Biodiversity Informatics, 8, 2013, pp. 185-197

DISCOVERY AND PUBLISHING OF PRIMARY BIODIVERSITY DATA ASSOCIATED WITH MULTIMEDIA RESOURCES: THE AUDUBON CORE STRATEGIES AND APPROACHES

ROBERT A. MORRIS(1)*, VIJAY BARVE(2), MIHAIL CARAUSU(3), VISHWAS CHAVAN(4)*, JOSÉ CUADRA(4), CHRIS FREELAND(5), GREGOR HAGEDORN(6)*, PATRICK LEARY(7), DIMITRY MOZZHERIN(7), ANNETTE OLSON(8), GREGORY RICCARDI(9), IVAN TEAGE(10), AND GREG WHITBREAD(11)

(1) University of Massachusetts at Boston, MA, USA, email: ram@cs.umb.edu
(2) Foundation for Revitalisation of Local Health Traditions, Bangalore, India
(3) Danish Biodiversity Information Facility (DanBIF), Copenhagen, Denmark
(4) Global Biodiversity Information Facility Secretariat, Universitetsparken 15, DK 2100, Copenhagen, Denmark, email: vchavan@gbif.org
(5) Washington University in St. Louis, USA
(6) Museum für Naturkunde Berlin, Germany, email: g.m.hagedorn@gmail.com
(7) Encyclopedia of Life, Woods Hole, MA, USA
(8) US Geological Survey, Reston, VA, USA (under contract via Information International Associates)
(9) Florida State University, Tallahassee, USA
(10) Natural History Museum, United Kingdom
(11) Australian National Botanical Garden, Australia
**corresponding authors*

Abstract.—The Audubon Core Multimedia Resource Metadata Schema (simply “Audubon Core” or “AC”) is a representation-free vocabulary for the description of biodiversity multimedia resources and collections, now in the final stages as a proposed standard under TDWG Biodiversity Information Standards. By defining only four terms as mandatory, it seeks to lighten the burden for providing or using multimedia useful for biodiversity science. At the same time it offers rich optional metadata terms that can help curators of multimedia collections provide authoritative media that document species occurrence, ecosystems, identification tools, ontologies, and many other kinds of biodiversity documents or data. About half of the vocabulary is re-used from other relevant controlled vocabularies that are often already in use for multimedia metadata, thereby reducing the mapping burden on existing repositories. A central design goal is to allow consuming applications to have a high likelihood of discovering suitable resources, reducing the human examination effort that might be required to decide if the resource is fit for the purpose of the application.

INTRODUCTION

Discovery and access to primary biodiversity data, as defined by the Global Biodiversity Information Facility (GBIF, 2007) are critical components in ensuring informed decision-making on the sustainable use of biological resources and on the conservation of biodiversity at all levels. With an increasing need for a high volume of credible, quality data for research, instruction, and decision support, biodiversity information systems and networks must mobilize primary data associated with non-traditional sources including multimedia resources and their metadata.

Multimedia resources are digital or physical artifacts that normally comprise more than text. These include photographs, artwork, drawings, sound, video, animations, and presentation materials, as well as interactive online media such as species identification tools. A multimedia collection is an assemblage of such objects, whether curated or not and whether digitally accessible or not. Collections are included under the umbrella of resources, though they sometimes

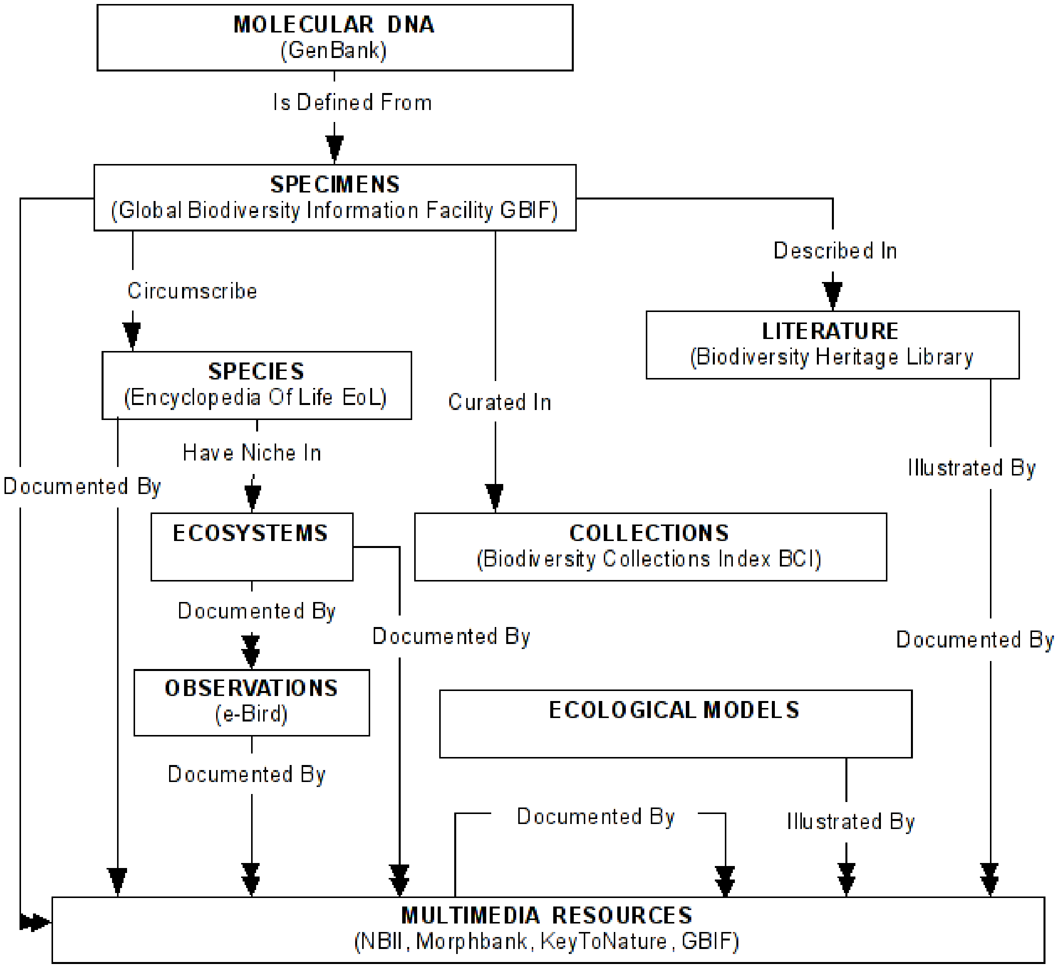

Figure 1: Relationships of multimedia resources to primary types of biodiversity resources, including some well-known example systems. The Figure is adapted from TDWG NCD Interest Group (2009).

need different kinds of treatment. Multi-media resources can provide reliable evidence for the occurrence of a taxon in a particular place and time, and there is a growing recognition that a biodiversity-related multimedia object could be used as a ‘primary biodiversity record’ if the metadata associated with the object is available and of high quality. As such, mobilizing such metadata for network access is an extension of one of GBIF’s central activities, the marshaling of occurrence data from its data contributors. Metadata on multi-media resources, and those resources themselves, can also enhance other biodiversity informatics applications such as species and specimen descriptions, glossaries, and image processing.

Because the potential quantity and quality of biodiversity multimedia resources are at least as great as that represented by observational data and have widespread potential uses, multimedia data merit special consideration. As depicted in Figure 1, applications exploiting a wide range of digital and physical biodiversity objects sometimes require the use of multimedia resources to document the objects. There is vast potential to channel the heterogeneous and distributed biodiversity-related multimedia resources through data publishers and partners. Unlike observation or specimen data, however, the network loads and latency for serving or acquiring multimedia resources may be so high that resource producers and consumers alike need mechanisms to determine the fitness-for-use of media upon discovery, before the media are fetched. To meet these and other goals described below, we describe the Audubon Core Multimedia Resource Metadata Schema proposal, now in the final stages of approval under the mechanisms of Biodiversity Information Standards (TDWG), http://www.tdwg.org.

GBIF MULTIMEDIA RESOURCES TASK GROUP

Recognizing the need for primary biodiversity data and information to extend beyond its current focus of specimen- and observation-based data records, in March 2008 GBIF asked members of the TDWG Image Working Group, and others whose work is related to images, audio, and video, to serve in the Multimedia Resources Task Group (MRTG), in order to suggest strategies to expand the types of primary biodiversity data that the GBIF network can discover and publish through the mobilization of multimedia resources (GBIF, 2008). MRTG was specifically asked to provide recommendations on (a) criteria for multimedia data sharing infrastructures, (b) best practices for multimedia resources metadata exchange/sharing, (c) estimation of the scale of multimedia resources in biodiversity, (d) metadata schema(s) for multimedia data management, and data exchange and/ sharing, (e) whether existing protocols for biodiversity data publishing services, such as DiGIR, TAPIR, or BioCASE will need to be altered, or new tools developed to handle these data types, (f) ways to encourage potential data providers to participate in the GBIF network for discovery of and access to multimedia resources, (g) ways to increase involvement of industry leaders, and (h) use of GPS-enabled mobile devices and other recording tools.

THE GBIF MRTG SURVEY

The MRTG conducted an online survey of multimedia resources in May 2008, with the objective to understand the extent of potentially useful, sharable biodiversity multimedia resources and the repositories that hold them. The survey revealed that a large quantity of biodiversity related multimedia objects are held in repositories with definite metadata recorded (such as scientific names and geo-references) indicating a huge potential for such resources to carry scientifically useful data. Many of the reported resources are managed at general-purpose repositories like Flickr (http://www.flickr.com) and PicasaWeb (http://picasaweb.google.com), and special purpose biodiversity image repositories such Morphbank (http://www.morphbank.net), Wildscreen (http://www.wildscreen.org.uk/).Their diversity highlighted the need for an infrastructure that can (1) leverage such collections for scientific analysis and (2) assist in the better management of these vast biotic resources. The survey further highlighted the need for annotation and attribution services to enhance the usability of objects and to recognize the efforts towards mobilization of such resources.

THE GBIF MRTG RECOMMENDATIONS

MRTG dealt with both social and technical issues related to the discovery, mobilization, and use of biodiversity-related multimedia resources. The principal recommendation of MRTG was that GBIF should facilitate the discovery and publishing of multimedia resources as primary biodiversity data (Morris et al., 2008). In particular, as a global information infrastructure, GBIF must reduce burdens on its stakeholders as a strategy for increasing access to high-quality resources.

Recommendations about social issues called on GBIF to (1) recognize the breadth and depth of information technology resources available to publishers of biodiversity media, (2) facilitate the publication of metadata with tools and training, (3) encourage free and open access and use of metadata, while increasing the ability to license resources, (4) support discovery and access of, at a minimum, thumbnails or other preview representation of resources, (5) encourage cultural change towards routine georeferencing of multimedia resources, and (6) encourage creation of national, regional, and thematic multimedia repositories across the GBIF network. Recommendations on technical issues focused on the development of georeferencing, annotation, and attribution services. Morris et al. (2008) listed 28 recommendations about social and technical issues with a rationale and with the possible burdens they may impose on GBIF or multimedia metadata publishers. The report further concluded that social and technical issues hampered the progress toward facilitating efficient discovery, publishing, and the use of biodiversity related multimedia objects or collections. Many valuable multimedia resources exist that have no documentation information stored in databases. Some may have a web presence and others not. Even those available on line may not be adequately discovered by search engines, or may be lost in the noise of images, audio and videos from unreliable sources. A brief descriptive record can act as the 'business card' for researchers, aggregators, decision makers, educators, or the general public to discover these resources. The development of a multimedia metadata schema for easy discovery, publishing and use of biodiversity-related multimedia resources was deemed helpful to address these issues.

GBIF MRTG Recommendation (Morris et.al., 2008)	**Facilitation through the Audubon Core**
R#3: Metadata about media resources is provided either without any restriction on its use or reproduction, or under a suitable open-content license.	Provide for copyright attribution and terms of use.
R#4: Publishers will be able to license their resources.	Specific terms can reference various versions of a multimedia resource including license and other attributes.
R#5: GBIF metadata and data sharing agreements should give the GBIF network the right to cache and display previews (e. g. thumbnails) if publisher grants the access.	Metadata identifies such resources.
R#12: Metadata should promote the ability of users of GBIF services to determine fitness-for-use without requiring the users to acquire underlying resources.	Ability to signal biologically relevant content metadata, such as Taxonomic and Geographic Coverage.
R#13: Ability to treat resource collections and objects uniformly.	Both resource collections and objects are described through a single schema.
R#14: Controlled vocabularies for metadata values should be encouraged and supported technically.	Specific values are suggested or required, particularly where arising from other vocabularies.
R#15: Specify that the copyright owner or available licenses are unknown when this is the case.	'Unknown' is an accepted value for the terms specifying these.

R#16: Support the identification of resources with publisher- defined GUID schemes in resource or collection level metadata.	An identifier is required for collections (strongly recommended for media), but the scheme for such identifiers is up to the provider, or to implementers of the representation- neutral form of the specification.
R#17: Support the ability to specify relations among described objects.	A generic relation 'relatedResourceID' is provided with no specified semantics. A small number of relations are provided for provenance, and a few for relations between different renderings of the same resource.
R#18: Services for georeferencing and scientific name recognition.	All the georeferencing predicates of the Darwin Core are accepted by inclusion. A collection of terms designated as the 'Taxonomic Coverage Vocabulary’ supports use of several Darwin Core nomenclatural predicates.
R#19: Allow support for the ‘documents’ relation, which asserts that a multimedia object provides evidence for an assertion that something else (e. g. an observation) is a primary biodiversity datum.	Subsets of the terms facilitate this, including the Taxonomic, Geographic, and Temporal coverage vocabularies.
R#20: Lightweight metadata schema by combining existing schemata.	Accomplished by use of existing namespaces from other vocabularies where semantically reasonable.
R#21: Ability to specify media formats.	Service access points for different formats can be separately specified.
R#22: Allow specification of media manipulation by the Publisher after acquisition.	Service access points for variants are supported, along with limited terminology for provenance description.

Table 1: Recommendations of the MRTG met through the Audubon Core. “R#” designates the recommendation addressed (Morris et al., 2008, and Morris et al., 2009).

Table 1 provides a list of 14 of the 28 issues from Morris et al. (2008) that MRTG sought to address through the development of the Multimedia Resources Metadata schema (Morris et.al., 2009), now designated as the ***Audubon Core Multimedia Resources Metadata Schema (****“Audubon Core”).*

Development of the Audubon Core

A subset of MRTG began development of the Audubon Core in August 2008, and a slightly different subset continued in September 2009, developing the key terms for a new metadata schema for multimedia resources. Development of the schema included the participation of key stakeholders such as GBIF, Key to Nature (http://www.keytonature.eu), the U S Geological Survey, Morphbank (http://www.morphbank.net), and the Encyclopedia of Life (http://eol.org), as well as expressions of interest and inputs from the Biodiversity Heritage Library (http://biodiversitylibrary.org), the University of Massachusetts at Boston Electronic Field Guide

Project (http://efg.cs.umb.edu) , and the Atlas of Living Australia (http://www.ala.org.au/).

Further work has been conducted on the schema since that meeting, and in February 2010, still known as the "MRTG schema", version 0.9 was submitted for internal review to the Biodiversity Information Standards (TDWG). As the schema progressed to the final stages of approval by the standards body, the name "Audubon Core" was proposed for the schema in honor of the great natural history illustrator, John James Audubon. In November 2010, v1.0 the schema was submitted to the TDWG Executive Committee (EC). The submission included the response to an internal review and the proposal to officially name it "Audubon Core". A second review was completed and substantial changes made based on it. Responses to these and two more reviews have been completed and addressed, with further detailed changes. Based on those, the Executive Committee permitted a period of public review as required by the TDWG rules, which is now complete. Responses to that review will be submitted to the EC, including any changes arising from the response, with a request to accept the Audubon Core as a TDWG standard as may be revised based on the responses to the public review.

Several projects have been exploring the use of AC for their image management metadata in the form proposed for public review. Of these, the most central to GBIF's goals is a draft produced by the iDigBio (2011) project of an Audubon Core IPT Darwin Core Extension[1] now under testing. IPT denotes the GBIF Integrated Publishing Toolkit (GBIF 2011), the recommended tool for publishing biodiversity data for harvesting by GBIF and exposure through its portal. An IPT Extension is an XML file that allows IPT to drive user interfaces and map the data publisher's data to easily harvested data using the domain vocabulary, in this case a subset of Audubon Core. A recently commissioned Indo-Norwegian IPBES Capacity Building Pilot project aims to implement Audubon Core based dataflow to collate and publish the camera trap data through the GBIF network. It is planned to use MS Excel based 'AC data templates' to collate the multimedia data captured through camera traps in key protected areas.

CAPABILITIES

It is expected that Audubon Core will facilitate (1) the enhanced discovery of multimedia resources, (2) the evaluation of fitness-for-use prior to fetching a resource, (3) the use of metadata records as potential taxon occurrence evidence, or for other biological inferences such as evidence for species interactions, habitats, and phenotypic variations, (4) identification aids, and (5) the ability of multimedia resource producers and publishers to gather and serve resources contributed by a wide variety of producers and custodians, particularly those with little or no information technology expertise or support.

The Audubon Core facilitates the above by describing with consistent metadata either media resources themselves or a collection of them. Other existing standards present very little opportunity to provide media resource metadata that are specifically biologically relevant. For instance, although it can describe multimedia, the use of Dublin Core (DC); http://dublincore.org/) alone would not ease the discovery of media resources that require precision with respect to geolocation and identification. Similarly, Darwin Core (DwC; http://rs.tdwg.org/dwc/ supports some biological semantics (e. g. taxonomy) but offers little about important intellectual property rights issues, or ways to express relations between alternate versions of media resources (e. g. services for different pixel resolution). The Natural Collections Description (TDWG NCD Interest Group, 2009) provides useful metadata on object collections, but is missing some aspects relevant to biological media collections. Metadata compliant with technical schemes, such as EXIF (http://www.exif.org/specifications.html), are frequently embedded directly in the media files by the imaging systems themselves. Such embedded data often can be managed by tools such as Adobe Photoshop™ and the GIMP open source image editor (http://www.gimp.org/). However technical metadata typically describe only the acquisition parameters of the media (e.g. pixel size, exposure data, etc.). They present little opportunity to embed biologically relevant information. Furthermore, the combination of all of these standards still does not,

[1] Available at this writing at http://rs.gbif.org/sandbox/extension/audubon.xml

or does only in a limited fashion; address the concerns of a wide variety of multimedia contributors, especially those with limited access to software engineers and digital librarians. Among such concerns are various aspects of multimedia object provenance, intellectual property rights and attribution, access services, and the impact on service quality of large multimedia resources. Below we discuss four examples: transfer cost, discovery of fitness-for-use, intellectual property rights, and provenance.

Transfer Cost: Individual digital multimedia resources such as images, video and sound may have very large file sizes. As a result, multimedia metadata must support use cases where humans or software agents fetch the resource in a reduced size (e.g., for images, small thumbnails or screen-sized resolutions). The management of multiple access points returning the resource in different forms and resolutions is therefore essential.

Fitness-For-Use Discovery. Without specific examination of possibly many thousands of images, it can be difficult to determine whether a media resource carries all the biological context and technical properties required for the intended use. For example, it may be difficult to determine whether the resource depicts an organism in its natural habitat, a specific behavior, or particular morphological characters. Furthermore, the resolution of an image may be too low, or it may contain labeling in an unsuitable language. The Audubon Core combines metadata terms representing these things (as well as several others from other widely used vocabularies) into one schema. It does so in a standardized way that makes it unambiguous what is being described and how it is made available by the provider.

Intellectual Property Rights: Ownership of physical objects (e.g., specimens) is generally governed by property laws, while text and media resources are often subject to Intellectual Property Rights (IPR). However, factual descriptions of objects are usually not subject to IPR (Agosti and Egloff, 2009). Although similar considerations may apply to factual media representations of organisms, media have a history of being treated as creative works of art, not as expressions of facts of nature. Consequently, the Audubon Core provides attributes to describe IPR, including ownership and license restrictions (such as reproduction permission and attribution requirements).

Provenance: For any scientific data, it is important to know the methodology used as well as how and when the data may have been changed from its original gathering. This is particularly important for media, which are commonly edited for a variety of purposes. If carelessly done, this may destroy some of the modified object's utility, or provide false impressions of data and thus influence research results. No current or proposed TDWG standard provides much provenance information, in part because widely accepted standards for specimen provenance and governance already exist. However, the creative aspects of media resources result in conflicting goals. The Audubon Core records object derivation (one media item is the source of another) and introduces a term called Resource Creation Technique, for information about the technical aspects of the creation, digitization, and post-processing (like background blurring, background elimination, color adjustment, etc.).

RELATION TO OTHER STANDARDS

A number of organizations concerned with addressing biodiversity multi-media in particular have informally or formally published specification for describing their resources. Representatives of, or consultants to, several of these organizations are among the authors of this paper and architects of the Audubon Core. Much of those organizations' published metadata terminology has in one way or another been folded into AC (See http://terms.gbif.org/wiki/Audubon_Core_Term_List_(1.0_normative)#References.) Most of the more general well-known multi-media metadata vocabularies focus on technical metadata of the image acquisition, or on curatorial, provenance, and intellectual property attributes. (NISO 2008, IPTC 2010, DCMI 2011, XMP 2010). They have limited expressivity about content, but we adopt their terms where we can.

Two crowd sourcing biodiversity media collections are worth mentioning, in part because they illustrate some of the problems of insufficiently formal or too dynamic metadata. The first of these, Wikispecies, documents its image requirements at http://species.wikimedia.org/wiki/Help:Image_Guidelines. Most of the guidance is dedicated to licensing (Wikispecies requires open access to material on its pages) and layout. However, Wikispecies images are actually uploaded to the Wikimedia Commons (http://commons.wikimedia.org/wiki/Main_Page). As is generally the case for images supported in the Wikimedia Commons, image metadata per-se is limited to three sorts, mostly optional: a text caption, some specific image provenance and licensing text and the assignment of new or existing MediaWiki "Categories". The last of these can be considered as lightly structured attributes (or rather "classes") of the images, but at this writing, the overwhelming fraction of those are the names of geographically constrained taxa, e.g. ("Australia Arthropoda"). All that said, images on Wikispecies are associated with a taxon page, and *that* has somewhat more information about the taxon, principally its taxonomy and nomenclature. The fact that contributors to Wikispecies can add MediaWiki Categories at will could hold some promise for its contributor community to provide more organization to the website in ways that would provide more metadata to the embedded images. However, MediaWiki Categories are a typing mechanism and do not provide simple ways to place attributes of objects on wiki pages (as evidenced by the 330 categories of geographically constrained taxa such as mentioned above, and which reference fewer than 20 georegions.) Wikispecies could be augmented by the Semantic MediaWiki extensions (http://semantic-mediawiki.org/). Note that the design of the Audubon Core puts emphasis on attributes rather than categories. AC only models as a class the access mechanism for retrieving media, because such mechanisms are highly variable and with many attributes. Finally, we note that all MediaWiki installations provide a permanent URL for each version of a page. By the association of the image with a page version, this "permalink" can serve as a globally unique, persistent, dereferenceable URI for the image.

A second crowdsourced biodiversity image repository may be seen in the Encyclopedia of Life Image Flickr group (http://www.flickr.com/groups/encyclopedia_of_life/) with metadata provided by a small set of Flickr "Machine Tags" (http://www.flickr.com/groups/encyclopedia_of_life/discuss/72157612488733900). These are limited to taxonomy, georeference, and licensing information, but ownership, license metadata, and some technical metadata is available by Flickr APIs (http://www.flickr.com/services/api/). About 88,000 images are served this way by Flickr, of which about 78,000 are harvested and associated with EOL pages. EOL itself offers similar metadata for all of its images (http://wiki.eol.org/display/dev/data_objects)

The documentation supporting the submission to TDWG for ratification includes a normative specification of the Audubon Core as a set of multimedia resource metadata terms independent of any digital representation (http://terms.gbif.org/wiki/Audubon_Core_Term_List_(1.0_normative)). That document will be updated to reflect any changes accepted for the standard after the period of public comment. The normative document provides metadata specifications describing biodiversity-related multimedia resources or collections. While focused on biodiversity-related multimedia resources, the Audubon Core addresses some of the same concepts as the Dublin Core, Darwin Core and other standards that describe access to resources. These standards include the Adobe Extensible Metadata Platform (XMP 2010), the International Press and Telecommunications Council (IPTC 2010) the Metadata Working Group (MWG 2010) schema, the TDWG Natural Collections Descriptions (TDWG-NCD 2009) schema, and others. Where a particular term meets the same need met by the terminology within another standard, MRTG adopted that standard's globally unique identifiers and definitions. Where this is unsuitable, MRTG defined new. The design intends to ease the burden of holders using descriptions already specified either by DwC or DC, to allow them to use existing descriptions where appropriate. In other words, much of the Audubon Core may be viewed as a standard profile that defines the best practice use of certain terms from other metadata vocabularies, and provides further vocabulary for metadata that improves the ability to utilize multimedia resources for scientific research.

AUDUBON CORE RECORDS

An Audubon Core metadata record is a set of terms and term values that describe an underlying multimedia resource. Each term is identified by a Uniform Resource Identifier (URI). Each URI refers to the attribute, not the underlying resource; it simply specifies which term is being provided. There are many URI schemes, some of which have been registered with the Internet Assigned Names Authority (IANA). All Audubon Core URIs conform to the widely used http URI Scheme. MRTG chose this scheme because it uses the familiar Internet URL syntax. However, this familiarity gives rise to a common misconception that pasting the URI into a browser URL line, or providing it to some other application that understands the http protocol, should result in the application returning some information about the object identified by the URI. Such dereferencing[2] of the URI is in no way guaranteed for all Audubon Core terms. Where possible, Audubon Core terms are dereferenceable, with information returned for how the metadata attribute identified by that URI is defined or used. Human-centric Audubon Core applications, however, should not present the URIs to users, nor use them as linking mechanisms. One possible exception is a self-documenting application that assigns metadata to multimedia resources. In that case the application might dereference the URI to provide a glossary entry aiding the user in the semantics of the metadata term. However, since dereferencing is not required for other functionality and may not be guaranteed in the long term, we suggest caution using it. In fact, all the "native" AC terms (as distinguished from those borrowed from other vocabularies) do have dereferenceable URIs, presently pointing to the normative documentation.

Where borrowed terms have dereferencable URIs links to documentation are provided. Finally, note that some controlled vocabularies are defined in PDFs or other documents that do not have URL links directly to each defined term. In these cases, any dereference may only link to the beginning of the document, leaving it necessary to search in the document for the referenced definition.

The proposed Audubon Core schema consists of 77 terms (plus the Darwin Core georeferencing terms by inclusion). Every term has a plain text name, a normative URI, and a plain text normative definition. In addition, terms have a recommended English label for use in applications, the aforementioned Details, some non-normative commentary on usage, and a non-normative, somewhat spare, set of usage notes. The final normative definitions of the standard, with full URIs, will be found on the Audubon Core Wiki http://terms.gbif.org/wiki/Audubon_Core_Term_List_(1.0_normative). It is expected that "best practices" documents will be developed by various user communities. To ensure that the barriers to use are as low as possible, only four terms of an Audubon Core record are considered to be mandatory. These are summarized in Table 2, with abbreviated URIs in parentheses.

Associated with each Audubon Core term is its value, whose data type is also specified. When the Audubon Core (or any vocabulary it references) uses literals, it is important that any metadata interchange use the literals verbatim, even if the record is declared to be in a different natural language. An example is the "Type" metadata term, which is required to come from the corresponding vocabulary, Dublin Core. Agents answering Audubon Core metadata queries must be able to consume and respond to queries framed with that controlled vocabulary. The normative document does not prevent a metadata publisher from asserting it has no records with a given controlled term, nor from internally mapping between a controlled vocabulary and its internal attributes, whose names may well be in a language other than English. Only a small number of terms take values in a specific, English-based controlled vocabulary. Of the mandatory Audubon Core terms, only Type has any such requirements.

[2] Commonly called "resolution", but the two terms are importantly distinguished in the IETF specification http://www.rfc-editor.org/rfc/rfc3986.txt

Term	**Definition**
Identifier (dcterms:identifier)	An arbitrary code that is unique for the resource, with the resource being a collection or a media item. The draft requires an Identifier for collections and strongly recommends but does not require an Identifier for media items.
Type (dcterms: type)	Any DCMI type term from http://dublincore.org/documents/dcmi-type-vocabulary/ may be used. Recommended terms are Collection, StillImage, Sound, MovingImage, InteractiveResource, Text, PanAndZoomImage , 3DStillImage, and 3DMovingImage.
Metadata Language (ac:MetadataLanguage)	Language of description and other meta data (but not necessarily of the image itself) represented in ISO639-1 or -3.
Copyright Statement (dcterms:rights)	Information about rights held in and over the resource. A full-text, readable copyright statement, as required by the national legislation of the copyright holder. On collections, this applies to all contained objects, unless the object itself has a different statement.

Table 2: The Four Mandatory Terms of Audubon Core

It may seem odd that so few terms are mandatory. One reviewer suggested that there is no use for a metadata record that contains only the mandatory terms, because such a record would not assist in discovery or fitness-for-use evaluation. But this is definitely not the case in circumstances where the resource metadata and/or the resource data are themselves available from several disparate sources. The simplest example might be the case in which an extensive AC metadata service is offered by one server without any resource service, but with a reference to a service that holds the resource. In this case, the resource service is likely to need only the AC Identifier value and might well be motivated to hold and serve only the mandatory metadata. A related scenario is one for which a user or software agent desires to formulate an AC-based query to a distant server as to whether that server holds any resources meeting a specific set of criteria relative to a given image for which only the mandatory data (including the Identifier) is met. For example, entirely with mandatory data and a sufficiently expressive query language, a remote service can be asked for a list of resources it holds (or even simply knows about) that are known to have the same taxonomic coverage as the one in hand, even though the user doesn't know what that coverage is. The reviewer suggested that MRTG could propose one or more standard subsets of AC to provide for various communities of practice, e.g. taxonomists, ecologists, etc., but the authors feel that such "profiles" are best organized by the communities themselves. Thus, the architecture is meant to enable, rather than define such profiles. Indeed, doing so will likely involve social and organizational considerations, e.g. the IT resources available to organizations holding the media and metadata, and no single group is likely competent to provide several different profiles. Instead, at the final adoption or soon thereafter, the TDWG Annotation Interest Group (AIG), of which MRTG is a Task Group, will probably also recommend mechanisms by which self-organizing communities can build such profiles and choose among the several TDWG mechanisms for recognizing applicability and use of standards. (See the section "Sustainability" below.)

In some cases, metadata terms are necessarily related to others. For example, an image might have several variants that must remain related even if they have their own metadata in another Audubon Core record. However, many image contributors are constrained to record their metadata in spreadsheets or other flat structures, use of which makes it difficult to represent such structural relationships. Consequently the Audubon Core itself is primarily flat, the exception being a few structures designated as members of a ServiceAccessPoint class, which describe various ways to access the media resource and related resources. One consequence of the flat structure is that a metadata publisher might have to make several metadata records available about the same underlying resource. An important case surrounds multilingual metadata. Because each metadata record is in a fixed language specified by the Metadata Language term (this is the language of the metadata record, not of any language featured within the multimedia resource itself), a provider might have to offer one metadata record about the same multimedia resource for each available language. The mandatory terms must be provided in every metadata record, even if repeated in other metadata records. This and other cases requiring multiple metadata carrying the same mandatory terms and only a little more, provide a huge number of combinations wherein extremely minimal metadata is in play. At the date of this writing, the normative document does not provide a mechanism for singling out a metadata record that might be overarching, the optional terms of which may be regarded as defaults for other records about the same resource.

Finally, many terms may be repeated in a record, but some may not. For example the "Modified" term corresponds to a date at which the media resource was modified and may be repeated to reflect the history of the resource. By contrast, "Date Available" is a single date, or a single range of dates, at which the underlying resource became, or will become, available. Audubon Core designates terms that may be repeated.

For use by software, two recommended serializations for the digital representation of the Audubon Core metadata are under development and will be submitted to the TDWG standardization process via the sustainability model described below. One is based on the World Wide Web Consortium (W3C) Resource Description Framework (RDF), a model for data interchange on the semantic web (http://www.w3.org/RDF/). The other is based on the Extensible Markup Language (XML; http://www.w3.org/standards/xml/), a standard language for constraining the form of markup permitted in data interchange. Another serialization, yet to be specified, will be based on delimited text files, such as comma- or tab-separated text. One such serialization is implicitly provided by the Audubon Core IPT Extension mentioned earlier. Also of note, the language of the normative specification is English, but this in no way constrains applications from using labels or content of the metadata in other languages. A term is provided to denote in which language the metadata is recorded.

The question of how much structure to provide in a metadata specification for science application is complex (Beard, 1996). Our choice was generally to avoid the issue and leave it to specific implementations, particularly as media providers may wish to have service or exchange profiles specialized to more than one purpose. We expect that the (nearly flat) normative schema enables the specification of a variety of profiles aimed at such different applications as metadata exchange, intelligent image discovery, and services providing quality control on evidence for species occurrence.

SUSTAINABILITY AND FUTURE DEVELOPMENTS

In its submission to the TDWG Executive Committee, MRTG included a sustainability plan based on procedures similar to those of the Darwin Core Namespace Policy document (DwC 2011). This provides procedures for the introduction of new namespaces and new terms, for dealing with editorial errata, and for introducing semantic changes to terms. In addition, MRTG manages Audubon Core issue tracking with a Google Code project similar to that of the Darwin Core. An initial implementation is at http://code.google.com/p/audubonCore/

Although Google Code provides a wiki, it is deliberately minimalistic and the normative documentation on terms.gbif.org will remain on that platform. More importantly, the GBIF Terminology Platform implementation uses the Semantic MediaWiki extension (SMW 2011) that supports reasoning, RDF export, and semantically enhanced data. Several of the authors plan to explore the use of these facilities to support images with Audubon Core metadata published on the Semantic Web.

SUMMARY

The Audubon Core is a representation-neutral metadata vocabulary for the description of biodiversity multimedia resources. It is capable of implementation in various constraint languages, and with profiles that specify further constraints, best practices, or term subsets. Because it is representation-neutral, its URIs may be used across a number of technologies, such as namespaces in XML Schema-validated documents, RDF, and column headings in comma-delimited text files. Its use of existing namespaces and vocabularies for a number of terms eases mappings from existing metadata to Audubon Core compliant metadata. The breadth and diversity of participation in the development of this schema, by multiple organizations, causes us to expect that a ratified Audubon Core will become a de facto standard for exchanging multimedia data that describe biodiversity multimedia resources. The GBIF Secretariat has begun to implement the Audubon Core schema in its next version of Integrated Publishing Tools (IPT) as an IPT Extension of Darwin Core. Some tools have already explored use of the Audubon Core in XML-based implementations of the early normative document (e.g. Saraiva and Catalano, 2010).There is huge potential in the discovery, publishing, and usage of multimedia resources in scientific analysis, in addition to the interpretation that leads to informed decisions in the sustainable use of biodiversity resources. The need for such resources calls for service and access of biodiversity multimedia records/resources as robust and simple as for other primary biodiversity data. An early uptake of the Audubon Core by the stakeholder communities would not only ensure the mainstreaming of multimedia resources into biodiversity research, but would help engage citizen scientists and professional naturalists in creating, and sharing scientifically useful primary biodiversity data.

REFERENCES

Agosti, D. and Egloff, W. 2009. "Taxonomic information exchange and copyright: the Plazi approach", *BMC Research Notes* 2009, 2:53doi:10.1186/1756-0500-2-53

Beard, K. 1996. "A Structure for Organizing Metadata Collection" in *Proceedings, Third International Conference/Workshop on Integrating GIS and Environmental Modeling*, Santa Fe, NM, January 21-26, 1996. Santa Barbara, CA: National Center for Geographic Information and Analysis. Paper at http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/beard_kate/metadatapaper.html

DCMI. 2011. Dublin Core Metadata Initiative, DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/

DwC. 2011. Darwin Core Namespace Policy, http://rs.tdwg.org/ac/terms/serviceAccessPoint.

GBIF. 2007. GBIF Work Programme 2009-2010. 54pp. http://imsgbif.gbif.org/CMS_NEW/get_file.php?FILE=976c6c3015d7d2505d5ac1d45357c8

GBIF. 2008. http://www.gbif.org/communications/news-and-events/showsingle/article/multimedia-resources-task-group/.

GBIF. 2011. The Integrated Publishing Toolkit, http://www.gbif.org/informatics/infrastructure/publishing/

iDigBio. 2011. National Resource for Advancing Digitization of Biological Collections. https://www.idigbio.org/

IPTC. 2010. IPTC Standard, Photo Metadata (July 2010), Document Revision 1, International Press Telecommunications Council, http://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata-201007_1.pdf

Morris R, Olson A, Freeland C, Hagedorn G, Riccardi G, Carausu M, O'Tauma E, and V Chavan. 2009. Mobilising multimedia resources in biodiversity: 2nd report of the GBIF Multimedia Resources Task Group (MRTG), March 2009. Copenhagen: Global Biodiversity Information Facility. 22 pp. http://imsgbif.gbif.org/CMS_NEW/get_file.php?FILE=a9369cf39af07c00d0891f34fe667a

Morris, R., Olson, A., O'Tuama, E., Riccardi, G., Whitbread, G., Hagedorn, G., Teage, I., Heikkinen, M., Leary, P., Barve, V., and V. S. Chavan. 2008. Recommendations of the GBIF Multimedia Resources Task Group. September 2008. Copenhagen: Global Biodiversity Information Facility. 18pp. http://imsgbif.gbif.org/CMS_NEW/get_file.php?FILE=e61a4e0dde908609320b1fb7dfcf3b

MWG. 2010. Metadata Working Group Guidelines for Handling Image Metadata, Version 2.0, November 2010. http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf

NISO. 2008. NISO Metadata for Images in XML Schema (MIX) Official Web Site. http://www.loc.gov/standards/mix/.

Saraiva, A. and Cartolano, E. A. Jr. 2010. *Biodiversity Data Digitizer*. TDWG Annual Meeting, Woods Hole MA, 2010 http://www.tdwg.org/fileadmin/2010conference/slides/Saraiva_Biodiversity_Data_Digitizer.pdf,

SMW. 2011. Semantic Media Wiki. http://semantic-mediawiki.org/

TDWG NCD Interest Group. 2009. Natural Collections Descriptions: A data standard for exchanging data describing natural history collections. http://www.tdwg.org/standards/312/.

XMP. 2010. *XMP Specification Part 1; Data Model, Serialization, and Core Properties, Adobe Systems*, July 2010. http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf