Poster Proposal for Open Repositories 2009

High-Throughput Workflow for Computer-Assisted Human Parsing of Biological Specimen Label Data

Aliasgar Amin, Jane Huang, Zainab Arsiwala, Jason Best, William E. Moen

Introduction

Hundreds of thousands of specimens in herbaria and natural history museums worldwide are potential candidates for digitization, making them more accessible to researchers. An herbarium contains collections of preserved plant specimens created for scientific use. Herbarium specimens are ideal natural history objects for digitization, as the plants are pressed flat and dried, and mounted on individual sheets of paper, creating a nearly two-dimensional object. Building digital repositories of herbarium specimens can increase use and exposure of the collections while simultaneously reducing physical handling. As important as the digitized specimens are, the data contained on the associated specimen labels provide critical information about each specimen (e.g., scientific name, geographic location of specimen, etc.). The volume and heterogeneity of these printed label data present challenges in transforming them into meaningful digital form to support research. The Apiary Project[1] is addressing these challenges by exploring and developing transformation processes in a systematic workflow that yields high-quality machine-processable label data in a cost- and time-efficient manner. The University of North Texas's Texas Center for Digital Knowledge (TxCDK) and the Botanical Research Institute of Texas (BRIT), with funding from an Institute of Museum and Library Services National Leadership Grant, are conducting fundamental research with the goal of identifying how human intelligence can be combined with machine processes for effective and efficient transformation of specimen label information. The results of this research will yield a new workflow model for effective and efficient label data transformation, correction, and enhancement.

Research Methodology

The Apiary Project is in its formative stages; it seeks to investigate what workflow provides for a combination of machine-assisted and human-assisted procedures to most effectively and efficiently convert textual data on specimen labels into machine-processable parsed data which can then be ingested into a database and associate with the digitized specimen. The project will develop and test innovative workflows for the transformation of specimen label data. It will examine instances where Optical Character Recognition (OCR) and Natural Handwriting Recognition (NHR) do not provide satisfactory results. The workflow will start with image acquisition of the herbarium specimen followed by the layout analysis process where all regions of interest (ROI) of the label data or other data on the specimen will be segregated for transformation. Processes in the workflow will provide for human entry of data in the ROI as well as machine tools to assist in the process (e.g., magnifying the image, adjusting contrast for better legibility, spell-checking etc.). Human and automated semantic parsing will insert data into established, standard metadata elements (e.g., genus, species, collector name, locality etc.) and enhance the data through web services (e.g., geocoding services). Quality assurance processes will compare the results of the human process and machine process to benchmark data from a test dataset.

Using a Repository to Support Transformation Workflow

The workflow processes will be supported by a Fedora repository for storing digitally-captured label data, the in-process data as it is transformed by human and machine processes, and the finished, quality-controlled metadata. The final structured metadata records can be exported into information systems such as Atrium[2] or Specify[3]. We intend to use the Islandora module which enables the Drupal content management system to be integrated with the Fedora repository in a way that Drupal acts as a front-end to the Fedora repository thereby allowing the users to access and manipulate the label data as it is transformed through the workflows. The Fedora framework is scalable and accessible through a web service (XML Storage, SOAP/ REST) which enables average users to view and access the Fedora repository.

[1] Apiary Project: http://www.apiaryproject.org

[2] Atrium: http://www.atrium-biodiversity.org/

[3] Specify: http://www.specifysoftware.org/Specify