DRAFT DOCUMENT FOR COMMENT

THE PENNSYLVANIA BIODIVERSITY CONSERVATION PLAN
SHARING BIODIVERSITY INFORMATION

Definition: Biodiversity informatics is defined as the *sharing* of information in any format pertaining to the taxonomy, ecology, and genetics of organisms and biological communities in Pennsylvania.

Background: Effective conservation of the Pennsylvania's biodiversity resources requires that diverse segments of the state's population have ready access to the best biodiversity data available. Examples of such data include species and habitat distribution information; species morphological features and life-histories; ecological attributes at the population, community, and ecosystem levels; and genetic properties. The information also includes management practices, conservation measures, or strategies for control. Tools for taxonomic identification, data gathering, data archiving, and data analysis are also needed.

Systems for sharing biodiversity data have long been in place, generally involving the generosity of data-holders to make their information (usually in the form of notecards or archived records) available to those who ask. Agency officials and other professionals commonly prepare technical reports containing biodiversity data. As a matter of practice, scientists share their information in the form of peer-reviewed articles in professional journals.

The advent of computer technology has led to a revolution in the way that biodiversity data are collected, stored, disseminated, and analyzed. Data can now be directly recorded in a digital format by dataloggers, laptop computers, and pen-based devices like personal data assistants (PDAs). Data can be formatted into spreadsheets and databases for easy archival and analysis. Digital data can be shared by copying files onto removable drives, through email, or by posting to the Internet. Systems for retrieving data enable people to identify datasets of interest, download them to their computer, and integrate data from different sources. Much of the data can be analyzed by sophisticated tools, such as Geographic Information Systems (GIS) for geospatially-referenced information.

While dissemination of paper-based data is still a common practice, the sharing of digital data is rapidly becoming the standard on a worldwide basis throughout the sciences. During the 1980s, the term "informatics" was coined to describe the practice of sharing digital data. By the early 1990s, geneticists and molecular biologists regularly shared their findings in a digital format, forming a practice termed *bioinformatics*. By the late 1990s, taxonomists and ecologists began exchanging their data in a digital format under the general terms of *biodiversity informatics* and *ecoinformatics*.

Elements of biodiversity informatics have been in place within Pennsylvania for the past ten years. Examples include the digital databases of museums and arboreta, the Pennsylvania Natural Diversity Inventory (PNDI), and the Pennsylvania Spatial Data Access (PASDA). The Pennsylvania Biodiversity Partnership recognized the importance of data-sharing, and created a Biodiversity Informatics Task Force (BITF) to promote effective practices within the state.

As part of its work, the BITF conducted an analysis of biodiversity-related datasets within the state as part of Phase 1 of the Pennsylvania Biodiversity Conservation Plan. To promote effective data-sharing practices as part of that Plan, the BITF has developed a vision for biodiversity data-sharing and identified issues that should be addressed for that vision to be realized.

Vision: Create a system for biodiversity information exchange in Pennsylvania that:

1. Provides types of information needed by various users.
2. Includes taxonomic, ecological, and genetic information.
3. Covers all taxa and habitats.
4. Conveys information that is scientifically valid.
5. Relates to data exchange efforts taking place outside of Pennsylvania.
6. Is readily accessible to a diverse set of users.
7. Evolves as technology evolves.

8. Respects intellectual property, private property, and data sensitivity concerns.
9. Contains information from as many organizations and individuals having information to share as possible.
10. Meets "best data standards."
11. Uses standardized exchange protocols.
12. Educates stakeholders on effectively providing and using biodiversity information.

Issues to be Addressed to Achieve Vision and Recommendations

1. ***Different users have different data needs***. Nearly everybody has, at some time, the need to access biodiversity information collected by others. A park naturalist in Allegheny County may wish to know whether a particular diatom species is new to Pennsylvania. A 10th grade student in Dauphin County may wish to know the identity of a stonefly encountered in a creek, and its propensity for growing in water having a pH of 5.8. A developer in Wayne County may wish to know whether a parcel may contain species listed as endangered or threatened before deciding to purchase it. A post-doctoral associate at Bucknell may wish to access information on the population genetics of tall goldenrod throughout Pennsylvania and convert that information into a GIS layer. An effective system for providing biodiversity information should account for those varying needs and have sufficient flexibility to convey a range of information and analysis tools.

   Recommendations:
   A. Survey stakeholder groups to identify, assess, and prioritize information needs (e.g., taxonomy and nomenclature, geographic range, life history information, management, threats, and inventory gaps)
   B. The information available for sharing should attempt to cover all species and habitats, including common, invasive, and cultivated species as well as less well-known groups, such as invertebrates, protists, fungi, nonvascular plants, and prokaryotes.
   C. The system should attempt to contain information at all levels, including the specimen, species, and community levels.
   D. The system should contain web-based or other electronic tools for taxonomic identification (for example, on-line identification guides).
   E. The system should contain a diverse array of topical or taxon-specific data analysis tools to meet the needs of different categories of users (for example, wetlands, rare species, economically important species, forestry, wildlife management).
   F. The system should be sufficiently easy for the general public to use, yet have sufficient sophistication for a diverse array of technical needs.

2. ***Data are scattered across the state***. Biodiversity data are housed in many different institutions, governmental agencies, and by private individuals. Since no single repository for biodiversity data exists within Pennsylvania, individuals seeking data must somehow locate relevant datasets. Historically, that task has been done through painstaking research and contacting likely dataholders. More recently, datasets can be found by searching *metadata* documents that provide summary information about datasets. Unfortunately, datasets often lack accompanying metadata documentation. For those metadata documents that do exist, systems for making them available to potential data-users need to be improved. To address that need, the Pennsylvania Biodiversity Partnership has established a Pennsylvania Biodiversity Database Inventory (PBDI). That resource needs to be further developed, however, in order for it to serve as an effective index of metadata.

   Recommendations:
   A. Encourage parties holding biodiversity data to create summary information (metadata) for their datasets.
   B. Promote and expand the Pennsylvania Biodiversity Data Inventory (PBDI).
   C. Create a directory of all collections housing specimens collected from Pennsylvania using *Index Herbariorum* as a potential model.

D. Explore options for helping potential users find data and tools via the Internet. Those options can include:
    i. Forum for data-providers and users to meet and collaborate.
    ii. Gateway with links to providers (metadata focused).
    iii. Functionally integrated network (database focused).
    iv. Single overarching web resource.

3. ***Data are in various formats***. Pennsylvania biodiversity data are stored in a variety of forms, ranging from labels on specimen jars, to pages of records stored in filing cabinets, to tabular data published in peer-reviewed articles, to archived data stored on 5 1/4" floppy disks, to datasets stored in an Oracle database posted to the Internet. Moreover, some of the data are geospatially referenced (making them suitable for GIS analysis), whereas other data are not. Parties seeking to integrate data from different sources typically need to convert the data into a single format, a process that can be labor intensive, time-consuming, and costly. Since analyses are increasingly computer-assisted, a first step typically involves converting the data into a digital format or converting an obsolete digital format to one that is more current. Data format standards are developing worldwide and thus efforts to create standardized datasets in Pennsylvania should strive to abide to widely-accepted standards. Regardless of their ultimate form, the concept of "best data practices" should guide efforts to create useful datasets.

   Recommendations:
   A. Identify standard formats for biodiversity data as they are developed, particularly by working with national and international organizations (e.g., National Biological Information Infrastructure (NBII), Global Biodiversity Information Facility (GBIF), NatureServe)
   B. Promote efforts to convert non-digital data into a standardized digital format that can be more readily shared with others.
   C. Promote efforts to convert digital data in obsolete formats into formats that are current.
   D. Encourage all data providers to adopt "best data practices." especially as they relate to data collection and maintenance techniques, and acceptable Quality Assurance / Quality Control (QA/QC) practices.
   E. Encourage organizations to devote sufficient financial and human resources to ensure that data can be converted and maintained in useful standardized formats.
   F. Recognizing that some legacy data in non-digital format will never be converted into a digital format, practices for addressing such data will need to be developed.

4. ***Data are of varying quality and levels of completeness***. Because few standards exist for biodiversity data, information on the state's biological resources vary greatly in quality. Some data are subject to careful collection and scrutiny by professional biologists with decades of experience, while other data are based on quick field surveys by amateurs. Taxonomic references may be based on the best expertise available, or may suffer from faulty identification or outmoded classifications. Species distributions may be current or decades out of date. Data may have precise geospatial referencing, enabling incorporation into GIS frameworks, or may entirely lack any precise information on geographic location. Users may not be in a position to evaluate data quality, especially where the identity of the collectors is not clear. Also, some categories of data, particularly for ecological and genetic dimensions of biodiversity may not be as complete as those for taxonomic dimensions. An ideal data-sharing system will have some way of enabling users to identify data quality, and will strive to be as complete as possible.

   Recommendations:
   A. Encourage the incorporation of quality indicators with data; such indicators may accompany the actual data or may be included with the metadata.
   B. Encourage the incorporation of standardized geospatial referencing to biodiversity data wherever possible.
   C. Encourage the use of best data practices, as they pertain to data collection, data review, and data storage.

D. Examine holdings of biodiversity data to identify categories of data (e.g., ecological and genetic) that need development.

5. ***Data are often not accessible***. Biodiversity data may be in a digital form that can be readily transmitted and used by others. However, privacy or propriety concerns often preclude open sharing. For example, some private property owners only allow biologists to conduct surveys on their property as long as the data are kept confidential. Government officials and others charged with species conservation withhold data on sensitive species that are potentially exploitable (e.g., ginseng). Scientists may wish to keep data confidential until all of the analyses are completed and results published in a peer-reviewed journal. Agencies may restrict access to data because of internal policies and jurisdictional responsibility. An ideal data-sharing system will promote the free flow of information, while respecting intellectual property concerns and maintaining confidentiality in the minority of cases where such confidentiality is necessary.

   Recommendations:
   A. Explore approaches used by other states to deal with the issue of sensitive/confidential data (e.g., providing only general location information).
   B. Encourage scientists to promptly disseminate data.
   C. Examine policies of agencies pertaining to datasharing and encourage practices that facilitate widespread dissemination.

6. ***Insufficient education and manpower infrastructure.*** Since biodiversity data-sharing is relatively new, few individuals have sufficient training to serve as effective data providers or data-users. Training in biodiversity informatics is spotty, at best, often requiring that individuals cobble together courses that cover issues in biodiversity with other courses on data collection, database construction and management, and data acquisition and analysis. Workshops, courses, and entire curricula focusing on biodiversity informatics need to be developed and offered to Pennsylvania residents of all ages. Such curricula can follow the lead of similar efforts underway to teach concepts of molecular bioinformatics to students and adults. Once the educational infrastructure is in place, more individuals will be able to successfully serve as data providers and users. Conceivably, those individuals could be employed by museums and collections, academia, state and local agencies, non-profits, and the private sector. However, the need for personnel adept in biodiversity data-sharing is unknown and deserves to be investigated.

   Recommendations:
   A. Determine manpower needs for individuals having expertise in biodiversity data-sharing within the state for the next 5, 10, and 20 year periods.
   B. Determine funding requirements to reach desired levels of expertise within organizations involved in biodiversity-related activities.
   C. Develop an education infrastructure in the form of workshops, courses, and entire curricula devoted to biodiversity informatics and datasharing within the state. The infrastructure should be broad-based with offerings for primary and secondary students, undergraduates, graduate students, adult professionals, and the general (non-professional) adult population. Such courses will:
      a. Promote and educate students how to effectively locate, obtain, and utilize biodiversity data.
      b. Promote and educate students how to effectively structure, archive, and disseminate biodiversity data.
      c. Educate students on ethical and professional concerns relating to biodiversity data sharing.