VertNet: Creating a Data-Sharing Community

Authors: Guralnick, Robert, and Constable, Heather

Source: BioScience, 60(4) : 258-259

Published By: American Institute of Biological Sciences

URL: https://doi.org/10.1525/bio.2010.60.4.2

VertNet: Creating a Data-sharing Community

ROBERT GURALNICK AND HEATHER CONSTABLE

Vertebrate biodiversity data-sharing projects have grown by leaps and bounds over the last 10 years, and are now providing more than 50 million biodiversity data records from more than 70 different institutions. These records are accessed at a rate of approximately 2.5 million per week. Success has been achieved through a combination of distributed database technology and a revolutionary change in knowledge about and attitudes toward data sharing and open access (Constable et al. 2010). We argue that the best strategy to ensure long-term, sustainable data-sharing success is to develop a network of people with the knowledge, skills, and tool kits to publish their data and work collaboratively. Such collaborations are integral to the very nature of a data-sharing network, and are fundamental to increasing the quality and quantity of data available to the broadest possible base of biodiversity data users. We are poised to consolidate all of the vertebrate biodiversity projects onto a new platform, called VertNet, that will significantly enhance the ability of our human network to work collaboratively and efficiently.

Initial collaborations to build the networks

The Mammal Network Information System (MaNIS, *http://manisnet.org*; Stein and Wieczorek 2005), the first demonstrated working example of a distributed biodiversity database network, was started in 2001 and worked like a pebble dropped in a pond. The ripples have spread outward and led to a wider scope of projects—for amphibians and reptiles in 2002 (HerpNET; *http://herpnet.org*), and fishes (FishNet 2; *www.fishnet2.net*) and birds in 2004 (ORNIS; *http://ornisnet.org*)—all funded by the National Science Foundation, Global Biodiversity Information Facility (GBIF), and National Biological Information Infrastructure (NBII). What was once a small group of like-minded individuals from 17 institutions, who agreed to share their data in MaNIS, has grown to include 72 institutions across the vertebrate biodiversity networks. The quadrupling of the size of our network in only nine years has led to new training and collaboration opportunities and raised challenges for our community, which became the focal points for later projects such as HerpNET and ORNIS.

The technology used to develop the vertebrate biodiversity networks is a means to share data and to facilitate training and collaboration. Initially, a clear and feasible technological plan was essential in order to create confidence early on in our data-network development. The participation process to date requires contributors to set up local software that registers their data resources to the network and specifies how their data match a shared schema. For most institutions, such setups are often difficult and require a roving programmer-expert to help perform installations on site or remotely. As the network has grown, so have the challenges to keep it running, as any of the myriad hardware or software snafus can lead to provider downtime until an expert can troubleshoot and fix the problem. A major issue is that the technology is not yet easy enough for most contributors to install or even manage on their own.

We also very quickly found that contributors wanted to improve the quality of their data. However, such improvements could not be done in a willy-nilly fashion, and required coordination across the network of participants. An essential data-quality step was adding geospatial coordinates (e.g., latitude and longitude) and associated geographical uncertainty based on a textual locality description field (e.g., 5 miles south-southeast of Berkeley, California). The process of retrospective georeferencing (Wieczorek et al. 2004) represented a huge challenge for two main reasons. First, georeferencing takes a considerable amount of knowledge and training to ensure consistency across the network. Second, the most efficient approach to georeferencing was to do it not by collection but by consolidating the work by geographic region and developing full-scale collaborations among institutions.

An essential solution to the challenge of creating consistency and fostering collaboration was to develop training programs for participants. With the start of HerpNET, regular georeferencing training workshops were held across the world for participants, museum staff, and members of other biodiversity informatics projects, with the support of the GBIF. We have held 14 international georeferencing workshops that have trained 292 higher-level researchers from 169 institutions in 41 countries. In addition to increasing technical understanding and capacity among contributors, these workshops foster a sense of community across traditional disciplinary boundaries.

VertNet: Consolidation and enhanced collaboration across the network

The successes of the vertebrate biodiversity networks have created new challenges. We have made great progress in growing the network and increasing data quality. However, the pace of georeferencing lags dramatically behind the rate of new records contributed to the network, and since contributors often lack the expertise or resources to maintain their hardware and software, the upkeep of the providers and network is a Sisyphean task. The vertebrate networks have had one NBII-funded programmer since 2008, who has serviced 73 percent of all contributor installations. This level of effort means that our speed of adding new contributors to the network is very slow, thus explaining why we have a long wait-list of 31 interested institutions. We are developing VertNet to clear this backlog by simplifying the data-publication process and expanding our training missions.

A key part of simplifying the publication process is to consolidate all the data and technology tools onto the "cloud." Cloud computing leverages third-party, dynamically scalable, and often virtualized computing resources. Think of the cloud as a very large, always growing, Internet-based data center. Examples of cloud-based services include Amazon Elastic Cloud Computing and the Google App Engine, already widely in use. In the near future, VertNet will move all contributor data onto the cloud. Contributors will still have full control over their data resources and will be able to update their records easily using simplified software that connects their local resources to the cloud-based data snapshot. Because all data and many tools are consolidated onto, or connected to, the cloud, contributors do not need to maintain cumbersome and expensive hardware (i.e., servers) and software installations. Although there will be upfront expenses to reconfigure VertNet to be cloud based, we expect a 16-fold decrease in management costs once the system is up and running. As well, consolidation simplifies accessing all contributors' data simultaneously. These changes, in total, greatly facilitate the development of network-wide collaborative tools.

Contributor-mediated publishing in the cloud introduces the feasibility of adding data-quality improvement services to the data-publishing workflow. As data are accessed, VertNet on the cloud will provide annotation functions that allow users to alert primary data contributors of errors or omissions and make comments while maintaining data provenance and integrity at the source. Under this model, VertNet can also provide a workflow and platforms for collaborative georeferencing, which will accelerate the rate of geospatial data enhancement. Within this architectural and community framework, vertebrate biodiversity data can be reliably accessed, visualized across space and time, and constantly enriched as our global biodiversity knowledge base grows.

As VertNet moves into this new stage of development, we plan to dramatically diversify and grow our training process. Past workshops for the community focused mainly on researching historical collection data, georeferencing and data standards, geography, maps, Internet resources, and geographic information systems. Future courses will include previous training material plus education in research applications, basic database concepts, schemas, metadata collection, network technologies, and the technical architecture behind the resources. These workshops will help spread the knowledge necessary to more effectively maintain and grow our networks, thus greatly alleviating expertise bottlenecks, and will provide a much-needed but currently unavailable program in biodiversity informatics education. We will also include undergraduate internships to introduce students to the field of biodiversity informatics, with the hope that this burgeoning area of data and tool development for biodiversity can attract the best and brightest students.

Vertebrate-based biodiversity networks have proven that the natural history collections community can self-organize in order to work collectively for the common good. We are convinced that the best way forward is to use technology in the service of building expertise and capacity among the people and institutions that contribute to our networks, either now or in the future. Our data consumers can take active roles in the promotion of this goal, effectively leading to a network of "prosumers"—people who produce and consume new data, information, and knowledge. The promise of a global, collaborative infrastructure for biodiversity knowledge is well on its way to being achieved.

References cited

Constable HC, Guralnick RP, Wieczorek JR, Spencer C, Peterson AT, VertNet Steering Committee. 2010. VertNet: A new model for biodiversity data sharing. PLoS Biology 8: e1000309.

Stein BR, Wieczorek J. 2004. Mammals of the world: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics 1: 14–22.

Wieczorek JR, Guo Q, Hijmans RJ. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18: 745–767.

Robert Guralnick (robert.guralnick@colorado.edu) is with the Department of Ecology and Evolutionary Biology and CU Museum of Natural History, at the University of Colorado, Boulder. Heather Constable (hconstable@gmail.com) is with the Museum of Vertebrate Zoology at the University of California, Berkeley.

doi:10.1525/bio.2010.60.4.2