AN INFORMATION SYSTEMS STRATEGY FOR THE ENVIRONMENTAL CONSERVATION COMMUNITY

by
Kristin Barker

Master's project submitted in partial fulfillment of the requirements for
the Master of Environmental Management degree in
the Nicholas School of the Environment and Earth Sciences
Duke University 2008

ABSTRACT

As the cause of environmental conservation emerges as a global priority, the need for a practical information systems strategy shared among conservation organizations becomes imperative. Historically, researchers and practitioners in conservation have met their own information management and analysis needs with inevitable variation in methodology, semantics, data formats and quality. Consequently, conservation organizations have been unable to *systematically* assess conditions and set informed priorities at various scales, measure performance of their projects and improve practices through adaptive management. Moreover, the demands on conservation are changing such that the bottom-up approach to information systems will become an increasing constraint to effective environmental problem solving. Where we have historically focused on the protection of "important" places and species and more recently "biodiversity," conservation is moving to a systems view, specifically ecosystem-based management, where relationships and process are as important as the individual elements. In parallel, awareness of the human dependency on functioning natural systems is on the rise and with it the need to explicitly value ecosystem services and inform tradeoffs. Climate change requires conservation to develop dynamic adaptation scenarios at multiple spatial and temporal scales. Credible assessments and effective conservation action increasingly rely on collaboration from multiple organizations and disciplines. Finally, the business of conservation is under increased pressure to account for its spending and objectively measure outcomes of its strategies. All of these changes translate to growing, not shrinking, demands on information and information systems.

In response to these challenges, this research presents an information systems strategy for the environmental conservation community. It proposes the development of a distributed systems infrastructure with end-user tools and shared services that support standardized datasets. Key strategies include removing the barriers to information sharing, providing valuable tools to data producers and directly supporting heterogeneity in conservation datasets. The strategy concludes with a call for high-level management involvement in information systems strategy and collaborative investment in implementation by the conservation community, partners in government and donors. Without these steps, conservation as an industry may find itself ill-equipped to meet the changing needs of people and nature.

Table of Contents

Introduction

The project is intended to a compelling case for shared investment in an information systems platform for the conservation community. The ultimate goal of the project is to provide the foundation for more systematic conservation across the conservation community, from large non-profits and natural resources agencies to small land trusts.

The term "conservation information systems" refers to those information systems that address the information types, problems, and tasks specific to biodiversity conservation. Content areas include field-observations, species occurrence and habitat mapping, species viability analysis, ecological assessments and trends analysis, land and water restoration, management of stewardship activities, measurement of conservation project performance and improved practices through adaptive management. In all of these, researchers and managers need to collect, organize and manage raw information as well as query, visualize (e.g., via geographic maps), analyze, summarize, share results, and guide decision making at various scales. Generally speaking, information systems facilitate productivity, analysis, workflows, communications, process improvement and accountability and can have a transformative effect on business process.

The document begins with a presentation of the case for investment in conservation information systems in the context of a changing conservation agenda. I then describe the primary challenges in the development of effective information technology for conservation, most notability a fundamental heterogeneity in conservation datasets and analyses. The strategy follows beginning with the development of a distributed conservation information systems infrastructure in which cooperating data nodes provide secure hosting of standardized, spatially explicit datasets for their respective geographies. Companion tools allow local users to collect, aggregate, manage and analyze these datasets. I stress the importance of end-user applications and utilities that bring direct benefit to data producers and, at the same time, maximize interoperability of their datasets. The strategy then describes how emerging semantic mediation technologies can address the fundamental challenge of necessary variation in the structure of conservation data. Finally, to meet the practical challenges in implementation and evolving needs, the strategy proposes specific mechanisms to organize and galvanize investments by conservation organizations and their partners in conservation information systems.

I have designed the strategy to build alignment and foster investment in phases starting within large-scale conservation organizations such as The Nature Conservancy, Conservation International, and World

Wildlife Fund with increasing reach into the conservation community. The primary audience for the strategy is senior managers, including but not limited to IT managers. Science and conservation practice managers may also be interested as well as donors with an interest in the power of information technology to bring new levels of efficiency, efficacy and accountability to conservation. I have written this strategy for the non-technical reader who is interested to understand information systems and their relationship to the problems in conservation. Wherever necessary, I have attempted to explain technical concepts and relate them to examples from conservation.

The Case for Conservation Information Systems

Conservation of natural systems and their elements is dependant on an understanding of complex, interrelated biophysical and socioeconomic factors operating at multiple temporal and spatial scales. That understanding in turn depends on effective information collection, management, analysis and communication to support decision making. Yet the conservation community, as an industry, lacks the information systems infrastructure to *systematically* assess conditions, set informed priorities, measure performance of projects and improve practices through adaptive management. We lack even the most basic unified views of protected areas, species and ecological community distributions and conservation projects. John Wiens, lead scientist at the Nature Conservancy bemoaned the "data management problem" as a threat to science-based conservation and adaptive management (Wiens & Comendant 2005). In 2002, The Heinz Center released its report on ecosystem conditions in the United States, "The State of the Nation's Ecosystems." In this report, the authors declared the assessment incomplete due to the lack of data collection, reporting and systems infrastructure to sufficiently assess ecosystem condition (Clark 2006). The Millennium Assessment project identified significant information challenges in its analysis: incomplete, uneven coverage, incompatible collection methods, lack of metadata, and data reliability (MA 2005).

Moreover, the demands on conservation as a business are changing such that the bottom-up approach to information systems will become an increasing constraint to environmental problem solving. Conservation has historically focused on the protection of places and species on the basis of intuitive values and, more recently, "biodiversity" with more sophisticated analyses of critical species habitat, richness, rarity or irreplacibility. Yet, by necessity, the conservation agenda is in motion on at least five significant fronts, all with significant ramifications for information management:

1. Management of *ecosystems*: the systems-view of nature recognizes the need to move beyond individual species and places to address complex ecological relationships and process. Ecosystem-based management requires sophisticated modeling of ecosystem dynamics to, for example, predict cascade effects of species loss and entire shifts in regimes (Wu & Hobbs 2002).

2. Conservation biology increasingly recognizes that the geographic *scale* at which analyses are performed changes the questions asked and answered (Redford et al. 2003). As a result, multi-scale assessments are required to effectively inform decision making within a given region (Zermoglio et al. 2005).

3. Awareness of the human dependency on functioning natural systems is on the rise and with it the need to explicitly value the services provided by functioning natural systems. This view recognizes conservation's role in informing tradeoffs in the ongoing human domestication of nature (Kareiva & Marvier 2007). Valuation of *ecosystem services* depends on highly quantitative, spatially-explicit, multi-scaled analyses based on both biophysical and socioeconomic datasets (Nelson et al. forthcoming).

4. *Global climate change* requires conservation to look into the future. We must develop models of biodiversity response to changing conditions at a scale that can inform natural resource management and landscape planning (Root & Schneider 2006). These analyses themselves must respond to improved prediction algorithms and increasingly granular and refined datasets.

5. The "go it alone" strategy has reached its apex. Both assessments and action increasingly require conservation organizations to *collaborate* with each other, partners in government and the private sector. To effect decision making assessments across the spectrum of conservation subjects, from the condition of individual species to integrated regional land-use planning, increasingly require contributions from multiple organizations and disciplines. Similarly, implementation of conservation projects more often involves active participation of cooperating organizations (McShane 2003). These collaborations depend on information sharing and integration.

6. Finally, the business of conservation is under increased pressure from the donors and the public to *account for its spending* and *objectively measure outcomes* of its strategies (Christensen 2002; Ferraro & Pattanayak 2006). Adaptive management specifically requires that we do not "wait for science" but rather measure and respond to measurement of our actions themselves (Lovejoy 2006). We must systematically account for project costs, benefits and strategies.

All of these changes translate to growing, not shrinking, demands on effective information systems. Information technology, when effectively designed, implemented and maintained, reduces costs and creates opportunities. Moreover, information technology can fundamentally transform business processes. Private sector organizations pursuing a specific information technology strategy have on average 20% higher profits (Weill & Ross 2004). Conservation must adapt to the changes underway in its core business. We risk relevancy to society if we do not invest in the information systems capacity required to credibly inform human impacts and dependencies on nature.

Challenges

There are three fundamental challenges to the development and maintenance of information systems for conservation. First and foremost is the basic heterogeneity of conservation information and analysis followed by the accessibility, transparency and sensitivity of current conservation datasets and finally the organizational challenges to understanding, prioritizing, funding and controlling investments in conservation systems. Each of these is described in more detail below.

1. Conservation information is fundamentally heterogeneous.

The primary challenge to the development of conservation information systems arises from the problem domain: natural systems. Natural systems are characterized by enormous variability of actors, that is, species and natural communities, each with their own complexity, as well as complex ecological relationships and processes. What matters depends on the focus, location, and scale of a conservation effort. In addition, human understanding of natural systems, their components and processes is incomplete, constantly evolving and described from diverse perspectives. Assessment and management of natural systems, as the core functions of environmental conservation, must therefore accommodate this fundamental variability, complexity and evolution in their subject matter. Consider that business domains such as finance or manufacturing are human-engineered and can therefore follow a top-down model of information systems design. In contrast, modern conservation, if we wish it to be based in science, must somehow represent its diversity of semantics (meaning), schema (data structure) and analyses in a bottom-up fashion (Ives et al. 2005). These issues form the central challenge to conservation information systems. Either by reduction or explicit support, information systems for conservation must somehow accommodate a basic heterogeneity in their subject matter.

Apart from schematic and semantic variation in conservation datasets, a related challenge arises from syntactic variation: the diversity of data format and transparency. The vast majority of data to inform conservation is collected, managed and analyzed for a specific purpose and without consideration for its utility to other inquiries or different spatial and/or temporal scales. Data format is almost always a function of available tools. Description of the data itself (i.e., metadata) is a low priority to data producers who have little to benefit from the added cost of annotation. This leaves other potential users with the expensive task of interpreting meaning and methods, and therefore reliability, from the data themselves.

Variability in schema, semantics and syntax also frustrates the development of general-purpose information systems to capture, manage and analyze conservation datasets. When information systems are developed for conservation, they support a specific methodology and therefore embed specific schema and semantics. To be functional to end-users, each system provides to varying degrees its own user interface, reporting, mapping, security, and import/export functionality. When the subject matter, focus or methodology varies, new systems are often developed including all of the supporting functionality (IABIN 2004). It is as if each dissertation, because of its unique content, required the development of a new word processor. The tight coupling of methodology and information systems combined with the considerable costs of information systems development has meant that much of conservation research and practice go unsupported by information systems. When tools are developed, usually on shoe-string budgets, end-users frequently suffer insufficient usability and functionality as well as a system that cannot keep up with changing practices. The diversity in conservation's domain has thus far limited the potential return on investments in custom information systems development.

In response to these challenges, the Taxonomic Data Working Group (TADWG), the Global Biodiversity Information Facility (GBIF), the Conservation Commons and other groups have developed data sharing policies and data format standards. "Data standards" are lauded as the cure for the menace of variation in conservation information. However, in the absence of good tools for the data producer, standards conformance benefits only consumers of conservation information and at sometimes considerable cost to data producers. Moreover, standards efforts are effective when variation can be resolved with communication and negotiation. Standard data models cannot address situations where variation is *irreducible*. Given the lack of incentives to implement standards as well as the enormous variation inherent in the problem domain, data producers' conformance to standards is unlikely to be realized. Until standards conformance is either passive or heavily incented, its ability to facilitate broad-scale integration of conservation datasets will be limited.

2. Conservation information is frequently isolated.

Apart from the practical integration issues that arise from conservation data's variability, potential users are simply unable to access many datasets of interest for several reasons. While data collectors and managers decreasingly collect their datasets in paper notebooks, the spreadsheet on a personal computer remains a popular data storage tool for conservation data. In other cases, researchers and practitioners develop personal or shared databases. Even these digitized datasets, however, remain offline, inaccessible to researchers and practitioners who would put these data to important use in conservation.

Online access requires data hosting services and uploading tools which may be unavailable and cost-prohibitive to producers of valuable conservation information. Some reluctance to sharing is explained by risk exposure. Field researchers are often obligated not to publicize the locations of rare species or ecological inventory of private lands (Barker 2008; McShane 2003). In academic settings, the considerable cost and research value field data also prevents liberal sharing, especially prior to but even after publication of results. Finally, data producers frequently have little incentive to share their datasets. Even if a given data producer believes sharing is "good," sharing generally benefits someone else somewhere else and usually comes at considerable cost and sometimes risk (Barker 2008). When incentives do arise, researchers will share datasets but most often on an individual request basis that does not scale to broad accessibility. Overall, conservation as an industry lacks the mechanisms, including hosting, publishing tools and security, as well as the incentives for data producers to share their datasets. Consequently, the industry is generally unable realize the long term value of its information assets.

3. Conservation organizations have underinvested in information systems.

Determining the overall spending by on conservation information systems is outside the scope of this research. Consider, however, that average private sector spending in information systems is 4.2% of annual revenues and rising (Weill & Ross 2004). Given the pressure on non-profit organizations in general to keep total overhead at or below 12%, it is hard to imagine that any conservation organizations are devoting one third of that total to information systems.

Donors are not inclined to specifically fund information systems development over direct conservation action (Barker 2008). Academia is sometimes looked to for tools to assist in conservation. Yet very few solutions originating from academic institutions have seen wide-spread adoption in conservation. As in other disciplines, academia is generally a good source for methodology and algorithms but a poor source of large scale, deployable solutions. Finally, the combination of complexity and market size has thus far prevented for-profit companies from developing solutions specifically for conservation. So it appears that we in the conservation community are on our own to provide the strategy, requirements and implementation of conservation information systems.

Leadership of conservation organizations may be generally skeptical of IT investment and with some good reason. IT is expensive to develop and can fail to demonstrate promised returns (Barker 2008). Conservation planning tools in particular may have been oversold, ultimately limited by the demands of sophisticated software engineering and lack of compatible sources to inform their data-hungry analyses

(Barker 2008). Yet, to-date, information technology to address these problems, that is, facilitate data development, integration and online access, has not been a priority.

Additionally, conservation organizations have been hesitant to develop expertise in information systems development; their IT leaders frequently have a science, not IT background (Barker 2008). Information systems analysis and design is its own rich discipline involving the development of user-friendly, fast, reliable, interoperating, extensible and durable software architectures. This discipline thus has a central role in the success of conservation IT especially in light of the significant challenges inherent in the subject matter and growing challenges in requirements.

The Information Systems Strategy

While the challenges presented above may seem daunting, there is a way forward. It will take time, expertise and commitment to realize. By combining principles and technologies from the software industry with insights into the problem domain, I will present here an information systems strategy for the conservation community, one that can be implemented overtime, within and among organizations, delivering value along the way.

There are four principals that inform the six sections of the strategy described below. First, because a) so much of conservation analysis is characterized by complexity, variability and dynamism and b) coverage of interoperable data is frequently the limiting factor in credible, multivariate analysis, therefore key information systems for conservation must be architected from the *bottom up*. For many conservation problems, users must be able to assemble their own solutions from compatible components, whether those components are interoperable datasets, applications and utilities or fine-grained definitions of data. That said, for other conservation problems, including some in climate change and ecosystem services valuation, datasets and methodologies are sufficiently standardized to warrant investment in end-user tools. We simply have to understand the conditions for success. Second, while standards play a role in interoperability, they are by themselves not useful. Only when standards are supported by *powerful, usable tools, especially for data producers*, do they have value. Third, *user-driven iterations* play a crucial role in the development of successful information systems. For instance, this strategy recommends against the upfront development of a comprehensive set of standards for conservation datasets. Rather, we begin with those that promise immediate benefit to data producers, bolster them with tools that fulfill the promise and build from there. Thriving information systems, like ecosystems, are the product of continuous iteration. Finally, effective information systems strategies depend on *organizations*, specifically on technical expertise and ongoing engagement of executive leadership.

Consistent with the “bottom up” principal, the strategy begins by proposing a distributed information infrastructure for conservation data, follows with prioritized standards selection or development which are then supported by end user applications. The core architecture is completed with the addition of variety of interoperable utilities to make the system practical and useful including data import, dynamic transform, analysis, aggregation, workflow facilitators, and more. Then, based on emerging semantic web technologies, the strategy poses an approach and suite of tools to provide direct support for heterogeneity of conservation data and analysis. The strategy concludes with specific steps for organization and community engagement.

1. Online Access to Conservation Data

As mentioned above, potential contributors of valuable conservation information such as land trusts, small research institutions and species-focused conservation non-profits often lack the funding and expertise to enable access to information they have developed or assembled. To address the isolation of conservation datasets, we propose the deployment of distributed data infrastructure as a service to the full spectrum of conservation data contributors. Regionally-based data nodes can act as shared repository for contributors in the area. The nodes must be configured enable secure access as defined by the contributors. Information assets specific to the region should be cataloged and listed in global directories such as the Conservation Geoportal (www.conservationmaps.org). The ability to share their information with partners in a secure hosted environment with high availability and reliability along with end-user applications and analytical tools, described later in the strategy, will encourage contributors' participation. Data management nodes will also provide a systems platform for these data and analytical services and end-user applications.

The development of this network of nodes is partially underway. The core of the Nature Conservancy's Conservation Information Systems Strategy is the deployment of "Data Management Node" network to be configured along ecological boundaries (TNC 2006). These nodes will act as a well-managed repository for the Conservancy's strategic datasets in their local geography and allow GIS analysts their locale to perform ad-hoc queries and analysis, including but not limited to conservation priority setting, based on these core conservation datasets. Node managers will work with Conservancy offices at the state and local level to aggregate datasets, normalize them to standards for use by other local parties, as well as regional, national and world-wide purposes. The Conservancy's strategy specifically identifies the opportunity and benefits to hosting data from other conservation organizations (TNC 2006).

Conservation International is developing a network of tropical forest field stations worldwide. The Tropical Ecology Assessment and Monitoring Network (TEAM) is monitoring long-term trends in biodiversity using standardized protocols in Central and South America. These field stations, currently numbering 14, may provide a natural data hosting and access role for their contributors and partners.

Large conservation organizations and academic institutions, having both the interest and the capacity, are currently the only potential providers of data hosting services. The systematic investments described here have begun and will prove (or disprove) the viability of this strategy. However, hardware, software and

administration costs may ultimately require a more formal approach including shared funding (e.g., through the Conservation Commmons) or donor-sponsored funding.

Summary of Information Benefits

A distributed network of data hosting providers would enable systematic sharing of conservation information, reducing organization IT capacity as a barrier. The network would provide all developers of conservation information that otherwise lack this capacity with a reliable, available and secure location for online access to their datasets. These datasets can then be more easily shared with partners and aggregators such as the Global Biodiversity Information Facility. Cataloguing capabilities of the hosting service would build awareness in the conservation community of datasets (leveraging such systems as the Conservation Geoportal). The distributed data centers will provide a foundation for data integration, analysis and end-user applications described later in this strategy. Indeed, a data hosting capacity, online access and systems platform are all pre-conditions for the rest of strategy proposed here. Finally, as understanding of the economic and emissions costs of data centers increases (Robb 2007), shared hosting services will enable conservation to share costs, power consumption and leverage associated best-practices.

Sample Conservation Benefits

The direct benefits to conservation of widely-available, shared hosting are limited. Many barriers to effective information sharing remain and will be addressed later in this strategy. At a basic level, data hosting services enable access and therefore improve participation and thus the coverage of conservation information resources. Systematic registration of datasets in directories such as the Conservation Geoportal may improve the efficiency of conservation research and assessments. As a more specific example, online access to biological monitoring information can improve researcher collaborations, improve secondary analysis such as ecosystem-based management and climate change and highlight information gaps.

Challenges

Costs of hardware and software for new data nodes will be significant as will the ongoing maintenance of these assets. Node administration costs will similarly be significant and ongoing. In addition, this strategy may enable a “tragedy of the data commons” wherein contributors may take advantage of hosting services and data provided by others while refrain from contributing their own datasets and/or enabling access. Finally, online access to datasets may not be sufficient incentive for potential contributors to participate.

Where to Start?

- Formalize the criteria for access, availability, reliability and security for conservation data nodes.
- Publicize through presentations and whitepapers a Data Management Node "recipe" consisting of hardware, software and data standards for implementation by those organizations with sufficient capacity.
- Identify existing institutions with the interest and capacity to provide hosting services for conservation data providers in their area.
- Identify funding for the enhancement of candidate nodes to meet the standard as well as the establishment of new nodes.
- Develop simple tools for the maintenance of node inventories and registration in the Conservation Geoportal

2. Interoperability of Conservation Datasets

While heterogeneity of conservation datasets is a challenge, it is neither universal nor debilitating. This strategy addresses the basic challenge in conservation data interoperability by proposing the formalization of information exchange standards for core conservation datasets. The emphasis is on a tiered approach to information exchange, that is, tiered, secure data *interfaces*, and less on the standardization of conservation data itself. This enables data managers to maintain datasets in the most appropriate format while enabling secure access according to the exchange standard. The tiered approach enables standards designers to structure complex domains into hierarchies. For example, ecological assessments may be decomposed into an array of vegetation classes and target species, each with a description of extent and condition that is optionally further decomposed. By structuring complexity into tiers, the exchange standard will give data producers the flexibility to publish at the most appropriate level of detail. Security is paramount to the pragmatic sharing. Many essential contributors of conservation data will not share those data unless they can completely control access.

Open Conservation Exchange

This strategy proposes the development of an Open Conservation Exchange standards program. The program would prioritize conservation data themes and develop exchange protocols based on Web Services technologies. Information categories such as “protected areas”, “ecological observations”, and “ecosystem services” would be prioritized for their strategic value, dependencies and practicality. Using an iterative and tiered approach to manage complexity, standards would be supported by reference implementations by key stake holders as well as utilities and end-user applications (both described later in this strategy) to ensure the practicality of the standard. The data exchange is considered “open” because anyone can participate by producing data to the standard or consuming data through the standard. However, access to specific datasets and even data within datasets is subject to owner control.

The key to a successful standard is establishing the triumvirate of a well-designed protocol and then a critical mass of producers and consumers. Together, these will attract the rest of the industry. For instance, when the Dolby noise reduction standard was developed, both data producers (media suppliers) and data consumers (media players) had to get on board for the standard to take off. Take off it did. In the financial sector, the Open Financial Exchange (OFX) has enabled independent banking institutions to exchange transaction information with customers, other banks and vendors all possessing highly diverse technologies and capacity. The lesson for conservation is that the exchange standard is not enough; we

must also ensure a critical mass of data producers and consumers that make use of the standard. With these in place, the standard attracts secondary producers, consumers, and tool developers resulting in a "network effect" wherein the addition of each new participant improves the value of the network to all participants.

The Services Architecture

This strategy recommends that the Open Conservation Exchange standards are implemented via Web Services. From an information systems point of view, data and analyses can be thought of as "services." Services, if they are compatible, can be combined to produce new data and analysis, just like "Legos." What makes Legos work is their interfaces. The IT industry has been developing the equivalent of Legos, Internet-based protocols that enable interoperability of disparately developed data and analysis, for the past decade and calls them "Web Services." Web Services, as systems integration technology, are now pervasive in both private and public sector information systems (Manes 2003). An individual web service is a component that can stand on its own and provides a "service" to other components. It does not have a user-interface but rather provides a programmatic interface so that other information systems can make use of its capabilities. Appendix A: shows how the Web Services approach can be applied to individual applications and a suite of solutions. From web-based applications to integration infrastructures and now even mobile devices, Web Services enable the owners of information assets to centrally maintain the integrity, security and business logic while enabling access to both internal and external clients. This data owner may even dramatically change its underlying structure, format and implementation technology, yet so long as the Web Service interface is maintained, dependant clients remain operable. Thus, the Web Services interface is a technical contract that allows both data producer and consumer to evolve according to their own needs while maintaining their beneficial relationship. Because they can support multiple interfaces or contracts, a substantial benefit of Web Services protocols is support for an evolving standard, that is, a standard that responds to changing business and technical needs. Web Services can also be added to existing information assets, so-called "legacy systems" to enable their participation in modern data interoperability. Numerous case studies have demonstrated the transformative nature of the Web Services approach to individual companies and even industries (Manes 2003).

Conservation information, characterized by its diversity of geography and providers and by the need to scale across time and space, is an excellent fit for a Web Services architecture. By implementing the Open Conservation Exchange on Web Services protocols (specifically REST), conservation information asset managers will enjoy all of the associated interoperability, productivity, and durability benefits.

Leverage Points

Standards for conservation datasets have already been developed with varying degrees of adoption. These include the SEEK protocols developed at NCEAS for ecological datasets, DiGIR protocols for the aggregation of biodiversity collections and field observations and marine observations standards authored under OBIS. A national effort led by GreenInfo Network in cooperation with USGS and the Conservation Biology Institute is developing a comprehensive protected areas inventory (Barker 2008). The Conservation Measures Partnership, the Nature Conservancy and Defenders of Wildlife are all investing in interoperability of spatially-explicit conservation project information (Barker 2008). NatureServe has developed Web Services access to its species and ecological systems profile information as well as specific location data (NatureServe 2008). The development of a set of Open Conservation Exchange standards should strongly leverage existing protocols and data standards, adopting highly successful standards and learning the lessons of those less successful.

There are a number of candidate datasets that will provide substantial value to conservation if standardized in an Open Conservation Exchange. The Heinz Foundation's report on the State of the Nation's Ecosystems, identified critical data gaps that prohibited systematic assessment of ecosystems in the United States including landscape pattern; distribution and condition of key habitat elements, at-risk species and communities, non-native species; stream and riparian condition; nitrogen yield and load, carbon storage and ground water levels (Clark 2006). The Nature Conservancy has identified four datasets as fundamental to their information strategy: protected areas, assessments (including conservation targets such as habitats and ecosystems), conservation projects and threats to biodiversity (TNC 2006). Studies in the automated valuation of ecosystem services require a different list of datasets including land use/land cover, soil types, hydrology, biomass, timber production, age, use and market value, agricultural production, costs and market value, species habitat and more (Nelson et al. forthcoming).

As stated above, data exchange standards and their underlying implementation through Web Services should be developed using an iterative approach We must begin with those that are the most practical to identify or develop, most likely to be adopted and provide the highest overall value to conservation. Candidate starting points include already standardized taxonomies in conservation including the IUCN Threats Classification Scheme (IUCN-SSC), the Conservation Measures Partnership conservation actions standard (CMP-Actions), and vegetation and ecosystems classifications (Grossman 1998; Comer 2003). This strategy proposes the development of an analytical framework that ranks the suitability of a conservation dataset for standardization. The framework would evaluate benefits (e.g., necessity to

priority conservation analyses, existence of a successful data standard that can be used or leverage, size of the producer sector) and contrasts them to costs (e.g., the heterogeneity of contributing datasets, risk exposure). A sample framework is given in Appendix B.

Standards development requires sustained commitment overtime. This strategy therefore recommends that conservation organizations allocate resources for the ongoing development and maintenance of exchange standards. However, I caution against large-scale, upfront standards efforts. Standards are best when strategic, light-weight and fully supported by tools (see next strategy).

Summary of Information Benefits

Information standards, if they are widely used, enable interoperability between systems. A conservation data exchange standard would increase the wealth of datasets available for analysis and save conservation researchers and practitioners time in data preparation. The Web Services approach enables sustainability of the overall architecture by a) facilitating control of datasets by their rightful owners and b) mediating implementation variance through explicit data exchange contracts. This in turn improves the potential for data currency, reliability and accuracy. Conservation datasets can evolve, growing in representation and variety, without disrupting existing clients. If the same Web Service interface is implemented to access data in different geographies, third-party clients can aggregate data at a higher scale. Similarly, two organizations collecting similar data can provide a common Web Service interface, again allowing third-party clients to aggregate based on the standard. Finally successful standards will attract new producers, consumers and tool developers improving the value to all participants. The realization of the benefits stated here is dependent on adoption of standards by a critical mass of participants. This in turn depends on the strategies described later in this document.

Sample Conservation Benefits

Assuming the conservation industry is able to achieve critical mass participation in the production of standards-compliant data, many conservation activities may benefit. For instance, broad-scale adoption of a protected areas data standard, a goal that has long frustrated conservation, would enable authoritative gap analysis, inform the reliability of protection in light of development pressures and help us understand the degree to which existing protected areas meet the needs of wide-ranging species and/or ecosystems responding to climate change.

Challenges

Because of inherent heterogeneity in many categories of conservation data, there may be numerous conservation datasets that should not be standardized, that is, they do not pass the cost/benefit analysis. Where standards development is justified, leaders of standards efforts are often challenged to enlist the required expertise. Standards development requires a special mix of expertise and communication skill and yet is not a highly-regarded activity in conservation (Barker 2008). More fundamentally, standards are challenged to respond to legitimate exceptions in their application. Eager adopters can lose their enthusiasm when faced with non-trivial information loss. Emerging semantic web technologies may offer a realistic means to support both heterogeneous datasets and, at the same time, interoperability (see strategy number four). But for the time being, standards development and conformance for highly heterogeneous datasets may require fundamental compromise. Finally, in the face of competing priorities, conservation will need to explicitly fund standards development, promotion and maintenance.

Where to Start?

- Formalize an analytical framework that evaluates the costs and benefits of selecting or developing standards for each conservation information domain. For datasets characterized by heterogeneity, hierarchical standards may be practical.
- Start small and enable success: identify information domains within conservation that are strong candidates for standardization. Support these efforts with end-user tools and transformation utilities (see the next two strategies)
- Engage key data producer organizations on the development of Web Services to produce datasets that adhere to the Web Services protocols specified by the standards. Where cost-effective, "wrap" existing systems with web service protocols.
- Engage key data consumer organizations on the development of Web Services clients that can demonstrate the value of the standards
- Build awareness of Open Conservation Exchange standards through case studies, live demonstrations, papers and conference presentations
- Encourage conservation organizations to adopt the Web Services architecture for all conservation information systems development. Based on the content of an information system, coordinate relevant data exchange standards through the Conservation Commons.

3. Tools for the Producers

Aggregators and analysts of conservation information generally agree on the imperative for data standards. They are united by their common interest in synthesis, analysis and decision support to address critical questions in natural resources management and conservation. This group recognizes both the unmet need and the missed opportunity in each conservation dataset (e.g. observations, conservation projects) that remains in simple spreadsheet form, digitized but nonetheless disconnected and undescribed. They are thus motivated to convene and deliberate, producing standards that reflect their interests in the data and then cajole and/or coerce another group, the conservation data producers, to go out of their way to conform, convert, reformat, translate, crosswalk, describe and their data then upload it to shared data servers. Yet the benefits of this additional effort are abstract, realized in another place and time and by someone else. Not surprisingly, conservation data that is undescribed, disconnected and highly variable in format and semantics continues to accumulate, the unfortunate outcome of mismatched incentives. The result is a wealth of information whose potential to inform and direct the understanding, effective management and conservation of natural systems is never realized.

This strategy proposes that we meet the needs of information producers with data entry, management and analysis tools that 1) outperform all alternatives in solving the user's problem in terms of ease of use, ease of learning, productivity, scalability and applicability to the problem at hand, 2) embed standards conformance in the data they produce and 3) leverage other information and computation services in a service-oriented architecture. The strategy is simple: "no rules without well-behaved tools." Versus the overt standards-promotion approach, this strategy insists on powerful and usable tools for data produces that happen to produce (and potentially consume) standards-conformant data. With sufficient adoption, these tools thus define de facto data standards for their data inputs and outputs. Only by providing direct and immediate benefit to data producers will we achieve the broad standardization required by data consumers.

By taking a Web Services approach, tool developers will a) enjoy increased efficiency by not having to reinvent shared services and b) contribute their unique components to the pool of services available to other developers. Complexity in analysis can be handled through services hierarchies, in a manner that parallels the tiered approach to information standards described above. For instance, in a tool under development by the Natural Capital Project to support systematic valuation of ecosystem services, the

tiered approach will allow users to use the built-in carbon sequestration models or supply their own as long as it conforms to the carbon sequestration data standard (Nelson et al. forthcoming).

The predicate of each centrally developed end-user application is the existence of specific methodology on which that application can be based. In addition, the detailed methodology must be applicable in a wide variety of situations, enough to warrant investment in a software system.

For example, the Conservation Measures Partnership (CMP), a consortium of conservation organizations, has developed a methodology to guide conservation projects that defines "targets," "threats," "actions" and "steps" as a means to abstract the highly custom approaches used by conservation practitioners. Importantly, CMP has followed up with the development of a desktop software tool, "Miradi," to support users in employing the methodology. If Miradi brings productivity to conservation practitioners within and beyond the CMP member organizations, conservation may have a powerful means, perhaps the only practical means, to produce standardized conservation project data. This in turn will enable systems that aggregate and analyze project-level data. The Nature Conservancy is actively investing in ConPro, a web-based tool for managing conservation projects that employs the CMP standards and will aggregate data from individual Miradi instances. Furthermore, an XML standard for conservation projects is under development that will at first enable data exchange between ConPro and Miradi but also enable other organizations (e.g. Defenders of Wildlife) to develop their own conservation project tools, aggregation and analysis systems. This is an example of the full realization of the "no rules without tools" strategy. Given the emergence of practical and powerful tool that truly meet end user needs, that tool's data format becomes a de facto data standard. Providing the data standard is open and shared, other tools, producers and consumers, can emerge around the original tool thus enabling a new level of efficiency.

Existing and potential end-user applications to support conservation include:

- Field observations management
- Collections management
- Species and ecosystem distribution modeling, including climate change scenarios
- Species population viability
- Vegetation classification automation from remote sense data
- Watershed, wetland and benthic modeling
- Watershed pollution, erosion and runoff for water quality and quantity analysis
- Food-web analysis
- Natural resources supply/demand analyses (e.g. timber, fisheries)

- Urban growth modeling and development impacts analysis
- Regional or even landscape scale climate change conditions and response analysis
- Invasive species observation, distribution, trend analysis and control
- Stewardship activity management
- Easement and preserve compliance monitoring
- Conservation project management
- Protected areas data management
- Conservation planning (i.e., ecological assessments)
- Policy and compliance (e.g. NEPA) analysis
- Ecosystem services valuation and mapping including modeling of regulation, provision, cultural and supporting services
- Integrated landscape planning

The inter-relatedness of these domains points to the significant benefits that may be derived by a standards-based, leveraged approach. These may be especially powerful when supported by analytical units in a workflow context (see the next strategy). For instance, an integrated landscape planning tool might combine ecosystem services valuations with socioeconomic development analyses. The ecosystem services valuation might combine natural resources supply and demand analyses with water quality and quantity, carbon sequestration, biodiversity for ecosystem resilience and other services. Each of these might be further broken down. From this perspective, the effect on conservation of an efficient structured approach to information and information services may be transformative.

It is important to acknowledge that many valuable end-user applications for conservation have already been developed. The Ecosystem-based Management Tools Network provides a database of tools, many originating in academia, that address a wide-range of issues in ecosystem-based management (see www.ebmtools.org). Yet the tool development has thus far been characterized by reinvention and isolation. While recognizing the value of site and issue specific problem solving, the EBM Network managers acknowledge the limited utility of many of these tools either because they are simply too difficult to use and/or too tailored to a specific thematic or geographic problem (Barker 2008). Many of the tools require users to prepare custom datasets for input, learn new methodologies and/or learn new user interface conventions. Tools rarely leverage functionality from other tools. The format and semantics of their outputs are frequently non-standard. Users pay a high price to derive the promised value of these tools.

Those tools that facilitate high productivity of data producers and are based on a widely shared methodology will be most successful. In the beginning phases of this strategy, developers may lack the expertise or incentives to make their tools widely applicable, interoperable and usable. Here the value of standards efforts supported by tools and an informed community can be realized. This strategy recommends that the Conservation Commons develop a certification and funding program based on a) adherence to usability and interoperability standards[1] and b) provision of data and computational accessible via Web Services. The program would seek donor investments and then fund new tools or certify existing tools according to a cost/benefit analysis similar to the one shown above for data standards. For instance, the framework would analyze 1) the strategic value of output datasets or analyses, 2) size of the data producer sector, 3) potential gains in data producer productivity and/or compliance 4) likelihood of successful generalized automation, more specifically, the developer's ability to manage inherent heterogeneity of the data and methodology and 5) conformance and contribution to the services architecture.

Summary of Information Benefits

The focus of tool development on conservation researchers and practitioners, as the providers of primary productivity within this information ecology, will first and foremost directly improve their efficiency. By encoding standards conformance into these tools, in a Trojan-horse fashion, we remove idiosyncratic (i.e., syntactic) divergence of datasets. In domains where data producers are successful, the secondary beneficiaries are the data consumers whose ability to aggregate and analyze standardized information is dramatically improved. This strategy, at a basic level, just makes the second strategy, "Operational Conservation Data," a practical possibility. Additionally, authors of tools in a Web Services architecture are able develop more efficiently by leveraging standard components.

Sample Conservation Benefits

In domains unburdened by semantic heterogeneity or complex workflows, the combination of data hosting services, data standards, end-user tools to enable success of the standard are enough to enable direct benefits to conservation. For instance, if the conservation project management methodology forwarded by the Conservation Measures Partnership is inherently sound, widely applicable and has the potential to substantially improve the productivity of conservation practitioners, an end-user application may not only manifest productivity improvement but also enable new capacity for accountability, measurement of strategy effectiveness and understanding patterns of conservation focus. As another

[1] A useful starting point for this standard may be the "Best Practices for Developing Interoperable EBM Software Tools" defined by the Ecosystem-based Management Tools Network (see http://www.ebmtools.org/node/150).

example, for those organizations with direct responsibility for preserves and easements, land stewardship activities may be supported by tools and integrated with biotic information using Web Services, thus improving the compliance, productivity and capabilities of land managers.

Challenges

Because of the inherent variation of the underlying information or required analyses, there may be many conservation activities that cannot be effectively supported by tools. Similarly, in dynamic domains such as detailed ecological assessments or ecosystem process analyses, tools development will lag behind science and usability needs leaving researches to their own – non-standard - devices. Finally, even where tool development or enhancement is practical, funding may be limited.

Where to Start?

- Invest in tools for high-value conservation analysis that are based on already standardized datasets and methodologies.
- For each dataset warranting standardization, provide end-user support in the form of tools for the production of data that complies with the standard. Tools must be easy to learn, use, consume standardized datasets wherever possible and clearly improve productivity of their users. In addition, they should support access to their datasets and computation through web-service protocols.
- Develop a cost/benefit analysis framework to prioritize funding and support for the most strategic and cost-effective tools. Investigate existing tools, successful and otherwise, to inform the analysis.
- Develop a certification and funding program that encourages the development and success of usable interoperable tools

4. Computation Web Services

Data management nodes will provide secure, reliable hosting of conservation datasets. Data web services will enable maintenance by their owners and data exchange standards will enable secure access by outside parties. Finally, end-user applications enable data editing, management, reporting, mapping and some forms of analysis. With the addition of computation services, the basic Web Services architecture becomes truly powerful. Computation services are utility components that transform, crosswalk, connect, analyze and synthesize conservation data.

Computation frameworks have been developed that allow users to assemble complex workflows or staged analyses based on raw inputs, transformations, syntheses and analyses. NCEAS' Kepler is one such environment (Ludäscher et al. 2006) as is ESRI's geoprocessing framework (ESRI 2008). These frameworks can complement the tiered approach to standards so that workflows may automate complex analyses with many intermediate yet still standardized results. In such a workflow, users can substitute the best suited analyses for individual components based on availability of inputs and expertise.

Summary of Information Benefits

Computational services complete the support for standardized datasets in conservation by enabling

- **Access to and conformance of non-standard datasets.**
  For instance, to increase coverage of the protected areas dataset, some existing data repositories, perhaps in the form of stand-alone files, may be uploaded to a shared server where attributes can be cross-walked to the standard and augmented with required metadata. Alternatively, protected areas data may be maintained in an existing repository and transformed on-the-fly in response to the standardized Web Services queries. As another example, proscribed burn data may be maintained by various custom applications or simple spreadsheets. These data can be combined by computation web services into a common standard enabling aggregation and analysis of proscribed burning within and across organizations. Finally, both GBIF and NatureServe have developed taxonomic reconciliation services, thus allowing the aggregation of biodiversity data captured using diverse taxonomies. Computation services therefore significantly enhance the size of the pool of datasets accessible through standard interfaces.

- **Aggregation, synthesis and analysis**
  Computation services can provide users with simple aggregations of similar datasets, synthesis of aggregated datasets into tabular reports, maps or other visualizations, and/or intensively analyzed

to produce fundamentally new information. Aggregation services access any number of data sources through Web Services protocols and combine that information. Some computation services may stop there. For instance, the Global Biodiversity Information Facility (GBIF) aggregates collections and field observation data using Web Services protocols and makes these aggregations available in turn to other services. At the next stage, aggregated information may be mapped or otherwise synthesized. Indeed, GBIF's observation data can be viewed on a web-based map or transformed into KML for presentation in Google Earth. Finally, the information may be intensively analyzed by a service. For instance, climate change models may combine regional biodiversity and climate data provided by users to produce climate change scenarios at regional scales.

Computation services that combine aggregation from common sources with intensive analysis will be significantly improved with the benefits of standardized inputs and the service architecture proposed here. For example, MARXAN is a widely-used tool in conservation for optimized reserve selection based on biodiversity and cost data. The current implementation requires users to assemble and transform data inputs into text files. Web computation services can automate this process based on a variety of common biodiversity and cost data. An updated version of MARXAN might support interoperable connections to Web Services directly, both in terms of its inputs and outputs. As another example, once an ecosystem services effort identifies or develops a water quality data standard based on elevation, soil, land cover, vegetation, temperature and precipitation, software developers can supply companion computation web services that connect to established data service providers. Based on these inputs, the service would then produce water quality data conformant to the standard, perhaps tuned by end-user parameters.

- **Connection and standardization of non-conservation datasets.**
  Computation web services may be developed that harvest these datasets from various sources and transform them to a common standard. For instance, to inform ecosystem services valuation, spatially explicit recreation data may be required. These datasets, however, are non-standard and managed outside the conservation domain. Where the value of the dataset warrants investment, computation services can use whatever means are available, perhaps employing multiple techniques for very similar datasets, to access and transform the information for interoperability with conservation Web Services.

- **Publishing and export.**
  Outputs of analysis and synthesis can be optimized for high availability and performance or exported into formats conformant with competing or external formats. For instance, the detailed delineation of spatially-explicit of conservation priorities, perhaps informed by ecosystem services valuation or high-probability climate change scenarios, can be shared with conservation partners, government agencies and the public in the media that will maximize their impact for each audience.

The value of the service-oriented architecture is a function of the number of participating services. That is, as the number of shared services and interoperating applications increases, each new solution is that much more leveraged. The diagram below (figure 1) depicts a high level view of interoperating services and applications at a data center sponsored by the Conservation Commons. Services shown may be distributed world-wide and among several cooperating organizations.

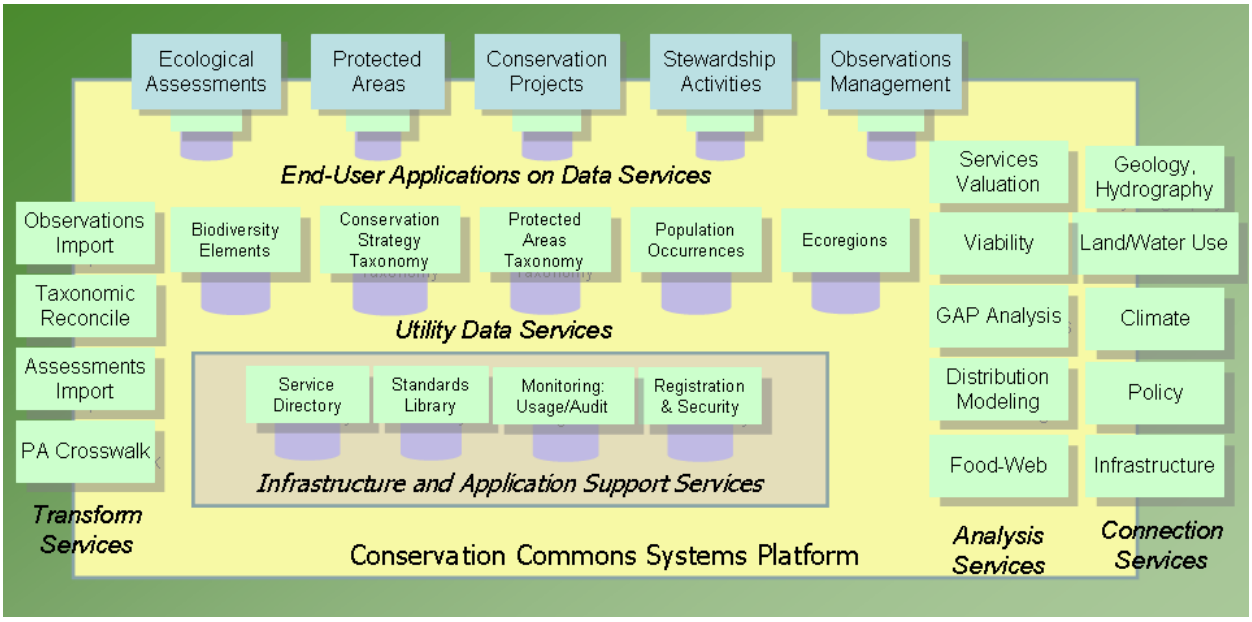

Figure 1: The Conservation Commons Systems Platform.
This figure shows a hypothetical mix of services on a virtual systems platform hosted by the conservation commons. Any given service may provide its functions by leveraging other services. Services need not be collocated. End-user applications provide data entry, management and analysis to their underlying datasets. Utility data services provide reference datasets for standardization. Transform services provide real-time transformation of datasets hosted outside the platform to meet the Commons' standards. Similarly connection services connect and aggregate externally hosted (and non-conservation) datasets for use by the Commons' solutions. Analysis services can be combined with data, connection and transform services to provide rich analytic workflows. Finally, infrastructure and application support services provide security and administration functions.

Sample Conservation Benefits

Computation services give us the ability to transform and thus incorporate non-standard datasets. Moreover, they can provide cost-effective support for complex analyses and workflows. The ramification for conservation may be dramatic. For instance, the ability to aggregate biodiversity data across taxonomic systems improves the scope and coverage enables more accurate indicators for biodiversity health such as the "Living Planet Index" (Loh et al. 2005). Conservation assessments have increased transparency, scope and coverage and can therefore more easily inform policy at multiple scales. Shared view of conservation priorities makes for less redundancy and better issue coverage among conservation organizations. Complex relationships and processes in ecosystems can be modeled dynamically based on online data sources thus enabling change prediction that is closer to real-time. Computation services offer a realistic means by which spatially-explicit ecosystems services valuations can be created and maintained based on a wide variety of biotic, geophysical and socioeconomic data, thus enabling tradeoffs-based decision making. The practicality, completeness and durability of climate change analyses may be improved because analysis can be hosted on shared, high-end servers and separated from data transformation and management. Conservation is therefore in a better position to forward climate change adaptation scenarios.

Challenges:

Again, the solutions proposed here will be limited to those datasets and analysis unburdened by semantic heterogeneity. For instance, detailed ecological assessments are based on raw observation data from the field. These inherently diverse datasets are difficult to standardize. In import and dynamic transformation services, those datasets that diverge semantically from the target standard will lose information in the conversion processes. Similar to the other strategies, funding for computation services may be limited. However, because computation services leverage so many other information assets, their investment should be easier to justify.

Where to Start?

- Develop computation services that support the priority datasets identified in step one and enable interoperability to valuable but inaccessible datasets. For instance, data adaptors that transform non-standard datasets may be essential to the success of the standards.
- Workflow frameworks will enable end-users to dynamically assemble solutions to meet their specific needs. Investigate existing frameworks such as NCEAS Kepler and ESRI's geoprocessing framework.

In some ways, computation services that aggregate, synthesize, analyze and publish conservation information are the pay-off for investment in the first three strategies. Without these strategies, the development of such capabilities is nearly cost-prohibitive. However, with availability of a shared computation platform, the standardization of input data sources supported by end-user data entry and management tools and a library of other computation utilities, the trend will reverse. Conservation will be a position to systematically assess conditions, inform priorities and steward its holdings.

5. *Kestrel: Supporting Variation with Standards*

To address the unavoidable complexity, variation and evolution in conservation practice, we must provide a solutions platform in conjunction with applications for end-users. The imperative to respond to diverse and changing conditions requires that users be able to *assemble their own solutions* from a library of components and within an overall framework. Our goal is to design these components and framework to provide both end user productivity, solving the problem at hand, and, *transparent to the end-user*, produce standards-compliant data that can be aggregated and analyzed elsewhere. Dynamic assembly from standardized components can operate on two levels. First, as described above, users may combine datasets with transformation and workflow tools to perform a specific analysis. At a deeper level, to respond to semantically novel situations, users may define schema itself from component parts.

The field of computer science is actively researching the area of semantic mediation wherein data from diverse, heterogeneous sources can be reconciled, effectively combined and queried (Green et al. 2007; Halevy et al. 2005; Ives et al. 2005). Not surprisingly, scientific datasets and especially bioinformatics, are the proving ground for these ideas (Ives et al. 2005). Researchers have developed semantic web languages such as Resource Description Framework (RDF) and Ontological Web Language (OWL) to facilitate a full ontological description of data and enable software systems to discover relationships between arbitrary entities and reason based on inferences (Wang et al. 2004). Accordingly, the burden of semantic description is high. By limiting the scope of the problem space (entities in a specific domain such as conservation) this research suggests a less ambitious approach: the *definition of entities is limited to that which will enable information systems to provide useful data entry, management, and query*. NatureServe has developed an initial implementation of this system, codenamed "Kestrel," to manage observation data (NatureServe). This project has proved the basic viability of the core concept: a system that operates on independently-defined schema can be usable and provide aggregation across diverse datasets. Projects in the open source community may be investigating similar approaches (Alon Halevy, personal communication.

While the strategies presented above are based on widely-used technologies in the private sector and government, this strategy relies on emerging solutions. While it is valuable to develop the concepts presented here for conservation information systems, neither conservation organizations nor their donors will or should fund the realizations of these concepts. This research presents an approach to supporting semantic variation with community-driven standards so that we may see its benefits and work with

dedicated information systems developers, including and perhaps especially those in the open-source software community towards an implementation that will meet the needs of conservation.

A Closer Look at Schema Variation

As noted above, representations of complex problem spaces such as biodiversity conversation are dominated by heterogeneity. That is, while the core concepts in conservation such as field observations, species and ecological community occurrences, stewardship activities, managed areas and biodiversity threats can be modeled in an information system, the relevant attributes required to adequately characterize them significantly vary, a function of scientific understanding and the specific purpose of a conservation activity. In much if not most of the problems of conservation, a generalized data model cannot be defined to adequate capture all of the details required to address the problem at hand.

Conservation Example: Species Observations in the Field

In characterizing a field observation of a bird, relevant attributes might include the observer identity, location, time and date, the species, gender, life-stage, nest characteristics etc. In characterizing a plant observation, however, attributes in common include observer identity, location, time and date, and the species but also include phonological stage, coverage extent, as well as soil type, acidity, and moisture. While these are both "species observations," the information required to describe them differs as a function both of taxonomy and the purpose of the survey.

Not surprisingly, most field researchers use generic tools such as spreadsheets to manage highly variable conservation information. These solutions are easy to use and extremely flexible and thus meet the basic needs of data producers to capture their information. However, they offer very little in the way of data validation or analyses specific to conservation. Furthermore, the resulting datasets are completely non-standard and thus of little or no value to data consumers.

The alternative is a custom data management system suited to a given conservation problem. Developers of such systems have three possible approaches. First, a core data model or schema may be standardized and provided support in the end-user application. For instance, in developing a species observation data management system, the core schema of an observation may consist of a species identifier, an observer, with the date and location of the observation. Yet such a system will not meet the needs of any sophisticated exercise in species observation for the data model is far too limited. Second, a generalized schema may be developed that attempts to abstract variation. However, meaningful variation is not always subject to successful abstraction, as in the case of the observation data described above. Finally, developers might take the "kitchen sink" approach. In this approach, we allow the schema to expand in order to meet the needs of a wide, and potentially infinite, variety of data producer problems. While such a system suffers less the problems of compromise in schema, it may be very difficult for data producers to

use and for developers to maintain. No matter the approach, the return on investment in custom data management systems that operate in a domain characterized by schema variation is inherently limited.

The Solution: Directly Support Semantic Variation

The solution to the problem of supporting semantic variation in conservation information systems arises from (a) the recognition that the structures of data entities in conservation, such as observations, conservation projects, stewardship activities, share a common core and components, (b) the similar recognition of a high degree of overlap in conservation systems functionality (e.g., supporting data entry, data validation, data search, data browse, reporting, mapping, import, and export) and (c) the opportunity provided by the maturity of certain information technologies such as XML and the processing speed of widely available computers which together enable systems to take a much more dynamic approach to basic application functionality.

We address the dichotomy of leveraged support for variable schema by separating schema from systems functionality. In traditional software applications, the schema is enmeshed with systems functionality. That is, application source code precisely and explicitly references the structure of data. This research proposes a system wherein application functionality, including data entry, management, and analysis, is *independent* of schema. To provide functionality to end-users, the system acts on schema that is externally defined. This separation of schema from system enables the fundamental innovation in this approach: *standardization of data is accomplished by sharing schema definitions among users*.

There are seven components to our shared-schema system, referred to for simplicity's sake as "Kestrel." These components are: (1) a mechanism for describing conservation schema, that is, the entities and attributes common to conservation, (2) information systems that operate on these schema descriptions and provide functionality to end users for data entry, management, reporting and more, (3) the heart of the system, an authoritative library of schema components owned and managed by the conservation user community (4) a solution designer to instantiate and customize existing schema and if necessary create new schema (5) aggregation and analysis tools that can harvest datasets based on shared schema, and (6) data adaptors as a bridge to existing datasets and conservation information systems and, finally, (7) community support tools.

Conservation Example: Species Observations in the Field

The central power of Kestrel lies in the outcomes it facilitates. Users of the system meet their own information management needs and yet contribute to the wealth of standardized data for use by others. Returning to our species observation example, this system enables the aggregation and subsequent analysis of diverse observation datasets. For instance, a researcher may create a species inventory of a watershed in as a first step to understanding its ecological processes. In doing so, she may document the location of several plant species. Sometime later, a second researcher documents the extent of a specific rare plant in a neighboring watershed. Both surveyors document the date, location and species, yet each captures a variety of additional attributes. If both researchers use Kestrel to gather their observations, the second researcher can easily harvest observations from the first researcher's survey that reference his target rare plant, thus further informing him of the extent of his target species. Later, a third researcher may combine both surveys' documentation of the rare plant species with her own in another geography and intersect all of these observations with soil type, acidity, moisture, aspect and temperature to understand the habitat characteristics of the rare plant and/or model its full distribution.

1. Describing Conservation Schema

The commonality of schema components allows us to make upfront investments in their definition that can be leveraged across a wide-variety of uses. For instance, if a species observation commonly consists of the attributes "who," "what," "where," and "when," we can provide the individual definitions of each along with the entity "species observation core" consisting of all four attributes. The definition of entities (i.e. data records) and attributes (i.e., data fields) as components of schema (i.e., data structure) must contain sufficient information to enable systems logic to provide a rich user interface for data entry, management, reporting, etc. Accordingly, attribute definitions include the basic data type (e.g., string, integer, date, and location), description, a reference to the component's author, a short label, and basic help text. The definition language, based on XML, allows attributes to reference other entities, grouping of attributes, and default or fixed attribute values, attribute measurement types (e.g., area, length, volume, mass, temperature, etc.) and default units, validation rules, end-user annotations, confidence ratings and more.

Entity definitions, as descriptions of things that have an independent, though not necessarily material existence, consist of a set of attributes. Examples of common entities in conservation include a species, ecological system, field observation, species population occurrence, protected area, restoration activity, burn event, etc. Further, entity definitions may constrain or otherwise customize the attributes they contain including specifying default and fixed values, augmenting validation rules, guiding usage, etc.

To facilitate reuse, the definition language supports inheritance so that new schema components can be expressed as an extension of existing components. The schema language gives full support to the global conservation community: all display text associated with entities and attributes can be localized into any language. Finally, both entities and attributes are uniquely identified via a unique namespace (e.g., "entities.standards.iucn.org") and name (e.g., "speciesObservation"). Appendix C: provides the definition of a sample observation entity and associated attributes.

2. Schema-Independent Information Systems

Overlap in basic systems functionality allows us to make leveraged investments in the development of systems which are independent of detailed schema. In these systems, schema is treated as data that prescribes what have traditionally been static system behaviors. In the system diagram below, a data entry form for observations is presented and populated with a specific observation record in four steps. The data management system begins by (1) reading the schema from the schema repository and (2) passing this description to the user interface component. The user interface code parses the schema (3) and generates the data entry form. For instance, the definition of the "Number Observed" attribute is translated in to a label and textbox on the form. With the form in place, the data management system (4) reads the data from the data repository and (5) passes this to the user interface component (data shown in italics) which in turn (6) populates the data entry form with the data.

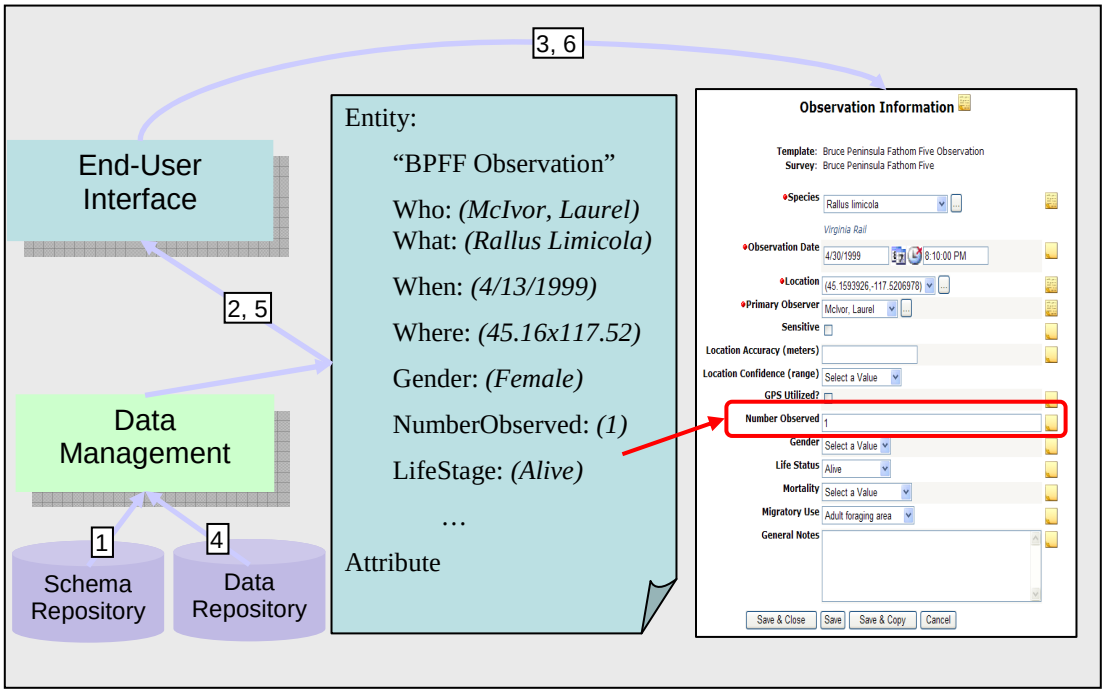

Figure 2: Dynamically Generating the User Interface from Schema.
The data management system generates the data entry user interface by (1) reading schema, (2) passing it to the user interface component where it is (3) translated into data entry forms. The system then (4) reads the entity's data record from the data repository and (5) passes this to the user interface component which (6) populates the form with the data attributes. The system provides basic data entry functionality such as data validation, intuitive entry based on data type, error messages, help text and annotation all based on the schema read from the schema repository.

Besides data entry, the system can also provide data management capabilities such as add, modify, and delete with entity records as the unit of operation. Entity records can be queried and reported using common attribute and entity definitions, a foreshadowing of the independent aggregation and analysis tools described below.

An attribute's simple data type allows the system to provide a significant amount of functionality for end-users. The system shown in figure 2 provides intuitive data entry based on data type (e.g. drop down lists for list-value selection, calendar selection for dates, and location specification through maps). The system can support attributes that are references to other entities with data entry tools such as rich search and browse capabilities. For instance, the system may support data entry for a species-reference attribute by leveraging search and browse capabilities offered by the referenced entity's data source, a species database, perhaps in the form of Web Services. Reporting functionality using date and location can include visual representations of time and space.

Data entry and management systems can be created for a variety of platforms including web-applications, personal computers, and even field devices. As long as these systems can consume the standardized

definition of schema components and support the standardized protocols for query and aggregation of resulting data, "competitive" offerings differentiated on usability, computer and networking resource demands, and even price serve to enhance, not detract, from the overall success of the system.

3. Schema of, by and for the Community: the Community Schema Library

Separation of schema from systems functionality is not without precedence. Intuit's QuickBase, GoogleBase and Microsoft's SharePoint are all examples of systems that allow users without database expertise to define schema and offer basic data entry, management, reporting and data exchange with other formats. However, in each of those systems, the schema and schema components are independently authored based on simple data types (e.g., string, integer, date). This research proposes that semantically rich schema components are defined by the user community using easy-to-use tools and managed by the community in a common library. It is the community-owned and managed conservation schema library that distinguishes this strategy for it enables productivity for data producers and, most importantly, *the ability to aggregate resulting datasets based on shared schema*.

The Community Schema Library is the central repository for schema components. In contrast to the development of existing systems for conservation, the developers of schema, as the contributors to this library, are conservation researchers and practitioners: subject matter experts, not software engineers. In the figure 3, two separate observation data entry applications reference a common set of attributes. The datasets that result from these systems can be aggregated with respect to the intersection of their schema.

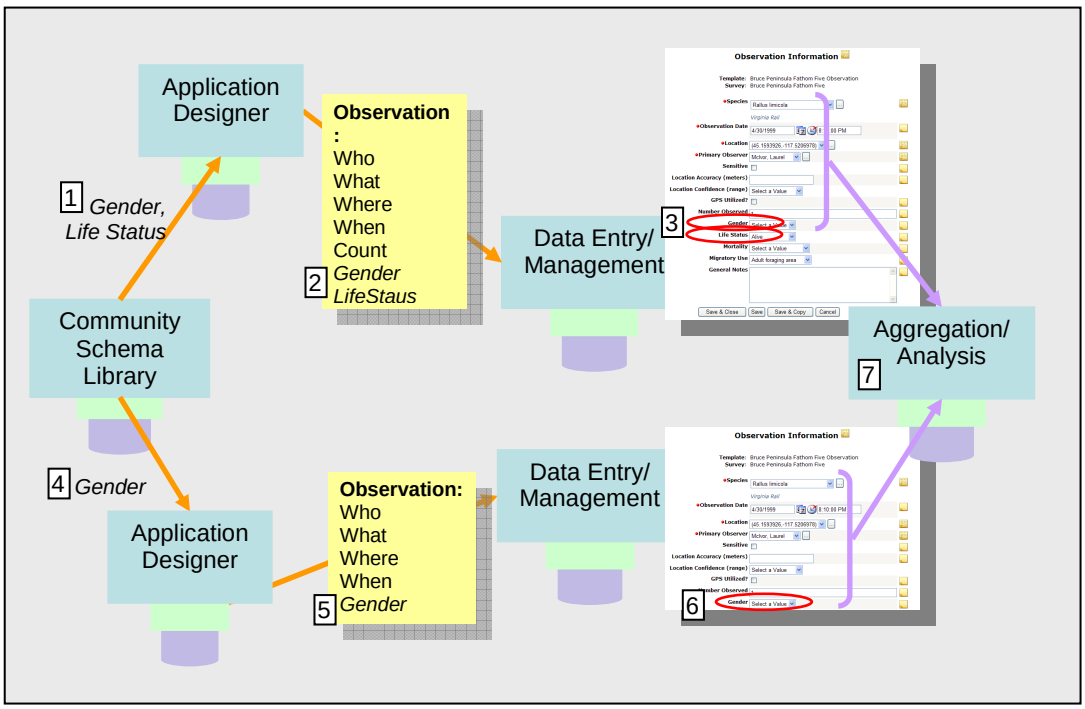

Figure 3: How Shared Schema Enables Standardized Data.
Observation data from two independent data entry and management systems can be aggregated with respect to the intersection of their common attributes. A users may (1) identify attributes "Gender" and "Life Status" in the Community Schema Library and (2) add them to an existing observation entity definition where they will (3) be presented by the data entry and management system in the user interface. Should a different user begin with the same entity definition, (4) identify the same "Gender" attribute in the Library and (5) add it to the entity definition for (6) presentation in the user interface, then observation data from all of the common attributes including "Gender" can be (7) aggregated and analyzed.

The Community Schema Library, as the central repository for schema components, must facilitate community stewardship and evolution of its content. To that end, there are several requirements.

a. Users must be able to find the entity and attribute definitions they are looking for. Because aggregation of datasets relies on reuse of schema components, rich search and browse capabilities are critical to the success of the system.

b. Schema must be allowed and encouraged to evolve. Users must be able to provide feedback and suggest enhancements to schema components. Component authors, as self-declared subject matter experts, must be able to provide updates to schema components based on their own experience and feedback from users. Users must know how to judge the fitness and quality of schema components. Here we can take a queue from community-driven resources such as sites for downloadable "shareware." Component descriptions must include objective usage counts and subjective user-ratings and comments. Author profiles should include credentials such as institutional affiliations and a list of schema component contributions, and perhaps publications

and other information systems contributions. In this way, users may gage the appropriateness and reliability of schema components and authors may esteem themselves with successful schema contributions. Emulating natural evolution, usage counts, ratings and comments will enable the "fittest" schema components to rise to the top.

Uncontrolled changes to schema components, as entity and attribute definitions, will dramatically compromise the reliability of datasets which reference these definitions. Consequently, in contrast to completely open community resources such as Wikipedia, this research proposes that the evolution of schema components is controlled by component authors, an approach similar to that taken by the open-source software community.

c. The library should be "seeded" with the schema definitions reflecting existing standards from conservation, bioinformatics and the IT industry. For example, GBIF, Conservation International, The Nature Conservancy, NatureServe, IUCN and the Conservation Measures Partnership all have detailed standards that can be defined in the library. Early adopters of the system will thus enjoy a strong foundation of schema components for meeting their own needs. If the system succeeds, translating these standards to entity and attribute schema components may constitute a dramatic act in support of their practical application.

The Community Standards Library has dramatic implications for the standards definition efforts described in the second strategy of this document. First and foremost, the standards definitions, as schema components, are *directly useful* to their constituents. They are immediately usable, testable and can be offered as viable alternatives to existing standards. Second, because users can submit their own fine-grained schema components, a given "standard" need not be perfect in order to be viable. The user community can address minor defects. As problems arise or enhancements are requested, schema component authors can adapt their solutions to meet the need, providing crosswalks from earlier standards as necessary. Finally, the suitability of individual schema components can be measured in terms of usage counts and ratings. This is valuable information to standards developers who previously have limited means of obtaining feedback on their work.

4. Application and Schema Designers

Users may draw schema components from the library for inclusion in an application. Users must be able to start with an existing entity or family of entities and add customizations, including new schema components. In an easy-to-use user interface, users may query the Community Schema Library for the

schema components that generally meet their needs and, in a "drag and drop" fashion, instantiate and customize them as necessary. With a schema defined and customized to the need at hand, users may "deploy" the application for data collection in a companion data entry and management system.

Conservation Example: Conservation International's Rapid Assessment Protocol

A researcher for Conservation International (CI) may be charged with leading a team to assess the biodiversity value of an area under immediate threat. Assuming CI's Rapid Assessment Protocol has already been translated into corresponding schema components in the library, the researcher may find and "drop" the observation and associated schema components into a workspace. From there, she may customize these to her particular assessment needs by providing names, constraining the location and date of observations, provide default or fixed values for the observation purpose, etc. She may also annotate the existing protocol, as defined in the entity schema, with additional guidance. Finally, she may preview the data entry forms, reports, and maps. At this point, she may transfer the customized schema to any number of data entry and management systems, including those hosted on handheld devices. Resulting data captured in the field can be fully aggregated with all survey data captured using the same schema components.

Either in the same application design tool or in a separate schema design tool, users must be able to author new attributes and, less frequently, entities. Attribute authorship consists of filling out a form with basic information such as a name, purpose, the basic data type, measurement type, default units of measure, localized labels and usage guidance text. Attribute authors may elect to begin with a pre-existing attribute and extend it. Entity authorship will most often be required in order to accommodate a newly defined attribute. That is, when a new attribute is defined, an entity definition must be updated or originated that references that attribute.

Note that entity and attribute definitions have a secondary usage: they describe the data they collect. Consumers of diverse datasets can attest to the importance of quality metadata, yet again because the data producer has little or no inherent incentive to describe their own datasets ("I know what I mean"), data consumers situated elsewhere in place and time are almost always frustrated. Because entity and attribute definitions rigorously define the authorship, purpose, usage and validation of schema components, for these are all necessary to the workings of the system, interpretation of data is substantially improved.

Like the data entry and management system, competitive implementations of application and schema designers can happily coexist. Implementations need only consume standardized schema components and conform to the same standard in their customizations and authoring of new components.

5. Aggregation and Analysis Tools

The community-driven standards approach enables aggregation tools to harvest standards-conformant data from semantically heterogeneous datasets. Data may be harvested across geographic scales and across organizational boundaries. Aggregation tools may query repositories using Web Services protocols. So long as all entity and attribute data is accompanied by its schema reference, the data can be properly interpreted and combined. Analysis tools can be constructed based on attributes common across the resulting datasets. The extent of shared entity and attribute usage directly defines the extent of data available for aggregation. All rests upon broad adoption, not of specific data entry, management, aggregation or analysis tools, but of the Community Schema Library and the schema standards it requires.

Conservation Example: Broad-Scale Cost/Benefit Analysis of Proscribed Burns

Proscribed burns may involve a wide variety of specific techniques and measurements across organizations and geographies. Using the systems proposed here, the extent of adoption of a core set of attributes for burn events will determine the extent of data available for analysis. At broad geographic and temporal scales, a variety of analyses can be conducted. For instance, data from proscribed burns may be combined with that of fire-dependant species distributions to determine overall efficacy of burn treatments in ensuring their viability.

6. Data Adaptors and Semantic Mediation

The approach to standardized conservation datasets taken here may appear to suffer from a significant drawback: only datasets originated in the system are subject to its benefits. However, a final category of tools can be developed to address this drawback: data adapters. Data adapters can be used to bring existing datasets and even tools into the overall system. Data adapters can be created to crosswalk the schema from existing datasets to schema components in the Community Schema Library, thus making their data available for aggregation.

Indeed, a core tool category in the overall suite may be the "data adaptor factory." A data adaptor factory can interpret one or more vendor-specific database schema definitions such as SQL, Oracle's implementation of SQL, or, as a de facto standard, Microsoft Excel. Given an existing dataset captured in a compatible database technology, the adapter can present the native schema to the user and walk the user through an easy-to-use process of mapping data record and field to schema components in the Community Schema Library. A sophisticated adaptor may assist the user in specifying transformations of the input data to match existing schema components. Direct crosswalks will not always be possible. In that potentially common situation, users can choose between losing information and creating new entities and/or attributes.

Data may be adapted to the system as a single event for storage in a Kestrel-compatible data management system (i.e., import) or, when the dataset must be maintained in its original format, as a dynamic transformation. Excel spreadsheets are a strong candidate for one-time import thus "bringing into the fold" a potentially vast amount of previously inaccessible and nonstandard datasets. In the conservation community, there are numerous large biodiversity datasets hosted by sophisticated data management applications with a sizable user community and supported by advanced processes and companion systems. In these cases, the benefits of transitioning to a dynamic-schema system are unlikely to be worth the costs. However, a run-time data adaptor can respond to Web Services requests Kestrel-compatible aggregators and dynamically transform these datasets, augmenting them with schema component references and transforming the raw data where necessary and as defined by the adaptor.

Finally, data adaptors themselves can be registered in the Community Standards Library in association with the entities and attributes they support. Again, while tolerating competition, the fittest adaptors will rise to the top. Whether as a single event or dynamically through time, the Kestrel approach has the potential to mediate semantic differences between disparate datasets.

7. Community Support

An obvious and final addition to this suite of systems is a community web site that defines the overall system, clarifies the ownership and community management of the schema definition standard. For end-users, the site should provide tutorials on the Community Schema Library as well as a directory of tool implementations. For tool developers, the site may provide documentation of the schema definition language and a more technical view of Kestrel.

As stated above, the only system that must be shared is that of the Community Schema Library. Competition in all but this can be supported without compromising the integrity of the overall system. However, it should also be clear that the community should encourage substantial investments in shared systems. Indeed, because investments can be leveraged across widely varying conservation domains, for systems are no longer tied to specific schema and thus a specific problem domain, users may enjoy remarkable richness of functionality, usability and performance in all of the systems described here.

<u>Summary of Information Benefits</u>

By separating schema from systems functionality and facilitating shared schema components, the Kestrel approach encourages standards while not stifling variation. Users can take advantage of rich software functionality for data defined on their terms and need no longer wait several months (or years) for the new

version of the software to come out that includes their requested changes to the data structure. Thus ends the tyranny of the software developer over user schema. Software developers build functionality; the user community builds, maintains and retires schema. The open-source approach to the schema library allows disagreement and facilitates selection of the fittest schema components and adaptors. Open source projects are most successful when users are contributors, as will be the case with the Community Schema Library. Finally, data created or adapted from common schema components can be aggregated. While imperfect, this approach creates community-driven, sustainable and adaptive data standards. In this way, it is far superior to the explicit standards efforts described in the second strategy.

In combination with the first three strategies, an implementation of the Kestrel approach completes the potential of this overall strategy. At this point, conservation datasets, even those that are substantially heterogeneous in semantics and structure, are online, secure, reliable, maintained by their rightful owners, self-describing and available for aggregation and analysis by the conservation community through web services protocols. As the basic functions of data entry, management, reporting and access are enabled by shared, schema independent applications, tool development can focus on more specialized needs in analysis or high performance aggregation.

Sample Conservation Benefits

Overall, the information systems strategy proposed here enables science-based conservation. By tolerating variability in the context of evolving standardization, the strategy supports the productivity of researchers and practitioners in the field with rich tools and enables downstream aggregation and analysis of their data at larger geographic and temporal scales. The implications for specific data domains are vast. This approach makes operational monitoring of ecosystems of the kind called for in the "State of the Nations Ecosystems" practical. A "Living Planet Index" might summarize observation data with far more scope and coverage. Analysts can assess ecological information dynamically to inform decisions and multiple geographic scales, generalizing and summarizing as a function of the input datasets. Ecologists can similarly assemble climate change scenarios from heterogeneous data sources to summarize specific biotic responses to changing abiotic conditions. Land-use planners can similarly value ecosystem services based on a combination of generalized calculations and analysis tailored to local conditions. All conservation analysis, including that of ecosystems, ecosystem services and climate change, can more responsive, complete, and scientifically credible.

Similarly conservation practice need not be constrained by widely accepted methodologies, standards and tools. Conservation practitioners can manage their restoration and ecological stewardship activities in

fine detail, while still enabling managers to aggregate common elements. In this way, the system provides practical support for adaptive management.

Challenges

The problem with this approach is that the technologies on which it is based are not fully developed. While systems developers have created the initial implementations, full-scale development is several years away and is largely in the hands of technologists outside of conservation. We will share these ideas and collaborate on requirements. In the meantime, conservation must work with static data models and build infrastructure. Even with implementations available, divergence of schema will occur in community factions. There will always be a role for standards bodies and a need for negotiation.

6. Sustainable Information Systems: Organizations and Community

Because information technology (IT) reduces costs and creates fundamentally new opportunities, private sector organizations pursuing a specific information technology (IT) strategy have on average 20% higher profits. Yet IT is expensive. Average private sector spending is 4.2% of annual revenues and rising (Weill & Ross 2004). Senior managers must therefore understand and control IT spending to ensure return on investments. Furthermore, infrastructure investments such as those proposed in this strategy are crucial for they allow managers to balance near-term returns with flexibility to support future needs. Finally, in information-intensive organizations, IT must be explicitly supportive of and integrated into organizational strategies (Weill & Ross 2004). As has been shown here, conservation is increasingly information-intensive.

Success of IT strategies directly correlates to top management engagement (Weill & Ross 2004). To realize the benefits of systems, development, deployment and maintenance of effective information systems is necessary but not sufficient for overall return on investment. The user community must be committed, whether by carrot or stick, to the required changes. Furthermore, centrally developed and managed IT is no longer possible or desirable. In successful distributed organizations, IT spending originates all over the enterprise and is coordinated with central investments. Central IT investments focus on infrastructure and shared services while business units develop utilities and end-user applications to meet more specific needs (Weill & Ross 2004).

In response to the information-intensive demands of its business, this strategy proposes that members of the conservation industry invest in information technology consistent with private sector, specifically 4.2% of annual revenues. In tandem, we propose that leadership in conservation organizations a) prioritize their own understanding and development of the role IT plays in conservation strategies, b) set clear objectives for both IT as well as user communities, c) invest in infrastructure and shared services development and maintenance, d) establish stable and effective governance and e) closely monitor progress.

To encourage data sharing, especially given sufficient deployment of hosting services and supporting tools, conservation organizations should explore requirements and rewards such as convincing research journals in related disciplines to require “publication” of datasets (i.e., their availability on accessible servers) as part of publication of peer-reviewed research articles.

To support coordination efforts and facilitate industry-wide learning, this strategy proposes an annual Conservation Information Systems Conference hosted by the Conservation Commons. At this meeting, conservation organizations, their public and private partners and donors can gather to ensure alignment of their strategies and leverage funding. Specific tracks may inform the management of regional data nodes, build initiatives around specific conservation data themes, negotiate data exchange standards (in lieu of solutions like Kestrel), provide incentives for data sharing and systems development and share lessons learned from both successes and failures.

Finally, the Conservation Commons website should be significantly enhanced to support the ongoing development and implementation of an information systems strategy for the conservation community. "We will lose the race to conserve nature unless we can establish systematic collaboration among conservation groups. This cooperation could set the stage for reaching a consensus about a set of conclusions and metrics for measuring and achieving global conservation. These could then be used to obtain broad society support for the conservation mission." (Redford et al. 2003)

Conclusion

This research has proposed an information systems strategy for consideration by the conservation community. The foundation is basic access to data on which we can build an infrastructure for interoperable datasets. We will foster the development of standardized datasets in a ground-up, iterative fashion, sustained on all sides by powerful, usable tools. As stated above, users in conservation must be able to assemble their own solutions from interoperable components. In the initial phases of the strategy, these interoperable components are web-service enabled datasets supported by entry and management tools as well as computational services to transform, reconcile, aggregate and analyze. In the later phase, bolstered by emerging semantic web technologies, the interoperable components are the very definitions of data, the entities and attributes defined and shared in a common library. The fine-grained leverage of data definitions, authored by the user community itself, will enable unprecedented aggregation and analysis of associated datasets. Finally, none of these advances will occur without the commitment of leadership to engage and invest, integrating information management and systems into their core strategies.

The game has changed in conservation. We can no longer rely on intuitive valuation of species, places or even biodiversity and the idea that "more is better" to tell us where to work. Rather, the systems-view is taking hold, one that requires intricate understanding of ecological relationships and dependencies and, most significantly, explicitly incorporates human impacts and dependencies on nature. Conservation strategies are changing in parallel. We are increasingly dependant on partners to incorporate conservation objectives and detailed plans into their work. We must inform policy and market prices based on social, economic and ecological principals and detailed analyses.

In the new game, information and information systems play a crucial role. With this strategy in place, we will be able to readily assess priorities, support methodologies and measurement requirements with tools. Productivity of researchers and practitioners can be significant enhanced. We will know what works and where we coordinate or need to improve. We can inform policies such as State Wildlife Action plans and payment systems for ecosystem services. We can inform tradeoffs and change societal priorities with hard, credible data. Accordingly, senior managers must understand and manage the role of information technology in accomplishing their goals.

Proven technologies and recent developments from the technology industry are available and can transform the business of conservation to become much more information driven, as it must be. Strategic

investments that solve real problems have a way of taking off. As a participant in the data management workshops on the State of the Nation's Ecosystems project remarked, "Experience has shown that when a credible organization creates an architecture and guidelines for integration and publishes it (with opportunities for feedback to refine and evolve the architecture and guidelines) the community will adopt and adapt to the architecture and guidelines as a matter of course" (Clark 2006). The collective missions of conservation organizations, whether to conserve specific species, places or the conditions of a healthy planet, demand that we dramatically improve our management of information and seize the opportunities provided by information systems. We must prioritize this need and convene the conversations to move forward. This strategy hopefully clarifies the need for that critical conversation and offers some useful approaches to consider.

Acknowledgements

I thank my advisor, Pat Halpin, for his candid feedback and high standards for this work. To all my Duke professors and support staff, those that taught me so well and provided such remarkable examples of both brilliance *and* kindness, I wish to thank Norm Christensen, Deb Gallagher, Dean Urban, Sara Aschenburg, and Don Wells. I thank all those who generously took the time to give me their insights on this as a science and bioinformatics research subject, especially Nick Salafsky, Silvio Olivieri, Ewen Eberhardt, Will Allen, Gina LaRocco, Alison Rowles-Anobile and Jaime Cavelier. And from the techie side, I have enjoyed inspiring collaborations with Alon Halevy, Paul Allen (yes, the one from the Cornell Bird Lab!), and Jason Roberts. I am grateful for the dedication, thoughtfulness and patience of my NatureServe colleagues, specifically Rob Solomon, Dave Hauver, and Bruce Stein, in the development of Kestrel. My colleagues at the Nature Conservancy, especially Paul Angelino, Mario Huipe, Dan Salzer, Peter Hage, Jonathon Adams and Michael Parker, have provided me with a wonderful mixture of support, inspiration and grounding in the completion of this work. Finally, to my boss at The Nature Conservancy, Dennis Fuze, immeasurable thanks for your leadership and support.

References

Barker, K. 2008. Various IT managers from large conservation organizations were surveyed from December 2007 to April 2008 for this research. I volunteered not to share individual or organization names.

Christensen, J. 2002. Fiscal accountability concerns come to conservation.(conservationists want more for the money)(Science Pages). The New York Times:D2(N) pF2(L).

Clark, W. 2006. Filling the Gaps: Priority Data Needs and Key Management Challenges for National Reporting on Ecosystem Condition. A Report of the Heinz Center's State of the Nation's Ecosystem Project. Heinz Center for Science, Economics and the Environment, Washington, DC.

CMP-Actions. Proposed Classification of Conservation Actions The Conservation Measures Partnership Group http://conservationmeasures.org/CMP/IUCN/browse.cfm?TaxID=ConservationActions

CMP. Conservation Measures Partnership www.conservationmeasures.org

Comer, P., D. Faber-Langendoen, R. Evans, S. Gawler, C. Josse, G. Kittel, S. Menard, M. Pyne, M. Reid, K. Schulz, K. Snow, and J. Teague. 2003. Ecological Systems of the United States: A Working Classification of U.S. Terrestrial Systems. NatureServe, Arlington, Virginia.

ESRI. 2008. Geoprocessing. Environmental Systems Research Institute, Redlands, CA http://support.esri.com/index.cfm?fa=downloads.geoprocessing.matrix

Ferraro, P. J., and S. K. Pattanayak. 2006. Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments. PLoS Biology **4**:e105.

GBIF. Global Biodiversity Information Facility.

Green, T. J., G. Karvounarakis, E. T. Nicholas, O. Biton, Z. G. Ives, and V. Tannen. 2007. ORCHESTRA: facilitating collaborative data sharing. Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, Beijing, China.

Grossman D.H., Faber-Langendoen D., Weakley A.S., Anderson M., Bourgeron P., Crawford R., Goodin K., Landaal S., Metzler K., Patterson K.D., Pyne M., Reid M., and Sneddon L. 1998. International classification of ecological communities: terrestrial vegetation of the United States. Volume I, The National Vegetation Classification System: development, status, and applications. The Nature Conservancy: Arlington, VA.

Halevy, A. Y., N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka. 2005. Enterprise information integration: successes, challenges and controversies. Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, Baltimore, Maryland.

IABIN. 2004. Biodiversity Tools - Systematics and Software Applications http://old.iabin.net/english/bioinformatics/guide/systematics.shtml

IUCN-SSC. Threats Types Authority File. IUCN Red List of Threatened Species. IUCN Species Survival Commission http://www.iucnredlist.org/info/major_threats

Ives, Z., N. Khandelwal, and A. Kapur. 2005. ORCHESTRA: Rapid, Collaborative Sharing of Dynamic Data. Conference on Innovative Database systems Research (CIDR), Asilomar, CA.

Kareiva, P., and M. Marvier. 2007. Conservation for the People. Scientific American **297**:50-57.

Loh, J., R. E. Green, T. Ricketts, J. Lamoreux, M. Jenkins, V. Kapos, and J. Randers. 2005. The Living Planet Index: using species population time series to track trends in biodiversity. Philosophical Transactions of the Royal Society B: Biological Sciences **360**:289-295.

Lovejoy, T. E. 2006. Glimpses of Conservation Biology, Act II. Conservation Biology **20**:711-712.

Ludäscher, B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. 2006. Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience **18**:1039-1065.

MA. 2005. Ecosystems and Human Well-being: Biodiversity Synthesis. Millennium Ecosystem Assessment. World Resources Institute, Washington, DC.

Manes, A. T. 2003. Web Services: A Manager's Guide. Addison-Wesley Longman Publishing Co., Inc.

McShane, T. O. 2003. The Devil in the Detail of Biodiversity Conservation. Conservation Biology **17**:1-3.

NatureServe. Kestrel Observations Data Management System. NatureServe, Arlington, VA http://kestrel.natureserve.org

NatureServe. 2008. NatureServe Web Services http://services.natureserve.org/

Nelson, E., G. Mendoza, J. Regtz, S. Polasky, H. Tallis, D. Cameron, K. Chan, G. Daily, J. Goldstein, P. Kareiva, E. Lonsdorf, R. Naidoo, T. Ricketts, and R. Shaw. forthcoming. Modeling Multiple Ecosystem Services and Tradeoffs at Landscape Scales.

OFX. Open Financial Exchange. http://www.ofx.net/

Redford, K. H., P. Coppolillo, E. W. Sanderson, G. A. B. Da Fonseca, E. Dinerstein, C. Groves, G. Mace, S. Maginnis, R. A. Mittermeier, R. Noss, D. Olson, J. G. Robinson, A. Vedder, and M. Wright. 2003. Mapping the Conservation Landscape. Conservation Biology **17**:116-131.

Robb, D. 2007. Data Centers: The heat is on. EPA's Energy Star program could focus on power consumption in the server room. Government Computer News, Falls Church, Va.

Root, T. L., and S. H. Schneider. 2006. Conservation and Climate Change: the Challenges Ahead. Conservation Biology **20**:706-708.

TADWG. Taxonomic Data Working Group http://www.tdwg.org

TEAM. Tropical Ecology Assessment and Monitoring Network. Conservation International, Alexandria, VA http://www.teaminitiative.org

TNC. 2006. Conservation Information Systems Strategy for Achieving the 2015 Goal. The Nature Conservancy, Arlington, VA.

Wang, X. H., D. Q. Zhang, T. Gu, and H. K. A. P. H. K. Pung. 2004. Ontology based context modeling and reasoning using OWL. Pages 18-22 in D. Q. Zhang, editor. Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on.

Weill, P., and J. W. Ross 2004. IT governance : how top performers manage IT decision rights for superior results. Harvard Business School Press, Boston.

Wiens, J., and T. Comendant. 2005. Conservation Information Needs and Priorities in the Nature Conservancy: Are We Managing Our Information Assests Effectively? A Report from the Science Office. The Nature Conservancy, Arlington, VA.

Wu, J., and R. Hobbs. 2002. Key issues and research priorities in landscape ecology: An idiosyncratic synthesis. Landscape Ecology **17**:355-365.

Zermoglio, M. F., A. S., V. Jaarsveld, W. V. Reid, J. Romm, R. Biggs, Y. Tianxiang, and L. Vicente. 2005. The Multiscale Approach. Pages 61-83 in D. Capistrano, M. Lee, C. Raudsepp-Hearne, and e. C. Samper, editors. Ecosystems and human well-being : multiscale assessments volume 4: findings of the Sub-global Assessments Working Group of the Millennium Ecosystem Assessment. Island Press, Washington, DC.

Appendix A: Web Services Application Architecture

Web Services provide a stable interface to functionality and data, protecting client applications that use them from the details of their implementation. Web Services therefore provide a means of integrating previously incompatible technologies. In a technique sometimes called "wrapping", a Web Service interface can be added to an existing application, potentially developed decades earlier thus allowing this legacy applications' data and functionality to interoperate within a modern Service-Oriented Architecture.

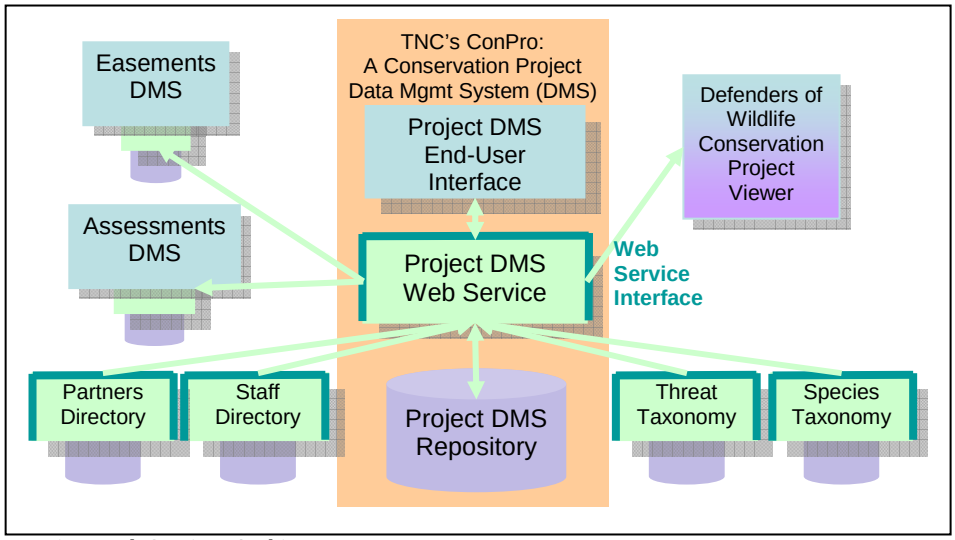

Figure A-1: ConPro in a Web Services Architecture.
The Nature Conservancy's web-based application for managing conservation projects, ConPro, was not designed from the perspective of a Service-Oriented Architecture. However, its core functionality and data can be "wrapped" in a Web Service. This will allow other applications, including those tracking the activity on the Conservancy's easements, to "pull in" conservation project data from ConPro and integrate that data into the presentation of easements.

In figure A-1, a Defenders of Wildlife's Conservation Project Viewer application can read conservation project information from multiple organizations across geographies providing that participating servers offer the same Web Service interface. Web Services can also provide functionality such as taxonomic reconciliation, mapping one species definition to another, thus allowing client systems to interoperate without adopting a common standard.

Appendix B: Evaluating Investments in Data Standards

To evaluate investment in the standardization of a dataset category such as "protected areas" or "ecological assessments," a number of criteria may be considered. I offer here a sample framework to evaluate investments in a given data standard. Each criterion is assigned a value ranging from "Very Low" as 0 to "Very High" as 5, the value is weighted and all criteria are summed to give the final cost/benefit score. The focus of this research is at a higher level. Future research should refine this framework by testing it against a variety of conservation datasets.

Sample Criterion	Definition	Sample Weight
Benefits		
Value to Conservation	Value to conservation of analysis based on aggregated datasets. Value may be a function of both urgency and strategic importance.	1.0
Dependency on completeness	Dependency of the value to conservation on the completeness of producer participation	-0.2
Successful Data Standard	If there are existing data standards in place, score their adoption by data producers	0.6
Producer Rewards	Direct value to producers by participating (i.e., contributing data)	0.5
Value of Tools to Producers	Direct value of any data authoring and management tools to the producers of contributing datasets	0.6
Producer Signers	Percentage of potential data producer organizations committing to the standard. This value might be weighted by each organization's percentage of the producer sector.	0.6
Size of Producer Sector	Relative size of data producer (providers) sector	0.4
Size of Consumer Sector	Relative size of data consumer (users) sector	0.4
Costs		
Heterogeneity	Complexity, variation and dynamism of contributing datasets	-0.7
Publishing Complexity	Complexity and controversy of reduction transformations required to publish to the standard	-0.3
Tier-able	Heterogeneity and publishing complexity can be offset if the dataset can be represented by a less complex, less controversial hierarchy of complexity	0.1
Shared Methodology	Degree of acceptance of a shared methodology to produce standard-conformant data. The methodology must lend itself to automation in a software application.	0.2
Cost of producer influence	Cost to conservation organizations of influencing data producers to participate. Lack of influence translates to higher costs.	-0.2
Risk Exposure	Sensitivity of data records will increase the costs of securing data exchange protocols and exposure should security systems fail	-0.3

Sample Criterion	Definition	Sample Weight
Dependant Datasets	Cost analysis of datasets on which the value of this dataset depends. Each dependant dataset is scored and weighted separately and then added to this dataset's score. More and costly dependencies reduce the overall value.	-Σ (0.1 * -Cost(dataset))

Note that the cost/benefit analysis for the development of an exchange standard is malleable. Specifically, it may be improved by finding ways to reward data producers, provide them with inherently valuable, structuring the dataset's complexity, or developing shared methodologies.

In the sample analysis below, that is, not based on rigorous research, the overall value of a protected areas exchange standard is significantly higher than that of ecological assessments.

Cost/Benefit Analysis of Exchange Standard		Protected Areas		Ecological Assessments	
Criterion	**Weight**	**Value**	**Weighted Value**	**Value**	**Weighted Value**
Benefits					
Value to Conservation	1.0	5	5	5	5
Dependency on completeness	-0.2	4	-0.8	3	-0.8
Success of Data Standards	0.6	2	1.2	0	0
Producer Rewards	0.5	1	0.5	3	0.5
Value of Tools to Producers	0.6	1	0.6	4	0.6
Size of Producer Sector	0.4	3	1.2	2	1.2
Producer Signers	0.5	4	2.0	3	2.0
Size of Consumer Sector	0.4	5	2.0	5	2.0
Costs					
Heterogeneity	-0.7	1	-0.7	5	-0.7
Publishing Complexity	-0.3	1	-0.3	5	-0.3
Tier-able	0.1	4	0.4	4	0.4
Shared Methodology	0.2	3	0.6	1	0.6
Risk Exposure	-0.3	2	-0.6	3	-0.9
Cost of producer influence	-0.2	3	-0.6	5	-0.6
Dependant Datasets	-Σ (0.2 * -Cost(dataset))		0		-6.0
	Total		**10.5**		**2.2**

Appendix C: Sample Field Observation Schema

The schema below describes a field observation entity ("observationCore") in the Conservation Commons domain in terms of the attributes that define it. Some sample attribute definitions follow including species identification, evidence type and evidence resource.

```
<?xml version="1.0" encoding="utf-8"?>
<obs:schema  xmlns:obs="http://services.conservationcommons.org/observations">
  <EntityDefinition name="observationCore" namespace="entities.datastd.conservationcommons.org"
xmlns="http://services.conservationcommons.org/observations">
    <AttributeRef uri="speciesIdentification.attributes.datastd.conservationcommons.org"
required="true" supportsComments="true" >
    </AttributeRef>
    <AttributeRef uri="dateRange.attributes.datastd.conservationcommons.org" required="true"
supportsComments="true">
      <LabelText xmlns="">
        <Text lang="en-US">Observation Date</Text>
        <Text lang="fr-CA">Date de l'observation</Text>
      </LabelText>
      <HelpText xmlns="">
        <Text lang="en-US">Day, month, and year when observation was made. Use Date Range if
precise date is not known.. Use Date Range if precise date is not known.</Text>
        <Text lang="fr-CA">Jour, mois et année au cours duquel l'observation a été réalisée.
Entrez une plage de date si la date précise est inconnue.</Text>
      </HelpText>
    </AttributeRef>
    <AttributeRef uri="location.attributes.datastd.conservationcommons.org" required="true"
supportsComments="true">
      <LabelText>
        <Text lang="en-US">Location</Text>
        <Text lang="fr-CA">Emplacement</Text>
      </LabelText>
      <HelpText>
        <Text lang="en-US">place where the observation was made</Text>
        <Text lang="fr-CA">L'endroit où l'observation a été réalisée</Text>
      </HelpText>
    </AttributeRef>
    <AttributeRef uri="person.attributes.datastd.conservationcommons.org" required="true"
supportsComments="true">
      <LabelText>
        <Text lang="en-US">Primary Observer</Text>
        <Text lang="fr-CA">Observateur principal</Text>
      </LabelText>
      <HelpText>
        <Text lang="en-US">Person who is made the observation or who is the primary contact for
information about it</Text>
        <Text lang="fr-CA">Personne qui a fait l'observation ou du principal contact au sujet de
celle-ci</Text>
      </HelpText>
    </AttributeRef>
    <AttributeRef uri="sensitive.attributes.datastd.conservationcommons.org" required="false"
supportsComments="true" />
    <AttributeRef uri="evidenceType.observation.attributes.datastd.conservationcommons.org"
required="false" supportsComments="true"/>
      <Behavior>
        <AllowMultiple/>
      </Behavior>
    </AttributeRef>
    <AttributeRef uri="evidenceResource.observation.attributes.datastd.conservationcommons.org"
required="false" supportsComments="true">
      <Behavior>
        <AllowMultiple/>
      </Behavior>
    </AttributeRef>
```

```
    <AttributeRef uri="habitatDescription.observation.attributes.datastd.conservationcommons.org"
supportsComments="true" required="false" />
  </EntityDefinition>

  <AttributeDefinition onEntity="observation.entities.datastd.conservationcommons.org"
name="speciesIdentification" namespace="attributes.datastd.conservationcommons.org">
    <DataType>
      <CoreType>
        <EntityReference
toEntity="speciesIdentification.entities.datastd.conservationcommons.org" />
      </CoreType>
    </DataType>
    <Confidence required="true" type="interval5percent"/>
    <LabelText>
      <Text lang="en-US">Species Identification</Text>
      <Text lang="fr-CA">Espèce Identification</Text>
    </LabelText>
  </AttributeDefinition>

  <AttributeDefinition name="evidenceType" namespace="attributes.datastd.conservationcommons.org"
onEntity="observation.entities.datastd.natureserve.org">
    <DataType>
      <CoreType>
        <PickList>
          <ListValue value="sighting">
            <DisplayValue lang="en-US">sighting</DisplayValue>
            <DisplayValue lang="fr-CA">observation</DisplayValue>
          </ListValue>
          <ListValue value="specimen">
            <DisplayValue lang="en-US">specimen</DisplayValue>
            <DisplayValue lang="fr-CA">spécimen</DisplayValue>
          </ListValue>
          <ListValue value="capture">
            <DisplayValue lang="en-US">capture</DisplayValue>
            <DisplayValue lang="fr-CA">capture</DisplayValue>
          </ListValue>
          <ListValue value="photograph">
            <DisplayValue lang="en-US">photograph</DisplayValue>
            <DisplayValue lang="fr-CA">photographie</DisplayValue>
          </ListValue>
          <ListValue value="sound">
            <DisplayValue lang="en-US">sound</DisplayValue>
            <DisplayValue lang="fr-CA">son</DisplayValue>
          </ListValue>
          <ListValue value="soundRecording">
            <DisplayValue lang="en-US">sound recording</DisplayValue>
            <DisplayValue lang="fr-CA">enregistrement sonore</DisplayValue>
          </ListValue>
          <ListValue value="dna">
            <DisplayValue lang="en-US">DNA</DisplayValue>
            <DisplayValue lang="fr-CA">ADN</DisplayValue>
          </ListValue>
          <ListValue value="tracks">
            <DisplayValue lang="en-US">tracks</DisplayValue>
            <DisplayValue lang="fr-CA">pistes</DisplayValue>
          </ListValue>
          <ListValue value="scat">
            <DisplayValue lang="en-US">scat</DisplayValue>
            <DisplayValue lang="fr-CA">excrément</DisplayValue>
          </ListValue>
          <ListValue value="otherSign">
            <DisplayValue lang="en-US">other sign</DisplayValue>
            <DisplayValue lang="fr-CA">autre signe</DisplayValue>
          </ListValue>
        </PickList>
      </CoreType>
    </DataType>
    <LabelText>
      <Text lang="en-US">Evidence Type</Text>
      <Text lang="fr-CA">Type de preuve</Text>
    </LabelText>
```

```
        <HelpText>
          <Text lang="en-US">The type of information on which the observation record is based.
      </Text>
          <Text lang="fr-CA">Type d'information sur lequel s'appuie l'enregistrement de
      l'observation</Text>
        </HelpText>
      </AttributeDefinition>

      <AttributeDefinition name="evidenceResourse" namespace="attributes.datastd.natureserve.org"
      onEntity="speciesIdentification.entities.datastd.natureserve.org">
        <DataType>
          <CoreType>
            <DigitalAsset/>
          </CoreType>
        </DataType>
        <LabelText>
          <Text lang="en-US">Evidence</Text>
          <Text lang="fr-CA">Evidence</Text>
        </LabelText>
        <HelpText>
          <Text lang="en-US">Provide the digital audio, video or image</Text>
        </HelpText>
      </AttributeDefinition>

</obs:schema>
```