Data Heritage in the Biologist's Field Notebook

Ron B. Yeh, Scott Klemmer
Stanford University HCI Group
Computer Science Department
Stanford, CA 94305-9035
{ronyeh, srk}@cs.stanford.edu

ABSTRACT

In this paper, we describe our progress toward building a system for field biologists to collect, organize, and share their information. We describe interactive prototypes that support a concept we call DATA HERITAGE, and how this will help biologists in the field and the lab. With data heritage, the system tracks the transformation and transportation of data between software, devices, and users. The heritage will be authored implicitly by the actions of biologists using our ButterflyNet system.

Author Keywords
Biology, augmented field notebook, Anoto digital pen, capture and access, interactive paper, sensor networks.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI).

INTRODUCTION

Biologists who work in the field face an increasingly difficult task of managing and searching through vast amounts of information. While today's technology is great at capturing data, it is not well suited to organize and search through the unstructured data.

Field Study

To determine the needs of biologists, we conducted in-depth interviews with biologists from Stanford University, the Jasper Ridge Biological Preserve, and the California Academy of Sciences. We also spent nine days in the tropical rainforest (in Los Tuxtlas, Mexico), living with nine field biology students, two TAs, and a professor. In Los Tuxtlas, we participated in four field experiments, and observed field research practices. We have also deployed digital pens to biologists, and are now analyzing the digital note pages to help inspire designs for our ButterflyNet system.

Results

From our need finding, we distilled a set of properties to direct our design. First, we realized that the paper notebook was the central organizing artifact of field biology research [4]. We found that for the notebooks we analyzed, much of the content was in the form of quantitative, tabular data. Other than numbers, notebooks also contained textual descriptions, images, references to computer files, and pasted-in procedures and visualizations.

Through further investigation, in the Jasper Ridge Docent Training class, and the Los Tuxtlas Field Research class, we found that field biologists are very mobile, as they may hike long distances to get to a remote site. Moreover, out in the field, biologists do not have free hands to operate extra equipment. Back at the lab, they have more time and can make use of tabletops to supplement their two hands.

COMBINING ADVANTAGES OF PAPER AND DIGITAL

Biologists use paper notebooks for the multitude of reasons that paper is better than its digital counterpart. Paper is portable, robust, and easy to manipulate. It is readable outdoors, has high resolution, and infinite battery life. However, paper notebooks lack desirable qualities that digital media afford, such as text search, flexible data organization, the ability to store vast amounts of data, and ease of sharing.

In our system, we combine the benefits by leveraging the Anoto digital pen/notebook, a palm-sized mobile device (OQO), and the Tablet PC. In our work, we have deployed the digital pens to the Los Tuxtlas students, and several Stanford biologists. We use the results from these deployments to direct our prototypes—pages displayed in our prototypes are all from real notebooks.

DATA HERITAGE

Our concept of DATA HERITAGE enables tracking of:

- Data transformations (*e.g.*, notes → spreadsheet)
- Data evidence from multiple streams (*e.g.*, notes + audio + photos → hypothesis)

- Data sharing between devices and users
- Data links between the physical and digital world

Transformations include transcribing from paper to spreadsheet, analyzing spreadsheet data, generating visualizations from the data, and creating publications out of the statistics and visualizations. Tracking where data goes throughout this process will enable a biologist to backtrack if necessary, to make observations or changes.

Evidence from multiple streams of data (such as audio, photos, notes, and sensors) allows biologists to obtain a more complete picture of their observations. Since these streams are correlated in time and space, they can help corroborate observations or support hypotheses that biologists make in the field.

Data heritage in sharing will allow biologists to keep track of where data travels, whether it is to another device the biologist uses, or to a colleague. This will allow a biologist to find the original collector of any piece of data, which in turn enables the biologist to better understand the limitations of the data. This also gives credit to the data collector.

Giving credit for data collection has a couple of benefits. First, it rewards better data collectors, by giving credit to those who contribute more, better-documented, well-structured, and usable data. Second, it will encourage sharing, because if everyone contributes a small bit of their work, they gain a large database of useful data. Plus, our physical + digital system facilitates sharing, which is easy with digitized notes.

Links between the physical and digital world allow the biologist to track the fact that many pieces of data are derived from physical samples. When these physical samples contribute to numerical data on paper, the system will track which sample contributed to which row (or rows) of data.

We hypothesize that these components of data heritage will have several benefits. They will encourage sharing between biologists. They will allow biologists to backtrack and verify data or observations. They will help biologists keep track of where samples and data are.

MOTIVATION FOR DATA HERITAGE

The idea of supporting data heritage was inspired by our interactions with biologists. A current practice that suggests that data heritage will be useful is the practice of writing file names in notebooks. With today's sensors, it is easy to generate lots of data. However, this data cannot all fit in the paper notebook, so biologists simply write the filename that corresponds to where the data is stored. This linking between the physical and the digital world will be part of our data heritage concept.

Second, biologists frequently need to tag biological samples with unique numbers, and record these numbers in their notebook. This number serves as the only link between the physical sample and the data generated from that sample. Our data heritage system will facilitate this within-experiment tracking.

Third, when biologists pull data from their colleague or from a publication, they will write down the source of the data (e.g., taken from Karen Whitmore's procedure, or D. Janzen '99). Data heritage will track the sharing between colleagues.

Sharing Data

In our interviews, we determined that there were several reasons why field biologists currently hesitate to share data (*e.g.*, notebook, photos, environmental sensor data). First, data from notebooks is not easy to share. With a paper notebook, the only way to share a page of notes is to make a copy of the page (through photocopy, scanning, or digital photography). Second, current tools for sharing digital data (*i.e.*, the Web) do not support giving credit to the biologist(s) who collected the data. Third, biologists express concern that even if it were easy to share data, other biologists who use the data may not understand the limitations of the data, or the context of the collection process. Finally, some data is proprietary. In the industry, biology or chemistry labs must protect their intellectual property with patents. In academia, intellectual property leads to publications, which leads to jobs and tenure. While we can solve the first three issues, we must provide flexibility, to allow a user to disable sharing for proprietary data.

INTERACTIVE PROTOTYPE

We created a Flash prototype[1] to demonstrate several scenarios where data heritage will help the biologist. The interface allows biologists to visualize the heritage of the data they are using. Our proposed system will automatically update the data heritage information when necessary. These updates will occur implicitly as a result of user actions. For example, when Roger borrows a photo from Karen's notebook, the system will track that the data came from Karen.

Scenario One: Visualization → Notes → Photos/Audio

While editing a visualization intended for publication, Jenny observes that the concentration of Acorn Woodpeckers near the dam is low for the month of July (see Fig. 1). She reveals the data heritage to backtrack through the excel spreadsheet and field notes. While reviewing the notes, she notices associated photos and a description of a fire, which had ruined an area of land near the Acorn Woodpecker's habitat. She plays the audio

[1] http://butterflynet.stanford.edu/~ronyeh/dataheritage/

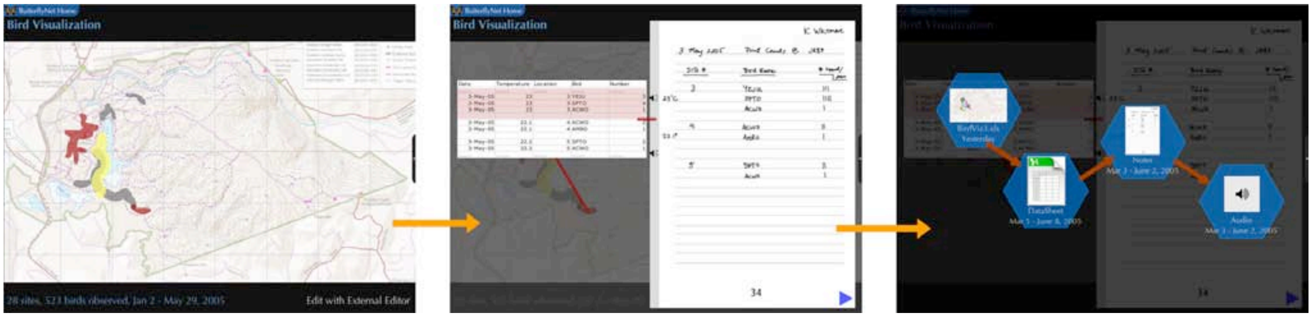


Figure 1. Left) Jenny's Bird Visualization. Middle) Data heritage revealed. Right) A graphical overview of the data heritage.

associated with those field notes, and hears the presence of a competitive species, the Dark-Eyed Junco. These observations will contribute to the discussion in her paper.

Similarly, Karen is tracking the spread of Argentine ants throughout Jasper Ridge. She has created a visualization describing the location of the native ant populations versus the invasive species (Argentine). To find handwritten notes and photos of Site B, near the Fire Road, she selects the area of the map she is interested in, and reveals the data heritage. The interface reveals spreadsheets that contributed to this visualization, and also points to ancestor field notes. She navigates to the page of interest, and finds several photos she took while observing that site on that day.

Scenario Two: Between Colleagues

Roger hypothesizes that flight patterns of the Chalcedon Checkerspot butterfly (*Euphydryas chalcedona*) changes with the humidity in the lakeside area. He downloads the humidity data for the time he is interested in, to see if there are any interesting correlations. He reveals the data heritage of a section of data that looks different from the rest. He sees that Karen collected that data, but entered a note in her notebook to remember to check if the sensors were broken. Roger can now contact Karen to ask about this. As heritage is revealed to both parties, Karen's system will also notify her that Roger has used her humidity data in his work.

Scenario Three: Mobile Photo Browsing → Notes

While Jenny collects samples in the pygmy forest, she notices increased herbivory rates in one of the trees (see Fig. 2). She browses her mobile devices for a photo of the same tree two months ago, and notices in the associated field notes that she had observed fewer ant colonies in the vicinity. She later decides to conduct a formal test to see whether the ant colonies defend the trees they live under from other herbivores.

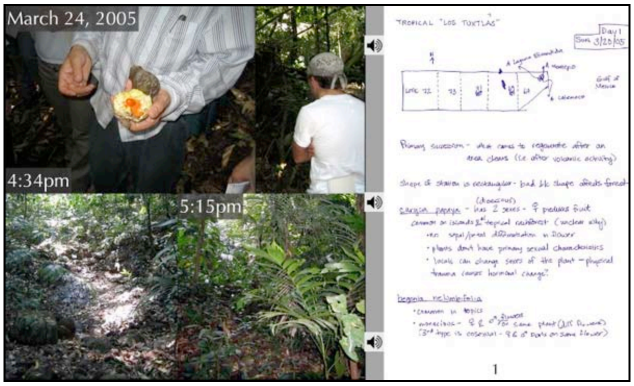


Figure 2. A mobile interface for browsing photos, notes, and audio, targeted for the OQO's screen resolution.

NOTE ON PROPOSED SYSTEM DESIGN

DATA HERITAGE can be stored as an XML-formatted metadata field attached to each file. Each application would then add to or remove from the data heritage in a standard way. For prototyping purposes, we will store the metadata in a separate file for simplicity. For any file X, the associated DATA HERITAGE metadata file is X.dhxml. Our system will handle X and X.dhxml as one logical file.

RELATED WORK

The Elephant File System provided users a way to reference previous versions of files on a file system that would automatically keep old versions of files [2]. While their implementation was robust, the user interface lacked the ability to see how many versions of a file existed, or which user or device created or contributed to the data file. Our system will provide a direct manipulation interface to navigate the data heritage, to select from available file versions, snapshots, or related documents.

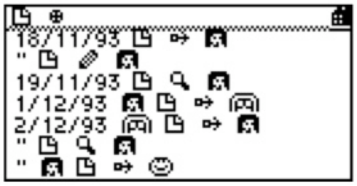
Data heritage is also maintained in version-control systems, such as CVS for source code, or the system used by Wikipedia. For example, one can browse Wikipedia's entry for "Windows XP," and see a list of the users who have contributed to the article. The oldest entry (of more than 850) for "Windows XP" was created on 13 Nov 2001 at 20:53, by user Dmerrill. While it is probably true that few people will ever look at this metadata, it is useful to be able to know where data comes from, in the case that (as we have learned happens in biology) people want to use data (or repeat experiments) 30 years after it was first collected (or reported).

Today, people use multiple web search results as a form of data evidence. For example, one can use multiple websites, or text, images, video, and audio to verify a finding. Our system will extend this metaphor to allow users to verify observations via different streams of data, such as audio, photos, and handwritten notes.

With the Flamenco system, users can search for items using multiple categories of metadata [3]. Our system will extend Flamenco's faceted metadata metaphor by including (as a metadata category) the identity of the users and devices that generate and operate on the shared data. With this extra type of metadata, users can search for data by user or device name.

The LabScape system introduced a graphical abstraction to represent laboratory procedures [LabScape 2002]. In LabScape, the cell biologist authors the experiment's flow graph (with the help of the ubiquitous computing infrastructure). Our system will not require the user to author the data heritage. In addition, our system will include sharing between colleagues, and will be used in both the field and the lab.

Forget-Me-Not [1] employs a visual interface for mobile devices that allows users to visualize the "biography" of a document. One can see on which dates the document was edited, or passed from one person to another. Our system's data heritage concept extends the document "biography" to handle data transformations. When a user pastes a data table from an Excel document A into a Word document B, and adds a textual interpretation of that data, we consider B to be a child of A in the data heritage.



Biography of a document.

Figure 3. Forget-me-not employs a visual representation of a document's biography.

Adobe Photoshop includes an undo feature that enables a user to take explicit snapshots of the current state of the document. However, snapshots are not saved with the file, so the within-file data heritage is lost once a user quits Photoshop. Our system can improve on this design by retaining snapshots, and allowing users to visualize the state and relation between various snapshots. To enable sharing, while still retaining snapshots, we can share the most recent snapshot, but retain knowledge that snapshots exist on another user's system, to prevent sharing of arbitrarily large files.

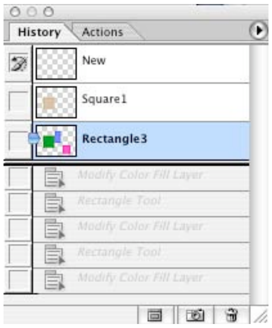


Figure 4. Adobe Photoshop's Snapshot system.

FUTURE WORK

While we have shown our current Flash prototypes to three Stanford biologists, and a class of Jasper Ridge docents, we have not yet conducted any user study to evaluate the usability/utility of our designs for data heritage. We plan to conduct studies once our prototypes handle more complete scenarios.

The concept of data heritage has broader impact in domains where people stand to benefit from knowing where data comes from. In product design, one can leverage data heritage to give credit to designers. In health, one can backtrack to verify diagnoses or prescriptions, and see if and why they were wrong.

ACKNOWLEDGMENTS

We thank the biologists who participated in our studies, and Andreas Paepcke, Terry Winograd, and Hector Garcia-Molina for providing support and insightful comments. This research is supported by NSF Grant IIS-0430448 and a technology donation from Nokia.

REFERENCES

1. Lamming, M. and Flynn, M. Forget-me-not: intimate computing in support of human memory. *FRIEND21: International Symposium on Next Generation Human Interface*, Meguro Gajoen, Japan, 1994.
2. Santry, D.S., Feeley, M.J., Hutchinson, N.C., Veitch, A.C., Carton, R.W. and Ofir, J. *Deciding when to forget in the Elephant file system*. ACM Press, Charleston, South Carolina, United States, 1999.
3. Yee, P., Swearingen, K., Li, K. and Hearst, M., Faceted Metadata for Image Search and Browsing. in *SIGCHI*, (2003).
4. Yeh, R. and Klemmer, S. Field Notes on Field Notes: Informing Technology Support for Biologists, Stanford University Computer Science Department, 2004.