

# Show me the numbers: what data currently exist for non-native species in the USA?

Alycia W Crall<sup>1\*</sup>, Laura A Meyerson<sup>2</sup>, Thomas J Stohlgren<sup>3</sup>, Catherine S Jarnevich<sup>3</sup>, Gregory J Newman<sup>1</sup>, and Jim Graham<sup>1</sup>

Non-native species continue to be introduced to the United States from other countries via trade and transportation, creating a growing need for early detection and rapid response to new invaders. It is therefore increasingly important to synthesize existing data on non-native species abundance and distributions. However, no comprehensive analysis of existing data has been undertaken for non-native species, and there have been few efforts to improve collaboration. We therefore conducted a survey to determine what datasets currently exist for non-native species in the US from county, state, multi-state region, national, and global scales. We identified 319 datasets and collected metadata for 79% of these. Through this study, we provide a better understanding of extant non-native species datasets and identify data gaps (ie taxonomic, spatial, and temporal) to help guide future survey, research, and predictive modeling efforts.

*Front Ecol Environ* 2006; 4(8): 414–418

Invasion by non-native species has adversely affected many ecosystems in the United States, threatening biodiversity, ecosystem function, human health, and the economy (Vitousek *et al.* 1997; Wilcove *et al.* 1998; Mack *et al.* 2000; Pimentel *et al.* 2000). Organisms continue to be introduced from other countries through trade and transportation, increasing the need for early detection and rapid response to new invaders (Vitousek *et al.* 1997). Synthesizing existing data on non-native species abundance and distributions can assist with this effort. When this is accomplished, new data from multiple sources can be integrated with existing data to provide the most up-to-date and accurate information on non-native species locations, creating a proactive control strategy (Ricciardi *et al.* 2000). However, the extent of existing non-native species data is not well known, and there have been few efforts to improve collaboration and data synergy among governmental agencies, non-governmental organizations, industry, academic researchers, and other non-native species networks (Crosier and Stohlgren 2004).

### In a nutshell:

- There are more than 300 existing non-native species datasets in the United States
- Most non-native species datasets cover plant species, leaving large gaps in our knowledge of other biological groups
- Better non-native species data integration should be undertaken to improve our ability to monitor and control the spread of harmful invaders

Although data for non-native species are collected using various research methods, spatial and temporal scales, and data quality procedures, the combination of these disparate datasets would have several benefits: (1) data sharing will help improve species lists for a particular area (Crosier and Stohlgren 2004); (2) data sharing and integration could provide watch lists to managed lands adjacent to currently invaded areas, to identify and prevent potential invasions before expensive control methods become necessary (Rejmànek and Pitcairn 2002); (3) combining species presence and distribution data would improve spatially predictive models on current and potential invasions, thus identifying data gaps to guide future surveys and research (Crosier and Stohlgren 2004); (4) datasets with varying temporal coverage will improve our ability to determine patterns of invasion over long periods of time; (5) combining multiple datasets will expand the extent and resolution of spatially limited datasets across multiple ownership boundaries; and (6) combining available datasets could leverage limited resources (ie time, money, and personnel) with minimal additional cost and effort.

Although some steps have already been taken to facilitate data sharing, these efforts have not been successful over large scales or are still in their infancy (Ricciardi *et al.* 2000; Simpson 2004). For example, the Global Invasive Species Information Network (GISIN) is developing a registry of online non-native species datasets worldwide to provide an outlet for easily obtaining non-native species information (Simpson 2004). However, many additional electronic datasets (eg spreadsheets, GIS) are not available online, perhaps because they are privately held or because the owners do not have the capacity to put their datasets online. In addition, many

<sup>1</sup>Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523-1499 \*(mawaters@nrel.colostate.edu);

<sup>2</sup>Department of Natural Resources Science, University of Rhode Island, Kingston, RI 02881; <sup>3</sup>Fort Collins Science Center, US Geological Survey, Fort Collins, CO 80526

non-native species data differ in how they are collected and the way they are handled following collection. Further, non-native species datasets may contain data that only meet specific research objectives. Metadata should therefore be collected and evaluated on each of these datasets to determine the type, quality, and availability of the data they contain.

### ■ Our search strategy

The Natural Resource Ecology Laboratory of Colorado State University, in collaboration with The Heinz Center ([www.heinzctr.org](http://www.heinzctr.org)), conducted a thorough review of existing non-native species datasets in the US from county, state, multi-state region, national, and global scales. The aim of this effort was to provide a better understanding of what data currently exist for non-native species and to determine where gaps exist (taxonomically, spatially, and temporally) to guide future survey, research, and predictive modeling efforts.

We began our search by reviewing existing non-native species datasets from both online and published sources. Several search strategies were implemented, including gathering pre-existing lists of datasets, conducting a comprehensive web search, and carrying out a literature review of related publications. To locate additional data sources, we contacted approximately 1500 experts in the field of non-native species science to determine if they had relevant data not found through our previous efforts. We were primarily concerned with information on species presence and distribution.

We then sent a request to these contacts to take our online survey in order to provide metadata for each of their datasets (Figure 1). Metadata included information related to the geographic scope, data collection methods, taxonomic focus, spatial extent, temporal coverage, and data quality of each dataset. Survey responses were automatically entered into a database linked to each survey question. Once the survey was closed, we were able to generate simple statistics to determine general patterns in the metadata we collected.

### ■ Conclusions

Our survey proved to be an effective way of collecting metadata on existing datasets. It was completed by a diverse group of researchers and land managers from a wide range of state and federal agencies, and so provides a

strong indication of what information is currently available. In our initial web-based and literature review efforts, we identified 188 datasets and 169 contacts. Of these 169 contacts, 43 did not respond to our survey participation request, giving a 75% response rate. After we sent out a request for survey participation to 1500 additional contacts, we added 155 datasets to the initial list, resulting in a total of 343 datasets and 315 dataset owner contacts. The number of total datasets dropped from 343 to 319 after being informed that 24 of our initial contacts did not in fact have a dataset.

This number represents 227 datasets beyond those found through the GISIN effort. The primary reason for this could be that the GISIN list deals specifically with online datasets; of the 319 datasets found, 43% were not available online. This demonstrates the importance of looking for additional data sources offline. To improve species lists and our modeling capabilities, we will need to make use of these additional data sources. Online tools

**THE HEINZ CENTER NON-NATIVE SPECIES DATABASE SURVEY THE HEINZ CENTER**

The Heinz Center, in collaboration with the USGS National Institute of Invasive Species Science, is working to build an exhaustive list of non-native species databases within the United States. The following survey has been developed to collect essential metadata for each of these databases. If any survey question does not specifically apply to your database, please address this in the comments field. If you have any questions about the survey, or if you prefer to answer the survey questions person-to-person, please feel free to contact [Alycia Wattuz](mailto:Alycia.Wattuz@niiss.org) (970) 481-2502.

Please enter the following information:

What is the name of your data set?

(If your data set does not have a name, please create one and enter it here.)

What is your data set's acronym?

(Leave blank if not applicable)

If your data set is online, give the URL. If not online, give its physical location.

Example URL: <http://www.nriis.org>

Example Physical Location: The National Institute of Invasive Species Science, A219 NEISS/NREL, Colorado State University, Fort Collins, CO 80523-1429

What was the purpose of data collection? (Please be concise)

Enter contact information for a person to contact regarding your data set.

Last Name

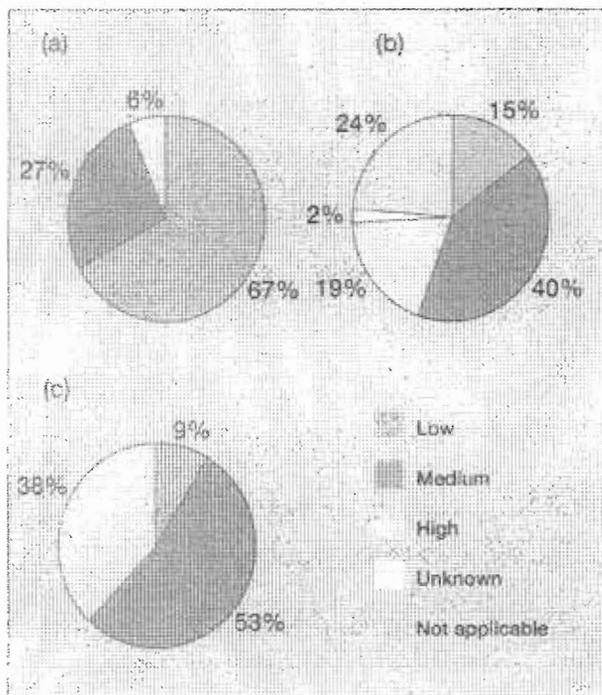
First Name

Email

Work Phone

Affiliation

**Figure 1.** A sample page from the online survey that was used to generate the results of this study. The full survey and report findings are available at [www.heinzctr.org/ecosystems](http://www.heinzctr.org/ecosystems) and [www.niiss.org](http://www.niiss.org).



**Figure 2.** Percentage of datasets in low, medium, and high categories for (a) taxonomic, (b) geographic, and (c) temporal completeness.

other than data harvesting services will need to be developed to provide data holders with easy methods for uploading their data online.

We closed the survey with 252 dataset entries from 214 survey participants (WebTable 1). We therefore collected metadata for 79% of the existing datasets found through our research. The remainder of this review will deal specifically with these 252 datasets entered into the online survey. An important caveat is that our results are solely dependent on our survey responses.

Each dataset was classified by dataset type to determine how each one could be used for various research and management objectives. Of the eight dataset types that we assigned, a majority had data on species locations and distribution (137). Species information, general species lists, and non-native species lists were the next most common dataset types, with 77, 64, and 58 datasets, respectively. There were also 41 datasets that tracked control of non-native species, nine distributed

datasets, 13 bibliographic datasets, and eight datasets that did not fit into any of the other categories.

#### Identification of gaps in non-native species datasets

To evaluate data completeness, we classified survey responses for each dataset into low, medium, and high classes for taxonomic, geographic, and temporal completeness categories using specified criteria. Although these classes were subjective, they were necessary to make the datasets comparable for these various fields.

We used information on the taxonomic focus of the datasets (ie plants, fungi, vertebrates, invertebrates, pathogens) to identify gaps in non-native species information specific to particular taxa. Low taxonomic completeness was defined as a dataset covering only one taxon, while high taxonomic completeness was defined as a dataset covering all taxa groups (Figure 2). A majority of datasets covered plants, twice as many as the next most prevalent taxonomic group (ie vertebrates; Table 1; WebTable 1; Figure 2a). Of all the datasets entered into our survey, 124 covered plants only, while just 60 did not include plants and focused solely on other taxa. These results suggest that more information should be collected on a wider range of taxa.

There are many reasons why plants tend to be more studied than other groups. Primarily, plants are easier to observe and record. Also, there are a greater percentage of non-native plants in the total plant species pool relative to other taxa (Pimentel *et al.* 2001). Greater numbers lead to greater attention in terms of control and monitoring efforts. Weedy plants also cause the greatest economic losses to crops and pastures in the US, whereas environmental losses are thought to be much greater for other taxonomic groups (Pimentel *et al.* 2001), but have not yet been adequately quantified. This could indicate that non-native species research efforts, and therefore datasets, may be driven primarily by economics rather than conservation.

It is also important to examine which systems are most vulnerable to invasion to fully assess the invasion patterns of all taxa (Stohlgren *et al.* 2006). The datasets collected through our survey covered all ecosystem types relatively well (general classifications included coasts and oceans, farmlands, forests, freshwaters, grasslands and shrublands, urban and suburban). In fact, the number of datasets for each ecosystem type was roughly proportional to the area of land that each of these systems covers within the US, using Bailey's ecoregions (Bailey 1980).

#### The influence of spatial and temporal scale

A full assessment of the impact of non-native species on a system requires that the spatial variability of the study area be adequately captured by the sampling process. To classify the geographic completeness of each dataset, we looked at survey responses related to how well the study area had been sampled, whether the study crossed all

**Table 1.** Number of datasets within each taxa group

Taxa group	Number of databases
Plants	193
Vertebrates	96
Invertebrates	77
Pathogens	36
Fungi	22

The total number of datasets listed here (424) adds up to more than the total number of datasets entered into our survey (252) because many datasets cover more than one taxonomic group.

environmental gradients within the study area, and the sampling design that was used in data collection. Low geographic completeness was defined as an opportunistic survey with many data gaps, and data not collected across all major environmental gradients within the study area. High geographic completeness was defined as a well surveyed study area, an appropriate sampling design, few data gaps existing in the study area, and data collected across all major environmental gradients within the study area (Stohlgren and Schnase 2006). It was determined from these classifications that 15% and 19% of the datasets fell into the low and high categories, respectively (Figure 2b).

Spatial scale is a critical factor, as any processes discovered may vary with the spatial resolution and extent at which observations are made. We focused primarily on spatial extent for this study. The non-native species datasets recorded in our survey covered a broad range of spatial extents, with a fairly even distribution among all categories, including smaller than county, county, state, multi-state, national, and global scales (Figure 3; WebTable 1). This is beneficial because invasion patterns are influenced by different factors at different scales. Although smaller scale studies can provide greater detail about the physiological mechanisms that control patterns of invasion, larger scale studies can provide a means to form broad generalizations about landscape-scale patterns (Wiens 1989). For land managers, surveys conducted at multiple spatial scales can account for all these various patterns and will prove most helpful when managing invasions (Stohlgren *et al.* 2002). For agencies tracking the effectiveness of prevention and control efforts, large-scale studies over time assist with assessments of success (Heinz Center in press).

Our understanding of ecological dynamics is also directly related to the temporal scale at which system attributes are measured. A full understanding of the nature of an ecological process may often only be gained after several years or decades of study. Systems that seem highly variable or chaotic over short time scales may reveal more stable dynamics when observed over longer periods, as has been found in many studies (Jackson and Jones 1999; Olabarria and Chapman 2002).

Information pertaining to the time period data were collected and how often data were updated was used to classify datasets by temporal completeness. Low temporal completeness was defined as data collected at one point in time or data collected for 5 or fewer years, with data collection not ongoing; high temporal completeness was defined as data collected continuously for more than 10 years, with data collection completed or ongoing. Although not as equally distributed as spatial scale, datasets did cover a range of temporal scales (Figure 2c). Only 9% of the datasets from our survey had low temporal

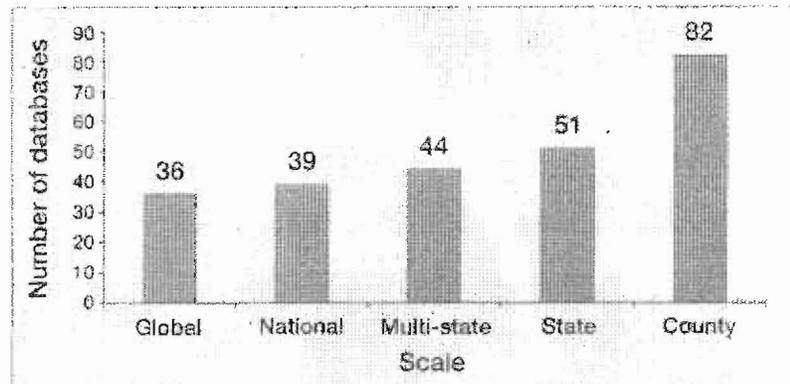


Figure 3. Number of datasets classified by spatial extent.

completeness. It was surprising to find that 38% of datasets have been generating new data for over 10 years, as historically there have been few long-term studies.

### The importance of data quality

No matter how many datasets are discovered, how complete the data, or how many records there are in a dataset, poor-quality, non-native species data are not very useful. Data quality is tightly linked to data analysis because it determines the importance and value of the results that are gathered through mining them. Poor data quality can affect findings, produce inaccuracies in spatial predictive models, and misguide management efforts, costing land managers both time and money. Data quality must therefore be monitored and managed from the very beginning to encompass data gathering, delivery, storage, integration, retrieval, analysis, and publication.

Information obtained from the survey related to the skill level of those who participated in data collection, the presence of a quality assurance/quality control procedure, and the description of that procedure, was used to place each dataset into a data quality category. A majority of data were collected by people with some field and/or taxonomic experience. However, only 55% of the datasets had a quality assurance/quality control procedure. It is necessary to emphasize the importance of establishing a standardized and rigorous quality control procedure for the many non-native species datasets currently in existence. If data sharing is to improve, this will be a necessary step in the immediate future.

### ■ Where do we go from here?

We are now aware of the major datasets that exist, and those that could be accumulated, formatted, and synthesized to address research issues. We are also aware of the inherent limitations of the datasets in terms of geographic, taxonomic, and spatial completeness. "Summing" information from various taxa and at various scales to meet research objectives will require great care since data quality varies, and there are noticeable data gaps in key taxa

and in particular areas of the country. Still, we are hopeful that our survey will allow for a formal investigation of data quality and completeness before a national assessment of non-native species patterns is undertaken.

The reported datasets are sometimes large and in archaic formats. However, of the 252 datasets entered into our survey, 46% were available to the public, 22% were available with conditions on access, 16% will be available in the future, 12% will be available in the future with conditions on access, and only 4% were unavailable. Pooling data into a standardized dataset will be a challenge, but it is not impossible. Many current computer systems and standard software packages (eg SQL-Server, Oracle) can handle this volume of data. Efficient online tools will be required to allow researchers, land managers, and other stakeholders who lack solid online data management systems to easily contribute their data to a centralized, global invasive species data management system. To address this need, our team at the National Institute of Invasive Species Science is developing a Global Organism Detection and Monitoring system ([www.niiss.org](http://www.niiss.org)), which will allow end users to perform powerful cross-dataset spatial analyses to make better use of existing information while guiding surveys, research, and control activities to strategic areas.

#### ■ Acknowledgments

Many people volunteered their time to complete our survey. C Thomas and K Searle assisted in our dataset search. The Heinz Center ([www.heinzctr.org](http://www.heinzctr.org)) provided funding for the project, as part of the State of the Nation's Ecosystems project national report on non-native species, *Tracking non-native species: indicator design and data assessment for the United States*. The Natural Resource Ecology Laboratory at Colorado State University and the USGS Fort Collins Science Center provided logistic support.

#### ■ References

- Bailey RG. 1980. Description of the ecoregions of the United States. Washington, DC: US Department of Agriculture.
- Crosier CS and Stohlgren TJ. 2004. Improving biodiversity knowledge with data set synergy: a case study of nonnative plants in Colorado. *Weed Technol* 18: 1441–44.
- Heinz Center. Tracking non-native species: indicator design and data assessment for the United States. Meyerson LA, Carroll IT, Cremer CF, and Fallon SM (Eds). Washington, DC: The H John Heinz III Center for Science, Economics and the Environment.
- Jackson G and Jones GK. 1999. Spatial and temporal variation in nearshore fish and macroinvertebrate assemblages from a temperate Australian estuary over a decade. *Mar Ecol-Prog Ser* 182: 253–68.
- Mack RN, Simberloff D, Lonsdale WM, et al. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecol Appl* 10: 689–710.
- Olabarria C and Chapman MG. 2002. Inconsistency in short-term temporal variability of microgastropods within and between two different intertidal habitats. *J Exp Mar Biol Ecol* 269: 85–100.
- Pimentel D, Lach L, Zuniga R, and Morrison D. 2000. Environmental and economic costs of nonindigenous species in the United States. *BioScience* 50: 53–65.
- Pimentel D, McNair S, Janecka J, et al. 2001. Economic and environmental threats of alien plant, animal, and microbe invasions. *Agr Ecosyst Environ* 84: 1–20.
- Rejmánek M and Pyšek M. 2002. When is eradication of exotic pest plants a realistic goal? In: Veitch CR and Clout MN (Eds). *Turning the tide: the eradication of invasive species*. Zurich, Switzerland and Cambridge, UK: IUCN SSC Invasive Species Specialist Group.
- Ricciardi A, Stienner WWM, Mack RN, and Simberloff D. 2000. Toward a global information system for invasive species. *BioScience* 50: 239–44.
- Simpson A. 2004. The global invasive species information network: what's in it for you? *BioScience* 54: 613–14.
- Stohlgren TJ, Chong GW, Schell LD, et al. 2002. Assessing vulnerability to invasion by nonnative plant species at multiple spatial scales. *Environ Manage* 29: 566–77.
- Stohlgren TJ and Schnase JL. 2006. Risk analysis for biological hazards: what we need to know about invasive species. *Risk Anal* 26: 163–73.
- Stohlgren TJ, Barnert D, Flather C, et al. 2006. Species richness and patterns of invasion in plants, birds, and fishes in the United States. *Biol Inv* 8: 427–47.
- Vitousek PM, D'Antonio CM, Loope LL, et al. 1997. Introduced species: a significant component of human caused global change. *New Zeal J Ecol* 21: 1–16.
- Wiens JA. 1989. Spatial scaling in ecology. *Funct Ecol* 3: 385–97.
- Wilcove DS, Rothstein D, Dubow J, et al. 1998. Quantifying threats to imperiled species in the United States. *BioScience* 48: 607–15.