Extraction and Parsing of Herbarium Specimen Data:
Exploring the Use of the Dublin Core Application Profile Framework

William E. Moen
College of Information
University of North Texas
Denton, TX
940-565-2473
william.moen@unt.edu

Jane Huang
College of Information
University of North Texas
940-565-2473
jqhuang@verizon.net

Melody McCotter
College of Information
University of North Texas
940-565-2473
melodymcotter@gmail.com

Amanda Neill
Botanical Research Institute of Texas
Fort Worth, TX
817- 332-4441
aneill@brit.org

Jason Best
Botanical Research Institute of Texas
Fort Worth, TX
817- 332-4441
jbest@brit.org

ABSTRACT
Herbaria around the world house millions of plant specimens; botanists and other researchers value these resources as ingredients in biodiversity research. Even when the specimen sheets are digitized and made available online, the critical information about the specimen stored on the sheet are not in a usable (i.e., machine-processible) form. This paper describes a current research and development project that is designing and testing high-throughput workflows that combine machine- and human-processes to extract and parse the specimen label data. The primary focus of the paper is the metadata needs for the workflow and the creation of the structured metadata records describing the plant specimen. In the project, we are exploring the use of the new Dublin Core Metadata Initiative framework for application profiles. First articulated as the Singapore Framework for Dublin Core Application Profiles in 2007, the use of this framework is in its infancy. The promises of this framework for maximum interoperability and for documenting the use of metadata for maximum reusability, and for supporting metadata applications that are in conformance with Web architectural principles provide the incentive to explore and add implementation experience regarding this new framework.

General Terms
Standardization

Keywords
Metadata application profiles, Darwin Core, Dublin Core Application Profile, Singapore Framework, biodiversity information, herbarium specimen

1. INTRODUCTION AND RESEARCH PROBLEM
Millions of specimens in museums and herbaria worldwide need to be digitized to be accessible to scientists. Digitizing collections in a well-planned and standard way can increase use and exposure of collections to a more heterogeneous audience while simultaneously reducing physical handling and producing a permanent digital archive [1]. Digitizing the specimen is a necessary but insufficient step to provide effective access and use of the specimen. Converting the specimen metadata into machine-processible form is essential for semantic searching via search engines, distributed databases, and other data portals. A key challenge faced by all natural history collections is determining a transformation process that yields high-quality results in a cost- and time-efficient manner.

The Texas Center for Digital Knowledge in the College of Information at University of North Texas and the Botanical Research Institute of Texas are exploring workflow and metadata issues to design and implement a high-throughput system that exploits computer-assisted human parsing and transformation into structured metadata of herbarium specimen label data. This two-year (December 2008 – November 2010) research projects is funded through a National Leadership Grant awarded by the U.S. Institute of Museum and Library Services. This paper addresses our work to date on metadata issues, and in particular, the use of new frameworks for metadata application profiles.

Herbaria are special natural history collections of preserved plant specimens created for scientific use. Holmgren et al. estimated approximately 3,000 herbaria in 145 countries, containing nearly 300 million specimens [2]. Ongoing collection activities continue to add specimens to existing herbaria. Herbarium specimens are ideal natural history objects, as the plants are pressed flat and dried, and mounted on individual sheets of paper of standard size creating a nearly two-dimensional object. Each specimen is accompanied by a range of information contained on the specimen sheet: attached label with data about the specimens themselves, including the scientific name, where they were collected and by whom and when, and who identified them, as well as other associated data, such as the name of the owning institution or collection, history of ownership, and information added during curation including geocoordinates, as well as measures of data quality [3]. Thus, the specimen sheet contains a wealth of information of interest to researchers, and our project is working to take this unstructured information and transform and enhance it into structured metadata records.

The volume and heterogeneity of the data are challenging for the digitization effort. For example, the Botanical Research Institute of Texas holds over one million plant specimens from around the globe. A survey was made of the complete holdings of one genus, *Artemisia* (sagebrushes and wormwoods), in the Asteraceae (Sunflower plant family). *Artemisia* represented an average holding for the herbarium in terms of size (1179 specimens, or slightly over one cabinet-full), range of localities (worldwide but mostly North America and Europe) and ages of specimens (1805-2007). Only 41% of the *Artemisia* specimen labels were found to be easily machine-readable with off-the-shelf optical character recognition (OCR) software. These specimens were generally North American in origin and collected after 1950. The remaining 59% of specimen labels when processed through OCR resulted in

text containing numerous errors (34%) or were handwritten and impossible to digitize without human processing (25%). Figure 1 presents a sample of the variation in the specimen labels and indicates the challenges to machine-only processes for transformation.

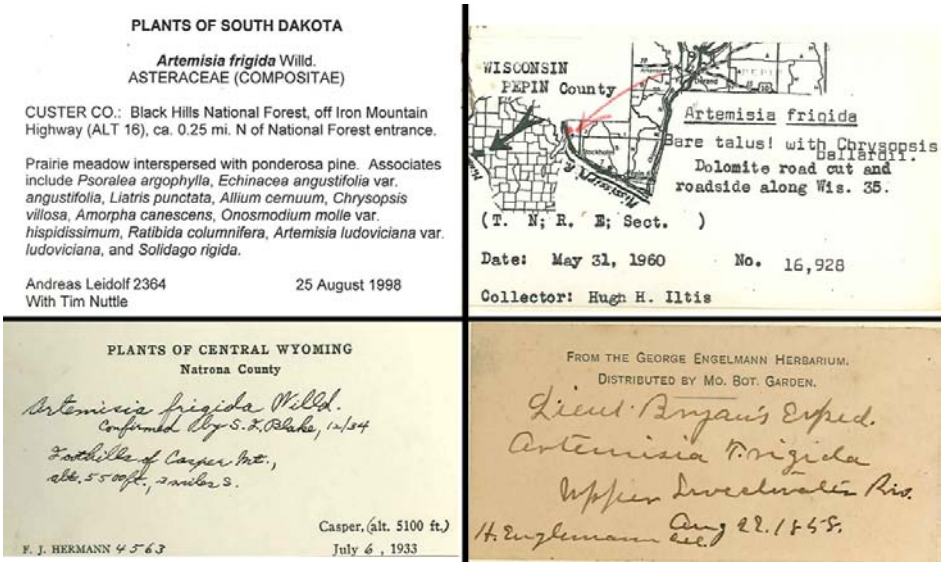

Figure 1. Typical herbarium specimen labels for *Artemisia frigida* from 1998, 1960, 1933, and 1858

2. PROJECT GOAL AND RESEARCH QUESTIONS

The current project has the following overarching goal: **Determine a workflow that provides for a combination of machine-assisted and human-assisted procedures to most effectively and efficiently convert textual data on specimen labels into machine-processible parsed data to ingest in a database and associate with the digitized specimen?** The study is examining how machines and humans can assist each other to yield high-quality and efficiently transformed specimen label data. The central focus of the research is the workflow processes for the transformation of the label data.

Three research questions are addressed in the project: to be addressed are:

- RQ1: To what extent can machine-processes accurately transform label data from a test set of specimen labels that represents variation in label types, quality, and other characteristics (e.g., handwritten versus typescript)?
- RQ2: Which human processes can be incorporated into a robust workflow to further transform, correct, and enhance label data?
- RQ3: What user interfaces are most effective and suitable to the tasks and users in supporting human processes in the workflow?

The results of this research will yield a new workflow model for effective and efficient label data transformation, correction, and enhancement that can be replicated, adapted, and transferred to herbaria and other natural history collections. Project activities are underway to address these research questions, and reports and papers generated later in the project will provide our answers to these questions. The remainder of this paper discusses metadata aspects of this project.

Metadata plays several important roles in this project. Of primary importance is capturing all relevant information from the specimen sheet, structuring that information into a specimen record in a format that uses appropriate terms from existing metadata vocabularies to ensure the metadata is shareable and enables interoperability and integration with other systems and applications. Metadata is also being used to support workflow processes and manage the digitized specimen image and derivative objects as they move through the workflow.

The following objectives related to metadata are being addressed in this project:

- Determining the metadata requirements to support the workflow and the specimen label data
- Identifying appropriate metadata vocabularies to use
- Formalizing and documenting the metadata used in the project to increase the shareability and interoperability of the resulting metadata records

3. METADATA FOR BIODIVERSITY INFORMATION

The communities working with biodiversity information, which includes botanists and herbaria, have been evolving metadata and other standards over the 8-10 years. Early work on metadata for biodiversity was initiated as part of the Species Analyst Project (http://xml.coverpages.org/speciesAnalyst.html). Emerging from

that work was a metadata scheme called Darwin Core. Since that time, the Taxonomic Database Working Group (TDWG, now referred to as the Biodiversity Information Standards group) has evolved the Darwin Core (DwC), and in October 2009 ratified a new version of DwC as a TDWG standard; DwC was developed to facilitate the sharing of information about biological diversity. (http://rs.tdwg.org/dwc/terms/index.htm). DwC terms focus on taxa, their occurrence in nature as documented by observations, specimens, and samples, and related information (http://rs.tdwg.org/dwc/index.htm).

The Dublin Core Metadata Initiative (DCMI) served as a model for DwC, and DwC can be viewed as a general extension to Dublin Core (DC) metadata terms. DwC uses a number of DC terms and also defines a list of terms that address the information needs of the biodiversity information community. With the ratification of the DwC standard, the community now has a solid basis on which to develop metadata records describing a broad range of naturally occurring organisms whether at the macro or micro level (e.g., animals to genes).

Since our current research project deals with herbarium specimens and associated data, the project is using DwC as the foundational metadata vocabulary. Specifying the use of DwC metadata as well as accommodating the needs of the project for metadata beyond what DwC offers requires a method for using and documenting metadata terms from various namespaces (i.e., from other metadata vocabularies).

4. METADATA APPLICATON PROFILES

The concept of application profiles has evolved in the past 10 years. Heery and Patel [4] first proposed profiles as a method for documenting the use, in a single application, of metadata elements from various namespaces. Application profiles can specify the use of, and constraints on, metadata elements in particular applications. In 2003, a European Committee for Standardization Workshop resulted in the *Dublin Core Application Profile Guidelines* (ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf). The form of these application profiles were typically documents that could be used by both producing and consuming applications. On the producing side, the application profile guided the input requirements for the creation of metadata records. For example, the application profile indicated the elements that would be in the metadata record, obligations and constraints on individual elements (e.g., whether an element was mandatory, repeatable, and/or used data values from specific controlled vocabularies). For those consuming or using the metadata records, the application profile provided the details to a system developer to know what to expect in the metadata record and thus develop programs to ingest and make sense of the metadata. The limitation of this approach to application profiles was that the profile document was not machine-actionable. It typically took the form of a text document.

More recently, the Dublin Core Metadata Initiative (DCMI) proposed a Dublin Core Application Profile (DCAP) framework "for maximum interoperability and for documenting such applications for maximum reusability". DCAPs developed using this new framework are intended to support metadata applications that are in "conformance with Web-architectural principles," and in particular, serve the needs of the Semantic Web (http://dublincore.org/documents/singapore-framework/).

The following sections describe the work to date in our project to develop and implement an application profile using this new framework.

5. THE NEW DUBLIN CORE APPLICATION PROFILE FRAMEWORK

Just as DwC metadata has evolved over the years to meet the needs of the biodiversity information community, the DCMI has also evolved along several dimensions: terminology, concepts, models, and support for emerging semantic web technologies. A key moment in this evolution was the adoption in 2005 of the Dublin Core Abstract Model (DCAM) with a status "Recommended". The abstract model was intended to "specify the components and constructs used in Dublin Core metadata… [and define] the nature of the components used and describes how those components are combined to create information structures (http://dublincore.org/documents/abstract-model/). The resulting information model was not tied to a particular encoding syntax, and instead was intended to assist understanding of the kinds of descriptions being created.

The DCAM defines three related model: Resource Model, Description Set Model, and Vocabulary Model. For example, the Resource Model is represented in Figure 2 with text explanation following.

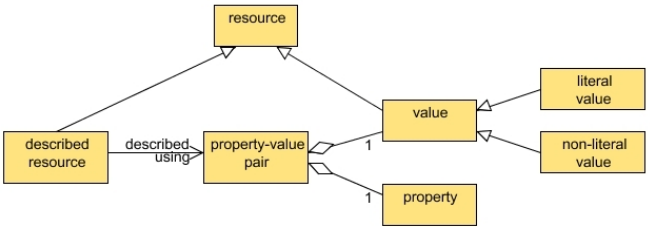

Figure 2. DCAM Resource Model

The abstract model of the resources *described by* descriptions *is as follows:*

- *Each* described resource *is described using one or more* property-value pairs.
- *Each* property-value pair *is made up of one* property *and one* value.
- *Each* value *is a* resource - *the physical, digital or conceptual entity or* literal *that is associated with a* property *when a* property-value pair *is used to describe a* resource. *Therefore, each* value *is either a* literal value *or a* non-literal value*:*
  - *A* literal value *is a* value *which is a* literal.
  - *A* non-literal value *is a* value *which is a physical, digital or conceptual entity.*
  - *A* literal *is an entity which uses a Unicode string as a lexical form, together with an optional language*

tag or datatype, to denote a resource *(i.e. "literal" as defined by RDF).*
(http://dublincore.org/documents/abstract-model/).

At the 2007 *International Conference on Dublin Core and Metadata Applications* in Singapore a new framework for application profiles was proposed. The framework, *Singapore Framework for Dublin Core Application Profiles* (http://dublincore.org/documents/singapore-framework/) builds upon the concepts described in the DCAM, and defines a formal mechanism for creating DCAPs "for maximum interoperability and for documenting such applications for maximum reusability". The framework defines the necessary components of an application profile:

- ***Functional Requirements (mandatory)*:** Describes the functions that the application profile (AP) is designed to support, as well as functions that are out of scope.
- ***Domain Model (mandatory)*:** Describes the objects metadata will describe and the relationships between those objects.
- ***Description Set Profile (DSP) (mandatory)*:** Defines a set of metadata records that are valid instances of an AP.
- ***Usage Guidelines (optional)*:** Describes how to apply the AP, how the properties are intended to be used in the application context, etc.
- ***Encoding syntax guidelines (optional)*:** Describes AP-specific syntaxes and/or syntax guideline, if any.

Figure 3 indicates the relationship of the components of the application profile with other related resources.

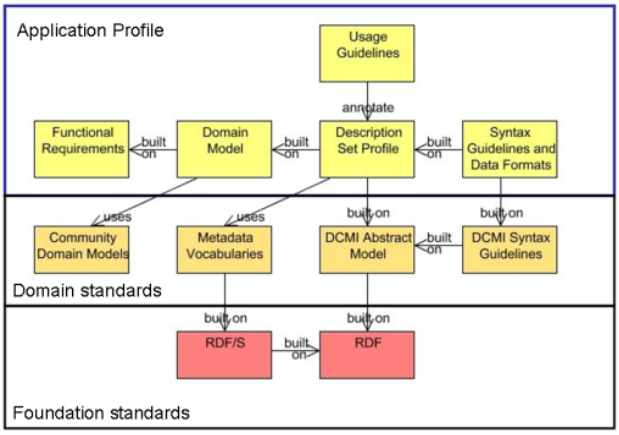

Figure 3. Application Profile Model

(http://dublincore.org/documents/singapore-framework/)

The development of application profiles in the new DCAP framework can be considered to be in its infancy with few examples from which to draw (e.g., the Dryad Project, see Greenberg, et al., 2009 [5]; Scholarly Works Application Profile, http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile).

6. THE PROJECT'S APPLICATION PROFILE RENDEREDIN THE DACP FRAMEWORK

Exercising the DCAP framework in the context our project provides a way of better understanding just what this framework requires as well as the benefits that can accrue from this approach. The following discussion is based on our work so far and may provide clarification to others who are considering using the DCAP framework. In developing the DCAP we are relying on three key documents:

- *The Singapore Framework for Dublin Core Application Profiles*: http://dublincore.org/documents/singapore-framework/
- *Guidelines for Dublin Core Application Profiles*: http://dublincore.org/documents/profile-guidelines/
- *Criteria for the Review of Application Profiles*: http://dublincore.org/documents/profile-review-criteria/

The Singapore Framework for Dublin Core Application Profiles (described above) identified several component parts of a DCAP; the *Criteria for the Review of Application Profiles* document adds one more component that addresses the objectives and scope of the application. The following sections address the three mandatory components of a DCAP.

6.1 DCAP Functional Requirements

The preliminary workflow for extracting and parsing of specimen label data is represented in Figure 4.

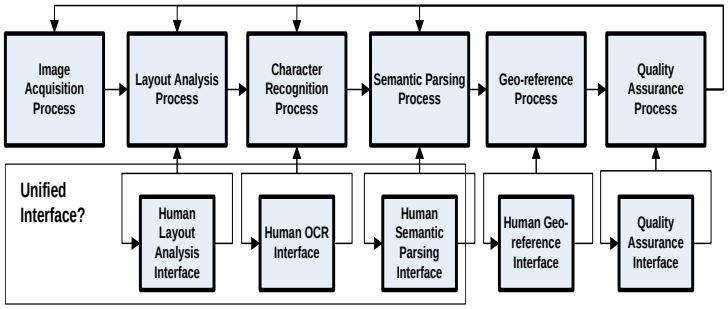

Figure 4. Transformative Process Workflow

This provides a starting point for thinking about the metadata needs that the application profile will address.

The project's functional requirements include high-level system requirements and goals (e.g., optimizing the workflow, system integration, and reusability of code) as well as more detailed requirements, especially in terms of metadata needed for various objects that move through the workflow in Figure 4. We used processed-centered use case modeling to identify key objects, subprocesses, and tasks for each of the processes outlined in

Figure 4. Specific metadata requirements that have been identified relate to: types of metadata; standard vocabularies (e.g., DwC); consistency and comprehensiveness; interoperability/shareability; granularity; reusability; and specific constraints on metadata terms.

6.2 DCAP Domain Model

This aspect of the DCAP relates to the objects of interest to the application profile and the metadata associated with each. The domain model for the project includes and defines four objects within the workflow that require metadata and shows the relationships/derivations of the separate objects. The four objects are:

- **Specimen Object:** This will have metadata derived from all of the information from the specimen sheet.
- **Specimen Image Object:** A scan of the herbarium specimen sheet and the source from which ROIs are derived.
- **Region of Interest Object (ROI):** A ROI is derived from the specimen image object and can include separate ROIs for primary label, first annotation, and other textual or graphical information on the herbarium sheet.
- **Digital Text Object:** This object results from OCR processing of a ROI or manual transcription of data from an ROI.

Relationships between these objects can be one-to-one (1…1) or one-to-many (1…n). Figure 5 shows the four objects and the relationships between the objects.

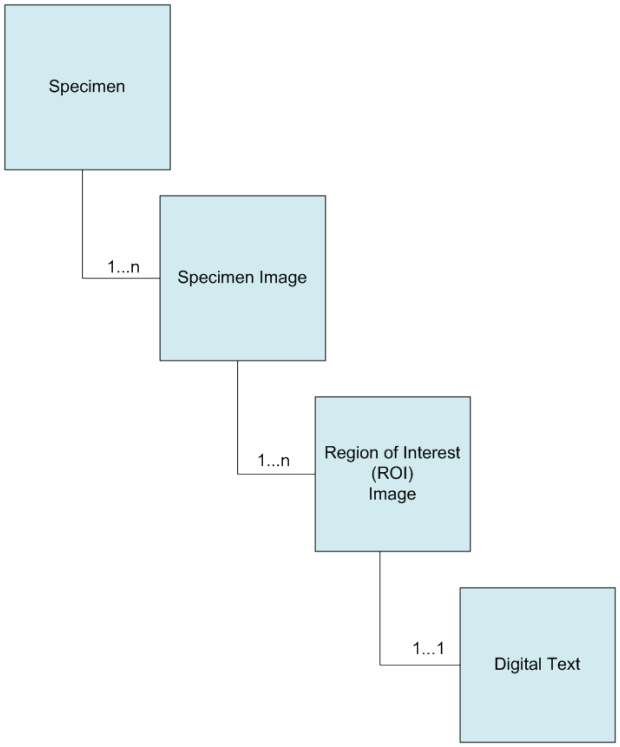

Figure 5. Objects in the Project's Domain Model

6.3 DCAP Description Set Profile

The description set profile (DSP) serves a key function in the application profile for defining the metadata terms that will be used and constraints on the use of the terms. Figure 2 above shows that the DSP is built upon domain standards that include metadata vocabularies. We see the development of the DSP is at least a two-step process:

- Determining the metadata terms required
- Formalizing the use of the terms in a structured document.

6.3.1 Determining the metadata terms

Darwin Core provides a community and domain standard for metadata terms that will be used in the application. However, from an analysis of the information on the herbarium specimen sheets we are using in the project (approximately 1,000 type specimens), DwC does not appear to accommodate all the information that appears on the sheets and that need to be recorded in the specimen metadata record. In the early phase of the project (Spring 2009), we identified a set of elements needed to accommodate the information needs. We then did a mapping to the existing DwC terms. Since the DwC was approved in October 2009 as a ratified TDWG standard, we are again investigating the DwC terms that can be used for the project's needs. For those that are not available in DwC, we need to locally define in a new namespace the terms needed.

Although the specimen label data are the focus of the specimen metadata record, the workflow also requires some technical and other metadata to help manage the objects as they move through the workflow. Two likely sources of terms are the Metadata for Images in XML (http://www.loc.gov/mix/) and Preservation Metadata (http://www.loc.gov/standards/premis) vocabularies.

6.3.2 Formalizing the use of the terms in a structured document

The concept of a description set profile model was first articulated in the DCAM document. In 2008, the DCMI published a more complete articulation of the concept in *Description Set Profiles: A Constraint Language for Dublin Core Application Profiles* (http://dublincore.org/documents/dc-dsp/). According to this document, a "DSP is a way of describing structural constraints on a description set. It constrains the resources that may be described by descriptions in the description set, the properties that may be used, and the ways a value surrogate may be given."

In the tradition "mix and match" approach for application profiles described by Heery and Patel [4] the metadata terms used in an application and constraints could be represented as in Table 1. We will use two metadata terms from our project to illustrate.

Term URI	http://rs.brit.org/ap/terms/barcode
Defined by	http://rs.brit.org/ap/
Name	Barcode
Source definition	The verbatim supplemental text associated with a barcode imprinted or affixed to the

	specimen.
Local definition	The verbatim supplemental text associated with a barcode imprinted or affixed to the specimen.
Type of term	n/a
Refines	n/a
Has encoding scheme	No
Obligation	Optional
Occurrence	Non-repeatable
Datatype	String

Term URI	http://rs.tdwg.org/dwc/terms/#scientificName
Defined by	http://rs.tdwg.org/dwc/terms/
Name	scientificName
Source definition	The taxon name (with date and authorship information if applicable). When forming part of Identification, this should be the name in lowest level taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term.
Local definition	The taxon name (with date and authorship information if applicable). When forming part of Identification, this should be the name in lowest level taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term.
Type of term	n/a
Refines	Has domain: http://rs.tdwg.org/dwc/terms/#Taxon
Has encoding scheme	http://www.ipni.org/ or BRIT compilation
Obligation	optional
Occurrence	repeatable
Datatype	string

Table 1. Recording Information for Metadata Terms in Traditional Format

Using the new DCAP framework, the above specifications can be rendered in a Description Set Profile. For ease of reading, we present the terms and constraints in a human-readable format as follows:

DescriptionSet: SpecimenData

Description template: Specimen
minimum = 1; maximum = 1
Statement template: barcode
minimum = 1; maximum = 1
Property: http://rs.brit.org/ap/terms/barcode
Type of Value = "literal"
Statement template: scientificName
minimum = 0; maximum = unlimited
Property: http://rs.tdwg.org/dwc/terms/#scientificName
Type of Value = "non-literal"
Take list = yes
Value Encoding Scheme URI = http://www.ipni.org/

The above indicates that the Description Set is related to something called a specimen (i.e., the botanical specimen on the herbarium sheet). The Description Template provides information about each object in the Domain Model. The statement “minimum = 1; maximum = 1” means that the metadata record represents one and only one specimen. The Statement Template contains the statements about the properties (metadata terms) used to represent the Specimen, giving information about the number of occurrences a term can have in the record, the URI to the property (metadata term), type of value associated with the term, and other constraint information.

The DSP can be represented in XML as well as RDF. The following shows the above information represented in XML

```
<DescriptionSetTemplate>
<DescriptionTemplate ID="Specimen" maxOccur="1" minOccur="1" standalone="no">
<ResourceClass>http://rs.brit.org/ap/objects/SpecimenMetadata</ResourceClass>
<StatementTemplate ID=”barcode” minimum= “1” maximum= “1” type="literal">
<Property>http://rs.brit.org/ap/terms/barcode</Property>
<LiteralConstraint>
<SyntaxEncodingSchemeOccurance>disallowed</SyntaxEncodingSchemeOccurance>
<LanguageOccurance>optional</LanguageOccurance>
</LiteralConstraint>
</StatementTemplate>
<StatementTemplate ID=”scientificName” minimum= “0” maximum= “unlimited” type="nonliteral">
<Property>http://rs.tdwg.org/dwc/terms/#scientificName</Property>
<NonliteralConstraint>
<VocabularyEncodingSchemeOccurrence>optional
</VocabularyEncodingSchemeOccurrence>
<VocabularyEncodingSchemeURI>http://www.ipni.org/
</VocabularyEncodingSchemeURI>
<ValueStringConstraint maxOccurs="0"/>
</NonliteralConstraint>
</StatementTemplate>
</DescriptionTemplate>
</DescriptionSetTemplate>
```

A complete DSP will address each metadata term used in the application’s metadata record, indicating information about the term (e.g., URI to the definition of the term), and indicating specific constraints on using the term in this particular

application. The DSP representation in XML or RDF provides what was not possible in the earlier forms of application profiles, namely having a machine-actionable representation of the metadata used in a particular application.

7. SUMMARY AND CONCLUSION

The research and development being addressed in our project focuses on two major areas:

- Development and testing of optimal workflows to extract, parse, and enhance the data from herbarium specimen sheets
- Using a standards-based approach for the resulting metadata describing the specimen on the herbarium specimen sheet.

This paper has described how we are exploiting new developments, concepts, and formalisms of the Dublin Core Metadata Initiative to improve the shareability and interoperability of the metadata created through the workflow. Using the Dublin Core Application Profile framework enables machine-actionable application profiles. This should lead to more efficient and effective data integration among systems and applications. The project is connecting concepts and practices of standards-based metadata, shareability and interoperability with the needs and goals of a large herbarium to make the valuable specimen label data available to botanists and other researchers. Providing legacy specimen data in the form of structured botanical metadata records, along with high resolution images of plant specimens, will provide new research opportunities for the biodiversity information community.

8. REFERENCES

[1] National Science Board. 2005. Long-lived Digital Data Collections: Enabling research and education in the 21st century. NSF. (http://www.nsf.gov/pubs/2005/nsb0540).

[2] Holmgren, P.K., N.H. Holmgren and L.C. Barnett. 1990. Index herbariorum. Part I: The herbaria of the world. 8th edition. New York Botanical Garden. 693 pp.

[3] Morris. P.J., 2005. Relational database design and implementation for Biodiversity Informatics. *Phyloinformatics* 7:1-63. (http://www.athro.com/general/Phyloinformatics_7_85x11.pdf).

[4] Heery, Rachel, and Manjula Patel. (2000). Application profiles: Mixing and matching metadata schemas. *Ariadne, 25*. Retrieved November 15, 2009, from http://www.ariadne.ac.uk/issue25/app-profiles/.

[5] Greenberg, Jane, Hollie C. White, Sara Carrier, and Ryan Scherle. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, 9:3, 194-212.