "}"
VALUES     ::= "values" "=" "{" ( value_name ":"
                ("[" TYPE (space_URI | min max step | enumeration)
                                  "," unitURI "]")+  )+ "}"
TYPE       ::= ("<"|">")?("c"|"b"|"i"|"d"|"f")
AUTOMATAS  ::= "automatas" "=" "{" ( container_name "="
                "{" MACHINE? FEATURE? VALUE? "}" )+ "}"
MACHINE    ::= "machine" "=" "[" RULE+ "]"
RULE       ::= "{" TYPE "," FUNCTION "," ARGUMENTS "," OUTPUT "}"
TYPE       ::= "type" ":" ("REGEX" | "DOM" | "ALG" | "ARITH")
FUNCTION   ::= "function": "FNAME"
ARGUMENTS  ::= "Arguments" ":" "[" container_name? "]"
OUTPUT     ::= "Output" ":" container_name
FEATURE    ::= "Feature" ";" verb_name
VALUE      ::= ("{" "Type" ":"Constant" "," "Value" ":" NUMBER "}")| value_name
```

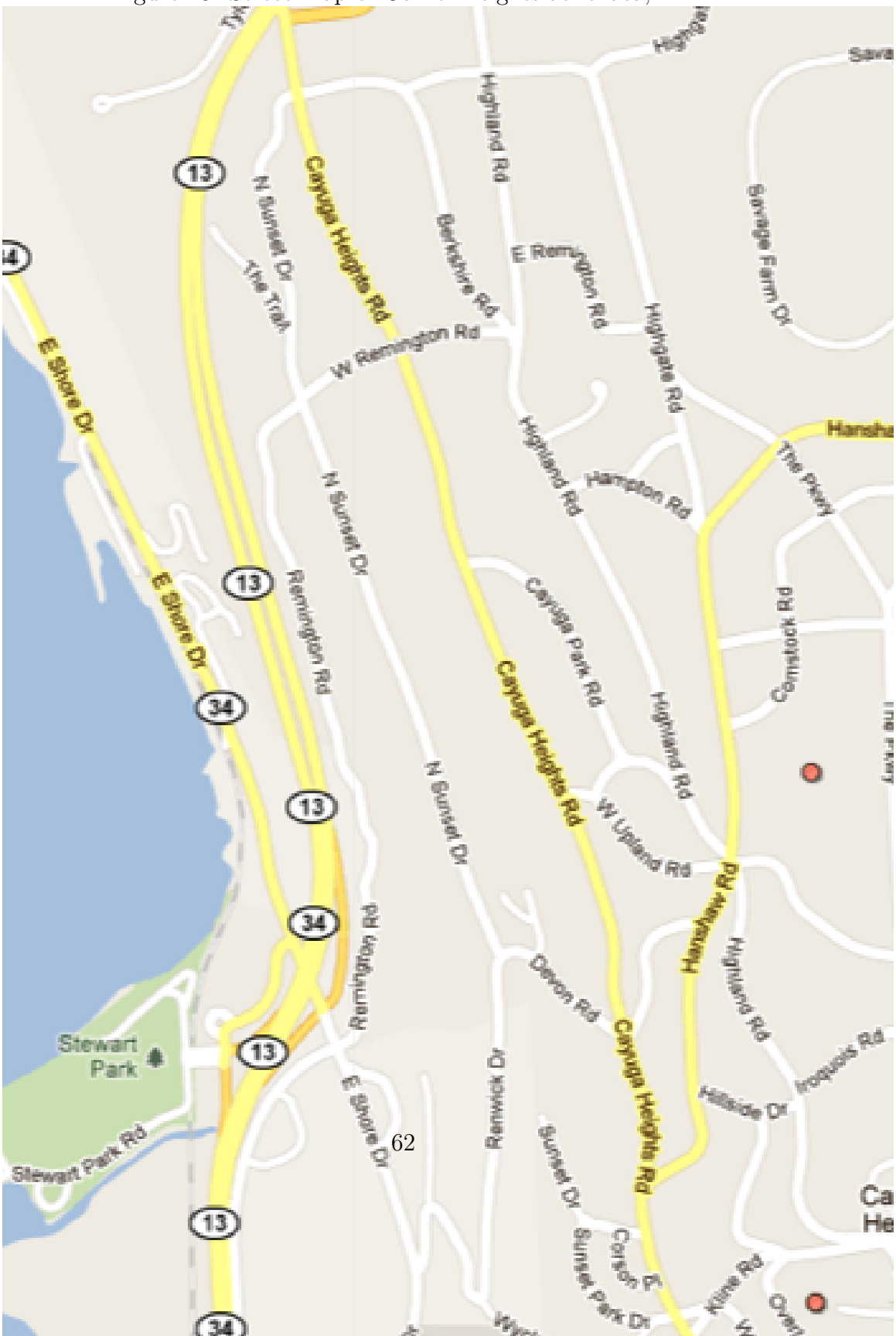
[3]JavaScript Object Notation is a lightweight data-interchange format easy to read and write, and built in common data structures found virtually in all modern programming languages.

[4]a collaborative and more interactive Web platform with many scalable services

2.7.4.6 Example: Metadata File for a Compressed Shapefile with Tree Data The compressed archive with tree data was downloaded from the website of Cornell University [65]. The location of the trees is in Cornel Heights at Ithaca, New York; a map of the trees can be seen in in Figure(15) and a street-map of the area can be seen in Figure(16) . The archive has 2561 records of trees and for each tree the data set has twenty-three attributes plus the spatial attributes. Compressed archives are very common containers to exchange data-centric documents since they use a lossless compression schema which help to reduce the size of the archive to save in transmission time through the network.

Figure 15: Tree Location Map in Corner Heights at Ithaca, NY

Figure 16: Street Map of Cornel Heights at Ithaca, NY



The metadata file in Listing (10) shows how a a very complex compressed archive can be easily defined using the "a posteriori" approach. This type of compressed archive is very common and it is used to exchange geographical point-data with attached attributes. The geographical information is stored in the shape file and the non-geographical information is stored in a DBF file. Both files along with another index file are compressed in a zip file. By working "a posteriori", the complex specifications rules of the zip file can be bypassed, and using this approach we can inspect the zip archive and calculate the offsets and lengths of the compressed streams, and used the RECORD function to define the DBHC and SHPC container directly from the ROOT container . Afterwards, the DBF and SHP containers that represent both the DBF and shape files are generated using the classical decompress[5] algorithm. As can be seen in Listing (11) the lexical description of the rest of the containers is pretty straight-forward. Since the shape file is semantically grounded, the lexical parsing can be constructed "a priori", i.e., the same lexical rules in the metadata description can be used for all shape files with point data. On the other hand, since a DBF file is semantically non-grounded, an "a posteriori" approach should be used to relax and ease the lexical rules in the metadata description. Therefore the containers DBFRECORDS, DBFID, SPECIE, GENUS and DBFID can be easily extracted using the RECORD lexical-function given the fact that we can know the offsets and lengths of all these records by studying the archive before creating the metadata information. The logical description of the structure can be seen in the railroad diagram from Figure (17) that represents a list of sentences where the "id" represent the reference to the tree, "dbh", "specie" and "genus represent the verbs/properties of the sentences and the "dbh_value", "specie_value" and "genus_value" represent the value of the properties. Besides, the name of the properties link unambiguously the property with the "Verbs" structure in Listing, and the name of their respective values are linked with the "Values" structure in Listing (10)which contain the semantic annotations that point to the controlled vocabularies. Therefore, the data from the file can be read, allocated in a logical object and defined semantically. Moreover the logical object MySON created by the engine can be easily serialized for exchanging purposes since MySON objects are by definition JSON objects, and the binary values of the properties have

[5]DEFLATE Compressed Data Format Specification version 1.3: http://tools.ietf.org/html/rfc1951

a self-describing structure.

Listing 10: Metadata File for a Compressed Shapefile With Tree (point) Data

```
metadata={
    "Subject":"http://wikipedia/database/tree",
    "Verbs":{
        "ID"       :"http://wikipedia/database/ID",
        "X"        : "http://wikipedia/database/longitude",
        "Y"        :"http://wikipedia/database/latitude",
        "GEOPOINT" :"http://wikipedia/database/geopoint",
        "GENUS"    :"http://wikipedia/database/genus",
        "SPECIE"   :"http://wikipedia/database/specie",
        "SPECIE"   :"http://wikipedia/database/dbh",
    },
    "Values":{
        "NUMBER"   :["f", "http://wikipedia/database/number_line" ,
                          "http://wikipedia/database/degrees"],
        "ID"       :[">i", "http://wikipedia/database/number_line" ,
                          "http://wikipedia/database/degrees"],
        "STRING"   :["c", "http://wikipedia/database/string"     ,],
        "GENUS"    :["c", "http://wikipedia/database/list_genus" ,],
        "SPECIE"   :["c", "http://wikipedia/database/list_species",],
        "GEOPOINT" :[
                     ["f", "http://wikipedia/database/number_line" ,
                           "http://wikipedia/database/degrees"],
                     ["f", "http://wikipedia/database/number_line" ,
                           "http://wikipedia/database/degrees"],
                     ],
    },
    "Automatas":{
        "ROOT":{
            "Machine":
                [
{"Type":"REGEX","Function":"RECORD","Arguments":["43","38325"],"Output":"SHPC"},
{"Type":"REGEX","Function":"RECORD","Arguments":["43783","149039"],"Output":"DBFC"},
                ],
            },
        "DBFC":{
            "Machine":[
{"Type":"REGEX","Function":"DECOMPRESS","Arguments":["-15"],"Output":"DBF"}
                ],
            },
        "DBF":{
            "Machine":[
{"Type":"REGEX","Function":"RECORD","Arguments":["513"],"Output":"TRASH"},
{"Type":"REGEX","Function":"WORDLENGTH","Arguments":["2198"],"Output":"DBFRECORDS"},
                ],
            },
        "DBFRECORDS":{
            "Machine":[
{"Type":"REGEX","Function":"RECORD","Arguments":["18"],"Output":"DBFIDV"},
{"Type":"DOM","Function":"NEWOBJECT","Arguments":["DBFID"],"Output":None},
{"Type":"REGEX","Function":"RECORD","Arguments":["15"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
```

```
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD", "Arguments":["18"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["32"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["32"],"Output":"SPECIEV"},
{"Type":"DOM","Function":"VALUE","Arguments":["SPECIE"],"Output":None},
{"Type":"DOM","Function":"PROPERTY","Arguments":["SPECIE"],"Output":None},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["32"],"Output":"GENUSV"},
{"Type":"DOM","Function":"VALUE","Arguments":["GENUS"],"Output":None},
{"Type":"DOM","Function":"PROPERTY","Arguments":["GENUS"],"Output":None},
{"Type":"REGEX","Function":"RECORD","Arguments":["18"],"Output":"DBHV"},
{"Type":"DOM","Function":"VALUE","Arguments":["DBH"],"Output":None},
{"Type":"DOM","Function":"PROPERTY","Arguments":["DBH"],"Output":None},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD","Arguments":["254"],"Output":"TRASH"},
{"Type":"DOM","Function":"OBJECT", "Arguments":["DBFID"],"Output":None},
                ],
            },
         "DBHV":{
            "Machine":[
{"Type":"REGEX","Function":"WORD","Arguments":["NUMBER"],"Output":"DBH"},
                ],
            },
        "DBH":{
            "value":"NUMBER",
            "feature":"DBH",
        },
         "DBFIDV":{
            "Machine":[
{"Type":"REGEX","Function":"WORD","Arguments":["ALFANUMERIC"],"Output":"DBFID"},
                ],
            },
        "DBFID":{
            "Value":"STRING",
            "feature":"ID",
        },
         "SPECIEV":{
            "Machine":[
{"Type":"REGEX","Function":"WORD","Arguments":["ALFANUMERIC"],"Output":"SPECIE"},
                ],
            },
        "SPECIE":{
            "Value":{"Type": "SPECIE"},
            "feature":"SPECIE",
        },
         "GENUSV":{
            "Machine":[
{"Type":"REGEX","Function":"WORD","Arguments":["ALFANUMERIC"],"Output":"GENUS"},
                ],
            },
        "GENUS":{
            "Value":{"Type":"GENUS"},
            "feature":"GENUS",
        },
        "SHPC":{
            "Machine":[
```

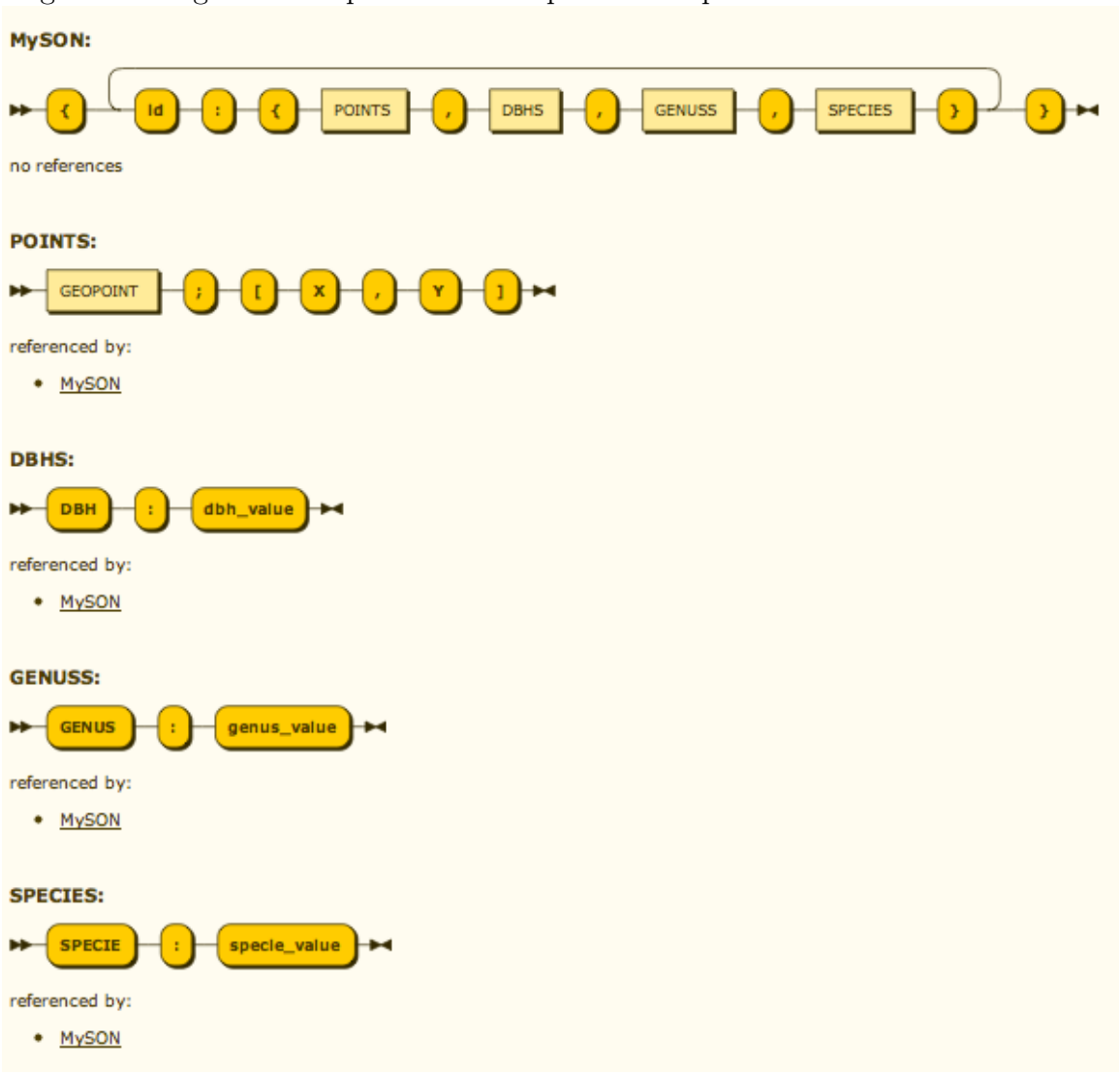
```
{"Type":"REGEX","Function":"DECOMPRESS","Arguments":["-15"],"Output":"SHP"}
                ],
            },
        "SHP":{
            "Machine":[
{"Type":"REGEX","Function":"RECORD","Arguments":["100"],"Output":"TRASH"},
{"Type":"REGEX", "Function":"WORDLENGTH","Arguments":["28"],"Output":"POINT"},
                ],
            },
        "POINT":{
            "Machine":[
{"Type":"REGEX","Function":"RECORD",   "Arguments":["4"],    "Output":"SHPID"},
{"Type":"REGEX","Function":"RECORD",   "Arguments":["4"],    "Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD",   "Arguments":["4"],    "Output":"TRASH"},
{"Type":"REGEX","Function":"RECORD",   "Arguments":["8"],    "Output":"X"},
{"Type":"DOM",  "Function":"VALUE",    "Arguments":["X"],    "Output":None},
{"Type":"DOM",  "Function":"PROPERTY", "Arguments":["X"],    "Output":None},
{"Type":"REGEX","Function":"RECORD",   "Arguments":["8"],    "Output":"Y"},
{"Type":"DOM",  "Function":"VALUE",    "Arguments":["Y"],    "Output":None},
{"Type":"DOM",  "Function":"PROPERTY", "Arguments":["Y"],    "Output":None},
{"Type":"DOM",  "Function":"NEWOBJECT","Arguments":["SHPID"],"Output":None},
{"Type":"DOM",  "Function":"OBJECT",   "Arguments":["SHPID"],"Output":None},
                ],
            "value":"GEOPOINT",
            "feature":"GEOPOINT",
            },
        "SHPID":{
            "Value":"ID",
            "feature":"ID"
        },
        "X":{
            "Value":"NUMBER",
            "Feature:"LONGITUDE",
        },
        "Y":{
            "Value":"NUMBER",
            "feature":"LATITUDE",
        },
        "-15":{
            "Value":{"Type":"Constant",  "Default":-15},
            },
        "1":{
            "Value":{"Type":"Constant",  "Default":1},
            },
        "43":{
            "Value":{"Type":"Constant",  "Default":43},
            },
        "38325":{
            "Value":{"Type":"Constant",  "Default":38325},
            },
        "43783":{
            "Value":{"Type":"Constant",  "Default":43783},
            },
        "149039":{
            "Value":{"Type":"Constant",  "Default":149039},
            },
        "15":{
```

```
            "Value":{"Type":"Constant",  "Default":15},
            },
        "18":{
            "Value":{"Type":"Constant",  "Default":18},
            },
        "32":{
            "Value":{"Type":"Constant",  "Default":32},
            },
        "254":{
            "Value":{"Type":"Constant",  "Default":254},
            },
        "4":{
            "Value":{"Type":"Constant",  "Default":4},
            },
        "8":{
            "Value":{"Type":"Constant",  "Default":8},
            },
        "20":{
            "Value":{"Type":"Constant",  "Default":20},
            },
        "28":{
            "Value":{"Type":"Constant",  "Default":28},
            },
        "100":{
            "Value":{"Type":"Constant",  "Default":100},
            },
        "513":{
            "Value":{"Type":"Constant",  "Default":513},
            },
        "2198":{
            "Value":{"Type":"Constant",  "Default":2198},
            },
        "TRASH":{
                },
        "ALFANUMERIC":{
            "Value":{"Type":"Constant",  "Default":"\s"},
            },
        },
}
```

Listing 11: Lexical Description of a Compressed Shapefile with Tree Data

ROOT ::= SPHC DBFC
SPHC ::= SPH
DBHC ::= DBF
SPH ::= SHPRECORD+
SHPRECORD ::= SHPID POINT+
POINT ::= X Y
DBF ::= DBFRECORD+
DBFRECORD ::= DBFID DBH GENUS SPECIE

Figure 17: Logical Description of a Compressed Shapefile with Tree Data



2.8 Conclusions

There is a clear need for better data integration to reinforce environmental collaboration among researchers and managers which should be heard and satisfied by searching the path between our current technology and the paradigmatic Arcadia where data exchange can happen in a blink of an eye without any problem whatsoever.

During the last decade many researchers trying to look for the roadmap to reach this particular Arcadia have been sweeping the road of data standardization to discover a labyrinth of paths where it was difficult to advance in the right direction. In addition, this large group of researches that have been trying to sweep together the road but instead of using a common language, they have kept working using hundreds of particular tongues which has created a gigantic Babelian tornado, and a dense cloud that did not allow them to see the the rest of the road. This does not imply that there is not a straight road through standardization to reach the happy Arcadia; however, it has not been found.

If the standardization road is difficult to seek, the metatada road is another plausible way to reach this Arcadia or at least getting closer. Very few workers have been cleaning this road, but important improvements have been accomplished by going from the semantically narrow road of the metadata with only archival purposes to the wide freeway of the Web of open data where metadata is used to unambiguously define data. Furthermore, these crews have adopted a simple clear RDF tongue so that many people can easily talk to collaborate and unveil the pathways to this paradigmatic village.

this thesis has shown that the standardization of data using XML schemas is dead as a common standard for representing environmental data since after a decade from its publication most of the environmental researchers do not use it, and a significant amount of them even do not even know about it. This is a significant finding from this thesis since it is the first time that has been study the reach of the XML technology in the real world of science beyond the groups working with projects where XML has been researched.

On the other hand, environmentalists working with primary data often use files with a multitude of legacy formats very well known and supported by the researchers of the different disciplines. Therefore, if standardization of data is a difficult or impossible solution to attain nowadays, the other alternative

for sharing data in between environmental agents is to share the data along with enough metadata so that the sharing and exchanging of data-centric files can be automatized.

Description of a data-centric file requires one to understand the nature and architecture of the different archives from orthogonal points of view to recognize certain patterns that should be understood to design a good metadata language. Looking at a number of familiar files used by environmentalists and other type of researchers, it has been shown that the specifications of these archives can be classified based on the complexity of the grammar, the grounding of their structures, and the contiguity of the records.

Besides depicting the specification of the physical structure of the archives, it has been shown that looking at the ontological essence of the metadata creation process can provide a formidable relaxation to the problem of creating a simple metadata language because of the following tautology: metadata comes after data since there cannot be metadata from no data. Most of the computer scientists that have been working with metadata languages have approached the solution following a classical parsing scheme where the automata to scan and parse a given file is known "a priori", and has been designed based on a particular context-free grammar. This classical approach of creating automata cannot work for environmental data-centric files since many of them cannot be described using context free languages. However, due to the aforementioned tautological fact, the problem of defining the file structure can be relaxed by analyzing the data-centric files using an "a posteriori" approach. It has been shown that by using this "a posteriori" methodology any data-centric file can be processed with a deterministic hierarchical automata independently of the complexity of the specification of the format of the data-centric file.

The Hierarchical Engine for Metadata Processing (HEMP) is an extended pushdown transducer that takes a data-centric and metadata files and produces a semantically annotated hierarchical data model. The HEMP has a processing unit and a unique namespace with a determined number of containers which hold the information to process the data-centric files. The functionality of the containers extends the capabilities of a regular pushdown transducer to adapt its facilities to the requirements of the task, and the physical peculiarities of the physical data on the data-centric files.

Environmental primary data can be represented using triplets and collected

in hash array structures which are common to modern programing languages. MySON is a new object type developed to storage the triplets on the computer memory using common data structures used in modern computer languages, and to properly storage the required semantic annotations. To ease its usage in the future, MySON has been designed as a JSON sub-type since JSON is becoming one of the most popular notations used in the Internet. In addition, MySON values are self-describing binary streams that can be serialized without the need of metadata which simplify its creation and utilization. Therefore, MySON objects are serializable objects that can be shared across the internet using common tools already present in millions of computers. In addition MySON objects are semantically grounded using common controlled vocabularies as required by the W3C in the Linked Data project since a MySON object that represents a data-centric archive is a collection of annotated triplets that can be linked to the Web in a open data cloud.

As a result of this research, environmental primary data written in common legacy formats can be easily describe using metadata descriptors to bring this precious data into the XXI century into the open Web. Tim Berner-Lee's omnipresent, semantic layered graphic in Figure (18)to represent the different levels of information for the Semantic Web can be reshaped as shown in Figure (19) for representing the Semantic Web for data-centric documents since XML has been probed dead among researchers and managers using environmental datasets. Therefore, data and metadata should replace the XML in the second layer of the "semantic cake". In addition, metadata should be kept in good health as long they are needed: long live the metadata.

Figure 18: Tim-Berner-Lee's Layered Graph of the Semantic Web

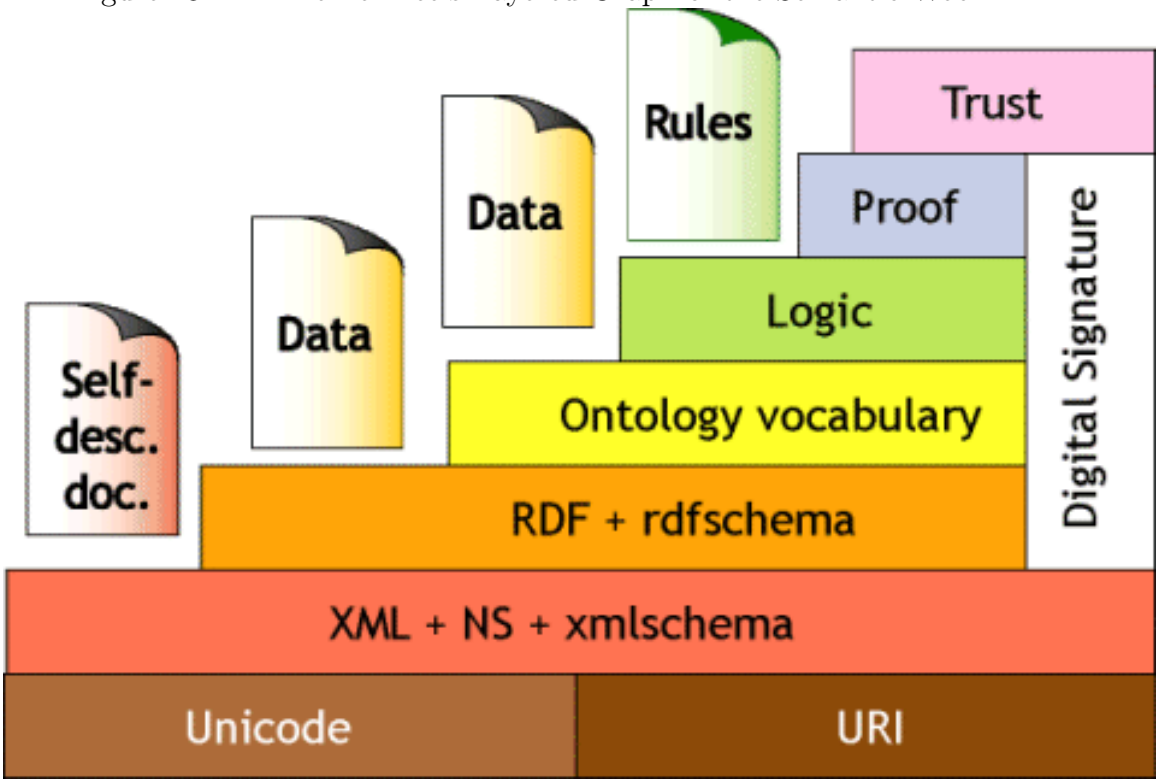
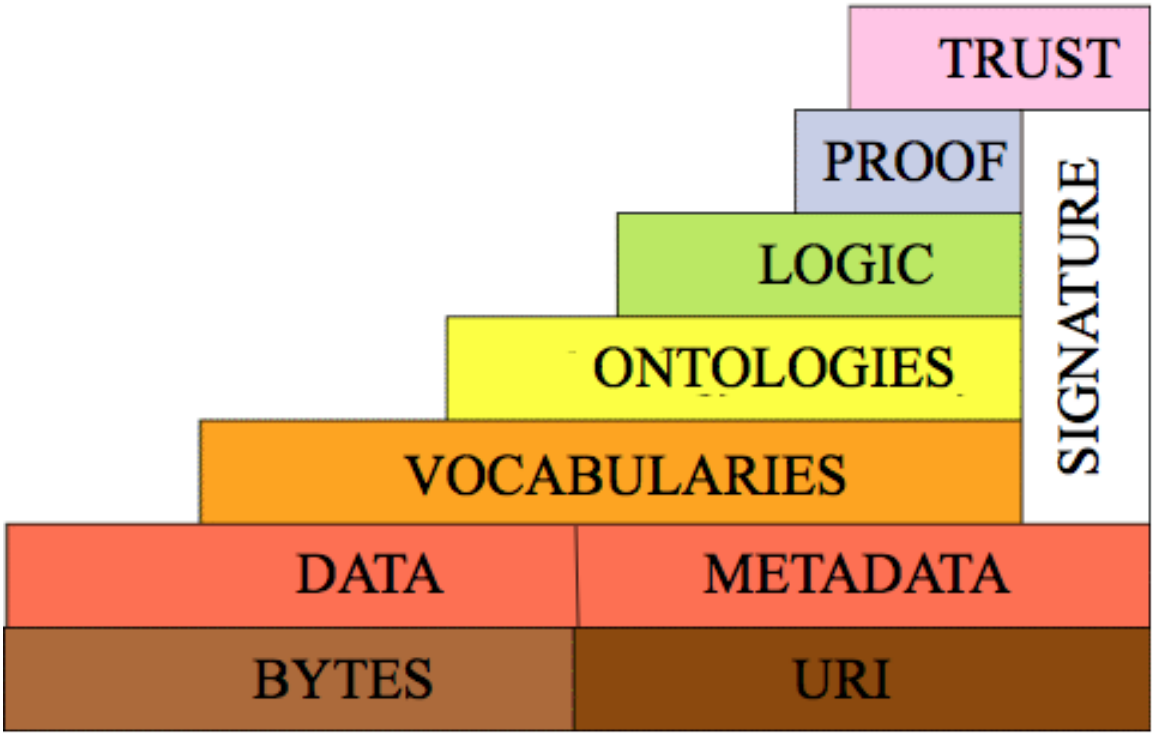


Figure 19: Modified Layered Graph of the Semantic Web for Data-centric Documents



Future work should:

- formally developed a metasyntax to describe data-centric files to automatize the metadata creation,
- find a way to reuse metadata descriptors
- create a library with metadata templates for common structures
- publish applications showing the capabilities of these extensions to a larger audience
- incorporate these extensions to the ecological modeling language and other metadata languages if they are well received.

The happy arcadia of the perfect data not only remains distant on the horizon, but also the pathways to it remains elusive:

> Ironically, the computer networks that make it easier to transmit data have made it more difficult to share data[224]

Therefore, scientists and managers have to persist on the quest until this battle with the data-monster is over, and we like Perseus, persist until we can slay the head of Gorgon forever, the terrible medusa with a thousand file formats in its tentacles and infinite semantics in its brain.

3 The Forest Cloud: Link between Local and Global Regimes

3.1 Introduction

As a sign of our times, concurrent and clashing events are creating scenarios where new tools should be developed to help the overall governance of natural resources. For instance, increasing population and energy requirements per capita are battled by growing concerns about global warming , and deforestation. Along these battle fields, increasing scientific knowledge and monitoring capacity should be used to bring sterile debates into more fruitful dialectics for finding the best solutions for the people and the environment. Along these lines, the Forest Cloud could be a constellation of little yet powerful tools which can positively contribute by linking local and global forest regimes under the transparent prism of the Internet [109] and the Cloud. As part of the Forest Cloud, the tree datastore should be a fundamental application where researchers, managers, officials and many other people interested in the governance of the forest of the world can participate taking action for the good of the environment and the trees of our forests.

For more than a century and since the beginning of scientific forestry at the end of the XIX century, forestry inventories have been the most basic tool to manage and control our forests. Hundreds of inventories have been developed at different scales through the world by public and private parties, yet little information from these archives is accessible through the Web that is nowadays our communication super highway [103].

Forestry databases on the Web are nothing new, and they have populated the Internet since the beginning of the Web to consolidate forestry data and facilitate access to information [204]. However these databases represent isolated exercises without any type of federation structure and/or communication protocol [191][222].

Urban Forestry, as a special case of the forestry discipline, has received a lot of importance during the last decade [42]. Urban foresters have developed a multitude of databases and tools to inventory urban trees and publish the tree information through Web-interfaces [233][9]. An innovative and valuable characteristic to these urban inventories is their community-based nature [235] which has been adapted to use consumer-grade digital devices with

online mapping tools and access to web services with databases where to store the surveyed data . For instance, a San Francisco based group has developed The Urban Forest Map website to collect tree data following the idea of its mantra "Creating a map of our urban forest one tree at a time ..." This groups is advocating the use of common mapping technologies on the Internet to feed its data store with tree information.

By promoting community-based participation and institutional collaboration, the infrastructure of the Forest Cloud will help to consolidate information about trees since offers unparalleled facilities to integrate the tree information from multiple sources regardless the format of the archives using an innovative metadata language based in a simple JSON schema and a Hierarchical Engine for Metadata Processing (HEMP). Furthermore, the Forest Cloud offers a simple yet powerful space for development of forestry applications by using MySON objects, which are JSON data streams with high density data to be used as a common object model to exchange information between the Forest Cloud applications.

3.2 Local forest governance

In many regions of the world the forest rights are managed by the local community following the traditional usages which have been in place for centuries. Even though in a few instances these traditional practices have been criticized because the soundness of them are more base on ideology than in science [144], empirical evidence has established a well documented body of sustainable practices when natural product are obtained following traditional methods [184]. For instance, modern fire suppression practices are based on lack a knowledge about fire ecology have brought Californian forests into unnatural stands subject to catastrophic fires [136]; on the other hand, Californian natives, using traditional knowledge have used fire more accordingly to sustainable practices [138].

There are many more documented examples of successful local governance of forestry resources in the world [186][11][95]. Nonetheless, theses local regimes have not always been as successful as every good mind would have desired. It is notorious the decimation of the forest in many developing countries such as Brazil, Indonesia, or Mozambique. Due to the lack of effective local forest governance regimes, global and international regimes to monitor and regulate forestry activities have been created to stop forestry abuses.

3.3 Global Forest Governance

Through the last two decades different actors have come into the scene to try to regulate forestry practices in developing countries to improve environmental and cultural aspects since there is a large conscience-based population all over the world worry about the ecology of our forest in particular and the wellbeing of nature in general.

During the Earth Summit in Rio, the world saw the first attempt where environmental groups tried to push for a global forest regulation to defend the forest and impede the rampant deforestation of the forest of the world . However, it was clear that the countries were not going to implement any of these petitions since these measurements would contravene the international commercial rules signed under the Word Trade Organization .

Because of this inaction, groups like the Forest Stewardship Council have implemented a global voluntary governance regime to control forest operations, and fight against illegal logging using market-driven strategies such as wood certification through a private chain of custody [214]. The irony of this private global regimes which often times attack public administration for their lack of transparency, is that their chain of custody data is out the reach of the general public. For instance, this author who assisted a seminar at UC Berkeley given by the CEO of private for-profit who certifies for FSC, was told that the data from the chain of custody is private and could not be make public for the general interested audience to oversee any of the certification processes.

After a decade of this market-driven, private global governance, there are not clear reports that can show the benefits of these private regimes; on the contrary, several studies show no improvements [155, 169].

3.4 International Governance

During the last decade and to overcome the limits imposed by the World Trade Organization , and answer to the public demand of fighting against the illegal trafficking of natural resources such as wood, developed countries have implemented legislation to respond to these requests.

The European Union (EU) signed the Forest Law Enforcement, Governance and Trade (FLEGT) to promote Voluntary Partnership Agreements (VPAs)

with other countries which supply forest products to the EU. The same kind of legislation was signed by the USA under the Lacey Act to combat illegal trafficking of natural resources [41]. Both legislations establish that any illegal wood products in origin are illegal products in the countries of destination; therefore, they became illegal products in Europe and USA. This approach which does not contravene the stipulations of the WTO has notorious inconveniences such as the lax definition of legal timber in the countries of origin given the the legal frameworks, and the technical impossibility to track wood[52].

3.5 The Tragedy of the Forest

Four decades ago, Garret Hardin opened a fertile academic debate after his publication "The Tragedy of the Commons" which has been kept alive since them [117]. In his controversial article, he used a fable to defend logically various thesis regarding logical acts and consequences of individuals interacting with common natural resources. In short, he stated the necessity of population control and the privatization of the commons to optimize their usage for us and future generations.

Insofar, during the next decades, some authors have been able to link population growth as the main cause of deforestation in developing countries since forestry lands has been claimed as agricultural fields to support the ever growing population needs [15]. This tragedy of the forest directly relates with tragedy of the commons that Hardin adverted in his tragedy of the commons since deforestation is causatively linked to population growth.

Beside the obvious increase in nutritional needs, the increasing energy footprint per capita have had a significant deforestation toll in some developing countries through conversions of forest lands into highly profitable energy crops [75]. This farming of highly profitable energy crops, in detriment of natural forests in developing countries, has exacerbated the tragedy of the forest during the last decade.

Not as important as the previous situations in terms of loss of forestry area, illegal logging of valuable woods in developing countries has also been reported as an important part of the tragedy of the forest and its degradation [154]. This degradation of the natural forest has not been uniform through the world and local conditions have rendered different patterns of forest degradation.

A decade later from Hardin's article, Ellinor Ostrom was able to describe after studying numerous cases of common pool resources through the globe a minimum set of eight guiding principles required to attain the a sustainable management of common pool resources [184]. Some of these principles are often violated in local, global and international forest regimes therefore contributing to the tragedy of the forest. Among them, the lack of effective monitoring, traceability, and identification of forest products plagues the regimes with plenty of latitude for illegal activities to pass unnoticed under the radar [56].

3.6 Tree Identification

For many years, a interdisciplinary body of scientists have been providing knowledge to be able to monitor our natural resources. Nonetheless, this growth has not being translated into practical applications to help forest enforcement agents to effectively monitor the trade of forest products [137].

The world forest can be defined as a large puzzle where each tree is a piece of the puzzle, and each piece can fit only in one place. Under this problem a custom agent of any country who faces how to identify a log basically has to place a piece in the puzzle. Given that one log of wood by itself gives very little or no information , any log will be represented as a square piece that perfectly can be placed in any area of the puzzle, as a result, any log can be smuggled through the border with no problem. If the agent would be able to pose some critical questions and get the correct answers from the tools available to build the puzzle, the pieces will show some peculiar angles and shapes that would limit the available number of slots for its placement. Ergo, under an hypothetical scenario of perfect information, the pieces of these puzzle will have enough singularities and will only fit in one place in the puzzle. Therefore, no player will be able either to fit a piece of the puzzle into the wrong spot or to smuggle a log through the border.

3.6.1 Where Is It From?

Once the log is far from the stump to answer which stump correspond to one given log is a pertinent question that the custom agent should be able to answer and to double check that the information given on the import papers about localization is indeed true.

For many years scientists using dendrochronological and chemochronological methods have been able to linked atmospherical events within the rings of the trees [26]. For instance, knowing the series of rings, scientists have been able to reconstruct preterite weather and atmospherical events. Therefore, it is a given that having extensive records of the weather and other atmospherical events around the globe for the last decades will be able to test that a chronology of rings could or not be linked to a given localization. However, this combinatorial problem has never been posed even at research level. Ergo, the custom agent will not be able to use the ring information from the logs to test the localization.

Furthermore, the custom agent could have tools that process remote sensing images to answer the question . The temporal and spatial resolution of satellite imagery from forest areas would give the custom agent enough information to test and verify the localization of the tree. However and due to non-technical reason, the custom agent has no tool that will facilitate his work of enforcement. Sadly, the agent will not be able to test the information.

3.6.2 Which Specie Is It?

In a second instance, the agent should resolve that the specie of the log is the one that figures in the official papers. Difficult problem indeed given the fact that there are only partial keys to classify woods and most of them required advance microcopy techniques [62]. Therefore, if the agent is not trained in wood anatomy and does not have experience using several keys with very technical operations, this agent will not be able to answer the question posed.

On the other hand, the scientific community knows that the answer to that question is feasible. For instance, several authors using different techniques were able to estimate the specie of the wood samples based on its physical and chemical properties. Near infrared technology has been successful in distinguishing the specie of the wood sample, and has been able to distinguish even subspecies [236][43]. Laser spectroscopy is being actively developed to fingerprint materials and could eventually distinguish tree species [167]. Even though these techniques have been positively tested to discriminate wood species accurately, there is a lack of a general comprehensive algorithm that could answer the question that the custom agent faces. Ergo, the custom agent does not have a basic tool too monitor and enforce our environmental laws.

3.6.3 Which One Is It?

Finally and perhaps the most obvious question to ask is if a given log really comes from the tree that the papers says. Unfortunately the custom agent once again will no have resources to try to answer this last question. Once again the agent will have to believe the information on the papers without being able to test its accuracy even though there are fingerprinting techniques available to be applied.

Genetic fingerprinting has been used as a forensic methods for many years, and specialized methods have been developed to study trees [187, 131]. Perhaps there are not still hand-held devices to help the custom agent to help her in her daily work, but we are not far away from that to happen (chapela).

Biometrics can also play an important role in fingerprinting logs [58]; it is not difficult to imagine that the ring series of the stump of the tree has to have a strong similarity with the ring series of the log from that stump. However, there is not information system that the custom agent can access to get the information from the stump, and even if there were a system with that information so far there is not tool that can resolve the problem posed; once again.

During the last decade methods have been developed to map the 3D information of the bole of trees using various methods to increase the information in forestry inventories [146][160][116][205]. Again, the information and methods to test information in the papers will not be available for the custom agent, inventory data often times is not consolidated within the Internet space and the ability to check any biometric measurement from the log will be unavailable.

True and sadly, the custom agent will not be able to enforce the laws since, the truth about the logs is beyond the reach of her abilities.

3.7 The Forest Cloud

At the core of any modern complex organization, information systems provide the power for taking care of the integrity of the business model for holding information and providing correct answers to pertinent questions. This forestry information systems as a collection of , data, tools and protocols is long over due since the tasks it can perform have been shown of capital importance

for the success of managing forestry resources. Therefore, global, and local governances should meet at the technical spaces to provide the regimes with the best tools to monitor the integrity of the regime [137]. The Forest Cloud has the potential capacity to fill the vacuum that nowadays reigns at the local and global monitoring level of the management of forestry resources which is one of the common cornerstones to effectively manage common pool resources [85].

3.7.1 Cloud Concept

The Cloud came to exist in the Internet as a realization of an old idea [204] whose time has finally come with an illusion of infinite resources available for those who provides data and services through the network [20].

Therefore cloud computing is a revolution in the architecture of the the internet and the personal computers that is starting to take ground in many different arenas. Among the many characteristics that the cloud has, some of them are very relevant for our forestry purposes in particular and environmental goals in general.

3.7.1.1 Software on Demand Classical software barriers and restrictions such as platform compatibility or software prerequisites to run an application will be transparently solved either by running the applications on the cloud or using rich internet applications that run in common browsers.

3.7.1.2 Computational Power The Cloud is a successor of grid computing, and as such, it provides supercomputing performances for data crunching application [44]. Many of our problem of interest are computationally complex such as genetic matching algorithms, or required to work with large archives such as remote sensing images. In either case, forestry managers and researchers can benefit for the computational power of the Cloud.

3.7.1.3 Illimitable Data Store As in many other fields, forestry scientists and managers are able to gather more information using digital devices. This large amount of digital information will need vast digital storages that the Cloud can facilitate. Therefore, the Cloud fulfills the idea of a hard drive with no restrictions in storage capacity.

Mobile Devices Mobile devices and mobile Internet is on the rise for the good of environmental research and management. The Cloud is there to help develop functional applications that run on the cloud for the users who can save valuable resources in mobile devices such as energy and memory space. Therefore the mobile devices become pure interfaces to access the functionality of the cloud. For instance, simple handheld devices can be used in the forest to sample data and send the data over the Internet to the appropriate place in the Cloud where it is store for ulterior use. This revolution of mobile devices can arm grass-root brigades to collect data anytime anywhere and to store it on the cloud for further processing.

3.7.1.4 Layer of Common Communications Communications between applications on the Cloud should be ingrained in its infrastructure and the adoption of common languages and protocols is a demand that has to be provided [149]. The idea of a big cloud where people can share data between peers and combine applications cannot exist without the concept of seamless communications among the applications [213]. Therefore, good common interfaces to applications have to be clearly defined, and data has to be fully qualified at the physical, logical, and semantic levels. Like in many other technological races from the past, we are seeing competitive projects with similar goals working trying to impose their respective protocols and data models as pictorially can be seen in figure(20). Nonetheless, as in previous technological races, there will be technological winners in the near future.

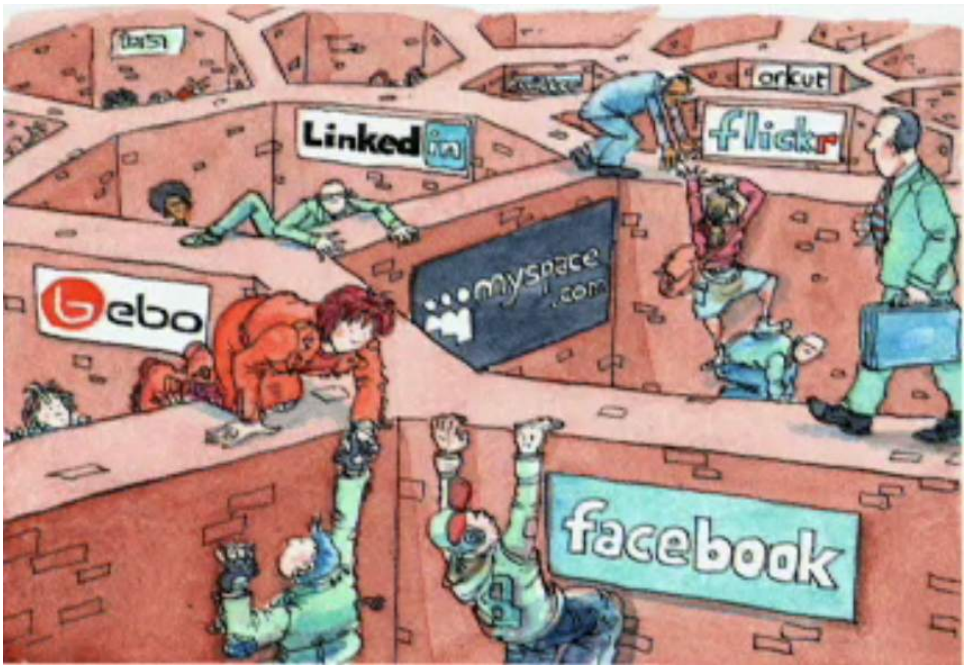


Figure 20: Walls in the Cloud created by private competitive communication protocols

3.7.2 Forest Cloud Philosophy

Following the successful unix philosophy that complex tasks can be achieved by using a set of simple tools with simple semantics and data structures [173][126], the Forest Cloud must be an Internet application that should be part of a constellation of Cloud applications that should be used to answer the demands of the society for a better management of forestry resources. Therefore, the Forest Cloud has to accomplish the following goals:

- The Cloud Forest will have to integrate under its umbrella any person from the globe with collaborative ideas who wants to inventory and defend the trees of the world acting locally and thinking globally.
- The Cloud-foresters will have to generate ideas and mobilize needed resources to expand its reach.

- The Cloud-foresters will have to feed the data store of the forest cloud with tree information with the ultimate goal of inventory all trees of the world.
- The tree data will have to be curated and consolidated to gain credibility and respect from the people working at locals and the global governances.
- The foresters of the Cloud Forest must be acredited with the organization and able to handle tree information properly.
- The data Forest Cloud will grant full access to the data store and other services to the acredited members of the Cloud while sensible information will no be disclosed in order to safeguard the integrity of the trees.
- The forest Cloud will incorporate a scientific tested set of tools to operate in tree data to answer important and basic questions to improve monitoring of forest products.

3.7.3 Forest Cloud Infrastructure

3.7.3.1 Metadata Templates Since the development of personal computers, forestry primary data collected by foresters are stored in common data files that contain the records for the individual trees. In order to make this information available to the Forest Cloud (or any other data store) a complementary metadata information which describes the tree data unequivocally has to be provided alongside to the engine of the Forest Cloud to scan and parse the tree information from the files at the physical, logical and semantic levels.

Chronologically and before the advent of worksheet programs, text-based files such as comma-separated files, or decadal files were the prevalent type of digital files used to keep tree information. Thereafter, big corporations such as Microsoft Corporation, or Environmental Systems Research Institute (ESRI) have developed very popular programs to manipulate environmental data using their own private file specifications to store data such as Excel files or Shapefiles. These few file types have become the "de facto" standard to store, manipulate, and exchange environmental primary data since their

software tools are the "de facto" standard programs to manipulate the data information.

Under this circumstances, for researchers and managers working with spatial data, Shapefiles were and still are the most popular format to convey information with spatial features. For researchers working with dendrochronological series, text-based decadal files are the most common files used to convey tree-ring information. On the other hand, Excel files have become the target file for all database and spreadsheet programs as they are able to read and write Excel files. Therefore, Excel files were and are the most popular archive among researchers and managers working with environmental datasets since almost all software programs can work with data store in Excel files.

All these types of popular files used by foresters can be processed by the Forest Cloud using the HEMP as long as they are described using the metadata language that HEMP uses. Because metadata development has been noted often as tedious [ref], and expensive[ref], the Forest Cloud provide a metadata templating system to develop metadata "a posteriori" for these common formats used by foresters and environmentalists.

3.7.3.1.1 Excel[6]Files Excel files are used by scientists and managers to store, transfer, exchange and manipulate environmental data because almost all actors manipulating environmental data have software tools in their computers to read, and manipulate Excel files. Therefore, Excel files have been the "de facto" standard data format for environmentalists. However, because Excel files contain much more information than the actual data from the worksheets, the physical structure of the file is rather convoluted [2] which has precluded the use of existent metadata languages to describe this type of archives. In addition, since Excel files are ungrounded files, they represent the quintessential archive for an 'a posteriori' metadata development.

There are different approaches to physically scan and parse Excel datasets because of the different internal treatments of numeric data. For instance, a real number can be stored in an Excel archive in more than one physical formats, and the user of the spreadsheet cannot control these variations since are treated by the program automatically and internally. In addition, the physical treatment of data for different cells of the same column in a single table for a given worksheet might vary.

[6]Excel is a Trademark belonging to Microsoft Corporation

Taking into account the physical treatment of data inside an Excel file, the more general approach to process a dataset stored in an Excel file following the "a posteriori" technique is to create a metadata container for every cell and row of the dataset as shown in figure (21). By doing so, the HEMP engine is able to read all the information from the dataset in an orderly way and create the correspondent logical MySON holding the data and metadata information for the given Excel archive.

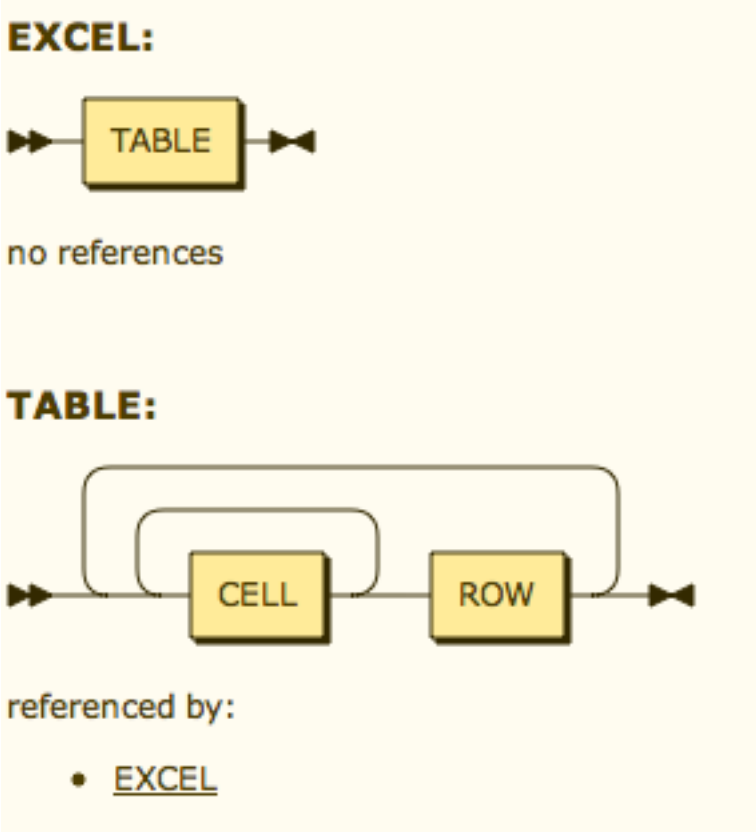


Figure 21: General Physical Description for Containers of an Excel Dataset

On the other hand, another more regular case for the physical treatment of data in a table of an Excel file can happen when all the the cells from the same column of the table receive the same physical treatment. In this case, a simpler declaration of containers as shown in figure (22) can be used to defined the information needed for the HEMP engine to process the Excel

file. In this case only one container per column plus two additional containers are needed.

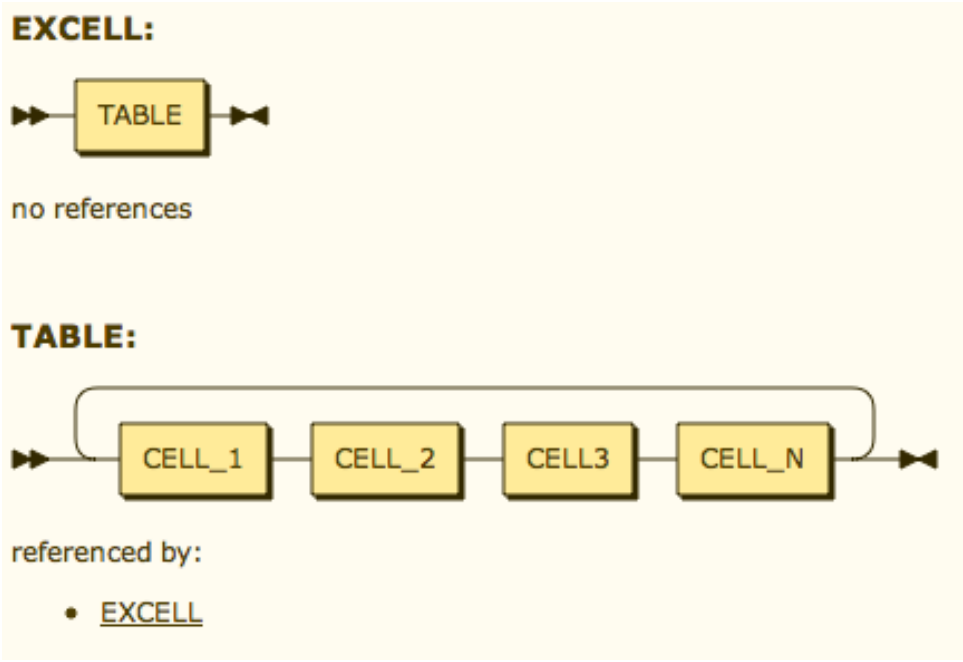


Figure 22: Physical Description for Containers of an Excel Dataset with regular structure

However, the most frequent case for large datasets stored in an Excel lays between these two aforementioned extremes where the physical treatment of numeric values for the cells of all columns for several rows are equivalent. In this case the declaration of containers to describe the file can be seen in figure(23). In this case the engine will need fewer containers than in the general scenario and more than in the regular scenario.

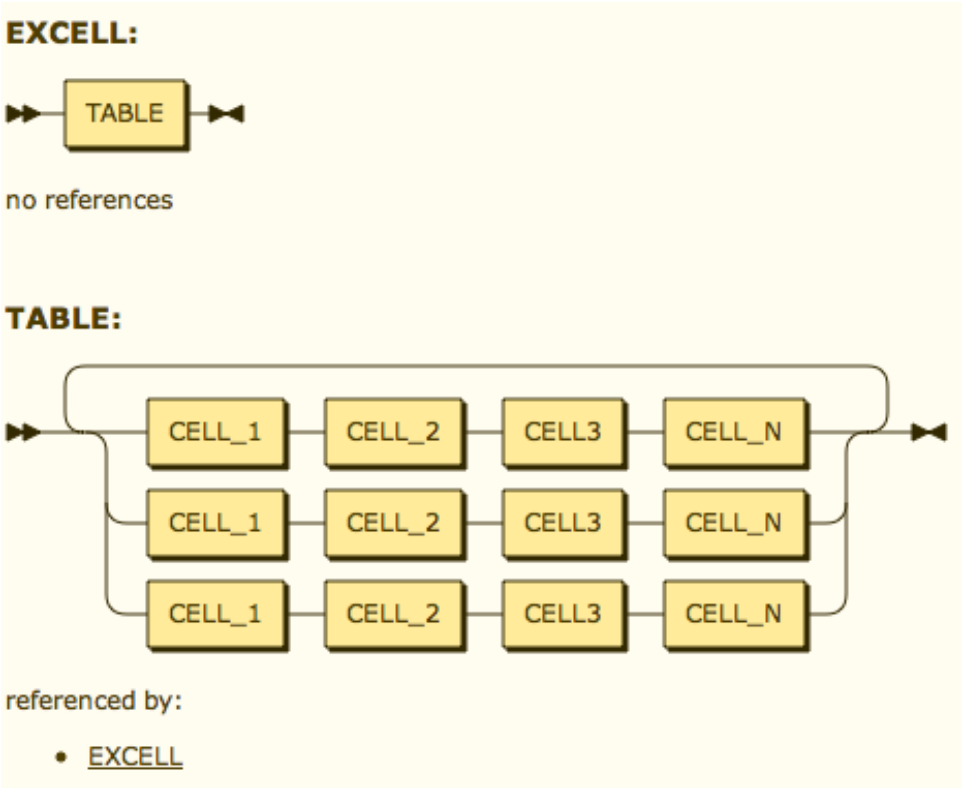


Figure 23: Physical Description for Containers of an Excel Dataset with a mixed regular structures

The Forest Cloud has a template system to 'a posteriori' process tabular data within an excel file with a single worksheet and a single table. The template takes besides the URL-location of the Excel file three arguments:

- position of the column that holds the horizontal geo-coordinate,
- postion of the column that holds the vertical geo-coordinate, and
- the name of the spatial geo-projection.

After an automatic, 'ad hoc' introspection of the Excel file, the Forest Cloud can generate a container grammar in algorithm(4) that describes the physical structure of the tabular data, and to map the data into a logical object with geospatial annotations. This approach has been used in the Forest Cloud to

store single-tree-information from California stored in a table inside an Excel file [7]; the trees can be observed in figure(24).

Algorithm 4 Excel 'a posteriori' Grammar of Containers

$$S \Rightarrow \sum < SECTION >$$

$$\sum \leq SECTION \geq \Rightarrow TABLE$$

$$|TABLE| \Rightarrow \sum_{ID=1}^{ID=n} ROW_k$$

$$\forall\, k\, |ROW_k| \Rightarrow \sum_{V=1}^{V=m} CELL_t$$

$$\forall\, t\, CELL_t \Rightarrow VALUE \Rrightarrow F(ID)[ID\,,\, V\,,\, VALUE]$$

[7] Excel file provided by Prof. Dennis Baldocchi

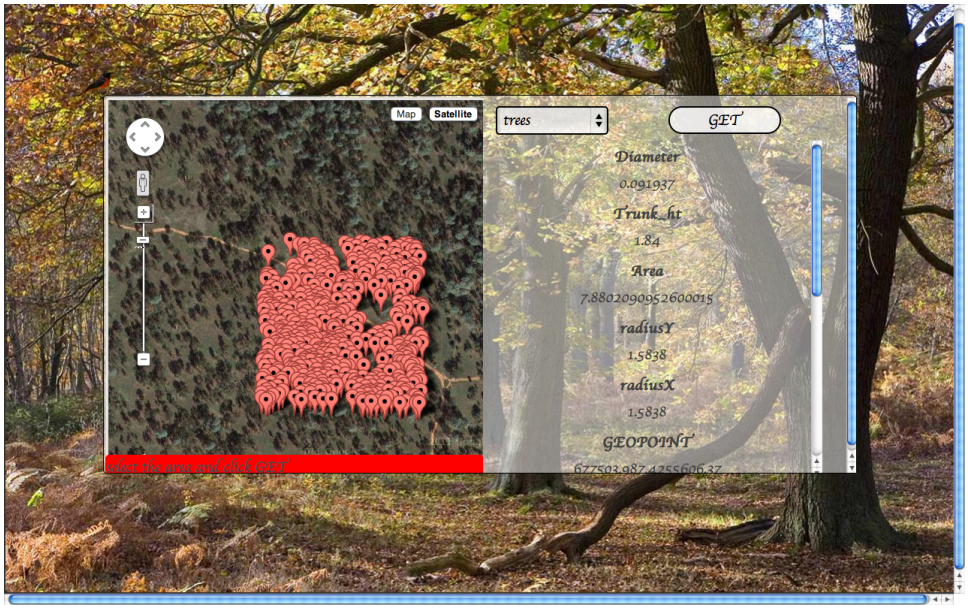


Figure 24: The Forest Cloud partial view of trees from research project in central California

3.7.3.1.2 Shapefiles Shapefiles have been the "de facto" file standard to present and exchange information using maps because almost all environmental agencies used ESRI products to treat datasets with spatial information. A Shapefile archive is a container with two basic files inside that represent the spatial and non-spatial information respectively [6]. Because the spatial information is collected in a grounded and non-context-free shape file, and the non-spatial information is stored in a non-grounded dBase file, the Shapefile is also a perfect candidate format for an "a posteriori" metadata development.

The overall specification of Shapefiles does not follow the rules of a context-free language; however, point-shape files can be described following the rules of a context-free language. Since trees are more often represented as points in maps, and point-shapefiles are context-free , and grounded files, the same template with containers can be used to define the spatial information of all files of this type. Therefore, following the classical "a priori" technique, the containers shown in figure(25) will be enough to define any point-shape file; moreover, polygon-shape files with a single ring per polygon which are

often used to represent tree-crowns can also be considered as context-free files; ergo, they could be treated using an “a priori” approach as well.

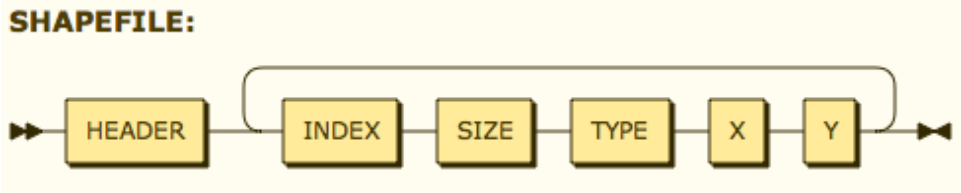


Figure 25: General Physical Description for Containers of a Shapefile with point data

The dBase file, however, due to its non-grounding and non-context-free nature can only be defined “a posteriori” for the HEMP engine to process it. The processing containers needed to process ’a posteriori’ a dBase file can be seen in figure(26). Therefore, Shapefiles with spatial and non-spatial information about trees are also a good candidate for metadata templating “a posteriori”.

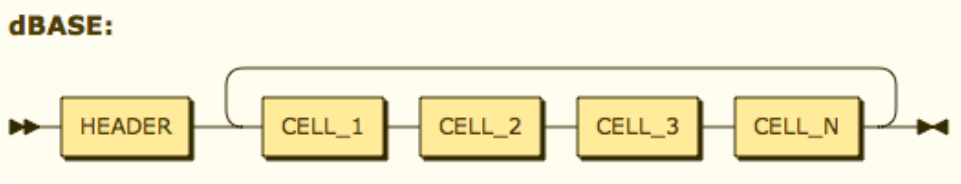


Figure 26: General Physical Description for Containers of a DBase file

The Forest Cloud following the same pattern of an ’a posteriori’ instrospection can develope automatically a container grammar like the algorithm(5) that completely describes the structure of the archive and translates the data items into a logical object with spatial annotations. This technique was used to store in the Forest Cloud the tree information provided by researchers at Cornel University in a shapefile[8]; the trees can be observed in figure(27).

[8]http://cugir.mannlib.cornell.edu/bucketinfo.jsp?id=8025

Algorithm 5 Shapefile 'a posteriori' Grammar of Containers

$$S \Rightarrow < SHAPE >< DBF >$$

$$|SHAPE| \Rightarrow RECORDS$$

$$|RECORDS| \Rightarrow HEADER + RECORD$$

|

$$HEADER| \Rightarrow ID$$

$$|RECORD| \Rightarrow < TREEPOSITION >$$

$$|TREEPOSITION| \Rightarrow X + Y \Rrightarrow \digamma(ID)[ID\ ,\ POSITION\ ,\ (X,Y)]$$

$$|DBF| \Rightarrow < RECORDS >$$

$$|RECORDS| \Rightarrow \sum_{i=1}^{i=n} RECORD$$

$$|RECORD| \Rightarrow ID + \sum_{V=1}^{V=n} FEATURE_i$$

$$\forall\, i\, |FEATURE_i| \Rightarrow VALUE_i \Rrightarrow \digamma(ID)[ID\ ,\ V\ ,\ VALUE]$$

Figure 27: The Forest Cloud partial view of trees from Cornel University

3.7.3.1.3 Decadal Decadal files are text-based and the "de facto" standard file to present and exchange dendrochronological information or environmental information grouped in a series of decadal sequences. The Tree-ring Database holds thousand of files with tree-ring data from all over the world. However, the spatial information from these files has very low resolution (1-minute spatial resolution) that will not allow to position the trees accurately on maps; therefore, the Forest Cloud has not implemented a template to incorporate this tree information into its data store even though is feasible.

Since decadal files are grounded and context-free, a single template defined "a priori" in the algorithm(6) will be sufficient to describe all files of this type.

Algorithm 6 Decadal 'a posteriori' Grammar of Containers

$$S \Rightarrow HEADER + RECORDS$$

$$HEADER \Rightarrow FIRSTLINE + SECONDLINE + THIRDLINE$$

$$FIRSTLINE \Rightarrow < SITE >< SPECIES >$$

$$SECONDLINE \Rightarrow < ELEVATION >< LATTITUDE >< LONGITUDE >$$

$$THIRDLINE \rightarrow < INVESTIGATOR >< DATE >$$

$$RECORDS \Rightarrow \sum_{i=1}^{i=n} RECORD$$

$$RECORD \Rightarrow \sum_{i=1}^{i=m} (ID + DECADE)_i$$

$$DECADE \Rightarrow \sum_{i=YEAR} GROWTH_i \Rrightarrow F(ID)[ID\,, YEAR\,, GROWTH]$$

3.7.3.1.4 Keyhole Markup Language(KML) The keyhole markup language was developed by researchers at Google Corporation to be used in the application Google Earth; because of its popularity, KML has been taken as an standard for the Open Geospatial Consortium. Therefore, this language with intrinsic spatial features has been used to encode spatial tree-data information on the Web. For instance, the environmental department of the autonomous Goverment of Andalucia has provided the information

about singular trees from of Andalucia using this type of file[9].

Because this type of file is tex-based, non-grounded and context-free, it is a clear candidate for an 'a posteriori' templating system to create the metadata. The algorithm(7)shows the grammar that can be used to describe 'a posteriori' KML files.

Using this 'a posteriori' approach, the Forest Cloud provides a template that produces the metadata file once it knows the KML file with the tree information. The forest cloud has incorporated the information from singular trees from Cadiz, Andalucia whose information is stored in a KML file; these trees can be observed in figure(28).

Algorithm 7 KML 'a posteriori' Grammar of Containers

$$S \Rightarrow \sum_{i=1}^{i=n} < RECORD >$$

$$|RECORD| \Rightarrow < TREEPOSITION >< ID > + \sum_{V=1}^{V=n} < FEATURE_i >$$

$$|TREEPOSITION| \Rightarrow LATITUDE + LONGITUDE$$

$$\forall\, i\, |FEATURE_i| \Rightarrow VALUE \Rrightarrow F(ID)[ID\ ,\ V\ ,\ VALUE]$$

[9]http://www.juntadeandalucia.es/medioambiente/

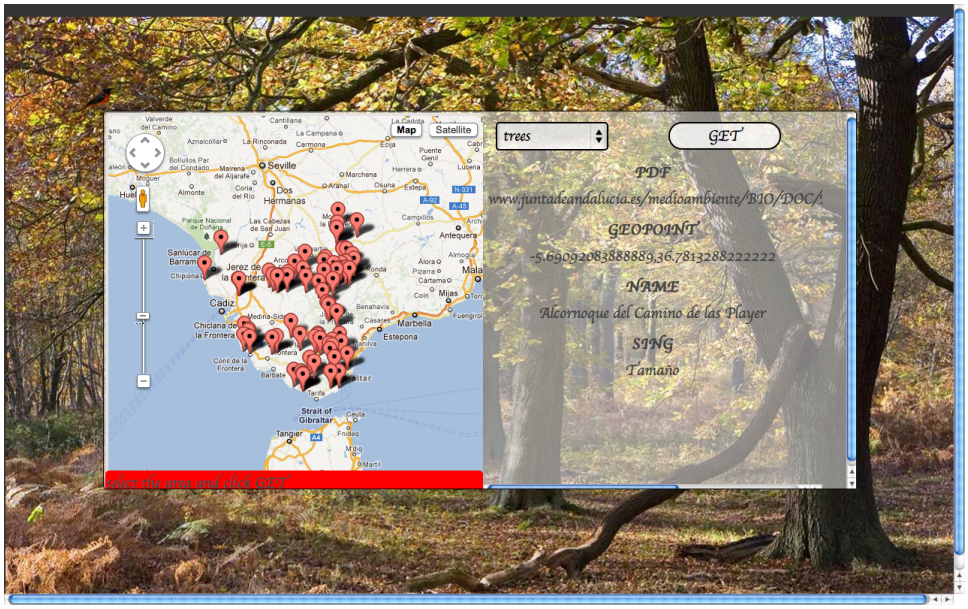


Figure 28: The Forest Cloud partial view of singular trees from Cadiz province, Spain

3.7.3.1.5 Comma Separated Values (CSV) Comma-separated-value files are often used in environmental projets to share data. This type of file is a classical text-based, context-free, non-ground file. Becaue it is non-grounded but context-free, it is a clear and easy target for an 'a posteriori' templating system.

The forest cloud has developed a template system for CSV files that allows to input tree information from these files. In order to run this template in the Forest Cloud, three parameters have to be supplied to the templating routine:

- Position of the column that holds the horizontal coordinate,
- Postion of the colum that holds the vertical coordinate, and
- The name of the spatial projection.

Using the template, the Forest Cloud can acquire single-tree data from the Philly Tree Map website that stores information about the urban trees of

the Philadephia region. A small partial view of the trees from this site was stored in the Forest Cloud and can be seen in figure(29).

Algorithm 8 CSV 'a posteriori' Grammar of Containers

$$S \Rightarrow < RECORDS >$$

$$|RECORDS| \Rightarrow \sum_{ID=1}^{ID=n} RECORD$$

$$|RECORD| \Rightarrow \sum_{V=1}^{V=n} (FEATURE_i + ", " \mid \varepsilon)$$

$$\forall\, i\, |FEATURE_{i|} \Rightarrow VALUE \Rrightarrow \digamma(ID)[ID\,,\, V\,,\, VALUE]$$

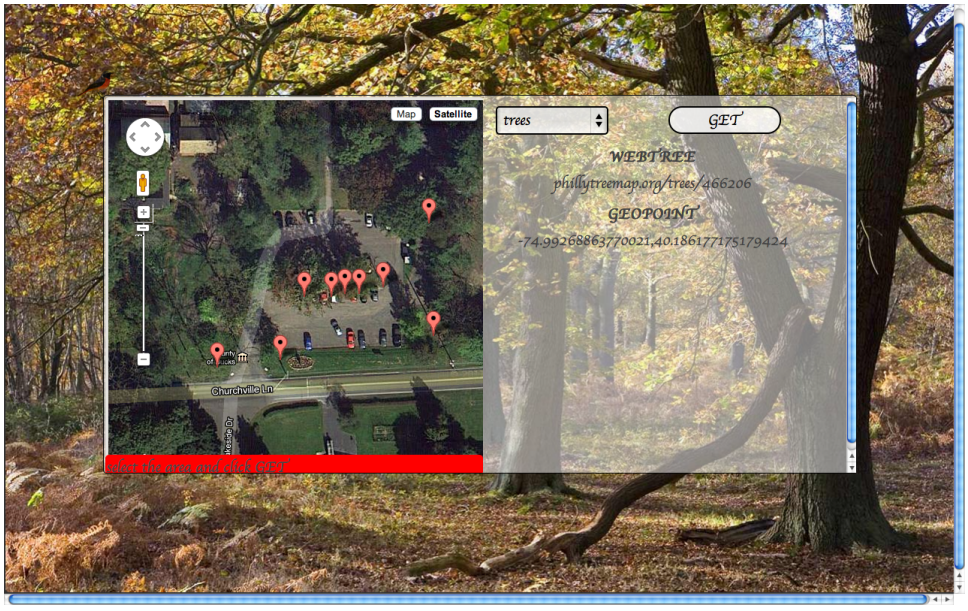


Figure 29: The Forest Cloud partial view of urban trees from Philladephia

3.7.3.2 Data Store The data store of the Forest Cloud is a database to consolidate tree information for a wide range of possible user applications. Given the global and collaborative nature of the Forest Cloud, the data store has an open character and a simple structure where the foresters from around the globe can upload their data about trees. The requirements to upload tree-data to the data store are twofold:

- the uploaded data is semantically described using values from common semantic spaces which are accessible from the Internet, and
- the dataset provides accurate geographical location values for the trees.

3.7.3.2.1 The Universe of Discourse and the Finite Nature of Tree Data The number of trees in the planet is finite and probably decreasing. Even though the ontological account of a tree is very clear, the biological definition of a single tree could bring difficulties at the time to identify single trees given into account that some trees species such as oaks and poplars can grow into colossal clonal colonies which form a large forestry area even though is only a single living organism. Therefore, for the Forest Cloud the tree concept is better understood as a main stem with its living crown that has standard marketable products. Under this construct, the number of trees of world is a finite number smaller than 100G[10] trees which is not a large number in computer-storage terms. Ergo, a data store for all trees of the world, beyond being of a finite nature, is plausible under the common current technological frameworks used to storage digital information. For instance, the World Data Centre for Climate has more than 6P[11] bytes of information stored in its servers. A simple arithmetic would say, that a similar data storage to the World Data Centre for Climate will provide 1K[12] bytes of storage per existent tree.

3.7.3.2.2 The triple nature of Tree primary data In the field of environmental sciences and management a datum represents the value of a property measured from one subject. Therefore every datum represents a conceptual logical sentence.

[10] Giga 10^9
[11] peta 10^{15}
[12] kilo 10^3

This ternary set often times has been modeled as subject-verb-object model. For instance, the RDF model widely used and recommended by the Open Web Consortium uses this semantic approach to conceptualize data.

Therefore, archives containing primary data can be seen as a simple collection of sentences which can be model using a sequence of triplets. As a clear example, a table of data is a concise way to represent sentences for primary data where the the first column represents the subjects, the first row represents the verbs, and a given cell represents the objects for a given subject and verb.

Therefore the data store of the Forest Cloud can be seen as a single big table of triplets where subjects are Internet resources that refer to trees, verbs are links to semantic definitions reachable on the Internet, and objects have a double structure: one that holds the value and the other links to a semantic space to describe the object-space. As the concise tables to represent sentences about data, the big table of tree-triplets can be seen as a big table with tree-records where each row represent a tree reference and the columns represents the properties that has been measured.

This conceptual approach to define trees create three disjunctive and general semantic spaces:

- the subject/thing space or space with references to real things ,
- the verb/conceptual space or space where properties are reachable from the Internet and are concisely described, and
- the object/ object space or space where the space and its dimensions are defined if required for units and accuracy

3.7.3.2.3 The Tree Properties and its Conceptual Spaces Since forestry is an old practice, many concepts have been defined to be measured on trees; for instance, the number of different approaches to estimate the volume of wood in standing trees is very large. Therefore, every one of the different approaches to estimate this volume should be the definition of different properties, and these definitions should be reachable using the internet link that identifies the property.

The big table of tree records has mandatory and non-mandatory properties. The mandatory properties, ergo, required for all rows in the big table are the

identification number, the geographical location , and the date the record was uploaded to the data store. In addition, since the domain of our data store is homogeneous and simple the semantic requirement are few but mandatory, and no data will be archived without metadata description. Moreover, many properties can be measured in a tree but not all are important to this data store based on the philosophy of the Forest Cloud; therefore, only properties that can help to identify individual trees are of importance for the Forest Cloud.

The semantic Spaces represent the network places where the definition of the properties reside. The characteristic required for the semantic spaces where the definitions resided are accuracy, persistence, Internet presence and well stablished.

For instance, Wikipedia represents a general, peer-reviewed, persistent, and popular dictionary on the Internet which can be extended to provided the needed semantic accuracy. Therefore, Wikipedia is a good semantic space to define verbs , objects and values needed for the Forest Cloud.

Other semantic spaces for the Forest Cloud that could be of utility to define some properties and which comply for the mentioned requirements are the Dublin Core space for authorship and related properties to characterize authorship.

The Spatial Reference Organization is an accurate, extensive, permanent and popular space to define geographical projections.

There are several semantic spaces to define biological taxa used to classify trees such as the Encyclopedia of Life which provide a RESTfull interface to link and define the taxon of the trees. This semantic space is supported for a large global community of researchers with trying to unify the often disperse and overlapping semiotics of taxonomic nomenclature.

3.7.3.2.4 The Objects and its Semantic Spaces The Objects are the last component of every triple in the data store which is represented by a pair: an object, and a structure that describes the dimensions of the object.

The objects used to represent the features of trees are common logical structures such as:

- categories,

- numbers ,
- structures,
- arrays
- polygons,
- matrices

In the forestry field, there are several common categorical spaces that are used to classify trees such as taxonomical or phytosociological spaces. The prolific Internet production of the biological community has populated the Internet with Internet resources that hold the definitions of all the names belonging to these categorical spaces; besides, they have provided these nominal values with permanent Internet links which facilitate the semantical definition of nominal values in metadata languages.

By the same token, numerical values are frequently used in forest inventories to represent tree features such as height, diameters or age. Structures can be used to define locational objects (latitude-longitude), color values (red-green-blue), or dates (year-month-day). Array structures are used to describe tree-ring series, bole taper, or genetic codes. Polygon structures are often used when using GIS tools to define location of trees; and matrix objects are used to represent images of trees, or cross-sections. The semantic definition of these abstract objects to be used to define trees attributes have been the object of discussion for ontologists who have provided a number of well known public ontologies that define them using permanent web spaces and resources. These internet resources can be used to semantically define the nature of the objects we need to describe the tree features.

All these former object types except the nominal ones, can be described further to provide a more accurate semantic definition of the values of the objects. In the first place, features measured with numerical objects often time have a possible range of values which is a subset of the general space such as the real line. Furthermore, the accuracy or resolution of the measurements also limits the possible space of the values of the object. Therefore, any numerical object has to be described not only in terms of its object space but also in terms of its possible range of values and its accuracy. For objects with several dimensions the same definitions have to to be declared for every dimension.

Finally, every object has to be provided with the appropriated units if the feature requires it. The Internet offers several public Web spaces where the common units used to measure tree primary data are defined. However, forestry is a field where derived features that are evaluated using functions that use values from other features are commonly used and the units of the derived features are complex and out of the scope of the common ontologies. Therefore, whenever derived data is used in the datasets, the forester should be able to provide or generate internet Links to web resources that define the units used for the derived features.

In general the process of semantic annotation is a fundamental operation in the open Internet, therefore, it is also an important process for the Forest Cloud. Fortunately, there are plenty of web resources that can be used to annotate the metadata section of MySON objects.

3.7.3.2.5 The Multiple, Spatial and Temporal Nature of Tree Records

The open and collaborative nature of the Forest Cloud has implications in the semantics of its data store and the underlying assumptions. Most data stores available work under the idea that every thing that is not stated on the database is false, and every name in the database that identifies a subject is unique. However, the data store of the Forest Cloud will not be able to work under these common assumptions because the Forest Cloud would be able to receive information from different agents about the same trees, and problably not having the exact same properties and values [208].

Record repetition is a classical error in database parlance, but in a collaborative data store multiple records characterizing the same tree are perfectly correct since they represent the different views that enriched the data store of the Forest Cloud. Therefore record redundancy is expected and should be treated differently than in normal databases since the redundancy is not going to be absolute and every record will have a minimum amount of unique information for the data store of the Forest Cloud.

Forest ground-based information is often collected repetitively under a time-based schedule because it is important to register the change in values of some tree properties such as DBH or tree height; also routinely airborne photogrammetrric inventories of trees are taken at different points of time to study forestry dynamics. Therefore, plausible records of the same tree must be expected in the data store of the Forest Cloud.

Spatial uncertainty is a common battle horse and an open source for complex errors which can puzzle experienced geographers since the number of measurement and operational errors are quite substantial. For instance, it is plausible that two tree records with similar geographical coordinates can be different trees, while other two records with significant disparities in their geographical coordinates could belong to the same tree.

Even though the common languages for querying and filtering tables with data could be used in this data store, the open nature of the information collected in the data store of the Forest Cloud will make the search for tree records more complex. Therefore, the development of expert systems would be a requirement to work with the data in the Forest Cloud.

3.7.3.2.6 Data Store Implementation The forest cloud has been implemented using a relational database with a relatively simple model as shown in figure(30). The data and metadata information for trees are stored as JSON streams. These two streams correspond with the both members of the MySON object. Because the relative simple structure of the model, simple classical SQL queries can be executed against the database. Nonetheless, complex general searches can be executed on the database for agents and expert systems that have knowledge to manipulate triplets from trees. However, this aspect of the data store has not being researched or implemented.

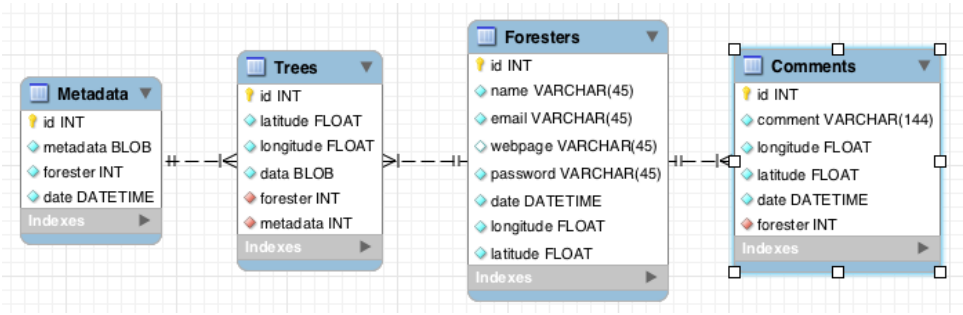


Figure 30: Data Store Entity-Relationship diagram

3.7.3.3 The Forest Cloud Services

3.7.3.3.1 Posting Tree Information Posting trees in the Forest Cloud has been made easy by using a metadata templating engine that can serve appropriate metadata based on the nature of the file that contains the tree information. By analyzing the structure and information of the data file, the Forest Cloud application can "a posteriori" serve a metadata template that describes the data file.

For grounded data files with context-free structures, the engine will be able to serve metadata templates that completely describe the data files. For non-grounded files, the engine will produce metadata templates that describe only the physical and logical level of the files while the forester will have to edit the semantic information of the metadata objects to completely describe the data files.

Once the files are completely described using the metadata object, the tree information from the data file can be uploaded and transformed into a MySON object that can be stored in the data store of the Forest Cloud.

This general process can be repeated for any type of file for which there could be a metadata template. Furthermore, because any data file in use by environmental scientists and managers has formal specifications, new metadata templates can be developed to suit the future requirements of the Forest Cloud for processing new types of files.

3.7.3.3.2 Getting Tree Information At this moment, the Forest Cloud offers a simple Web interface to retrieve the information of trees. However, the development of other more elaborated interfaces can be developed with ease since data has been well defined and stored. This new interfaces will have to deal with the development of expert systems that has not been the study of this work.

3.8 The Future of The Forest Cloud

The Forest Cloud has started with the goal of a collaborative effort to provide the mechanisms and ideas to create the first global open forest inventory which can help global and local managers to provide answers to key questions for a good governance of the trees. In order to achieve the goals the Forest Cloud, the Forest Cloud has to become popular, accurate, and versatile to get the respect of the general people and the forestry managers.

Given the technological advances, open Internet efforts, and the facilities of the Forest Cloud, collecting, transmitting and archiving tree data have been made easy. Therefore, the Forest Cloud will have to increase the number of on-the-ground local foresters around the globe to check and validate tree data with ground truth information.

In addition to on-the-ground foresters, the Forest Cloud has to have a large array of active scientists who should provide methods to facilitate key operations in the Forest Cloud. For instance, scientists with expertise in remote sensing and automatic tree delineation have to provide algorithms to analyze imagery, delineate tree crowns and even classified them by taxon. In addition, scientists with forensic expertise have to developed tree forensic methods which can be used to respond to key questions regarding tree identification which will help officials worldwide to enforce public rules regarding trees. Finally experts in information technology should create mechanisms to ease the connection between models and data by providing, for instance, a metadata language to describe models functionality and interfaces.

The Forest Cloud has to seek the collaboration of private and public organizations who can provide access to data from which the Forest Cloud scientists can retrieve tree information and feed the data store with tree information that could then be curated using ground truth evidence from the on-the-ground foresters of the Forest Cloud.

Given the fact that the open Internet requires common understanding of Internet standards, the Forest Cloud has to evolve to be able to link its data and application to the open Internet since the open access to data and applications is a general goal to pursue in collaborative environments. This will required that the scientists from the Forest Cloud have to be aware of the ever evolving open standards from the Internet.

As a collaborative and non-profit organization, it will have to find support from private and public groups that can provide resources that will support the operations and goals of the Forest Cloud. Financial, data, and technological resources will be required to equip the foresters and scientists of the Forest Cloud; for instance, to get access to satellite images, and hardware to collect biological fingerprints on the field will be of critical importance.

The forest footprint is decreasing by the day and it has been reported in many public forums; ergo, the habitat of many organisms which depend on the trees for their existence has been compromised to say the least. In addition,

deforestation and forest degradation have been regarded as an important source of CO2 which plays and important role in the general warming trend of the planets whose effects are for the most point of view non-desirable. Therefore, if we defend the trees, the trees will defend us.

References

[1]

[2] Excel binary file format (.xls) structure specification. This is an electronic document. Date of publication: [Date unavailable]. Date retrieved: November 29, 2011. Date last modified: [Date unavailable].

[3] Ogf standards: Data format description language (dfdl). This is an electronic document. Date of publication: [Date unavailable]. Date retrieved: August 10, 2011. Date last modified: [Date unavailable].

[4] Tree ring data description. This is an electronic document. Date of publication: [Date unavailable]. Date retrieved: November 29, 2011. Date last modified: [Date unavailable].

[5] Tiff revision 6.0. *Aldus Developers Desk, Seattle*, 1992.

[6] Shapefile technical description. White paper, Environmental Systems Research Institut, 1998.

[7] Pdf reference: Adobe portable document format version 1.3. 2000.

[8] R.S. Abate. Redd, white, and blue: Is proposed us climate legislation adequate to promote a global carbon credits system for avoided deforestation in a post-kyoto regime? *Tul. J. Int'l & Comp. L.*, 19:95–397, 2010.

[9] A.H. Abd-Elrahman, M.E. Thornhill, M.G. Andreu, and F. Escobedo. A community-based urban forest inventory using online mapping services and consumer-grade digital images. *International Journal of Applied Earth Observation and Geoinformation*, 12(4):249–260, 2010.

[10] F Achard, G Vaysseix, and E Barillot. XML, bioinformatics and data integration. *BIOINFORMATICS*, 17(2):115–125, FEB 2001.

[11] A. Agrawal and A. Chhatre. Explaining success on the commons: Community forest governance in the indian himalaya. *World Development*, 34(1):149–166, 2006.

[12] A.V. Aho, R. Sethi, and J.D. Ullman. Compilers: principles, techniques, and tools. *Reading, MA,*, 1986.

[13] M.S. Aktas, G.C. Fox, M. Pierce, and S. Oh. Xml metadata services. *Concurrency and Computation: Practice and Experience*, 20(7):801–823, 2008.

[14] H. Alani, D. Dupplaw, J. Sheridan, K. O'Hara, J. Darlington, N. Shadbolt, and C. Tullo. Unlocking the potential of public sector information with semantic web technology. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, pages 708–721. Springer-Verlag, 2007.

[15] J.C. Allen and D.F. Barnes. The causes of deforestation in developing countries. *Annals of the association of American Geographers*, 75(2):163–184, 1985.

[16] C. Anderson and M. Wolff. The web is dead. long live the internet. *Wired Magazine, August*, 17, 2010.

[17] D.P. Anderson. Boinc: A system for public-resource computing and storage. In *proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, pages 4–10. IEEE Computer Society, 2004.

[18] M. Antonioletti, M. Atkinson, R. Baxter, A. Borley, N.P. Chue Hong, B. Collins, N. Hardman, A.C. Hume, A. Knox, M. Jackson, et al. The design and implementation of grid database services in ogsa-dai. *Concurrency and Computation: Practice and Experience*, 17(2-4):357–376, 2005.

[19] M. Araujo, S. Kant, and L. Couto. Why brazilian companies are certifying their forests? *Forest Policy and Economics*, 11(8):579–585, 2009.

[20] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, et al. Above the clouds: A berkeley view of cloud computing. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28*, 2009.

[21] E. Aronova, K.S. Baker, and N. Oreskes. Big science and big data in biology: From the international geophysical year through the international biological program to the long term ecological research (lter) network, 1957-present. *Historical Studies in the Natural Sciences*, 40(2):183–224, 2010.

[22] T. K. Attwood. Genomics - the babel of bioinformatics. *SCIENCE*, 290(5491):471–473, Oct 2000.

[23] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[24] D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, and R. Evans. Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. 2001.

[25] H. .. Balen. Deconstructing babel: Xml and application integration. *Application Development Trends*, 7(12):19–23, December 2000.

[26] J.C. Balouet, K.T. Smith, D. Vroblesky, and G. Oudijk. Use of dendrochronology and dendrochemistry in environmental forensics: does it meet the daubert criteria? *Environmental Forensics*, 10(4):268–276, 2009.

[27] E. Barillot and F. Achard. Xml: a lingua franca for science? *Trends in Biotechnology*, 18(8):331–333, 2000.

[28] J. .. Beall. The death of metadata. *Serials Librarian*, vol.51, no.2:55–74, 2006.

[29] A.L. Beberg, D.L. Ensign, G. Jayachandran, S. Khaliq, and V.S. Pande. Folding@ home: Lessons from eight years of volunteer distributed computing. 2009.

[30] P.H. Beckman. Building the teragrid. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1833):1715, 2005.

[31] F. Berkes, D. Feeny, B.J. McCay, and J.M. Acheson. The benefits of the commons. *Nature*, 340(6229):91–93, 1989.

[32] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M.Q. Dang, and K. Pentikousis. Energy-efficient cloud computing. *The Computer Journal*, 53(7):1045, 2010.

[33] T. Berners-Lee. Www: Past, present, and future. *Computer*, 29(10):69–77, 1996.

[34] T. Berners-Lee. Axioms of web architecture: Metadata, january 1997. *URL http://www. w3. org/DesignIssues/Metadata. Updated Sept*, 2000.

[35] T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[36] T. Berners-Lee, D. Connolly, L. Kagal, Y. Scharf, and J. Hendler. N3logic: A logical framework for the world wide web. *Theory and Practice of Logic Programming*, 8(3):249–269, 2008.

[37] T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, and D.J. Weitzner. Creating a science of the web. *Science*, 313(5788):769, 2006.

[38] T. Berners-Lee and J. Hendler. Scientific publishing on the semantic web. *Nature*, 410:1023–1024, 2001.

[39] S. Bernstein and B. Cashore. Nonstate global governance: Is forest certification a legitimate alternative to a global forest convention? *Hard Choices, Soft Law: Voluntary Standards in Global Trade, Environment and Social Governance*, pages 33–63, 2004.

[40] FA Bisby. The quiet revolution: Biodiversity informatics and the internet. *SCIENCE*, 289(5488):2309–2312, SEP 29 2000.

[41] D. Brack and C. House. Controlling illegal logging: Lessons from the us lacey act. *Chatham House: Energy, Environment and Development Programme*, 2007.

[42] S. Bratkovich, J. Bowyer, S. Stai, K. Fernholz, J. Bratkovich, S. Howe, J. Bratkovich, J. Bowyer, S. Stai, K. Fernholz, J. Bratkovich, and J. Howe. Urban forestry: an evolving discipline. 2010.

[43] M. Brunner, R. Eugster, E. Trenka, and L. Bergamin-Strotz. Ft-nir spectroscopy and wood identification. *Holzforschung-International Journal of the Biology, Chemistry, Physics and Technology of Wood*, 50(2):130–134, 1996.

[44] R. Bryant, R.H. Katz, and E.D. Lazowska. Big-data computing: Creating revolutionary breakthroughs in commerce, science and society. *Computing Research Initiatives for the 21st Century. Computing Research Association*, 2008.

[45] P. Buffers. Protocol buffers. google's data interchange format.

[46] C. Burks. Molecular biology database list. *Nucleic acids research*, 27(1):1, 1999.

[47] A. Callado, C. Kamienski, G. Szabo, B.P. Gero, J. Kelner, S. Fernandes, and D. Sadok. A survey on internet traffic identification. *Communications Surveys & Tutorials, IEEE*, 11(3):37–52, 2009.

[48] JT CALLAHAN. LONG-TERM ECOLOGICAL RESEARCH. *BIOSCIENCE*, 34(6):363–367, 1984.

[49] F. Cappello, S. Djilali, G. Fedak, T. Herault, F. Magniette, V. Néri, and O. Lodygensky. Computing on large-scale distributed systems: Xtremweb architecture, programming models, security, tests and convergence with grid. *Future Generation Computer Systems*, 21(3):417–437, 2005.

[50] A. Carleer and E. Wolff. Exploitation of very high resolution satellite data for tree species identification. *Photogrammetric engineering and remote sensing*, 70(1):135–140, 2004.

[51] B. Cashore. Legitimacy and the privatization of environmental governance: How non–state market–driven (nsmd) governance systems gain rule–making authority. *Governance*, 15(4):503–529, 2002.

[52] B. Cashore and M. Stone. Can legality verfiication rescue global forest governance? assessing interacting effects of economic mechanism of forestry policy and governance: Lessons learnt from southeast asia. *Retrieved November*, 16:2010, 2010.

[53] A.D. Chandler. *Scale and scope: the dynamics of industrial capitalism*. Belknap Pr, 1994.

[54] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber. Bigtable: A distributed

storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.

[55] AD Chave, M. Arrott, C. Farcas, E. Farcas, I. Krueger, M. Meisinger, JA Orcutt, FL Vernon, C. Peach, O. Schofield, et al. Cyberinfrastructure for the us ocean observatories initiative: Enabling interactive observation in the ocean. In *OCEANS 2009-EUROPE*, pages 1–10. IEEE.

[56] A. Chhatre and A. Agrawal. Forest commons and local enforcement. *Proceedings of the National Academy of Sciences*, 105(36):13286, 2008.

[57] A.B. Chimeli, R.G. Boyd, and D.M. Adams. International timber markets and tropical deforestation: the evidence from prices. *Applied Economics*, 99999(1):1–12, 2011.

[58] S. Chiorescu and A. Gr
"onlund. The fingerprint method: Using over-bark and under-bark log measurement data generated by three-dimensional log scanners in combination with radiofrequency identification tags to achieve traceability in the log yard at the sawmill. *Scandinavian journal of forest research*, 19(4):374–383, 2004.

[59] G. Chirici, D. Giuliarelli, D. Biscontini, D. Tonti, W. Mattioli, M. Marchetti, and P. Corona. Large-scale monitoring of coppice forest clearcuts by multitemporal very high resolution satellite imagery. a case study from central italy. *Remote Sensing of Environment*, 2011.

[60] N.E. Clark. 3d reconstruction of a tree stem using video images and pulse distances. In *Symposium on Statistics and Information Technology in Forestry*, pages 8–12.

[61] The Federal Geographic Data Committee. Content standard for digital geospatial metadata content standard for digital geospatial metadata content standard for digital geospatial metadata content standard for digital geospatial metadata content standard for digital geospatial metadata content standard for digital geospatial metadata. *Concurrency and Computation: Practice and Experience*, 20(7):801–823, 1998.

[62] H.A. Core, W.A. Côté, and A.C. Day. Wood structure and identification.[syracuse wood science series 6.]. *Syracuse Univ. Press.*, 182:2–3, 1979.

[63] Mark J. Costello and Edward Vanden Berghe. 'Ocean biodiversity informatics': a new era in marine biology research and management. *MARINE ECOLOGY-PROGRESS SERIES*, 316:203–214, 2006.

[64] M.J. Costello. Motivating online publication of data. *BioScience*, pages 418–427, 2009.

[65] Fred Cowett. Village of cayuga heights street tree inventory (shapefile : 2009).

[66] M.H. Cragin, C.L. Palmer, J.R. Carlson, and M. Witt. Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1926):4023, 2010.

[67] D. Crockford. Json: The fat-free alternative to xml. In *Proc. of XML*, volume 2006, 2006.

[68] F. Dahdouh-Guebas and N. Koedam. Long-term retrospection on mangrove development using transdisciplinary approaches: A review. *AQUATIC BOTANY*, 89(2):80–92, 2008.

[69] R. A. Daley, E. A. Immer, B. I. Fortner, and M. B. Weiss. Scientific resource access system: A concept for getting "living with a star" information to do science. *JOHNS HOPKINS APL TECHNICAL DIGEST*, 26(1):22–35, 2005.

[70] O. Damette and P. Delacote. Unsustainable timber harvesting, deforestation and the role of certification. *Ecological Economics*, 2011.

[71] G.S. Dangl, J. Yang, D.A. Golino, and T. Gradziel. A practical method for almond cultivar identification and parental analysis using simple sequence repeat markers. *Euphytica*, 168(1):41–48, 2009.

[72] L.S. Davis and G. Liu. Integrated forest planning across multiple ownerships and decision makers. *Forest science*, 37(1):200–226, 1991.

[73] H.F. Deus, R. Stanislaus, D.F. Veiga, C. Behrens, I.I. Wistuba, J.D. Minna, H.R. Garner, S.G. Swisher, J.A. Roth, A.M. Correa, et al. A semantic web management model for integrative biomedical informatics. *PloS one*, 3(8):e2946, 2008.

[74] J. Domingue, C. Pedrinaci, M. Maleshkova, B. Norton, and R. Krummenacher. Fostering a relationship between linked data and the internet of services. *The future internet*, pages 351–364, 2011.

[75] J. Douglas and M. Simula. Deforestation: Causes and symptoms. *The Future of the World's Forests*, pages 81–104, 2010.

[76] S. Doveri, F. Sabino Gil, A. Diaz, S. Reale, M. Busconi, A. da Câmara Machado, A. Martín, C. Fogher, P. Donini, and D. Lee. Standardization of a set of microsatellite markers for use in cultivar identification studies in olive (olea europaea l.). *Scientia Horticulturae*, 116(4):367–373, 2008.

[77] S. Draghici, S. Sellamuthu, and P. Khatri. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *BIOINFORMATICS*, 22(23):2934–2939, Dec 2006.

[78] PD EAGAN and SJ VENTURA. ENHANCING VALUE OF ENVIRONMENTAL DATA - DATA LINEAGE REPORTING. *JOURNAL OF ENVIRONMENTAL ENGINEERING-ASCE*, 119(1):5–16, JAN-FEB 1993.

[79] JL Edwards, MA Lane, and ES Nielsen. Interoperability of biodiversity databases: Biodiversity information on every desktop. *SCIENCE*, 289(5488):2312–2314, SEP 29 2000.

[80] A. Eliassen. Vilhelm bjerknes and his students. *Annual Review of Fluid Mechanics*, 14(1):1–12, 1982.

[81] G. Erbach. Data-centric view in e-science information systems. *Data Science Journal*, 5(0):219–222, 2006.

[82] J. Esparza and T. Yamada. The discovery value of "big science". *The Journal of experimental medicine*, 204(4):701, 2007.

[83] LG Esteban, F.G. Fernández, P. Palacios de Palacios, RM Romero, and N.N. Cano. Artificial neural networks in wood identification: the case of two juniperus species from the canary islands. *IAWA Journal*, 30(1):87–94, 2009.

[84] D.C. Esty. Toward optimal environmental governance. *NYUL Rev.*, 74:1495, 1999.

[85] D.C. Esty and M.H. Ivanova. The road ahead: Conclusions and action agenda. *Esty, DC and Ivanova, MH Global Environmental Governance: Options & Opportunities*, pages 225–240, 2002.

[86] M.J. Falkowski, M.A. Wulder, J.C. White, and M.D. Gillis. Supporting large-area, sample-based forest inventories with very high spatial resolution satellite imagery. *Progress in Physical Geography*, 33(3):403, 2009.

[87] F. Fischer, G. Unel, B. Bishop, and D. Fensel. Towards a scalable, pragmatic knowledge representation language for the web. *Perspectives of Systems Informatics*, pages 124–134, 2010.

[88] K. Fisher and D. Walker. The pads project: An overview. In *Proceedings of the 14th International Conference on Database Theory*, pages 11–17. ACM, 2011.

[89] Food and Agriculture Organization of the United Nations. *Global forest resources assessment 2005: progress towards sustainable forest management*. Food and Agriculture Organization of the United Nations, 2006.

[90] JF FRANKLIN, CS BLEDSOE, and JT CALLAHAN. CONTRIBUTIONS OF THE LONG-TERM ECOLOGICAL RESEARCH-PROGRAM - AN EXPANDED NETWORK OF SCIENTISTS, SITES, AND PROGRAMS CAN PROVIDE CRUCIAL COMPARATIVE ANALYSES. *BIOSCIENCE*, 40(7):509–523, JUL-AUG 1990.

[91] J.E.F. Friedl. *Mastering regular expressions*. O'Reilly Media, Inc., 2006.

[92] L.M. Friedman. Business and legal strategies for combating grey-market imports. In *Int'l L.*, volume 32, page 27. HeinOnline, 1998.

[93] M. Gadgil, F. Berkes, and C. Folke. Indigenous knowledge for biodiversity conservation. *Ambio*, pages 151–156, 1993.

[94] C. Germain, V. Neri, G. Fedak, and F. Cappello. Xtremweb: building an experimental platform for global computing. *Grid Computing—GRID 2000*, pages 107–129, 2000.

[95] C. Gibson. Forest resources: Institutions for local governance in guatemala. *Protecting the commons: A framework for resource management in the Americas*, pages 71–89, 2001.

[96] C.C. Gibson. *People and forests: communities, institutions, and governance*. The MIT Press, 2000.

[97] C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–693, 2008.

[98] P. GODFREY-SMITH. Explanation in evolutionary biology: Comments on fodor. *Mind & Language*, 23(1):32–41, 2008.

[99] ME Goodchild. Geographic information science and systems for environmental management. *ANNUAL REVIEW OF ENVIRONMENT AND RESOURCES*, 28:493–519, 2003.

[100] ME Goodchild and JY Zhou. Finding geographic information: Collection-level metadata. *GEOINFORMATICA*, 7(2):95–112, JUN 2003.

[101] M.F. Goodchild. Different data sources and diverse data structures: metadata and other solutions. In *Geocomputation: a primer*, pages 61–73, 1998.

[102] D.L. Goodhue, J.A. Quillard, and J.F. Rockart. Managing the data resource: a contingency perspective. *MIS Quarterly*, pages 373–392, 1988.

[103] A. Gore. Bringing information to the world: The global information infrastructure. *Harv. JL & Tech.*, 9:1, 1996.

[104] V. Grimm and C. Wissel. Babel, or the ecological stability discussions: An inventory and analysis of terminology and a guide for avoiding confusion. *OECOLOGIA*, 109(3):323–334, Feb 1997.

[105] S. Gronas. Vilhelm bjerknes' vision for scientific weather prediction. *The Nordic seas: an integrated perspective: oceanography, climatology, biogeochemistry, and modeling*, 158:357, 2005.

[106] N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. *Towards Very Large Knowledge Bases Knowledge Building and Knowledge Sharing*, 1(9):25–32, 1995.

[107] N. Guarino and Istituto (Roma) Consiglio nazionale delle ricerc. Formal ontology in information systems, 1998.

[108] L.H. Gulbrandsen. Overlapping public and private governance: Can forest certification fill the gaps in the global forest regime? *Global Environmental Politics*, 4(2):75–99, 2004.

[109] A. Gupta. Transparency under scrutiny: information disclosure in global environmental governance. *Global Environmental Politics*, 8(2):1–7, 2008.

[110] R. Gyimah. Piloting wood tracking system in ghana. In *Precision Forestry Symposium Stellenbosch, South Africa 1 to 3 March 2010*, 2010.

[111] SS Hale, MM Hughes, JF Paul, RS McAskill, SA Rego, DR Bender, NJ Dodge, TL Richter, and JL Copeland. Managing scientific data: The EMAP approach. *ENVIRONMENTAL MONITORING AND ASSESSMENT*, 51(1-2):429–440, JUN 1998.

[112] SS Hale, AH Miglarese, MP Bradley, TJ Belton, LD Cooper, MT Frame, CA Friel, LM Harwell, RE King, WK Michener, DT Nicolson, and BG Peterjohn. Managing troubled data: Coastal data partnerships smooth data integration. *ENVIRONMENTAL MONITORING AND ASSESSMENT*, 81(1-3):133–148, JAN-FEB 2003.

[113] B. Hammersley. *Developing feeds with RSS and Atom*. O'Reilly Media, Inc., 2005.

[114] T. Hammond, T. Hannay, and B. Lund. The role of rss in science publishing: Syndication and annotation on the web. *D-Lib Magazine*, 10(12), 2004.

[115] B. Hanson, A. Sugden, and B. Alberts. Making data maximally available. *Science*, 331(6018):649, 2011.

[116] A.I. Hapca, F. Mothe, and J.M. Leban. A digital photographic method for 3d reconstruction of standing tree shape. *Annals of Forest Science*, 64(6):631–637, 2007.

[117] G. HARDIN. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

[118] S. Hastings, M. Ribeiro, S. Langella, S. Oster, U. Catalyurek, T. Pan, K. Huang, R. Ferreira, J. Saltz, and T. Kurc. Xml database support for distributed execution of data-intensive scientific workflows. *ACM SIGMOD Record*, 34(3):50–55, 2005.

[119] M Heemskerk, K Wilson, and M Pavao-Zuckerman. Conceptual models as tools for communication across disciplines. *CONSERVATION ECOLOGY*, 7(3), DEC 2003.

[120] S. Henley. The problem of missing data in geoscience databases. *Computers & geosciences*, 32(9):1368–1377, 2006.

[121] A. .. Hill. The xml revolution. *Financial Technology*, 1(14):5–6, April 2002.

[122] M. Hu, E.P. Lim, A. Sun, H.W. Lauw, and B.Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management*, pages 243–252. ACM, 2007.

[123] W. Huang and D. Webster. Intelligent rss news aggregation based on semantic contexts. In *ACM SIGIR 2004 Workshop on Information Retrieval in Context*, pages 40–43, 2004.

[124] B.A. Huberman and L.A. Adamic. Growth dynamics of the world wide web. *J. Reprod. Fertil*, 9:55–60, 1993.

[125] J.L. Innes. Madagascar rosewood, illegal logging and the tropical timber trade. *Madagascar Conservation & Development*, 5(1), 2010.

[126] M.B. Jadhav, V.J. Gaikwad, CV Patil, and GS Deshpande. Cloud computing applications in computational science. *International Journal of Computer Science and Applications*, 1(1):1–6, 2010.

[127] C. .. Jaenicke. Server-side xml: taming the tower of babel, 1998.

[128] P. Jain, P. Hitzler, P.Z. Yeh, K. Verma, and A.P. Sheth. Linked data is merely more data. *Linked Data Meets Artificial Intelligence*, 2009.

[129] CH Jakobsen, T Hels, and WJ McLaughlin. Barriers and facilitators to integration among scientists in transdisciplinary landscape analyses: a cross-country comparison. *FOREST POLICY AND ECONOMICS*, 6(1):15–31, JAN 2004.

[130] W. Janssen and P. Goldsworthy. Multidisciplinary research for natural resource management: conceptual and practical implications. *Agricultural systems*, 51(3):259–279, 1996.

[131] D.H. Janzen, W. Hallwachs, P. Blandin, J.M. Burns, J. CADIOU, I. Chacon, T. Dapkey, A.R. Deans, M.E. Epstein, B. Espinoza, et al. Integration of dna barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, 9:1–26, 2009.

[132] S. Jiang, L. Fang, and X. Huang. An idea of special cloud computing in forest pests' control. *Cloud Computing*, pages 615–620, 2009.

[133] Matthew B. Jones, Mark P. Schildhauer, O. J. Reichman, and Shawn Bowers. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *ANNUAL REVIEW OF ECOLOGY EVOLUTION AND SYSTEMATICS*, 37:519–544, 2006.

[134] E.C. Kansa and A. Bissell. Web syndication approaches for sharing primary data in" small science" domains. *Data Science Journal*, (0):1006280237, 2010.

[135] H. Karasti, K.S. Baker, and E. Halkola. Enriching the notion of data curation in e-science: data managing and information infrastructuring in the long term ecological research (lter) network. *Computer Supported Cooperative Work (CSCW)*, 15(4):321–358, 2006.

[136] J.E. Keeley, CJ Fotheringham, and M. Morais. Reexamining fire suppression impacts on brushland fire regimes. *Science*, 284(5421):1829, 1999.

[137] A. KEMPF. Advocating global forest issues on the internet. In *International online information meeting*, pages 99–105, 1996.

[138] R.W. Kimmerer and F.K. Lake. The role of indigenous burning in land management. *Journal of Forestry*, 99(11):36–41, 2001.

[139] B. M. Knoppers and M. Saginur. The babel of genetic data terminology. *NATURE BIOTECHNOLOGY*, 23(8):925–927, Aug 2005.

[140] L. Koivuranta, K. Leinonen, and P. Pulkkinen. Marketing of forest reproductive material: the use of microsatellites for identification of registered tree clones in finland. *Finnish For Res Inst, Vantaa*, 2008.

[141] G. Koloniari and E. Pitoura. Peer-to-peer management of xml data: issues and research challenges. *ACM Sigmod Record*, 34(2):6–17, 2005.

[142] S. Korten and C. Kaul. Application of rfid (radio frequency identification) in the timber supply chain. *Croatian Journal of Forest Engineering*, 29(1):85–94, 2008.

[143] A.T. Kralidis. Geospatial open source and open standards convergences. *Open Source Approaches in Spatial Data Handling*, pages 1–20, 2008.

[144] S. Krech. *The ecological Indian: myth and history.* WW Norton & Company, 2000.

[145] M. Krotzsch, D. Vrandecic, M. Volkel, H. Haller, and R. Studer. Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):251–261, 2007.

[146] J. Lalonde, N. Vandapel, and M. Hebert. Automatic three-dimensional point cloud processing for forest inventory. *Robotics Institute*, page 334, 2006.

[147] A. Langegger, W. W
"oß, and M. Bl

"ochl. A semantic web middleware for virtual data integration on the web. *The Semantic Web: Research and Applications*, pages 493–507, 2008.

[148] F. Laroui and BCC Van der Zwaan. Environment and multidisciplinarity three examples of avoidable confusion. *Integrated Assessment*, 3(4):360–369, 2002.

[149] C.A. Lee. A perspective on scientific cloud computing. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pages 451–459. ACM, 2010.

[150] S. Lélé and R.B. Norgaard. Practicing interdisciplinarity. *BioScience*, 55(11):967–975, 2005.

[151] R.M. Lerner. At the forge: syndication with rss. *Linux Journal*, 2004(126):8, 2004.

[152] RA Letcher, BFW Croke, and AJ Jakeman. Integrated assessment modelling for water resource allocation and management: A generalised conceptual framework. *Environmental Modelling and Software*, 22(5):733–742, 2007.

[153] J. Lewin. An introduction to rss news feeds: Using open formats for content syndication. *This article includes sample code that demonstrates elements of an RSS file*, 2000.

[154] D. Liddick. *Crimes Against Nature: Illegal Industries and the Global Environment*. Praeger Publishers, 2011.

[155] R.D. Lipschutz. Why is there no international forestry law: An examination of international forestry regulation, both public and private. *UCLA J. Envtl. L. & Pol'y*, 19:153, 2000.

[156] P. Longley. *Geographic information systems and science*. John Wiley & Sons Inc, 2005.

[157] F. López-Muñoz, C. Alamo, F.J. Quintero-Gutiérrez, and P. García-García. A bibliometric study of international scientific productivity in attention-deficit hyperactivity disorder covering the period 1980–2005. *European Child & Adolescent Psychiatry*, 17(6):381–391, 2008.

[158] F. López-Muñoz, J. Boya, F. Marín, and J.L. Calvo. Scientific research on the pineal gland and melatonin: a bibliometric study for the period 1966-1994. *Journal of Pineal Research*, 20(3):115–124, 1996.

[159] M. Lutz. *Programming python*. O'Reilly Media, Inc., 2011.

[160] H.G. Maas, A. Bienert, S. Scheller, and E. Keane. Automatic forest inventory parameter determination from terrestrial laser scanner data. *International Journal of Remote Sensing*, 29(5):1579–1593, 2008.

[161] A. Maedche, S. Staab, N. Stojanovic, R. Studer, and Y. Sure. Seal—a framework for developing semantic web portals. *Advances in Databases*, pages 1–22, 2001.

[162] JJ MAGNUSON. LONG-TERM ECOLOGICAL RESEARCH AND THE INVISIBLE PRESENT - UNCOVERING THE PROCESSES HIDDEN BECAUSE THEY OCCUR SLOWLY OR BECAUSE EFFECTS LAG YEARS BEHIND CAUSES. *BIOSCIENCE*, 40(7):495–501, JUL-AUG 1990.

[163] J. Mallet and K. Willmott. Taxonomy: renaissance or tower of babel? *TRENDS IN ECOLOGY & EVOLUTION*, 18(2):57–59, 2003.

[164] B.L. Markham, J.C. Storey, D.L. Williams, and J.R. Irons. Landsat sensor performance: history and current status. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(12):2691–2694, 2004.

[165] J. Markwell and D.W. Brooks. "link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology*. *Biochemistry and Molecular Biology Education*, 31(1):69–72, 2003.

[166] M. Marsik, F.R. Stevens, and J. Southworth. Amazon deforestation: Rates and patterns of land cover change and fragmentation in pando, northern bolivia, 1986 to 2005. *Progress in Physical Geography*, 2011.

[167] M.Z. Martin, S.D. Wullschleger, C.T. Garten, A.V. Palumbo, and J.G. Smith. Elemental analysis of environmental and biological samples using laser-induced breakdown spectroscopy and pulsed raman spectroscopy. *Journal of dispersion science and technology*, 25(5):687–694, 2005.

[168] F. MARTIN-RECUERDA. Application integration using conceptual spaces (cspaces). *Lecture notes in computer science*, pages 234–248, 2006.

[169] A. Marx and D. Cuypers. Forest certification as a global environmental governance tool: What is the macro-effectiveness of the forest stewardship council? *Regulation & Governance*, 4(4):408–434, 2010.

[170] S. Matsuoka, S. Shinjo, M. Aoyagi, S. Sekiguchi, H. Usami, and K. Miura. Japanese computational grid research project: Naregi. *Proceedings of the IEEE*, 93(3):522–533, 2005.

[171] William K. Michener. Meta-information concepts for ecological data management. *ECOLOGICAL INFORMATICS*, 1(1):3–7, JAN 2006.

[172] WK Michener, JW Brunt, JJ Helly, TB Kirchner, and SG Stafford. Nongeospatial metadata for the ecological sciences. *ECOLOGICAL APPLICATIONS*, 7(1):330–342, FEB 1997.

[173] S.P. Mirashe and NV Kalyankar. Cloud computing. *Arxiv preprint arXiv:1003.4074*, 2010.

[174] M. Mirtl. Introducing the next generation of ecosystem research in europe: Lter-europe's multi-functional and multi-scale approach. *Long-Term Ecological Research*, pages 75–93, 2010.

[175] Yasuko Mizoguchi, Akira Miyata, Yoshikazu Ohtani, Ryuichi Hirata, and Satoko Yuta. A review of tower flux observation sites in Asia. *JOURNAL OF FOREST RESEARCH*, 14(1):1–9, FEB 2009.

[176] B. Muys, J. Hynynen, M. Palahí, MJ Lexer, M. Fabrika, H. Pretzsch, F. Gillet, E. Briceño, GJ Nabuurs, and V. Kint. Simulation tools for decision support to adaptive forest management in europe. *Forest Systems*, 19:86–99, 2010.

[177] J. Myers and A. Chappell. Binary format description language. Technical report, Technical report, Pacific Northwest National Laboratory, 2003.

[178] LE Nave and C. Swanston. Big data for big questions: Global soil change and the national soil carbon network. In *AGU Fall Meeting Abstracts*, volume 1, page 06, 2010.

[179] J.R. Neilon. Communications in meteorology. *Geoscience Electronics, IEEE Transactions on*, 8(2):101–105, 1970.

[180] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. The eucalyptus open-source cloud-computing system. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, pages 124–131. IEEE Computer Society, 2009.

[181] K. O'Hara, T. Berners-Lee, W. Hall, and N. Shadbolt. Use of the semantic web in e-research. *Dutton, WH, and Jeffreys, PW (eds), World Wide Research: Reshaping the Sciences and Humanities*, 2010.

[182] A.T. OLIVEIRA FILHO and J.A. Ratter. Database: Woody flora of 106 forest areas of eastern tropical south america. *Occasional Monographs, Royal Botanic Garden Edinburgh, Edinburgh.[Links]*, 1995.

[183] E. Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, 1999.

[184] E. Ostrom, R. Gardner, and J. Walker. *Rules, games, and common-pool resources*. Univ of Michigan Pr, 1994.

[185] Y.O. Ouma, TG Ngigi, and R. Tateishi. On the optimization and selection of wavelet texture for feature extraction from high-resolution satellite imagery with application towards urban-tree delineation. *International journal of remote sensing*, 27(1):73–104, 2006.

[186] P.R. Oyono, J.C. Ribot, and A.M. Larson. Green and black gold in rural cameroon: natural resources for local governance, justice and sustainability. 2006.

[187] G. Paglia and M. Morgante. Pcr-based multiplex dna fingerprinting techniques for the analysis of conifer genomes. *Molecular breeding*, 4(2):173–177, 1998.

[188] MA Palmer, ES Bernhardt, EA Chornesky, SL Collins, AP Dobson, CS Duke, BD Gold, RB Jacobson, SE Kingsland, RH Kranz, MJ Mappin, ML Martinez, F Micheli, JL Morse, ML Pace, M Pascual,

SS Palumbi, O Reichman, AR Townsend, and MG Turner. Ecological science and sustainability for the 21st century. *FRONTIERS IN ECOLOGY AND THE ENVIRONMENT*, 3(1):4–11, FEB 2005.

[189] A. Pantze, A.H. Krantz, J.E.S. Fransson, H. Olsson, M. Santoro, L.E.B. Eriksson, and L.M.H. Ulander. Mapping and monitoring clear-cuts in swedish forest using alos palsar satellite images. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*, volume 3, pages III–589. IEEE, 2009.

[190] J.N. Parker. *Organizational collaborations and scientific integration: The case of ecology and the social sciences*. PhD thesis, ARIZONA STATE UNIVERSITY, 2006.

[191] J. Peacock, TR Baker, SL Lewis, G. Lopez-Gonzalez, and OL Phillips. The rainfor database: monitoring forest biomass and dynamics. *Journal of Vegetation Science*, 18(4):535–542, 2007.

[192] J. Phelps, E.L. Webb, and A. Agrawal. Does redd+ threaten to recentralize forest governance? *Science*, 328(5976):312, 2010.

[193] D.J.S. Price. Little science, big science. *New York*, pages 87–89, 1963.

[194] Hui-Je Qiao, Yan Han, Nuo Li, and Li-Qiang Ji. A model of biodiversity information integration. *Shengwu Duoyangxing*, 12(5), September 2004.

[195] I. Rafols and M. Meyer. How cross-disciplinary is bionanotechnology? explorations in the specialty of molecular motors. *SCIENTOMETRICS*, 70(3):633–650, 2007.

[196] R. Ramachandran, S. Graves, H. Conover, and K. Moe. Earth science markup language (esml):::: a solution for scientific data-application interoperability problem. *Computers & Geosciences*, 30(1):117–124, 2004.

[197] R. Ramakrishnan and J. Gehrke. *Database Management Systems: International Edition*. McGraw-Hill, 2003.

[198] D. M. Rapoport, I. Ayappa, R. G. Norman, and S. T. Herman. Npsg data interchange-dealing with the tower of babel. *SLEEP*, 29(5):599–600, 2006.

[199] S. Rhodes. Breaking down link rot: The chesapeake project legal information archive's examination of url stability. *Law Libr. J.*, 102:581–705, 2010.

[200] L. Richardson and S. Ruby. *RESTful web services*. O'Reilly Media, 2007.

[201] B.P. Rimal, E. Choi, and I. Lumb. A taxonomy and survey of cloud computing systems. In *2009 Fifth International Joint Conference on INC, IMS and IDC*, pages 44–51. IEEE, 2009.

[202] L. Roberts. The arpanet and computer networks. In *Proceedings of the ACM Conference on The history of personal workstations*, pages 51–58. ACM, 1986.

[203] L.G. Roberts. Beyond moore's law: Internet growth trends. *Computer*, 33(1):117–119, 2000.

[204] P.M. Rodkewich and J. Albrecht. Forestry databases on the internet. *Journal of agricultural & food information*, 3(3):19–23, 1996.

[205] J. Rossmann, M. Schluse, C. Schlette, A. Buecken, P. Krahwinkler, and M. Emde. Realization of a highly accurate mobile robot system for multi purpose precision forestry applications. In *Advanced Robotics, 2009. ICAR 2009. International Conference on*, pages 1–6. IEEE, 2009.

[206] O. Ruebenacker and M. Blinov. Using views of systems biology cloud: application for model building. *Theory in Biosciences*, pages 1–10, 2010.

[207] E. Russell and J. Kane. The missing link: Assessing the reliability of internet citations in history journals. *Technology and Culture*, 49(2):420–429, 2008.

[208] S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Prentice hall, 2010.

[209] S.J. Russell, P. Norvig, J.F. Canny, J.M. Malik, and D.D. Edwards. *Artificial intelligence: a modern approach*, volume 74. Prentice hall Englewood Cliffs, NJ, 1995.

[210] SK Saari, CD Campbell, J. Russell, IJ Alexander, and IC Anderson. Pine microsatellite markers allow roots and ectomycorrhizas to be linked to individual trees. *New phytologist*, 165(1):295–304, 2005.

[211] J. Saez-Rodriguez, A. Goldsipe, J. Muhlich, L.G. Alexopoulos, B. Millard, D.A. Lauffenburger, and P.K. Sorger. Flexible informatics for linking experimental data to mathematical models via datarail. *Bioinformatics*, 24(6):840, 2008.

[212] I. San Gil and K. Baker. The ecological metadata language milestones, community work force, and change. *Databits*, pages 4–7, 2007.

[213] P.A. Schad, L.R. Mobley, and C.M. Hamilton. Building a biomedical cyberinfrastructure for collaborative research. *American Journal of Preventive Medicine*, 40(5):S144–S150, 2011.

[214] D.H. Schepers. Challenges to legitimacy at the forest stewardship council. *Journal of business ethics*, 92(2):279–290, 2010.

[215] K. M. Schoeffler and W. E. Hammond. Standards for the electronic health record emerging from health care's tower of babel. *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, pages 388–392, 2001.

[216] E Schoenberger. Interdisciplinarity and social power. *PROGRESS IN HUMAN GEOGRAPHY*, 25(3):365–382, SEP 2001.

[217] B. Schueler, S. Sizov, and S. Staab. Management of meta knowledge for rdf repositories. 2007.

[218] A.R. Shaikh, P. Das, C.A. Vinson, and B. Spring. Cyberinfrastructure for consumer health. *American Journal of Preventive Medicine*, 40(5):S91–S96, 2011.

[219] R. Singh. Geoblogging: collaborative, peer-to-peer geographic information sharing. In *URISA Public Participation in GIS 3rd Annual Conference*, 2004.

[220] M. Slee, A. Agarwal, and M. Kwiatkowski. Thrift: Scalable cross-language services implementation. *Facebook White Paper*, 2007.

[221] B.C. Smiley. *Chemi-code: an innovative method for wood product tracking*. PhD thesis, The University Of British Columbia, 2008.

[222] B. Sohngen, C. Tennity, M. Hnytka, and K. Meeusen. Global forestry data for the economic modeling of land use. *Economic analysis of land use in global climate change policy*, pages 49–71, 2009.

[223] C. Song, M.B. Dickinson, L. Su, S. Zhang, and D. Yaussey. Estimating average tree crown size using spatial information from ikonos and quickbird images: Across-sensor and across-site comparisons. *Remote Sensing of Environment*, 114(5):1099–1107, 2010.

[224] J. Sowa. Ontology, metadata, and semiotics. *Conceptual structures: Logical, linguistic, and computational issues*, pages 55–81, 2000.

[225] R. Srinivasan. Xdr: External data representation standard. 1995.

[226] T.W. Steele and J.C. Stier. The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science*, 51(5), 2000.

[227] F. Steinheimer. The ornithological babel - nomenclatorial confusion in avian taxonomy. *JOURNAL OF ORNITHOLOGY*, 147:257–257, 2006.

[228] S.L. Stephens. Evaluation of the effects of silvicultural and fuels treatments on potential fire behaviour in sierra nevada mixed-conifer forests. *Forest Ecology and Management*, 105(1-3):21–35, 1998.

[229] M. Stonebraker. Too much middleware. *ACM Sigmod Record*, 31(1):97–106, 2002.

[230] B.J. Strasser. Genbank–natural history in the 21st century? *Science*, 322(5901):537, 2008.

[231] NWS Internet Services Team. Evolution of the national weather service.

[232] C. Tenopir, S. Allard, K. Douglass, A.U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. Data from: Data sharing by scientists: practices and perceptions. 2011.

[233] M. Thornhill, A. Abd-Elrahman, and M. Andreu. Urban forest inventory using open access web mapping services and photogrammetric solution. In *Geoinformatics, 2009 17th International Conference on*, pages 1–5. IEEE, 2009.

[234] Jing Yi Tou, Yong Haur Tay, and Phooi Yee Lau. Rotational invariant wood species recognition through wood species verification. *Intelligent Information and Database Systems, Asian Conference on*, 0:115–120, 2009.

[235] R. Tretheway, M. Simon, E.G. McPherson, and S. Mathis. Volunteer-based urban forest inventory and monitoring programs. In *Results from A Two-Day Workshop: February*, volume 17, page 27, 1999.

[236] S. Tsuchikawa, K. Inoue, J. Noma, and K. Hayashi. Application of near-infrared spectroscopy to wood discrimination. *Journal of Wood Science*, 49(1):29–35, 2003.

[237] A.J. Valentine. Comment on 'big science, little science'. *EMBO reports*, 11(3):152–152, 2010.

[238] P. Van den Besselaar and G. Heimeriks. Disciplinary, multidisciplinary, interdisciplinary: Concepts and indicators. In *8th International Conference on Scientometrics and Infometrics-ISSI2001, Sydney, UNSW*, 2001.

[239] EA Wheeler and North Carolina. Agricultural Research Service. *Computer-aided wood identification*. North Carolina Agricultural Research Service, 1986.

[240] T. O. H. White, R. P. Bruin, G. T. Chiang, M. T. Dove, R. P. Tyer, and A. M. Walker. Lessons in scientific data interoperability: XML and the eMinerals project. *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES*, 367(1890):1041–1046, MAR 13 2009.

[241] M. Wildgrube. Structured data exchange format (sdxf). 2001.

[242] J. G. Wu and R. Hobbs. Key issues and research priorities in landscape ecology: An idiosyncratic synthesis. *LANDSCAPE ECOLOGY*, 17(4):355–365, 2002.

[243] Z. Yang, S. Cai, Z. Zhou, and N. Zhou. Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Information & Management*, 42(4):575–589, 2005.

BIBLIOGRAPHY

References

[1]

[2] Excel binary file format (.xls) structure specification. This is an electronic document. Date of publication: [Date unavailable]. Date retrieved: November 29, 2011. Date last modified: [Date unavailable].

[3] Ogf standards: Data format description language (dfdl). This is an electronic document. Date of publication: [Date unavailable]. Date retrieved: August 10, 2011. Date last modified: [Date unavailable].

[4] Tree ring data description. This is an electronic document. Date of publication: [Date unavailable]. Date retrieved: November 29, 2011. Date last modified: [Date unavailable].

[5] Tiff revision 6.0. *Aldus Developers Desk, Seattle*, 1992.

[6] Shapefile technical description. White paper, Environmental Systems Research Institut, 1998.

[7] Pdf reference: Adobe portable document format version 1.3. 2000.

[8] R.S. Abate. Redd, white, and blue: Is proposed us climate legislation adequate to promote a global carbon credits system for avoided deforestation in a post-kyoto regime? *Tul. J. Int'l & Comp. L.*, 19:95–397, 2010.

[9] A.H. Abd-Elrahman, M.E. Thornhill, M.G. Andreu, and F. Escobedo. A community-based urban forest inventory using online mapping services and consumer-grade digital images. *International Journal of Applied Earth Observation and Geoinformation*, 12(4):249–260, 2010.

[10] F Achard, G Vaysseix, and E Barillot. XML, bioinformatics and data integration. *BIOINFORMATICS*, 17(2):115–125, FEB 2001.

[11] A. Agrawal and A. Chhatre. Explaining success on the commons: Community forest governance in the indian himalaya. *World Development*, 34(1):149–166, 2006.

[12] A.V. Aho, R. Sethi, and J.D. Ullman. Compilers: principles, techniques, and tools. *Reading, MA,*, 1986.

[13] M.S. Aktas, G.C. Fox, M. Pierce, and S. Oh. Xml metadata services. *Concurrency and Computation: Practice and Experience*, 20(7):801–823, 2008.

[14] H. Alani, D. Dupplaw, J. Sheridan, K. O'Hara, J. Darlington, N. Shadbolt, and C. Tullo. Unlocking the potential of public sector information with semantic web technology. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, pages 708–721. Springer-Verlag, 2007.

[15] J.C. Allen and D.F. Barnes. The causes of deforestation in developing countries. *Annals of the association of American Geographers*, 75(2):163–184, 1985.

[16] C. Anderson and M. Wolff. The web is dead. long live the internet. *Wired Magazine, August*, 17, 2010.

[17] D.P. Anderson. Boinc: A system for public-resource computing and storage. In *proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, pages 4–10. IEEE Computer Society, 2004.

[18] M. Antonioletti, M. Atkinson, R. Baxter, A. Borley, N.P. Chue Hong, B. Collins, N. Hardman, A.C. Hume, A. Knox, M. Jackson, et al. The design and implementation of grid database services in ogsa-dai. *Concurrency and Computation: Practice and Experience*, 17(2-4):357–376, 2005.

[19] M. Araujo, S. Kant, and L. Couto. Why brazilian companies are certifying their forests? *Forest Policy and Economics*, 11(8):579–585, 2009.

[20] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, et al. Above the clouds: A berkeley view of cloud computing. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28*, 2009.

[21] E. Aronova, K.S. Baker, and N. Oreskes. Big science and big data in biology: From the international geophysical year through the international biological program to the long term ecological research (lter) network, 1957-present. *Historical Studies in the Natural Sciences*, 40(2):183–224, 2010.

[22] T. K. Attwood. Genomics - the babel of bioinformatics. *SCIENCE*, 290(5491):471–473, Oct 2000.

[23] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[24] D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, and R. Evans. Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. 2001.

[25] H. .. Balen. Deconstructing babel: Xml and application integration. *Application Development Trends*, 7(12):19–23, December 2000.

[26] J.C. Balouet, K.T. Smith, D. Vroblesky, and G. Oudijk. Use of dendrochronology and dendrochemistry in environmental forensics: does it meet the daubert criteria? *Environmental Forensics*, 10(4):268–276, 2009.

[27] E. Barillot and F. Achard. Xml: a lingua franca for science? *Trends in Biotechnology*, 18(8):331–333, 2000.

[28] J. .. Beall. The death of metadata. *Serials Librarian*, vol.51, no.2:55–74, 2006.

[29] A.L. Beberg, D.L. Ensign, G. Jayachandran, S. Khaliq, and V.S. Pande. Folding@ home: Lessons from eight years of volunteer distributed computing. 2009.

[30] P.H. Beckman. Building the teragrid. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1833):1715, 2005.

[31] F. Berkes, D. Feeny, B.J. McCay, and J.M. Acheson. The benefits of the commons. *Nature*, 340(6229):91–93, 1989.

[32] A. Berl, E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M.Q. Dang, and K. Pentikousis. Energy-efficient cloud computing. *The Computer Journal*, 53(7):1045, 2010.

[33] T. Berners-Lee. Www: Past, present, and future. *Computer*, 29(10):69–77, 1996.

[34] T. Berners-Lee. Axioms of web architecture: Metadata, january 1997. *URL http://www. w3. org/DesignIssues/Metadata. Updated Sept*, 2000.

[35] T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[36] T. Berners-Lee, D. Connolly, L. Kagal, Y. Scharf, and J. Hendler. N3logic: A logical framework for the world wide web. *Theory and Practice of Logic Programming*, 8(3):249–269, 2008.

[37] T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, and D.J. Weitzner. Creating a science of the web. *Science*, 313(5788):769, 2006.

[38] T. Berners-Lee and J. Hendler. Scientific publishing on the semantic web. *Nature*, 410:1023–1024, 2001.

[39] S. Bernstein and B. Cashore. Nonstate global governance: Is forest certification a legitimate alternative to a global forest convention? *Hard Choices, Soft Law: Voluntary Standards in Global Trade, Environment and Social Governance*, pages 33–63, 2004.

[40] FA Bisby. The quiet revolution: Biodiversity informatics and the internet. *SCIENCE*, 289(5488):2309–2312, SEP 29 2000.

[41] D. Brack and C. House. Controlling illegal logging: Lessons from the us lacey act. *Chatham House: Energy, Environment and Development Programme*, 2007.

[42] S. Bratkovich, J. Bowyer, S. Stai, K. Fernholz, J. Bratkovich, S. Howe, J. Bratkovich, J. Bowyer, S. Stai, K. Fernholz, J. Bratkovich, and J. Howe. Urban forestry: an evolving discipline. 2010.

[43] M. Brunner, R. Eugster, E. Trenka, and L. Bergamin-Strotz. Ftnir spectroscopy and wood identification. *Holzforschung-International Journal of the Biology, Chemistry, Physics and Technology of Wood*, 50(2):130–134, 1996.

[44] R. Bryant, R.H. Katz, and E.D. Lazowska. Big-data computing: Creating revolutionary breakthroughs in commerce, science and society. *Computing Research Initiatives for the 21st Century. Computing Research Association*, 2008.

[45] P. Buffers. Protocol buffers. google's data interchange format.

[46] C. Burks. Molecular biology database list. *Nucleic acids research*, 27(1):1, 1999.

[47] A. Callado, C. Kamienski, G. Szabo, B.P. Gero, J. Kelner, S. Fernandes, and D. Sadok. A survey on internet traffic identification. *Communications Surveys & Tutorials, IEEE*, 11(3):37–52, 2009.

[48] JT CALLAHAN. LONG-TERM ECOLOGICAL RESEARCH. *BIOSCIENCE*, 34(6):363–367, 1984.

[49] F. Cappello, S. Djilali, G. Fedak, T. Herault, F. Magniette, V. Néri, and O. Lodygensky. Computing on large-scale distributed systems: Xtremweb architecture, programming models, security, tests and convergence with grid. *Future Generation Computer Systems*, 21(3):417–437, 2005.

[50] A. Carleer and E. Wolff. Exploitation of very high resolution satellite data for tree species identification. *Photogrammetric engineering and remote sensing*, 70(1):135–140, 2004.

[51] B. Cashore. Legitimacy and the privatization of environmental governance: How non–state market–driven (nsmd) governance systems gain rule–making authority. *Governance*, 15(4):503–529, 2002.

[52] B. Cashore and M. Stone. Can legality verfiication rescue global forest governance? assessing interacting effects of economic mechanism of forestry policy and governance: Lessons learnt from southeast asia. *Retrieved November*, 16:2010, 2010.

[53] A.D. Chandler. *Scale and scope: the dynamics of industrial capitalism*. Belknap Pr, 1994.

[54] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.

[55] AD Chave, M. Arrott, C. Farcas, E. Farcas, I. Krueger, M. Meisinger, JA Orcutt, FL Vernon, C. Peach, O. Schofield, et al. Cyberinfrastructure for the us ocean observatories initiative: Enabling interactive observation in the ocean. In *OCEANS 2009-EUROPE*, pages 1–10. IEEE.

[56] A. Chhatre and A. Agrawal. Forest commons and local enforcement. *Proceedings of the National Academy of Sciences*, 105(36):13286, 2008.

[57] A.B. Chimeli, R.G. Boyd, and D.M. Adams. International timber markets and tropical deforestation: the evidence from prices. *Applied Economics*, 99999(1):1–12, 2011.

[58] S. Chiorescu and A. Gr
"onlund. The fingerprint method: Using over-bark and under-bark log measurement data generated by three-dimensional log scanners in combination with radiofrequency identification tags to achieve traceability in the log yard at the sawmill. *Scandinavian journal of forest research*, 19(4):374–383, 2004.

[59] G. Chirici, D. Giuliarelli, D. Biscontini, D. Tonti, W. Mattioli, M. Marchetti, and P. Corona. Large-scale monitoring of coppice forest clearcuts by multitemporal very high resolution satellite imagery. a case study from central italy. *Remote Sensing of Environment*, 2011.

[60] N.E. Clark. 3d reconstruction of a tree stem using video images and pulse distances. In *Symposium on Statistics and Information Technology in Forestry*, pages 8–12.

[61] The Federal Geographic Data Committee. Content standard for digital geospatial metadata content standard for digital geospatial metadata content standard for digital geospatial metadata content standard

for digital geospatial metadata content standard for digital geospatial metadata content standard for digital geospatial metadata. *Concurrency and Computation: Practice and Experience*, 20(7):801–823, 1998.

[62] H.A. Core, W.A. Côté, and A.C. Day. Wood structure and identification.[syracuse wood science series 6.]. *Syracuse Univ. Press.*, 182:2–3, 1979.

[63] Mark J. Costello and Edward Vanden Berghe. 'Ocean biodiversity informatics': a new era in marine biology research and management. *MARINE ECOLOGY-PROGRESS SERIES*, 316:203–214, 2006.

[64] M.J. Costello. Motivating online publication of data. *BioScience*, pages 418–427, 2009.

[65] Fred Cowett. Village of cayuga heights street tree inventory (shapefile : 2009).

[66] M.H. Cragin, C.L. Palmer, J.R. Carlson, and M. Witt. Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1926):4023, 2010.

[67] D. Crockford. Json: The fat-free alternative to xml. In *Proc. of XML*, volume 2006, 2006.

[68] F. Dahdouh-Guebas and N. Koedam. Long-term retrospection on mangrove development using transdisciplinary approaches: A review. *AQUATIC BOTANY*, 89(2):80–92, 2008.

[69] R. A. Daley, E. A. Immer, B. I. Fortner, and M. B. Weiss. Scientific resource access system: A concept for getting "living with a star" information to do science. *JOHNS HOPKINS APL TECHNICAL DIGEST*, 26(1):22–35, 2005.

[70] O. Damette and P. Delacote. Unsustainable timber harvesting, deforestation and the role of certification. *Ecological Economics*, 2011.

[71] G.S. Dangl, J. Yang, D.A. Golino, and T. Gradziel. A practical method for almond cultivar identification and parental analysis using simple sequence repeat markers. *Euphytica*, 168(1):41–48, 2009.

[72] L.S. Davis and G. Liu. Integrated forest planning across multiple ownerships and decision makers. *Forest science*, 37(1):200–226, 1991.

[73] H.F. Deus, R. Stanislaus, D.F. Veiga, C. Behrens, I.I. Wistuba, J.D. Minna, H.R. Garner, S.G. Swisher, J.A. Roth, A.M. Correa, et al. A semantic web management model for integrative biomedical informatics. *PloS one*, 3(8):e2946, 2008.

[74] J. Domingue, C. Pedrinaci, M. Maleshkova, B. Norton, and R. Krummenacher. Fostering a relationship between linked data and the internet of services. *The future internet*, pages 351–364, 2011.

[75] J. Douglas and M. Simula. Deforestation: Causes and symptoms. *The Future of the World's Forests*, pages 81–104, 2010.

[76] S. Doveri, F. Sabino Gil, A. Diaz, S. Reale, M. Busconi, A. da Câmara Machado, A. Martín, C. Fogher, P. Donini, and D. Lee. Standardization of a set of microsatellite markers for use in cultivar identification studies in olive (olea europaea l.). *Scientia Horticulturae*, 116(4):367–373, 2008.

[77] S. Draghici, S. Sellamuthu, and P. Khatri. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *BIOINFORMATICS*, 22(23):2934–2939, Dec 2006.

[78] PD EAGAN and SJ VENTURA. ENHANCING VALUE OF ENVIRONMENTAL DATA - DATA LINEAGE REPORTING. *JOURNAL OF ENVIRONMENTAL ENGINEERING-ASCE*, 119(1):5–16, JAN-FEB 1993.

[79] JL Edwards, MA Lane, and ES Nielsen. Interoperability of biodiversity databases: Biodiversity information on every desktop. *SCIENCE*, 289(5488):2312–2314, SEP 29 2000.

[80] A. Eliassen. Vilhelm bjerknes and his students. *Annual Review of Fluid Mechanics*, 14(1):1–12, 1982.

[81] G. Erbach. Data-centric view in e-science information systems. *Data Science Journal*, 5(0):219–222, 2006.

[82] J. Esparza and T. Yamada. The discovery value of "big science". *The Journal of experimental medicine*, 204(4):701, 2007.

[83] LG Esteban, F.G. Fernández, P. Palacios de Palacios, RM Romero, and N.N. Cano. Artificial neural networks in wood identification: the case of two juniperus species from the canary islands. *IAWA Journal*, 30(1):87–94, 2009.

[84] D.C. Esty. Toward optimal environmental governance. *NYUL Rev.*, 74:1495, 1999.

[85] D.C. Esty and M.H. Ivanova. The road ahead: Conclusions and action agenda. *Esty, DC and Ivanova, MH Global Environmental Governance: Options & Opportunities*, pages 225–240, 2002.

[86] M.J. Falkowski, M.A. Wulder, J.C. White, and M.D. Gillis. Supporting large-area, sample-based forest inventories with very high spatial resolution satellite imagery. *Progress in Physical Geography*, 33(3):403, 2009.

[87] F. Fischer, G. Unel, B. Bishop, and D. Fensel. Towards a scalable, pragmatic knowledge representation language for the web. *Perspectives of Systems Informatics*, pages 124–134, 2010.

[88] K. Fisher and D. Walker. The pads project: An overview. In *Proceedings of the 14th International Conference on Database Theory*, pages 11–17. ACM, 2011.

[89] Food and Agriculture Organization of the United Nations. *Global forest resources assessment 2005: progress towards sustainable forest management*. Food and Agriculture Organization of the United Nations, 2006.

[90] JF FRANKLIN, CS BLEDSOE, and JT CALLAHAN. CONTRIBUTIONS OF THE LONG-TERM ECOLOGICAL RESEARCH-PROGRAM - AN EXPANDED NETWORK OF SCIENTISTS, SITES, AND PROGRAMS CAN PROVIDE CRUCIAL COMPARATIVE ANALYSES. *BIOSCIENCE*, 40(7):509–523, JUL-AUG 1990.

[91] J.E.F. Friedl. *Mastering regular expressions*. O'Reilly Media, Inc., 2006.

[92] L.M. Friedman. Business and legal strategies for combating grey-market imports. In *Int'l L.*, volume 32, page 27. HeinOnline, 1998.

[93] M. Gadgil, F. Berkes, and C. Folke. Indigenous knowledge for biodiversity conservation. *Ambio*, pages 151–156, 1993.

[94] C. Germain, V. Neri, G. Fedak, and F. Cappello. Xtremweb: building an experimental platform for global computing. *Grid Computing—GRID 2000*, pages 107–129, 2000.

[95] C. Gibson. Forest resources: Institutions for local governance in guatemala. *Protecting the commons: A framework for resource management in the Americas*, pages 71–89, 2001.

[96] C.C. Gibson. *People and forests: communities, institutions, and governance*. The MIT Press, 2000.

[97] C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–693, 2008.

[98] P. GODFREY-SMITH. Explanation in evolutionary biology: Comments on fodor. *Mind & Language*, 23(1):32–41, 2008.

[99] ME Goodchild. Geographic information science and systems for environmental management. *ANNUAL REVIEW OF ENVIRONMENT AND RESOURCES*, 28:493–519, 2003.

[100] ME Goodchild and JY Zhou. Finding geographic information: Collection-level metadata. *GEOINFORMATICA*, 7(2):95–112, JUN 2003.

[101] M.F. Goodchild. Different data sources and diverse data structures: metadata and other solutions. In *Geocomputation: a primer*, pages 61–73, 1998.

[102] D.L. Goodhue, J.A. Quillard, and J.F. Rockart. Managing the data resource: a contingency perspective. *MIS Quarterly*, pages 373–392, 1988.

[103] A. Gore. Bringing information to the world: The global information infrastructure. *Harv. JL & Tech.*, 9:1, 1996.

[104] V. Grimm and C. Wissel. Babel, or the ecological stability discussions: An inventory and analysis of terminology and a guide for avoiding confusion. *OECOLOGIA*, 109(3):323–334, Feb 1997.

[105] S. Gronas. Vilhelm bjerknes' vision for scientific weather prediction. *The Nordic seas: an integrated perspective: oceanography, climatology, biogeochemistry, and modeling*, 158:357, 2005.

[106] N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. *Towards Very Large Knowledge Bases Knowledge Building and Knowledge Sharing*, 1(9):25–32, 1995.

[107] N. Guarino and Istituto (Roma) Consiglio nazionale delle ricerc. Formal ontology in information systems, 1998.

[108] L.H. Gulbrandsen. Overlapping public and private governance: Can forest certification fill the gaps in the global forest regime? *Global Environmental Politics*, 4(2):75–99, 2004.

[109] A. Gupta. Transparency under scrutiny: information disclosure in global environmental governance. *Global Environmental Politics*, 8(2):1–7, 2008.

[110] R. Gyimah. Piloting wood tracking system in ghana. In *Precision Forestry Symposium Stellenbosch, South Africa 1 to 3 March 2010*, 2010.

[111] SS Hale, MM Hughes, JF Paul, RS McAskill, SA Rego, DR Bender, NJ Dodge, TL Richter, and JL Copeland. Managing scientific data: The EMAP approach. *ENVIRONMENTAL MONITORING AND ASSESSMENT*, 51(1-2):429–440, JUN 1998.

[112] SS Hale, AH Miglarese, MP Bradley, TJ Belton, LD Cooper, MT Frame, CA Friel, LM Harwell, RE King, WK Michener, DT Nicolson, and BG Peterjohn. Managing troubled data: Coastal data partnerships smooth data integration. *ENVIRONMENTAL MONITORING AND ASSESSMENT*, 81(1-3):133–148, JAN-FEB 2003.

[113] B. Hammersley. *Developing feeds with RSS and Atom*. O'Reilly Media, Inc., 2005.

[114] T. Hammond, T. Hannay, and B. Lund. The role of rss in science publishing: Syndication and annotation on the web. *D-Lib Magazine*, 10(12), 2004.

[115] B. Hanson, A. Sugden, and B. Alberts. Making data maximally available. *Science*, 331(6018):649, 2011.

[116] A.I. Hapca, F. Mothe, and J.M. Leban. A digital photographic method for 3d reconstruction of standing tree shape. *Annals of Forest Science*, 64(6):631–637, 2007.

[117] G. HARDIN. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

[118] S. Hastings, M. Ribeiro, S. Langella, S. Oster, U. Catalyurek, T. Pan, K. Huang, R. Ferreira, J. Saltz, and T. Kurc. Xml database support for distributed execution of data-intensive scientific workflows. *ACM SIGMOD Record*, 34(3):50–55, 2005.

[119] M Heemskerk, K Wilson, and M Pavao-Zuckerman. Conceptual models as tools for communication across disciplines. *CONSERVATION ECOLOGY*, 7(3), DEC 2003.

[120] S. Henley. The problem of missing data in geoscience databases. *Computers & geosciences*, 32(9):1368–1377, 2006.

[121] A. .. Hill. The xml revolution. *Financial Technology*, 1(14):5–6, April 2002.

[122] M. Hu, E.P. Lim, A. Sun, H.W. Lauw, and B.Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management*, pages 243–252. ACM, 2007.

[123] W. Huang and D. Webster. Intelligent rss news aggregation based on semantic contexts. In *ACM SIGIR 2004 Workshop on Information Retrieval in Context*, pages 40–43, 2004.

[124] B.A. Huberman and L.A. Adamic. Growth dynamics of the world wide web. *J. Reprod. Fertil*, 9:55–60, 1993.

[125] J.L. Innes. Madagascar rosewood, illegal logging and the tropical timber trade. *Madagascar Conservation & Development*, 5(1), 2010.

[126] M.B. Jadhav, V.J. Gaikwad, CV Patil, and GS Deshpande. Cloud computing applications in computational science. *International Journal of Computer Science and Applications*, 1(1):1–6, 2010.

[127] C. .. Jaenicke. Server-side xml: taming the tower of babel, 1998.

[128] P. Jain, P. Hitzler, P.Z. Yeh, K. Verma, and A.P. Sheth. Linked data is merely more data. *Linked Data Meets Artificial Intelligence*, 2009.

[129] CH Jakobsen, T Hels, and WJ McLaughlin. Barriers and facilitators to integration among scientists in transdisciplinary landscape analyses: a cross-country comparison. *FOREST POLICY AND ECONOMICS*, 6(1):15–31, JAN 2004.

[130] W. Janssen and P. Goldsworthy. Multidisciplinary research for natural resource management: conceptual and practical implications. *Agricultural systems*, 51(3):259–279, 1996.

[131] D.H. Janzen, W. Hallwachs, P. Blandin, J.M. Burns, J. CADIOU, I. Chacon, T. Dapkey, A.R. Deans, M.E. Epstein, B. Espinoza, et al. Integration of dna barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, 9:1–26, 2009.

[132] S. Jiang, L. Fang, and X. Huang. An idea of special cloud computing in forest pests' control. *Cloud Computing*, pages 615–620, 2009.

[133] Matthew B. Jones, Mark P. Schildhauer, O. J. Reichman, and Shawn Bowers. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *ANNUAL REVIEW OF ECOLOGY EVOLUTION AND SYSTEMATICS*, 37:519–544, 2006.

[134] E.C. Kansa and A. Bissell. Web syndication approaches for sharing primary data in" small science" domains. *Data Science Journal*, (0):1006280237, 2010.

[135] H. Karasti, K.S. Baker, and E. Halkola. Enriching the notion of data curation in e-science: data managing and information infrastructuring in the long term ecological research (lter) network. *Computer Supported Cooperative Work (CSCW)*, 15(4):321–358, 2006.

[136] J.E. Keeley, CJ Fotheringham, and M. Morais. Reexamining fire suppression impacts on brushland fire regimes. *Science*, 284(5421):1829, 1999.

[137] A. KEMPF. Advocating global forest issues on the internet. In *International online information meeting*, pages 99–105, 1996.

[138] R.W. Kimmerer and F.K. Lake. The role of indigenous burning in land management. *Journal of Forestry*, 99(11):36–41, 2001.

[139] B. M. Knoppers and M. Saginur. The babel of genetic data terminology. *NATURE BIOTECHNOLOGY*, 23(8):925–927, Aug 2005.

[140] L. Koivuranta, K. Leinonen, and P. Pulkkinen. Marketing of forest reproductive material: the use of microsatellites for identification of registered tree clones in finland. *Finnish For Res Inst, Vantaa*, 2008.

[141] G. Koloniari and E. Pitoura. Peer-to-peer management of xml data: issues and research challenges. *ACM Sigmod Record*, 34(2):6–17, 2005.

[142] S. Korten and C. Kaul. Application of rfid (radio frequency identification) in the timber supply chain. *Croatian Journal of Forest Engineering*, 29(1):85–94, 2008.

[143] A.T. Kralidis. Geospatial open source and open standards convergences. *Open Source Approaches in Spatial Data Handling*, pages 1–20, 2008.

[144] S. Krech. *The ecological Indian: myth and history.* WW Norton & Company, 2000.

[145] M. Krotzsch, D. Vrandecic, M. Volkel, H. Haller, and R. Studer. Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):251–261, 2007.

[146] J. Lalonde, N. Vandapel, and M. Hebert. Automatic three-dimensional point cloud processing for forest inventory. *Robotics Institute*, page 334, 2006.

[147] A. Langegger, W. W
"oß, and M. Bl

"ochl. A semantic web middleware for virtual data integration on the web. *The Semantic Web: Research and Applications*, pages 493–507, 2008.

[148] F. Laroui and BCC Van der Zwaan. Environment and multidisciplinarity three examples of avoidable confusion. *Integrated Assessment*, 3(4):360–369, 2002.

[149] C.A. Lee. A perspective on scientific cloud computing. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pages 451–459. ACM, 2010.

[150] S. Lélé and R.B. Norgaard. Practicing interdisciplinarity. *BioScience*, 55(11):967–975, 2005.

[151] R.M. Lerner. At the forge: syndication with rss. *Linux Journal*, 2004(126):8, 2004.

[152] RA Letcher, BFW Croke, and AJ Jakeman. Integrated assessment modelling for water resource allocation and management: A generalised conceptual framework. *Environmental Modelling and Software*, 22(5):733–742, 2007.

[153] J. Lewin. An introduction to rss news feeds: Using open formats for content syndication. *This article includes sample code that demonstrates elements of an RSS file*, 2000.

[154] D. Liddick. *Crimes Against Nature: Illegal Industries and the Global Environment*. Praeger Publishers, 2011.

[155] R.D. Lipschutz. Why is there no international forestry law: An examination of international forestry regulation, both public and private. *UCLA J. Envtl. L. & Pol'y*, 19:153, 2000.

[156] P. Longley. *Geographic information systems and science*. John Wiley & Sons Inc, 2005.

[157] F. López-Muñoz, C. Alamo, F.J. Quintero-Gutiérrez, and P. García-García. A bibliometric study of international scientific productivity in attention-deficit hyperactivity disorder covering the period 1980–2005. *European Child & Adolescent Psychiatry*, 17(6):381–391, 2008.

[158] F. López-Muñoz, J. Boya, F. Marín, and J.L. Calvo. Scientific research on the pineal gland and melatonin: a bibliometric study for the period 1966-1994. *Journal of Pineal Research*, 20(3):115–124, 1996.

[159] M. Lutz. *Programming python*. O'Reilly Media, Inc., 2011.

[160] H.G. Maas, A. Bienert, S. Scheller, and E. Keane. Automatic forest inventory parameter determination from terrestrial laser scanner data. *International Journal of Remote Sensing*, 29(5):1579–1593, 2008.

[161] A. Maedche, S. Staab, N. Stojanovic, R. Studer, and Y. Sure. Seal—a framework for developing semantic web portals. *Advances in Databases*, pages 1–22, 2001.

[162] JJ MAGNUSON. LONG-TERM ECOLOGICAL RESEARCH AND THE INVISIBLE PRESENT - UNCOVERING THE PROCESSES HIDDEN BECAUSE THEY OCCUR SLOWLY OR BECAUSE EFFECTS LAG YEARS BEHIND CAUSES. *BIOSCIENCE*, 40(7):495–501, JUL-AUG 1990.

[163] J. Mallet and K. Willmott. Taxonomy: renaissance or tower of babel? *TRENDS IN ECOLOGY & EVOLUTION*, 18(2):57–59, 2003.

[164] B.L. Markham, J.C. Storey, D.L. Williams, and J.R. Irons. Landsat sensor performance: history and current status. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(12):2691–2694, 2004.

[165] J. Markwell and D.W. Brooks. "link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology*. *Biochemistry and Molecular Biology Education*, 31(1):69–72, 2003.

[166] M. Marsik, F.R. Stevens, and J. Southworth. Amazon deforestation: Rates and patterns of land cover change and fragmentation in pando, northern bolivia, 1986 to 2005. *Progress in Physical Geography*, 2011.

[167] M.Z. Martin, S.D. Wullschleger, C.T. Garten, A.V. Palumbo, and J.G. Smith. Elemental analysis of environmental and biological samples using laser-induced breakdown spectroscopy and pulsed raman spectroscopy. *Journal of dispersion science and technology*, 25(5):687–694, 2005.

[168] F. MARTIN-RECUERDA. Application integration using conceptual spaces (cspaces). *Lecture notes in computer science*, pages 234–248, 2006.

[169] A. Marx and D. Cuypers. Forest certification as a global environmental governance tool: What is the macro-effectiveness of the forest stewardship council? *Regulation & Governance*, 4(4):408–434, 2010.

[170] S. Matsuoka, S. Shinjo, M. Aoyagi, S. Sekiguchi, H. Usami, and K. Miura. Japanese computational grid research project: Naregi. *Proceedings of the IEEE*, 93(3):522–533, 2005.

[171] William K. Michener. Meta-information concepts for ecological data management. *ECOLOGICAL INFORMATICS*, 1(1):3–7, JAN 2006.

[172] WK Michener, JW Brunt, JJ Helly, TB Kirchner, and SG Stafford. Nongeospatial metadata for the ecological sciences. *ECOLOGICAL APPLICATIONS*, 7(1):330–342, FEB 1997.

[173] S.P. Mirashe and NV Kalyankar. Cloud computing. *Arxiv preprint arXiv:1003.4074*, 2010.

[174] M. Mirtl. Introducing the next generation of ecosystem research in europe: Lter-europe's multi-functional and multi-scale approach. *Long-Term Ecological Research*, pages 75–93, 2010.

[175] Yasuko Mizoguchi, Akira Miyata, Yoshikazu Ohtani, Ryuichi Hirata, and Satoko Yuta. A review of tower flux observation sites in Asia. *JOURNAL OF FOREST RESEARCH*, 14(1):1–9, FEB 2009.

[176] B. Muys, J. Hynynen, M. Palahí, MJ Lexer, M. Fabrika, H. Pretzsch, F. Gillet, E. Briceño, GJ Nabuurs, and V. Kint. Simulation tools for decision support to adaptive forest management in europe. *Forest Systems*, 19:86–99, 2010.

[177] J. Myers and A. Chappell. Binary format description language. Technical report, Technical report, Pacific Northwest National Laboratory, 2003.

[178] LE Nave and C. Swanston. Big data for big questions: Global soil change and the national soil carbon network. In *AGU Fall Meeting Abstracts*, volume 1, page 06, 2010.

[179] J.R. Neilon. Communications in meteorology. *Geoscience Electronics, IEEE Transactions on*, 8(2):101–105, 1970.

[180] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov. The eucalyptus open-source cloud-computing system. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, pages 124–131. IEEE Computer Society, 2009.

[181] K. O'Hara, T. Berners-Lee, W. Hall, and N. Shadbolt. Use of the semantic web in e-research. *Dutton, WH, and Jeffreys, PW (eds), World Wide Research: Reshaping the Sciences and Humanities*, 2010.

[182] A.T. OLIVEIRA FILHO and J.A. Ratter. Database: Woody flora of 106 forest areas of eastern tropical south america. *Occasional Monographs, Royal Botanic Garden Edinburgh, Edinburgh.[Links]*, 1995.

[183] E. Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, 1999.

[184] E. Ostrom, R. Gardner, and J. Walker. *Rules, games, and common-pool resources*. Univ of Michigan Pr, 1994.

[185] Y.O. Ouma, TG Ngigi, and R. Tateishi. On the optimization and selection of wavelet texture for feature extraction from high-resolution satellite imagery with application towards urban-tree delineation. *International journal of remote sensing*, 27(1):73–104, 2006.

[186] P.R. Oyono, J.C. Ribot, and A.M. Larson. Green and black gold in rural cameroon: natural resources for local governance, justice and sustainability. 2006.

[187] G. Paglia and M. Morgante. Pcr-based multiplex dna fingerprinting techniques for the analysis of conifer genomes. *Molecular breeding*, 4(2):173–177, 1998.

[188] MA Palmer, ES Bernhardt, EA Chornesky, SL Collins, AP Dobson, CS Duke, BD Gold, RB Jacobson, SE Kingsland, RH Kranz, MJ Mappin, ML Martinez, F Micheli, JL Morse, ML Pace, M Pascual,

SS Palumbi, O Reichman, AR Townsend, and MG Turner. Ecological science and sustainability for the 21st century. *FRONTIERS IN ECOLOGY AND THE ENVIRONMENT*, 3(1):4–11, FEB 2005.

[189] A. Pantze, A.H. Krantz, J.E.S. Fransson, H. Olsson, M. Santoro, L.E.B. Eriksson, and L.M.H. Ulander. Mapping and monitoring clear-cuts in swedish forest using alos palsar satellite images. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*, volume 3, pages III–589. IEEE, 2009.

[190] J.N. Parker. *Organizational collaborations and scientific integration: The case of ecology and the social sciences*. PhD thesis, ARIZONA STATE UNIVERSITY, 2006.

[191] J. Peacock, TR Baker, SL Lewis, G. Lopez-Gonzalez, and OL Phillips. The rainfor database: monitoring forest biomass and dynamics. *Journal of Vegetation Science*, 18(4):535–542, 2007.

[192] J. Phelps, E.L. Webb, and A. Agrawal. Does redd+ threaten to recentralize forest governance? *Science*, 328(5976):312, 2010.

[193] D.J.S. Price. Little science, big science. *New York*, pages 87–89, 1963.

[194] Hui-Je Qiao, Yan Han, Nuo Li, and Li-Qiang Ji. A model of biodiversity information integration. *Shengwu Duoyangxing*, 12(5), September 2004.

[195] I. Rafols and M. Meyer. How cross-disciplinary is bionanotechnology? explorations in the specialty of molecular motors. *SCIENTOMETRICS*, 70(3):633–650, 2007.

[196] R. Ramachandran, S. Graves, H. Conover, and K. Moe. Earth science markup language (esml):::: a solution for scientific data-application interoperability problem. *Computers & Geosciences*, 30(1):117–124, 2004.

[197] R. Ramakrishnan and J. Gehrke. *Database Management Systems: International Edition*. McGraw-Hill, 2003.

[198] D. M. Rapoport, I. Ayappa, R. G. Norman, and S. T. Herman. Npsg data interchange-dealing with the tower of babel. *SLEEP*, 29(5):599–600, 2006.

[199] S. Rhodes. Breaking down link rot: The chesapeake project legal information archive's examination of url stability. *Law Libr. J.*, 102:581–705, 2010.

[200] L. Richardson and S. Ruby. *RESTful web services*. O'Reilly Media, 2007.

[201] B.P. Rimal, E. Choi, and I. Lumb. A taxonomy and survey of cloud computing systems. In *2009 Fifth International Joint Conference on INC, IMS and IDC*, pages 44–51. IEEE, 2009.

[202] L. Roberts. The arpanet and computer networks. In *Proceedings of the ACM Conference on The history of personal workstations*, pages 51–58. ACM, 1986.

[203] L.G. Roberts. Beyond moore's law: Internet growth trends. *Computer*, 33(1):117–119, 2000.

[204] P.M. Rodkewich and J. Albrecht. Forestry databases on the internet. *Journal of agricultural & food information*, 3(3):19–23, 1996.

[205] J. Rossmann, M. Schluse, C. Schlette, A. Buecken, P. Krahwinkler, and M. Emde. Realization of a highly accurate mobile robot system for multi purpose precision forestry applications. In *Advanced Robotics, 2009. ICAR 2009. International Conference on*, pages 1–6. IEEE, 2009.

[206] O. Ruebenacker and M. Blinov. Using views of systems biology cloud: application for model building. *Theory in Biosciences*, pages 1–10, 2010.

[207] E. Russell and J. Kane. The missing link: Assessing the reliability of internet citations in history journals. *Technology and Culture*, 49(2):420–429, 2008.

[208] S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Prentice hall, 2010.

[209] S.J. Russell, P. Norvig, J.F. Canny, J.M. Malik, and D.D. Edwards. *Artificial intelligence: a modern approach*, volume 74. Prentice hall Englewood Cliffs, NJ, 1995.

[210] SK Saari, CD Campbell, J. Russell, IJ Alexander, and IC Anderson. Pine microsatellite markers allow roots and ectomycorrhizas to be linked to individual trees. *New phytologist*, 165(1):295–304, 2005.

[211] J. Saez-Rodriguez, A. Goldsipe, J. Muhlich, L.G. Alexopoulos, B. Millard, D.A. Lauffenburger, and P.K. Sorger. Flexible informatics for linking experimental data to mathematical models via datarail. *Bioinformatics*, 24(6):840, 2008.

[212] I. San Gil and K. Baker. The ecological metadata language milestones, community work force, and change. *Databits*, pages 4–7, 2007.

[213] P.A. Schad, L.R. Mobley, and C.M. Hamilton. Building a biomedical cyberinfrastructure for collaborative research. *American Journal of Preventive Medicine*, 40(5):S144–S150, 2011.

[214] D.H. Schepers. Challenges to legitimacy at the forest stewardship council. *Journal of business ethics*, 92(2):279–290, 2010.

[215] K. M. Schoeffler and W. E. Hammond. Standards for the electronic health record emerging from health care's tower of babel. *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, pages 388–392, 2001.

[216] E Schoenberger. Interdisciplinarity and social power. *PROGRESS IN HUMAN GEOGRAPHY*, 25(3):365–382, SEP 2001.

[217] B. Schueler, S. Sizov, and S. Staab. Management of meta knowledge for rdf repositories. 2007.

[218] A.R. Shaikh, P. Das, C.A. Vinson, and B. Spring. Cyberinfrastructure for consumer health. *American Journal of Preventive Medicine*, 40(5):S91–S96, 2011.

[219] R. Singh. Geoblogging: collaborative, peer-to-peer geographic information sharing. In *URISA Public Participation in GIS 3rd Annual Conference*, 2004.

[220] M. Slee, A. Agarwal, and M. Kwiatkowski. Thrift: Scalable cross-language services implementation. *Facebook White Paper*, 2007.

[221] B.C. Smiley. *Chemi-code: an innovative method for wood product tracking*. PhD thesis, The University Of British Columbia, 2008.

[222] B. Sohngen, C. Tennity, M. Hnytka, and K. Meeusen. Global forestry data for the economic modeling of land use. *Economic analysis of land use in global climate change policy*, pages 49–71, 2009.

[223] C. Song, M.B. Dickinson, L. Su, S. Zhang, and D. Yaussey. Estimating average tree crown size using spatial information from ikonos and quickbird images: Across-sensor and across-site comparisons. *Remote Sensing of Environment*, 114(5):1099–1107, 2010.

[224] J. Sowa. Ontology, metadata, and semiotics. *Conceptual structures: Logical, linguistic, and computational issues*, pages 55–81, 2000.

[225] R. Srinivasan. Xdr: External data representation standard. 1995.

[226] T.W. Steele and J.C. Stier. The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science*, 51(5), 2000.

[227] F. Steinheimer. The ornithological babel - nomenclatorial confusion in avian taxonomy. *JOURNAL OF ORNITHOLOGY*, 147:257–257, 2006.

[228] S.L. Stephens. Evaluation of the effects of silvicultural and fuels treatments on potential fire behaviour in sierra nevada mixed-conifer forests. *Forest Ecology and Management*, 105(1-3):21–35, 1998.

[229] M. Stonebraker. Too much middleware. *ACM Sigmod Record*, 31(1):97–106, 2002.

[230] B.J. Strasser. Genbank–natural history in the 21st century? *Science*, 322(5901):537, 2008.

[231] NWS Internet Services Team. Evolution of the national weather service.

[232] C. Tenopir, S. Allard, K. Douglass, A.U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. Data from: Data sharing by scientists: practices and perceptions. 2011.

[233] M. Thornhill, A. Abd-Elrahman, and M. Andreu. Urban forest inventory using open access web mapping services and photogrammetric solution. In *Geoinformatics, 2009 17th International Conference on*, pages 1–5. IEEE, 2009.

[234] Jing Yi Tou, Yong Haur Tay, and Phooi Yee Lau. Rotational invariant wood species recognition through wood species verification. *Intelligent Information and Database Systems, Asian Conference on*, 0:115–120, 2009.

[235] R. Tretheway, M. Simon, E.G. McPherson, and S. Mathis. Volunteer-based urban forest inventory and monitoring programs. In *Results from A Two-Day Workshop: February*, volume 17, page 27, 1999.

[236] S. Tsuchikawa, K. Inoue, J. Noma, and K. Hayashi. Application of near-infrared spectroscopy to wood discrimination. *Journal of Wood Science*, 49(1):29–35, 2003.

[237] A.J. Valentine. Comment on 'big science, little science'. *EMBO reports*, 11(3):152–152, 2010.

[238] P. Van den Besselaar and G. Heimeriks. Disciplinary, multidisciplinary, interdisciplinary: Concepts and indicators. In *8th International Conference on Scientometrics and Infometrics-ISSI2001, Sydney, UNSW*, 2001.

[239] EA Wheeler and North Carolina. Agricultural Research Service. *Computer-aided wood identification*. North Carolina Agricultural Research Service, 1986.

[240] T. O. H. White, R. P. Bruin, G. T. Chiang, M. T. Dove, R. P. Tyer, and A. M. Walker. Lessons in scientific data interoperability: XML and the eMinerals project. *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES*, 367(1890):1041–1046, MAR 13 2009.

[241] M. Wildgrube. Structured data exchange format (sdxf). 2001.

[242] J. G. Wu and R. Hobbs. Key issues and research priorities in landscape ecology: An idiosyncratic synthesis. *LANDSCAPE ECOLOGY*, 17(4):355–365, 2002.

[243] Z. Yang, S. Cai, Z. Zhou, and N. Zhou. Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Information & Management*, 42(4):575–589, 2005.

Appendices

List of XML Languages

The following list represent the markup languages found in the Internet by March 2009 by researching the Internet. The list is not a complete listing, and many markup languages probably have not been reported. Moreover, many of the cited markup languages, have gone through several revisions, therefore, the actual number of XML languages is larger since most of the time posterior revisions are not compatible with previous versions.

1. AConML Automatic Control Markup Language
2. AGXML Agricultural Markup Language
3. AIML Artificial Intelligence Markup Language
4. AML Algebraic Markup Language
5. AML Automation Markup Language
6. AML Astronomical Markup Language
7. AML ASCII Markup Language
8. AML ARIS Markup Language
9. AML Accelerator Markup Language
10. AML Abstract Markup Language
11. AML Annotation Markup Language
12. AML Avatar Markup Language

13. AnatML Anatomical Markup Language
14. AniML Analytical Markup Language
15. AORSML Agent Object Relationship Markup Language
16. APML Attention Profiling Mark-up Language
17. ArcXML ArcGIS Markup Language
18. ATOM Atom Syndication Format
19. BDBXML Berkeley Database Engine Extensible Markup Language
20. BEML Body Expression Markup Language
21. BIOML The BIOpolyer Markup Language
22. BML Behavior Markup Language
23. BML Broadcast Markup Language
24. BML Bean Markup Language
25. BSML Binding Schema Markup Language
26. BSML Bioinformatics Sequence Markup Language
27. BXML Backbase eXtensible Mark-up Language
28. CAAML Canadian Avalanche Association Markup Language
29. CabosML carbohydrate sequence markup language
30. CBML Comic Book Markup Language
31. CCML Computational Chemical Markup Language
32. CCXML Control eXtensible Markup Language
33. CDML Climate Data Markup Language
34. CellML Cell Markup Language
35. CFML ColdFusion Markup Language

36. ChessGML Chess Game Markup Language
37. ChessML Chess Markup Language
38. ChordML ChordML
39. CHTML Compact Hypertext Markup Language
40. CML Chemical Markup Language
41. CML CoolMon Markup Language
42. CML Cinematographic Markup Language
43. CML Character Markup Language
44. CMML Continuous Media Markup Language
45. CMSML Content Management Systems Markup Language
46. ComicsML Comics Markup Language
47. CSML Cave Survey Markup Language
48. CSML chemical structure markup language
49. CSML Cell System Markup Language
50. CSML Climate Science Modelling Language
51. CSML Customizable Swing Markup Language
52. CSML Component/Constraint Specification Markup Language
53. CSML Chemical Style Markup Language
54. CVML Computer Vision Markup Language
55. CWML Cyclone Warning Markup Language
56. CWML Camera-Work Markup Language
57. CWML Compact Wireless Markup Language
58. DAML DARPA Agent Markup Language

59. DAML Directory Assertion Markup Language
60. DBML Data Base Markup Language
61. DC Dublin Core DCML datacentre Markup Language
62. DCML Disney Comics Markup Language
63. DCML Data Center Markup Language
64. ddXML duct design eXtensible Markup Language
65. DHTML Dynamic Hyper Text Markup Language
66. DML Document Markup Language
67. DML Dialog Markup Language
68. DML Database Markup Language
69. DML Development Markup Language
70. DML Disambiguation Markup Language
71. DML Dependency Markup Language
72. DRML DTD Data-entry and Reporting Markup Language
73. DSML Directory Services Markup Language
74. EACML Extensible Access Control Markup Language
75. ebXML Electronic Business using eXtensible Markup Language
76. ecgML Electrocardiogram Markup Language
77. ECTXML ECT's open extensible markup language
78. EDML Electronic Data Management Language EDT-ML
79. ETD-ML Electronic Thesis and Dissertation Markup Language
80. eLML eLesson Markup Language
81. EMI-ML Experiment Meta Information Markup Language

82. EML Election Markup Language
83. EML Expression Markup Language
84. EML Ecological Metadata Language
85. EML Educational Modelling Language
86. EPML EPC Markup Language
87. ESML EARTH SCIENCE MARKUP LANGUAGE
88. ETD-ML Electronic Thesis and Dissertation Markup Language
89. FDML Forms Data Markup Language
90. FITSML Flexible Image Transport System Markup Language
91. FlowML FlowML
92. FML Function Markup Language
93. FML Flash Markup Language
94. FML Form Markup Language
95. FML FAQ Markup Language
96. FOAF Friend of a friend
97. FpML Financial products Markup Language
98. FSML Financial Services Markup Language
99. GAME Genome Annotation Markup Elements
100. GAML Generalized Analytical Markup Language
101. gbXML green building extensible markup language
102. GCTML Glider Coordinated Trajectories Markup Language
103. GDML Geometry Description Markup Language
104. GEML Gene Expression. Markup Language

105. GeoXACML Geospatial Extensible Access Control Markup Language
106. GHML Gas Hydrate Markup Language
107. GIML Gastro Intestinal Markup Language
108. GML GUI Markup Language
109. GML Generalised Markup Language
110. GML Geography Markup Language
111. GnowML Gnowledge Markup Language
112. GovML Governmental Markup Language
113. GPML GPlates Markup Language
114. GSVML Genomic Sequence Variation Markup Language
115. HDML Handheld Device Markup Language
116. HGML Hyper Graphics Markup Language
117. HML Hybrid Mail Language
118. HML Histoimmunogenetics Markup Language
119. HML Human Markup Language
120. HTML HyperText Markup Language
121. HydroML Hydrologic Markup Language
122. IML Interface markup language
123. IML Information Markup Language
124. IML Instrument markup language
125. IML Image Markup Language'
126. InfoML information markup language
127. IRCML IRC Markup Language

128. JCML Job Control Markup Language
129. JDL Job Definition Language
130. JDML Job Description Markup Language
131. JML JSP Markup Language
132. JML Journalism Markup Language
133. JML Juggling Markup Language
134. JPML Jpartner Markup Language
135. JSDL Job Submission Definition Language
136. JSML Java Speech API Markup Language
137. JsonML JSON Markup Language
138. KML Keyhole Markup Language
139. KML Knowledge Markup Language
140. LDML Locale Data Markup Language
141. LitML Liturgical Markup Language
142. LML Lisp Markup Language
143. LML Lightweight Markup Language
144. LML List Markup Language
145. LML Lambda Markup Language
146. LMML Learning Material Markup Language
147. LogML Log Markup Language
148. MAGE-ML Microarray Gene Expression Markup Language
149. MAML Microsoft Assistance Markup Language
150. MAML MicroArray Markup. Language

151. MathML Mathematical Markup Language
152. MatML Materials Markup Language
153. MCML Media Center Markup Language
154. MDML Market Data Markup Language
155. MML Modest Markup Language
156. MML Medical Markup Language
157. MML Mashup Markup Language
158. MML Music Markup Language
159. MML Maker Markup Language
160. MNML Music Notation Markup Language
161. MODL Molecular Dynamics Markup Language
162. MOML Media Object Markup Language
163. MOML Media Server Call Control Extensible Markup Language
164. MPML Multimodal Presentation Markup Language
165. MPML MPEG Markup Language
166. MQML Message Queuing Markup Language
167. MRML Multimedia Retrieval Markup Language
168. MSCML Media Server Control Markup Language
169. MSML Media Sessions Markup Language
170. MusicML MusicML
171. MusicXML Music XML
172. MXML Macromedia FleX Markup Language
173. nanoML Nano Markup Language ncML

174. NetCDF Markup Language
175. NeuroML Neuron Markup Language
176. NLML Natural Language Markup Language
177. NML News Markup Language
178. NML Network Markup Language
179. NML Neighbourhood Markup Language
180. NNML Neural Network Markup Language
181. NVML NaVigation Markup Language
182. OIL Ontology Inference Layer OML
183. Ostensible Markup Language OML
184. Ontology Markup Language OML
185. Outline Markup Language OOML
186. Object-Oriented Markup Language
187. OpenXACML Open Source Extensible Access Control Markup Language
188. OPML Outline Processor Markup Language
189. OPML Open Planet Markup Language
190. OSML Optical Spectroscopy Markup Language
191. OWL Ontology Web Language
192. PapiNET PapiNET
193. PDBML Program Call Markup Language
194. PDBML Protein Data Bank Markup Language
195. PDML Panel Definition Markup Language

196. PDML Product Data Markup Language
197. PGML Precision Graphics Markup Language
198. PIML Pathogen Information Markup Language
199. PML Phone Markup Language
200. PML Palm Markup Language
201. PML Physical Markup Language
202. PML Proof Markup Language
203. PML Product Markup Language
204. PML Psychogeographical Markup Language
205. PML Polymorphism
206. Markup Language
207. PMML Predictive Model Markup Language
208. PNML Petri Net Markup Language
209. ProML Protein Markup Language
210. PVML PowerVista Markup Language
211. QML Quest Markup Language
212. QML ' Quantum Markup Language
213. RCML Rich Client Markup Language
214. RDDL Resource Directory Description Language
215. RDF Resource Description Framework
216. RDFS Resource Description Framework Schema
217. REML Recipe Exchange Markup Language
218. RevML Revision Markup Language

219. RiboML Ribonucleic Acid Markup Language
220. RML Regatta Markup Language
221. RML Requirements Markup Language
222. RML Relational Markup Language
223. RML Research Markup Language RML
224. Rule Markup Language RNAML
225. RNA Markup Language
226. RRL Rich Representation Language
227. RTML Role-based Trust-management Markup Language
228. RTML Remote Telescope Markup Language
229. RSS Really Simple Syndication
230. RSS RDF Site Summary
231. RSS Rich Site Summary
232. RuleML
233. Rule Markup Language
234. RWML Road Web Markup Language
235. SAML Security Assertion Markup Language
236. SAML Security Access Markup Language
237. SBML Systems Biology Markup Language
238. SecPAL Security Policy Assertion Language
239. SMIL Synchronized Multimedia Integration Language
240. SML Simple Markup Language
241. SML Spacecraft Markup Language

242. SML Steel Markup Language
243. SOOML Standoff Object-Oriented Markup Language
244. SpatialML Spatial Markup Language
245. SPML Service Provisioning Markup Language
246. SpotsML Sports Markup Language
247. SRML Simulation Reference Markup Language
248. SRML Simulation Reference Markup Language
249. SSML Speech Synthesis Markup Language
250. STMML Scientific Technical and Medical Markup Language
251. SubML Substitutionary Markup Language
252. SVG Scalable Vector Graphics
253. SWML SignWriting Markup Language
254. SWRL Semantic Web Rule Language
255. TML Transducer Markup Language
256. TML Thesaural Markup Language
257. TML Test Markup Language
258. TML Typographic Markup Language
259. TML Terence's Markup Language
260. TML Tutorial Markup Language
261. TML Taxonomic Markup Language
262. UIML User Interface Markup Language
263. UML Unified Markup Language
264. UsiXML USer Interface eXtensible Markup Language

265. VEML Vocal Expression Markup Language
266. VML Vector Markup Language
267. VML Voice Markup Language
268. VoiceXML Voice eXtensible Markup Language
269. VRML Virtual Reality Markup Language
270. VTML Visual Tools Markup Language
271. WAML Web Application Markup Language
272. WasabiXML WasabiXML
273. WeatherML Weather Markup Language
274. WML Wireless Markup Language
275. XACML Extensible Access Control Markup Language
276. XAL eXtensible Application Language
277. XAML Extensible Application Markup Language
278. XBRL Extensible Business Reporting Language
279. XCML Extensible Critter Markup Language
280. XDF Extensible Data Format
281. XGMML eXtensible Graph Markup and Modeling Language
282. XHTML eXtensible HiperText Markup Language
283. XMML eXploration and Mining Markup Language
284. XOML Extensive Object Markup Language
285. XrML eXtensible rights Markup Language
286. XSIL Extensible Scientific Interchange Language
287. XUL XML User Interface Language
288. YAML Yet Another Markup Language

Survey Questionnaire

Surveying Data and Metadata usage among Researchers and Managers Working with Natural Resources

About you Your Scientific Discipline

- The institution where you develop your research/management e.g. UC Berkeley, Lawrence Berkeley National Laboratory
- Multidisciplinary research projects where you are an active member e.g. Long Term Ecological Research, FluxNet, Tropical Atmosphere Ocean Project (one per line)
- Public Data Repositories that you use (or know about it) in your Discipline (one per line)
- Check the categories that describe your computer habits (check as many as necessary as long as they are true statements about you

About XML Archives

- XML stands for eXtensible Markup Language; have you hear about XML before?
- If you have, have you used XML to exchange data among your peers?
- If you have, what XML Languages have you used (you could select more than one language)
- If you use any XML language not included in the previous list, please write them below (one per line)

About Other types of Archives for Data or Metadata

- This icon stands for Rich Site Summary (RSS); have you seen it before?
- If you have, have you used RSS to collect and/or provide metadata?

- The Ecological Metadata Language (EML) is a metadata language for data-centric documents; have you heard about it before?
- If you have, have you used the EML to collect and/or provide metadata?
- The DCMI stands for Dublin Core Metadata Initiative; have you heard about it before?
- If you have, have you used the DCMI to collect and/or provide metadata?
- The Hierarchical Data Format (HDF) is a type of archive for data; have you heard about it before?
- If you have, have you used the HDF to archive or share data?
- The network Common Data Form (NetCDF) is a type of archive for data; have you heard about it before?
- If you have, have you used netCDF archives to archive or share data?
- The eXternal Data Representation (XDR) is a type of archive for data; have you heard about it before?
- If you have, have you used XDR archives to archive or share data?
- Spreadsheet is a type of archive for data; have you heard about it before?
- If you have, have you used spreadsheet archives to archive or share data?
- The DBase is a type of archive for data; have you heard about it before?
- If you have, have you used DBase archives to archive or share data?
- The Tab or Comma Separated Value (CSV) is a type of archive for data; have you heard about it before?
- If you have, have you used CSV archives to archive or share data?

- Yes No The Data Interchange Format (DIF) is a type of archive for data; have you heard about it before?
- If you have, have you used DIF archives to archive or share data?
- The Structured Data Exchange Format (SDXF) is a type of archive for data; have you heard about it before?
- If you have, have you used SDXF archives to archive or share data?
- The Flexible Image Transfer System (FITS) is a type of archive for data; have you heard about it before?
- If you have, have you used FITS archives to archive or share data?

About Another Prevalent Data Archives Used in your Discipline

- Can you name another prevalent 'filetypes' that YOU use in your discipline to archive or share data?(one per line)
- Can you name other prevalent 'filetypes' used in your discipline to archive or share data?(one per line)

XML documents in the Web Space

This table compilates the results from searching XML documents belonging to the different XML languages. To find these documents, the search engine was asked to return indexed files that they had a xml-extension and the URI of the location of their respective schemata; e.g., to estimate the number of GML-files the engine was asked to return the number of xml files which contain http://www.opengis.net/gml inside the file.

This surveyed was realized in May 2009.

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Really Simple Syndication	5330000	1470000

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
RDF Site Summary	16000000	1390000
Dublin Code	2370032	1370000
Friend of a friend	1180000	1190000
Resource Description Framework	834000	901000
Rich Site Summary	1040000	830000
Scalable Vector Graphics	960000	689000
Wireless Markup Language'	273000	289000
MusicXML	130000	177000
Synchronized Multimedia Integration Language	202000	157000
Vector Markup Language	206000	145000
Extensible Business Reporting Language	164000	137000
Security Assertion Markup Language	147000	134000
Geography Markup Language	113000	103000
Chemical Markup Language	90500	99200
Systems Biology Markup Language	80000	94500
Mathematical Markup Language	118000	82900
Document Markup Language	130000	81300
Outline Processor Markup Language	87900	74800
eXtensible Application Markup Language	108000	71200
XML User Interface Language	78200	61500
Voice eXtensible Markup Language	58100	59200
Extensible Access Control Markup Language	58700	54700
Extensible Access Control Markup Language	58700	54700
DARPA Agent Markup Language	54600	54000
ColdFusion Markup Language	56500	48000
Handheld Device Markup Language	78700	44300
Simple Markup Language	54400	42700
Financial products Markup Language	59000	42600
Keyhole Markup Language	67500	41600
Virtual Reality Markup Language'	45700	40600
Speech Synthesis Markup Language	48100	39900
Ontology Web Language	15500	32800
Generalised Markup Language	11200	30800
PapiNET	17400	29400

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Ontology Inference Layer	3190	25400
Interface markup language	15000	25200
Semantic Web Rule Language	2940	25000
eXtensible rights Markup Language	14200	24000
Attention Profiling Mark-up Language	2630	23900
User Interface Markup Language	13400	23400
Directory Services Markup Language	5920	21000
Educational Modelling Language	1630	20500
Rule Markup Language	2690	20200
Rule Markup Language	2690	20200
Physical Markup Language	3040	17500
Artificial Intelligence Markup Language	3120	16100
Predictive Model Markup Language	2100	15200
Service Provisioning Markup Language	2490	14200
Expression Markup Language	1780	13400
Resource Directory Description Language	1520	12700
Gene Expression. Markup Language	1640	12500
Data Center Markup Language	2360	12400
Electronic Business using eXtensible Markup Language	1270	12400
Broadcast Markup Language	1330	11500
Media Center Markup Language	7590	10600
Election Markup Language	1260	10300
Information Markup Language	1070	9950
information markup language	1070	9950
MusicML	917	9540
Petri Net Markup Language	883	9310
Form Markup Language	100	4870
Web Application Markup Language	293	3700
Psychogeographical Markup Language	2120	2830
Log Markup Language	384	2760
Lisp Markup Language	2520	2760
Knowledge Markup Language	316	2710
eXtensible Graph Markup and Modeling Language	322	2470
Financial Services Markup Language	351	2250

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Java Speech API Markup Language	229	2100
Music Markup Language	555	2000
Development Markup Language	573	1990
Transducer Markup Language	723	1750
eXploration and Mining Markup Language	157	1730
Remote Telescope Markup Language	1800	1660
JSON Markup Language	213	1570
Geometry Description Markup Language	190	1530
ChordML	127	1480
Test Markup Language	466	1450
Image Markup Language'	240	1440
Palm Markup Language	500	1390
Learning Material Markup Language	434	1380
Spacecraft Markup Language	312	1360
Abstract Markup Language	648	1240
Macromedia FleX Markup Language	467	1150
Network Markup Language	408	1120
EARTH SCIENCE MARKUP LANGUAGE	391	1080
Quest Markup Language	151	1080
Locale Data Markup Language	976	1080
Proof Markup Language	817	1050
Database Markup Language	541	1020
Road Web Markup Language	228	995
Precision Graphics Markup Language	1220	993
Ecological Metadata Language	1140	984
Voice Markup Language	918	963
Dialog Markup Language	401	951
Phone Markup Language	194	928
Ontology Markup Language	708	925
Human Markup Language	1220	912
Flash Markup Language	928	898
Rich Representation Language	346	893
Microsoft Assistance Markup Language	268	831
Microarray Gene Expression Markup Language	654	825

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Unified Markup Language	268	806
Tutorial Markup Language	330	798
List Markup Language	698	759
Job Definition Language	155	754
Avatar Markup Language	267	751
Bean Markup Language	1050	734
Dynamic Hyper Text Markup Language	8090	725
Astronomical Markup Language	357	709
Extensible Scientific Interchange Language	311	705
Annotation Markup Language	723	705
Control eXtensible Markup Language	1060	703
Simulation Reference Markup Language	193	683
green building extensible markup language	365	683
Simulation Reference Markup Language	193	683
Compact Hypertext Markup Language	592	682
Product Data Markup Language	263	674
Extensible Data Format	583	674
Sports Markup Language	225	654
Yet Another Markup Language	981	640
Medical Markup Language	549	636
Materials Markup Language	353	619
eLesson Markup Language	362	608
Behavior Markup Language	703	604
Multimedia Retrieval Markup Language	477	603
Continuous Media Markup Language	1020	596
USer Interface eXtensible Markup Language	341	594
Lightweight Markup Language	690	580
Climate Science Modelling Language	354	562
Research Markup Language	188	556
Security Policy Assertion Language	503	552
Media Server Control Markup Language	478	538
Program Call Markup Language	635	536
Multimodal Presentation Markup Language	393	527
Product Markup Language	518	522

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
News Markup Language	590	507
Cell System Markup Language	672	506
NaVigation Markup Language	238	457
EPC Markup Language	302	443
Instrument markup language	527	437
JSP Markup Language	494	420
Outline Markup Language	98	420
JSP Markup Language	494	420
GUI Markup Language	91	393
NetCDF Markup Language	224	389
eXtensible Application Language	166	386
Taxonomic Markup Language	152	382
Rich Client Markup Language	319	381
Maker Markup Language	143	379
Genome Annotation Markup Elements	210	378
Compact Wireless Markup Language	99	364
Hyper Graphics Markup Language	124	349
Pathogen Information Markup Language	189	347
Panel Definition Markup Language	263	334
Computer Vision Markup Language	511	318
Weather Markup Language	217	316
SignWriting Markup Language	186	313
Visual Tools Markup Language	343	313
Steel Markup Language	198	313
WasabiXML	332	307
Object-Oriented Markup Language	99	307
Binding Schema Markup Language	190	294
Comic Book Markup Language	206	293
Character Markup Language	340	290
Protein Markup Language	157	274
FlowML	174	265
Media Sessions Markup Language	202	265
Security Access Markup Language	128	263
Function Markup Language	448	258

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
carbohydrate sequence markup language	86	258
Cell Markup Language	105	246
Governmental Markup Language	137	243
Market Data Markup Language	135	234
Liturgical Markup Language	108	232
ASCII Markup Language	75	222
Automation Markup Language	77	220
Polymorphism Markup Language	138	206
Mashup Markup Language	203	203
Analytical Markup Language	85	199
MicroArray Markup. Language	268	195
Automatic Control Markup Language	32	190
ARIS Markup Language	178	187
Electronic Thesis and Dissertation Markup Language	189	179
Hybrid Mail Language	154	177
Accelerator Markup Language	99	172
Requirements Markup Language	63	172
Bioinformatics Sequence Markup Language	146	171
Terence's Markup Language	105	164
duct design eXtensible Markup Language	86	159
Recipe Exchange Markup Language	55	155
Molecular Dynamics Markup Language	153	149
Relational Markup Language	148	146
Generalized Analytical Markup Language	57	140
chemical structure markup language	147	136
Chemical Style Markup Language	218	136
Neural Network Markup Language	178	128
Job Description Markup Language	111	128
Genomic Sequence Variation Markup Language	37	127
Dependency Markup Language	98	124
Chess Markup Language	88	98
Data Base Markup Language	69	83
Chess Game Markup Language	73	83
Music Notation Markup Language	56	81

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Revision Markup Language	48	80
FAQ Markup Language	37	79
Message Queuing Markup Language	42	78
Geospatial Extensible Access Control Markup Language	133	76
Climate Data Markup Language	65	71
'Quantum Markup Language	50	70
Ostensible Markup Language	65	68
MPEG Markup Language	12	58
Thesaural Markup Language	55	55
Canadian Avalanche Association Markup Language	53	53
Hydrologic Markup Language	81	53
GPlates Markup Language	51	52
Anatomical Markup Language	67	46
Backbase eXtensible Mark-up Language	25	44
Gas Hydrate Markup Language	15	44
Cave Survey Markup Language	29	42
Comics Markup Language	47	41
Role-based Trust-management Markup Language	38	38
Spatial Markup Language	6	33
Job Submission Definition Language	46	30
Scientific Technical and Medical Markup Language	36	27
Standoff Object-Oriented Markup Language	21	26
Cinematographic Markup Language	9	25
Modest Markup Language	19	24
RNA Markup Language	15	23
Protein Data Bank Markup Language	14	19
Neighbourhood Markup Language	13	18
Juggling Markup Language	13	17
Algebraic Markup Language	16	10
Neuron Markup Language	32	10
Typographic Markup Language	15	9
Job Control Markup Language	5	9
Berkeley Database Engine Extensible Markup Language	31	8
ETD-ML Electronic Thesis and Dissertation Markup Language	14	8

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Ribonucleic Acid Markup Language	13	8
Histoimmunogenetics Markup Language	3	8
Natural Language Markup Language	17	8
Agricultural Markup Language	34	8
PowerVista Markup Language	280	7
Jpartner Markup Language	319	7
Journalism Markup Language	15	7
Nano Markup Language	8	6
Camera-Work Markup Language	14	6
IRC Markup Language	339	6
Gastro Intestinal Markup Language	9	5
Computational Chemical Markup Language	22	5
Substitutionary Markup Language	18	5
Open Planet Markup Language	18	5
Regatta Markup Language	15	5
Lambda Markup Language	17	5
Agent Object Relationship Markup Language	14	5
CoolMon Markup Language	17	5
Gnowledge Markup Language	5	4
Body Expression Markup Language	1	4
Vocal Expression Markup Language	1	4
Experiment Meta Information Markup Language	4	4
Cyclone Warning Markup Language	13	4
Forms Data Markup Language	50	4
Optical Spectroscopy Markup Language	18	4
Extensive Object Markup Language	12	3
Content Management Systems Markup Language	5	3
Flexible Image Transport System Markup Language	6	3
Open Source Extensible Access Control Markup Language	4	3
Customizable Swing Markup Language	2	2
The BIOpolyer Markup Language	7	2
Component/Constraint Specification Markup Language	1	2
Disney Comics Markup Language	5	2
datacentre Markup Language	49	2

Table 1: Number of XML documents found by Google and Yahoo Search Engines

Language	Yahoo	Google
Directory Assertion Markup Language	1	1
ECT's open extensible markup language	1	1
Glider Coordinated Trajectories Markup Language	3	1
ArcGIS Markup Language	3	1
Disambiguation Markup Language	5	1
Extensible Critter Markup Language	2	1
DTD Data-entry and Reporting Markup Language	0	0
Media Server Call Control Extensible Markup Language	0	0
Electronic Data Management Language	0	0
Media Object Markup Lenguage	0	0
Electrocardiogram Markup Language	0	0

Lifespan of XML Languages

The table below was compilated by searching the Internet through common search engines; in particular the following two sites were very useful:

- Internet Archive - Wayback Machine: http://www.archive.org/web/web.php
- Cover Pages: http://xml.coverpages.org

Table 2: Estimated Lifespan of XML Langauges

XML LANGUAGE	1st Reference	Last Reference
Agricultural Markup Language	2001	2008
Artificial Intelligence Markup Language	2001	2009
Astronomical Markup Language	2000	2005
Astronomical Instrument Markup language	2000	2003
Anatomical Markup Language	2001	2005
Analytical Information Markup Language	2003	2007
ArcGIS Markup Language	2002	2009
BIOpolymer Markup Language	1999	2002
Bioinformatic Sequence Markup Language	1999	2001

Table 2: Estimated Lifespan of XML Langauges

XML LANGUAGE	1st Reference	Last Reference
carbohydrate sequence markup language	2005	2005
Catalyst Markup Language	2003	2003
Compressed chemical markup language	2002	2003
Computational Chemical Markup Language	2002	2003
Climate Data Markup Language	2002	2005
Cell Markup Language	2000	2009
COLDFUSION markup language	2000	2009
Chemical Markup Language	1995	2008
Chemical Markup Language for Computational chemistry	2007	2008
CLP-ML	2006	2006
chemical structure markup language	1995	1997
Climate Science Modelling Language	2005	2009
Cave Survey Markup Language	1999	2001
Cell System Markup Language	1999	2009
Chemical Style Markup Language	1998	1998
CytometryML	2003	2007
Clustering Markup Language	2005	2005
discontinuous deformation analysis markup language	2001	2001
Electrocardiogram Markup Language	2003	2004
Environmental Markup Language.	2000	2000
Ecological Metadata Language	2001	2009
EMG-extensible markup language	2002	2002
EarthquakeML	2005	2005
Earth Science Markup Language	2001	2004
Field Markup Language	1999	2003
Fusion Simulation Markup Language	2005	2007
Flexible Image Transport System Markup Language	2000	2001
Field Markup Language	2007	2007
FUGE Markup Language	2006	2007
Fusion Simulation Markup Language	2005	2005
Herbal Medicine Markup Language	1999	1999
Gas Hydrate Markup Language	2007	2008
Gating-ML	2006	2008
Genome Annotation Markup Elements	1999	2000
Genomic Contextual Data Markup Language	2008	2008

Table 2: Estimated Lifespan of XML Langauges

XML LANGUAGE	1st Reference	Last Reference
Gene Expression Markup Language	1999	2000
GeneX Gene Expression Markup Language	1999	2001
Geospatial Extensible Access Control Markup Language	2002	2009
GeoScience Markup Language	2003	2005
Gas Hydrate Markup Language	2000	2009
Gastro Intestinal Markup Language	2001	2001
Geography Markup Language	1998	2009
Genomic Sequence Variation Markup Language	2006	2007
Health-checkup Data Markup Language	2000	2002
Histoimmunogenetics Markup Language	2005	2007
Hydrologic Markup Language	2003	2004
Human Proteome Markup Language	2002	2006
image markup language	2000	2000
Interactive Image Markup Language	1996	1996
Keyhole Markup Language	2001	2009
Materials Markup Language	1999	2005
MicroArray and Gene Expression Markup Language	2001	2009
Microarray Markup Language	2000	2002
Medical Markup Language	1999	2009
Molecular Dynamics Language	1999	2000
XML for Multiple Sequence Alignments	2000	2000
molecular spectroscopic simulations markup language	2008	2008
NetCDF Markup Language	2000	2007
NeuroML	2001	2009
Optical Spectroscopy Markup Language	2003	2003
PapiNET	2000	2009
Protein Data Bank Markup Language	2005	2009
Pathogen Information Markup Language	2005	2005
Physical Markup Language	2003	2009
Protein markup language	2002	2002
Protein Extensible Markup Language	1999	2001
RiboML	2002	2005
RNA Markup Language	2002	2002
SleepXML	2004	2004
survey markup language	2002	2004

Table 2: Estimated Lifespan of XML Langauges

XML LANGUAGE	1st Reference	Last Reference
Scientific Technical Medical Markup Language	2002	2003
Spectro ML-	2001	2004
Systems Biology Markup Language	2000	2009
Taxonomic Markup Language	2000	2000
UnitsML	2000	2009
Vector Markup Language	1998	2009
XyM markup language	1999	2005
Flow cytometry markup language	2002	2003
Cytometry markup language	2006	2009

Bibliometrics

The bibliographical corpus was developed by downloading references from UC Berkeley Libraries by keywords,

Number of References to Peer-reviewd Articles by Year and Keywords

For Collaboration the following keywords were used: inter-disciplinary,multi-disciplinary,trans-disciplinary, cross-disciplinary, and collaboration

Table 3: DemostrationTitle

year	Collaboration	XML	Metadata
1980	565	0	0
1981	540	0	0
1982	545	0	2
1983	541	0	5
1984	724	0	2
1985	689	0	7
1986	758	0	11
1987	937	0	10

Table 3: DemostrationTitle

year	Collaboration	XML	Metadata
1988	927	0	9
1989	812	0	8
1990	1071	0	19
1991	1340	0	22
1992	1369	0	23
1993	1390	0	32
1994	1677	0	59
1995	1925	0	52
1996	2074	2	110
1997	2047	5	157
1998	2261	53	203
1999	2434	91	317
2000	2611	166	411
2001	1179	11	10
2002	3019	0	646
2003	3189	297	806
2004	3481	248	969
2005	4029	230	1114
2006	4425	271	1198
2007	5134	215	1005
2008	5337	162	1103

Number of Authors per Article by Discipline

Table 4: DemostrationTitle

Authors	Environment	Genetics	Sociology	Physics	Mathematics
1	199	105	232	29	177
2	347	249	104	62	151
3	369	233	33	104	73
4	321	214	18	80	18
5	260	195	7	64	7
6	186	159	1	71	2
7	101	105	0	33	0

Table 4: DemostrationTitle

Authors	Environment	Genetics	Sociology	Physics	Mathematics
8	64	98	0	16	0
9	52	60	1	18	0
10	36	51	2	7	0
11	14	35	3	5	0
12	13	35	4	1	0
13	10	24	5	3	0
14	4	15	6	3	0
15	7	15	7	3	0
16	4	8	8	1	0
17	0	7	9	0	0
18	2	2	10	0	0
19	0	3	11	0	0
20	3	2	12	0	0
21	1	5	13	0	0
22	1	2	14	0	0
23	0	3	15	0	0
24	0	1	16	0	0
25	0	1	17	0	0
26	0	1	18	0	0
27	0	3	19	0	0
28	0	1	20	0	0
29	0	1	21	0	0
30	0	2	22	0	0
31	0	4	23	0	0
32	1	1	24	0	0
33	0	2	25	0	0
34	0	1	26	0	0
35	0	3	27	0	0
36	0	1	28	0	0
37	0	1	29	0	0
38	0	1	30	0	0
39	0	1	31	0	0
40	0	1	32	0	0
41	1	0	33	0	0

Compressed Shapefile with Tree data

File was downloaded from the following website from Cornell University:

- website at http://cugir.mannlib.cornell.edu/bucketinfo.jsp?id=8026

The metadata file for the aforementioned file, which does not have descriptors to describe the physical data of the archive, and follows the common standards from the Federal Geographic Data Committee, can be seen at the following URI:

- http://cugirmetadata.mannlib.cornell.edu/VCHtree.VCH.shp.08026.xml

The archive has 2529 records with the following fields: