Contemporary Challenges in Ambient Data Integration for Biodiversity Informatics*

David Thau[1], Robert A. Morris[2,3], and Sean White[4,5]

[1] Dept. of Computer Science, University of California Davis, CA
[2] University of Massachusetts Boston, MA
[3] Harvard University Herbaria, Cambridge, MA
[4] Dept. of Computer Science, Columbia University, NY
[5] Dept. of Botany, Smithsonian Institution, Washington, D.C.

Abstract. Biodiversity informatics (BDI) information is both highly localized and highly distributed. The temporal and spatial contexts of data collection events are generally of primary importance in BDI studies, and most studies are focused around specific localities. At the same time, data are collected by many groups working independently, but often at the same sites, leading to a distribution of data. BDI data are also distributed over time, due to protracted longitudinal studies, and the continuously evolving meanings of taxonomic names. Ambient data integration provides new opportunities for collecting, sharing, and analyzing BDI data, and the nature of BDI data poses interesting challenges for applications of ADI. This paper surveys recent work on utilization of BDI data in the context of ADI. Topics covered include applying ADI to species identification, data security, annotation and provenance sharing, and coping with multiple competing classification ontologies. We conclude with a summary of requirements for applying ADI to biodiversity informatics.

1 Introduction

Biodiversity informatics (BDI) applies information technology to the acquisition, storage, access, distribution, and analysis of data concerning organisms, populations, and biological taxa and interactions between them. BDI research is carried out in many places, from using sound to identify species in remote biological field stations [1], to identifying trees in urban environments [2], to completing all taxa biological inventories (ATBIs) in national forests [3].

Biodiversity studies increasingly rely on sensor networks and other small devices for data collection and dissemination [4]. The strong spatial and temporal components of the data lend themselves naturally to the application of pervasive

* Work supported by NSF awards IIS-0630033 (David Thau), DBI-0646266 (Robert A. Morris), and IIS-03-25867 (Sean White). The first author would like to thank Shawn Bowers and Bertram Ludäscher for many constructive conversations.

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 59–68, 2009.

computing techniques. This paper discusses elements of biodiversity informatics that can benefit from pervasive computing, shows ways in which the BDI context can inform research in pervasive computing, and discusses challenges in data integration that arise for pervasive computing in the BDI context.

Spatial and temporal contextualization. Biodiversity data are highly sensitive to spatial and temporal context. All aspects of data integration in biodiversity informatics are affected by this. When identifying specimens, the location and time of a study strongly constrain the types of biological taxa that may be found and their appearance. As discussed in Section 4, location and time may impact the integration of metadata about taxa. In addition, the geographic location of studies or species must often be protected, affecting how data are shared.

Challenging environments. Much biodiversity research by necessity takes place far from an internet connection and power sources. This places constraints on how much data are brought into the field and how data are taken from the field. In addition, it constrains the types of analyses that may be done on site, which impacts how data collection occurs. These constraints argue for a division of labor among devices, which in turn drives the need for integrating the data that the disparate devices collect.

Biodiversity studies also occur in environments that attenuate communication signals. For example, signals from GPS satellites are notoriously unreliable in rain forests and often too coarse in urban environments. In addition, certain environments preclude the use of specific frequencies for communication. All of these limitations point to the necessity for creative means of sharing data from sensors and other ambient-enabled devices.

Dynamic teams. Teams engaging in biodiversity studies frequently comprise individuals from different countries, institutions and levels of expertise. In National Geographic Bioblitzes,[1] e.g., thousands of volunteers and scientists gather for two days to complete an inventory. ATBIs of a region often span years and many groups of researchers. In all these cases, different individuals have different levels of knowledge and may bring different resources to the field. This kind of team-based data collection falls into the category of participatory sensing [5] where immediate data analysis and integration can drive additional collection behavior. In order to leverage the information stored on individual devices, data integration techniques must be applied to normalize differences in metadata. In addition, the contemporaneous existence of disparate user devices and on-site sensors requires sophisticated network security protocols. As described in Section 3, specific trust issues arise in biodiversity studies that may be less prevalent in other contexts. Finally, data sharing among independent teams requires a focus on the active management of data provenance and data ownership.

[1] http://www.nationalgeographic.com/field/projects/bioblitz.html

Data Complexity. Biodiversity data have some unusual properties that set them apart from many other types of data. Perhaps the most significant such property is the complexity of naming the fundamental entities of study: observations, specimens, species, and other taxa. The primary system currently used for naming biological taxa has evolved from a standard described by Linnaeus in the middle of the 18th century. Over the subsequent 250 years, as information about biological taxa has accumulated, the names for species and the taxonomies relating them to each other have steadily changed. This change means that a species name used today may mean something different than it meant 5 years ago. One way to mitigate the problems caused by taxonomy evolution is to be clear about *which* version of the taxonomic name is meant when it is applied. In biology this is called the taxon's "name authority," and current BDI data exchange standards (e.g., the Darwin Core[2]) all support (or insist on) inclusion of a name authority. However, as we discuss in Section 4, specifying the name authority is only a first step in supporting data integration.

Another challenge presented by biodiversity data is the amount and location of information that may be relevant to scientists while they perform their research in the field. Biodiversity data are highly distributed. For example, the Global Biodiversity Information Facility (GBIF)[3] indexes over 174 million specimen and other georeferenced species-occurrence records from over 7000 data sets at 285 different data providers. The fastest growing type of such data comprises field observations, often by experienced lay observers ("citizen scientists" and parataxonomists). For example, the Avian Knowledge Network(AKN) e-Bird project[4] provides nearly 23M bird occurrence observations of which 18M have geocoordinates, and AKN collects as many as 70 thousand North American checklists annually. By expanding its definition of what is a biodiversity datum (e.g., to include biodiversity multimedia metadata), GBIF has an ambitious plan to operate indexing and caching services for access to a billion biodiversity data items in a fully distributed fashion. The distribution and amount of biodiversity data that may be useful for data collection in the field, where connectivity may be limited, requires creative data management techniques.

Road Map. The remainder of the paper discusses specific aspects of biodiversity studies, and shows how pervasive computing techniques can be used to better collect and manage the data at these stages, as well as how the BDI context impacts the requirements of data integration in a pervasive computing context. Section 2 focuses on the need for ADI in data collection. Section 3 describes specific trust and provenance issues that must be addressed when integrating BDI data. Section 4 focuses on the metadata involved in integrating BDI information and shows how the context sensitivity of BDI data impacts critical aspects of ADI. We conclude in Section 5 by describing several requirements for integrating BDI data in a pervasive computing context.

[2] `http://www.tdwg.org/activities/darwincore/`
[3] `http://www.gbif.org/`
[4] `http://www.ebird.org/`

2 Identification and Data Collection

Novel field sensors and sensor systems have enabled unique access to information about the environment, bringing useful data to and from the field while greatly expanding the spatial and temporal resolution of data collection [4]. Complementary to this are advances in hand-held mobile devices, which support supervised sensing through human interaction in the data collection process and novel interfaces to vast stores of biodiversity information for real-time analysis and synthesis. These field sensor systems and mobile devices improve existing field research practices and create opportunities for new practices, such as participatory sensing [5] and citizen science [6].

For example, a collaboration amongst Columbia University, University of Maryland, and the Smithsonian Institution has developed a series of mobile electronic field guides that aid in the identification of botanical species, provide access to digitized species information, and support specimen collection in the field [7,2]. Successive iterations of the prototype system, LeafView, run on Tablet PC, Ultra Mobile PC (UMPC) and mobile phone platforms. The system works by first taking a photograph of a leaf specimen. The photo is then analyzed using a custom computer vision algorithm to extract leaf shape [8]. Based on the shape of the photographed leaf, the system provides a visualization of the best matching species so the botanist can make a final visual identification. Contextual information including geolocation, collector, time, and date are saved along with the sample image and associated identification and all of this data is aggregated over the course of a collection. Access to the entire digitized image collection of the Smithsonian Herbarium supports detailed comparison of new samples with existing voucher specimens. The system has been used by Smithsonian botanists on Plummers Island, MD, and at the 2007 Rock Creek Park National Geographic Bioblitz in Washington, D.C. Use of the system has uncovered a variety of challenges related to ambient data integration.

2.1 Management and Integration of Identification Data

Expanding data sets are used both for automated identification and assisted matching. In the current LeafView system, data sets for a region are loaded prior to entering the field. While this works on a small scale, for larger scales and multiple taxa, larger data sets need to be moved in and out of the system, retrieved and cached, based on specific regions and tasks. For example, current data sets for identification include:

- Flora of Plummers Island. 5,013 leaves of 157 species. Provides complete coverage of all vascular plant species of Plummers Island, MD, an island in the Potomac River near Washington, DC, which has long been studied by botanists.
- Woody Plants of Baltimore-Washington, DC. 7,481 leaves of 245 species. Provides complete coverage of all native woody plants (trees and shrubs) of the Baltimore-Washington, DC area.
- Trees of Central Park. 4,320 leaves of 144 species.

The computed feature distances necessary for automated identification are represented in an NxN matrix where N is the number of individual leaves in the data set. For the Woody Plants of Baltimore-Washington, D.C., this requires 400 MB of storage. Even with improvements to the algorithm, the feature sets for matching data promise to be large and grow with the number of species, requiring compartmentalization and filtering. In addition to these data sets, access to digitized images is necessary to visually match sample specimens with voucher specimens. The US National Herbarium Type Specimen Collection alone incorporates over 90,000 images, covering more than one quarter of all known plant species. Each specimen has been digitally photographed under controlled lighting to produce an 18 megapixel image. A decimated version of the voucher specimens for Woody Plants of Baltimore-Washington, DC (300K GIF images instead of 18 MB TIFF) requires 295 MB but a full resolution version of the data set would provide more detail and would require much more space. These data management issues are compounded when the data for an individual species is extended to alternative representations. For example, recent research in augmented reality uses situated visualization to superimpose relevant species information directly onto the physical scene [9].

In the presence of a robust network, processing and data necessary for identification and matching can reside on server systems. However, remote areas without connectivity require prediction about necessary data sets for identification so analysis and data sets can be moved to the device. Task and location context can help filter the search space and thus the data requirements. However, filtering and inaccuracies in matching can complicate use of the system. When a new specimen is not found through automated identification or keys, is it because the data is simply not in the current data set, is the identification tool failing, or is this a new species?

2.2 Collaborative Identification and Shared Collections

With similar issues to data management, collaborative identification requires sharing of collected specimen data and annotations in real-time. ADI issues arise in several situations. First, in the case of censuses, a shared collection list may be used. Synchronization of the collection list across multiple teams of collectors helps focus resources on finding species that have yet to be collected. Second, multiple sensing devices may be aggregated under a single processing unit. For example, in one collection, several cameras were connected to a single LeafView system, each able to send photographs for identification across a local ad-hoc wireless network. Third, the collected data itself may be shared to aid in identification. For example, collector A may be able to identify a particular species and share their history of collection with other team members. If the same species is observed by collector B, they can use the shared history of the collection to help identify the species. Finally, the data needs to be shared and used beyond any given field activity. In the current, non-networked device, data is simply exported at the end of a field study. In a networked version, collections should be opportunistically pushed to a proxy, mediator, or server.

2.3 Observation Driven Data Collection

Data collection, mediated through human agency, can also be driven by immediate observations in the field. For example, reviewing a map of the locations of collected specimens in a given geographic region may reveal areas that have not yet been inspected. By creating shared models of data that reflect spatial and temporal histories of observations, individuals and groups iteratively navigate locations for collection of species. Such iteration requires real time data curation incorporating explicit and implicit association of metadata.

3 Data Sharing

There are benefits to sharing data between sensors and other ambient-enabled devices throughout the data collection process. Before data are collected, devices must have access to information that will assist in the identification of species. As the data are collected, the devices can inform each other about what has been collected so far. In addition, sensors and other data sources at the study location can supply data to inform and drive collection events. While BDI shares many features with other participatory sensing scenarios, there are a few differentiating aspects. Two of these are particular details about what data may be shared with whom, and how an ambient data integrating system should deal with evolving information about the objects being studied.

3.1 Access Control Issues

Security and access control are common problems in pervasive computing scenarios [10,11]. BDI has some additional security requirements. The most widely mentioned of these is the protection of sensitive geographic information, for example to defend the exact location of organisms of rare or endangered species, or to protect landowners who have given permission to use their land for biodiversity surveys but do not want uninvited guests wandering around their property looking for rare organisms. Unfortunately, professional practices can complicate attempts to protect such data. For example, rigorous collection or observation protocols require that collection and observation events have unique identifiers. A standard practice is to assign sequential integers as part of an otherwise constant event identifier. This causes problems for database systems that try to suppress geographical information for sensitive specimens. For example, imagine three records, r_1, r_2, r_3 collected in the same location, the second of which is considered sensitive. A "smart" database system that suppresses information about r_2 but returns the coordinates for r_1 and r_3 would give away r_2s location. A number of strategies are in use for protecting the geocoordinates of occurrences of endangered species while still making full resolution available to authorized users for such things as predictive range modeling. Among them are one or another form of generalizing the geocoordinates, wherein the location is given either at a low geo-resolution (e.g., to a 10 km square on a fixed grid) or a named geopolitical entity, such as a town, county, or province.

One controversial reason sometimes given for biodiversity access control is that some class of users may make use of the data in a way that is inappropriate in the eyes of the data holder. See Chapman and Grafton [12] for a more extensive review. Morris et al. [13] provided a fully distributed XACML-based access control system whose control policies can be defined or enforced by the original data provider or a host to which they delegate those services, and which meets many of the needs expressed by networks of distributed taxon occurrence data. Any of the access control services can be centralized and slowly migrated to network nodes as and when their operators acquire sufficient IT skills and resources to support such services. The filters are defined by XPath expressions on the data interchange schema expressed in XML-Schema.

BDI access control per se does not give rise to different issues for ADI than for computing regimes that are not context aware. It is, however, an instance of challenges that arise in attempting to reason in dynamic contextual computing environments, whether that reasoning is statistical or logical; namely it may amplify imperfect context information. Henrickson and Indulska identify four types of imperfect context information: unknown, ambiguous, imprecise, and erroneous [14]. The first three of these correspond to examples of georeference access control mentioned above. The fourth, in the form of deception, is sometimes proposed for access control, but is notoriously subject to data mining techniques designed to find logical outliers. For example, a report of an arboreal animal swimming 100 km. off the coast of Portugal should usually be hypothesized to be erroneous.

3.2 Distributed Annotations for Quality Control

As in any scientific endeavor, the quality of the data acquired, stored and shared is of paramount importance. In general, data quality can be measured by comparison with similar data already collected. For example, Calder et al. describe a rule-based reasoning system targeted at sensor network data, that allows scientists to put forth hypotheses about possible explanations of their observations and have a reasoning engine select which of them are consistent with the currently accepted value of observation data [15]. Unfortunately, a substantial amount of primary biodiversity data that might drive reasoning about field or laboratory observations remains undigitized or is only partly digitized (e.g., to the level of scanned images with no OCR). There are estimates that the world's natural history museums hold 3 billion specimens, of which fewer than 200 million have any kind of digital record. The Biological Heritage Library[5] has scanned over 14 million pages of legacy taxonomic literature, much of which provides original taxonomic descriptions of newly discovered species over the last three centuries. Museum (and individual collector) specimen records and original literature represent part of the "ground truth" of species identification, but even after imaging, many of these documents are being incrementally made digitally useful by databasing, by rough machine-learning based automated markup, or by semi-automatic markup guided by humans[6]. Most of these strategies result in an ever

[5] `http://www.biodiversitylibrary.org/`

[6] e.g., `http://plazi.org/`

moving target of increasingly accurate and increasingly fine-grained knowledge content. This presents challenges and opportunities for individual or coupled ambient computing platforms to reason over the data and knowledge to which they have access for the purpose of assessing the quality of data they may hold, and the quality of data they may report. This post hoc analysis and digitization of historical biodiversity data adds special requirements to any system that attempts to collect, record and share new biodiversity data. First, provision should be made for data records to be annotated with record-level quality control metadata (or other annotations of interest). Second it must be possible for the annotations to circulate in communities of interest, along with notification mechanisms that attempt to provide the annotations and commentary upon them to human or software agents that express an interest. A team at Harvard and UMASS-Boston has designed and is implementing a "P2P Filtered Push (FP) Annotation Exchange" for such a purpose [16]. Its currently implemented prototype is dedicated to data of a special case, namely the digital form of accumulated annotations on related botanical specimens. (Conventionally, botanists collect multiple specimens from the same organism and circulate copies to multiple institutions for, usually, independent curation.) FP is built on the Apache Hadoop Map-Reduce framework together with the Apache ActiveMQ Java Messaging Service. FP is being extended to allow arbitrary workflows anywhere in the local community or the Cloud to generate and announce QC (or other) annotations.

4 Ontology-Based Data Integration

The importance of ontologies in pervasive computing is widely recognized [17]. When investigators from disparate organizations, nations, and levels of expertise collaborate in a BDI study, chances are they will bring with them a multitude of heterogeneous metadata standards. As we have seen, data collection and data sharing can be influenced by events that occur during and after a data collecting event. Before ambient-enabled devices can integrate their data, they must mitigate the differences in their metadata.

In BDI, metadata differences can appear in the standards used to describe measurements [18], as well as to describe the things being measured. One particularly salient metadata issue in BDI revolves around the difficulties in naming biological entities. As mentioned in the introduction, multiple taxonomies may be used to classify a given set of biological taxa. Two groups using different field guides may use different names to identify the same specimen. To minimize the difficulties this inevitably creates when trying to integrate biodiversity data, experts create mappings between well-known taxonomies [19,20]. These mappings can be reasoned over to discover inconsistencies and new mappings [21], and may be used to integrate data [22]. A great deal of uncertainty may occur when integrating data sets under multiple taxonomies. Often, this uncertainty can best be resolved at the time of data collection. A challenge for ambient data integration is to integrate data collected by heterogeneous devices rapidly enough to discover when the results of the integration are uncertain, and to notify the data collectors while they are still in the field so that the uncertainties can be resolved.

An interesting extension of the work on mapping biological taxonomies that has not been addressed is the context specificity of the mappings. For example, in one spatial context, such as North America, two taxonomic names A (mentioned in one taxonomy) and B (mentioned in a different taxonomy) may refer to identical biological entities, while in another spatial context, such as South America, one of the taxonomic names may refer to a subset of the second taxonomic name. This might arise if specimens of taxon B that are not also in taxon A have been identified in South America, but in North America all specimens of taxon B are also specimens of taxon A. The discovery of a specimen of B that is not also a specimen of taxon A in North America would be especially interesting, either because it is new (possibly publishable) information about the taxa involved, or because it is a misidentification. The interestingness of the identification of a B that is not an A arises from the taxonomic mapping, which itself may only come into play when ambient-enabled devices are expected to integrate their data in the field. This again points to a challenge for ambient data integration: it needs to be sensitive to the context (e.g., geographic context) under which the integration occurs.

5 Conclusion

Biodiversity informatics presents several interesting challenges for data integration in ambient computing. First, connectivity in the field is reduced, creating an emphasis on device provisioning of data and clever means for sharing data between devices. Second, the data themselves are complex. Although most ADI applications need to perform some semantic mediation for mismatched metadata, the 250 year history of evolving taxon names presents a particularly extreme situation. Third, data integration occurring in real time can have immediate impact on collecting events. This, along with the attenuated connectivity, argues for intelligent ambient-enabled devices that can analyze data as they are collected and distribute information from these analyses. Finally, all aspects of a biodiversity informatics study are affected by the spatial and temporal context of the study. This includes the identification of species, the protection of sensitive data, and the application of semantic metadata mediation. In the future, as sensors and devices brought into the field are increasingly capable (e.g., identification via on site DNA sequencing), this sensitivity to context will continue to influence analyses and data dissemination.

References

1. Gage, S.H.: Observing the acoustic landscape. In: Estrin, D., Michener, W., Bonito, G. (eds.) Environmental Cyberinfrastructure Needs for Distributed Sensor Networks, August 2003, p. 64 (2003)
2. Belhumeur, P.N., Chen, D., Feiner, S., Jacobs, D.W., Kress, W.J., Ling, H., Lopez, I., Ramamoorthi, R., Sheorey, S., White, S., Zhang, L.: Searching the world's herbaria: A system for visual identification of plant species. In: Forsyth, D.A., Torr, P.H.S., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 116–129. Springer, Heidelberg (2008)
3. Sharkey, M.J.: The all taxa biological inventory of the great smoky mountains national park. The Florida Entomologist 84(4), 556–564 (2001)

4. Porter, J.H., Nagy, E., Kratz, T.K., Hanson, P., Collins, S.L., Arzberger, P.: New eyes on the world: Advanced sensors for ecology. BioScience 59(5), 385–397 (2009)
5. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S.: Srivastava: Participatory sensing. In: WSW 2006: Mobile Device Centric Sensor Networks and Applications (2006)
6. Caruana, R., Elhawary, M., Munson, A., Riedewald, M., Sorokina, D., Fink, D., Hochachka, W.M., Kelling, S.: Mining citizen science data to predict revalence of wild bird species. In: KDD 2006, pp. 909–915. ACM, New York (2006)
7. White, S., Marino, D., Feiner, S.: Designing a mobile user interface for automated species identification. In: Rosson, M.B., Gilmore, D.J. (eds.) CHI, pp. 291–294. ACM, New York (2007)
8. Ling, H., Jacobs, D.W.: Using the inner-distance for classification of articulated shapes. In: CVPR (2), pp. 719–726. IEEE Computer Society, Los Alamitos (2005)
9. White, S., Feiner, S., Kopylec, J.: Virtual vouchers: Prototyping a mobile augmented reality user interface for botanical species identification. In: Proc. 3DUI 2006 (IEEE Symp. on 3D User Interfaces), pp. 119–126 (2006)
10. Walters, J.P., Liang, Z., Shi, W., Chaudhary, V.: Wireless sensor network security: A survey. In: Security in distributed, grid, mobile, and pervasive computing, p. 849. CRC Press, Boca Raton (2007)
11. Cuevas, A., Khoury, P.E., Gomez, L., Laube, A.: Security patterns for capturing encryption-based access control to sensor data. In: SECURWARE 2008, pp. 62–67 (2008)
12. Chapman, A.D., Grafton, O.: Guide to Best Practices For Generalizing Sensitive Species Occurrence, version 1. Global Biodiversity Information Facility (2008)
13. Dong, H., Wang, Z., Morris, R., Sellers, D.: Schema-driven security filter generation for distributed data integration. In: Hot Topics in Web Systems and Technologies, pp. 1–6 (2006)
14. Henricksen, K., Indulska, J.: Modelling and using imperfect context information. In: PERCOMW 2004, Washington, DC, USA, pp. 33–37. IEEE Computer Society, Los Alamitos (2004)
15. Calder, M., Morris, R.A., Peri, F.: Machine reasoning about anomalous sensor data (2009) (submitted for publication)
16. Wang, Z., Dong, H., Kelly, M., Macklin, J.A., Morris, P.J., Morris, R.A.: Filteredpush: A map-reduce platform for collaborative taxonomic data management. In: CSIE 2009. IEEE Computer Society, Los Alamitos (2009)
17. Ye, J., Coyle, L., Dobson, S., Nixon, P.: Ontology-based models in pervasive computing systems. Knowledge Engineering Review 22(4), 315–347 (2007)
18. Bowers, S., Madin, J.S., Schildhauer, M.P.: A conceptual modeling framework for expressing observational data semantics. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) ER 2008. LNCS, vol. 5231, pp. 41–54. Springer, Heidelberg (2008)
19. Koperski, M., Sauer, M., Braun, W., Gradstein, S.: Referenzliste der Moose Deutschlands, vol. 34. Schriftenreihe Vegetationsk (2000)
20. Peet, R.K.: Taxonomic concept mappings for 9 taxonomies of the genus ranunculus published from 1948 to 2004. Unpublished dataset (June 2005)
21. Thau, D., Ludascher, B.: Reasoning about taxonomies in first-order logic. Ecological Informatics 2(3), 195–209 (2007)
22. Thau, D., Bowers, S., Ludaescher, B.: Merging sets of taxonomically organized data using concept mappings under uncertainty. In: Meersman, R., Dillon, T., Herrero, P. (eds.) OTM 2009, Part II. LNCS, vol. 5871, pp. 1103–1120. Springer, Heidelberg (2009)