An Event Model for Herbarium Specimen Data in XML

William E. Moen
Director of the Texas Center for Digital Knowledge, University of North Texas

Amanda K. Neill
Director of the Herbarium

Jason Best
Director of Biodiversity Informatics, Botanical Research Institute of Texas

The Apiary Project is a collaboration of the Texas Center for Digital Knowledge at the University of North Texas and the Botanical Research Institute of Texas, funded by a National Leadership Grant from the Institute of Museum and Library Services. In this project, we are building a framework and web-based workflow for the extraction and parsing of herbarium specimen data. The workflow will support the transformation of written or printed specimen data into a high-quality machine-processable XML format. This poster describes an event model that informed the development of the Apiary XML Application Schema.

The Apiary Project is not developing an overall reference or data model for biological data. It has a more narrow focus, namely modeling the data on a herbarium specimen sheet to inform an extension to the Generic Darwin Core (DwC) XML Schema. The extension will need to accommodate all events that change specimen object metadata elements over time, including new identifications, ownership changes, and other non-taxonomic annotations. DwC is the foundational vocabulary for the Apiary Project, and we have defined additional Apiary-specific vocabulary terms to address the requirements of herbarium specimen sheets. The latter is now represented in a Generic Apiary XML Schema. Modeling the specimen data, however, was necessary to inform the development of the Apiary Application XML Schema, which imports the Generic DwC and Generic Apiary schemas and defines the structure of the Apiary XML record. Schemas are available at: http://www.apiaryproject.org/documents.

A herbarium specimen object and the data on the specimen sheet related to the specimen object result from a sequence of activities. These activities occur in some temporal order as follows:

- Collector collects an individual organism in the wild.
- The collector notes information about the collected organism (and information related to collecting the organism) and at some point this information is attached to a specimen sheet.
- Once the sample of the organism is in the herbarium, it takes the form of a Specimen Object.
- The herbarium, in the processing and preparing of the specimen object may place curatorial information (i.e., annotations) about the specimen on the specimen sheet (e.g., barcode, ownership stamp, etc.)
- At various points in time, other annotations (e.g. determinations, confirmations, etc.) may be added to the specimen sheet.

It is appropriate to view the information associated with the herbarium specimen object, including the sheet to which it is attached, as dynamic over time. While some information is unchanging (e.g., Collection and Occurrence information), other information can change after curatorial actions and research use. The changes are traditionally manifest as additional marks or labels on the specimen

sheet. The implication for the Apiary XML Schema is that it must accommodate the addition of information over time in a normative way.

For purposes of modeling, we propose the following two events:

- Collection event: data about this comes from the primary label and includes information about the collection (who/where/when) of the plant glued to the sheet. This event and associated data have a one-to-one relationship with the Specimen Object.
- Annotation event: data about these come from identifications (determinations), curatorial actions related to processing and ownership changes, and other additions of information modifying what we know about the plant glued to the sheet, or what we know about the sheet itself. The Specimen Object and Annotation Event have a one-to-many relationship.

We define annotation very broadly and Annotation Event accommodates multiple annotation events related to adding information to the specimen sheet (and in some cases information about the object that may not be present on the specimen sheet). These events can include the identification of the specimen object – the original identification and subsequent determinations, confirmations, and any other information modifying what we know about the specimen object. It can also include what can be considered curatorial and collection management information such as multiple events related processing and managing the specimen object.

The resulting Apiary Application XML Schema reflects primarily the focus on the annotations. Since there is only one collection event for a specimen object, the collection event *qua* event does not need to be represented in the XML Schema. However, given the different types of annotations, the schema uses separate vocabulary terms defined in the Apiary Vocabulary to distinguish taxonomic and non-taxonomic annotations.

Our conclusion is that an event model accommodates the dynamic nature of information associated with a specimen object over time and may have utility in consideration of filtered-push. The Apiary Application XML Schema provides the structure to capture the creation, evolution, and transition of specimen objects.

URLs:

The Apiary Project: http://www.apiaryproject.org/
The Apiary XML Schemas: http://www.apiaryproject.org/documents
A Framework and Workflow for Extraction and Parsing of Herbarium Specimen Data:
http://www.tdwg.org/proceedings/article/view/567

Acknowledgements:
The Apiary Project is funded by a National Leadership Grant (LG-06-08-0079) from the U.S. Federal Institute of Museum and Library Services.