



Developing an application profile for the **Art of Life**: the mixing and matching of art and biodiversity data

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session



What is Art of Life?

- Full title - *The Art of Life: Data Mining and Crowdsourcing the Identification and Description of Natural History Illustrations from the Biodiversity Heritage Library (BHL)*
- Grant given to Missouri Botanical Garden in St Louis
- Funded by National Endowment for the Humanities



- Runs May 2012-April 2014

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session



What is the Biodiversity Heritage Library (BHL)?

The Biodiversity Heritage Library (BHL) is a consortium of natural history and botanical libraries that cooperate to digitize and make accessible the legacy literature of biodiversity held in their collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons.”



VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

For those of you who may not have heard of the BHL let me give you some background. The Biodiversity Heritage Library (BHL) is a consortium of natural history and botanical libraries that cooperate to digitize and make accessible the legacy literature of biodiversity held in their collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons.”



www.biodiversitylibrary.org

Now Online
 73,536 titles
 132,998 volumes
 42,879,506 pages
 Recent Additions

BHL
 Biodiversity Heritage Library

Inspiring discovery through free access to biodiversity knowledge.
 The Biodiversity Heritage Library works collaboratively to make biodiversity literature openly available to the world as part of a global biodiversity community.
 BHL also serves as the foundational literature component of the Encyclopedia of Life (EOL).

Search across books and journals, scientific names, authors and subjects

Search

ADVANCED SEARCH

Browse Our Collection By:

Titles Authors Date Collection

Help Support BHL
 BHL's existence depends on the financial support of its patrons. Help us keep this free resource alive!
 Donate Now

New on the BHL Blog
 Alexander Wilson and the Catbird
 A tiny corner of grass in a hunting city landscape, the osprey's of Philadelphia's Gloria Dei...

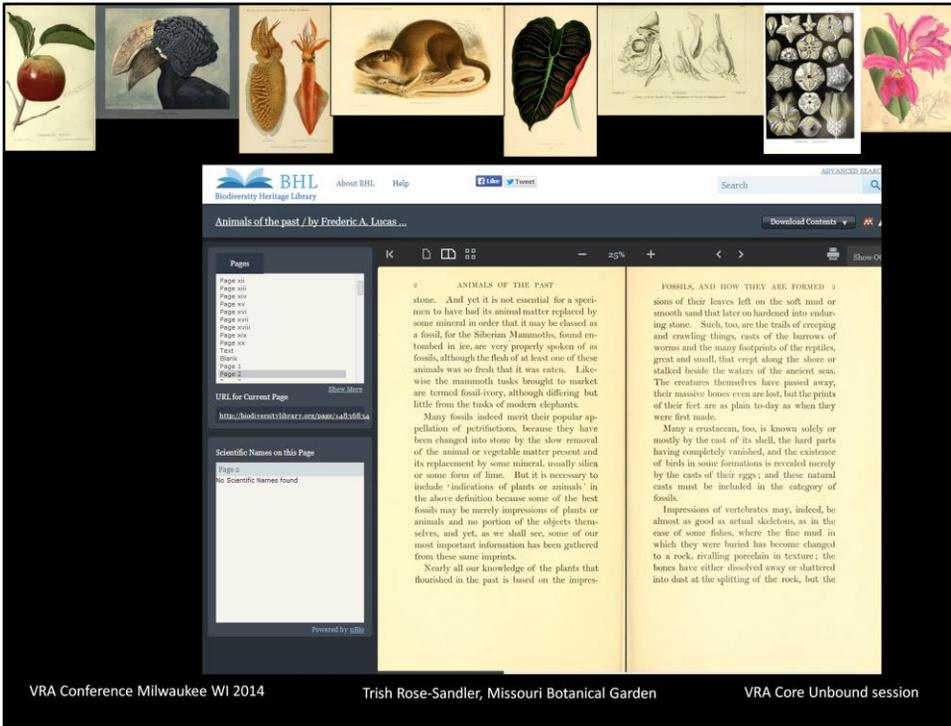
BHL's Venture into New Territory
 We really love trying new things and using open tools to help provide more access to the literature...

Today's Picks Flickr Stream

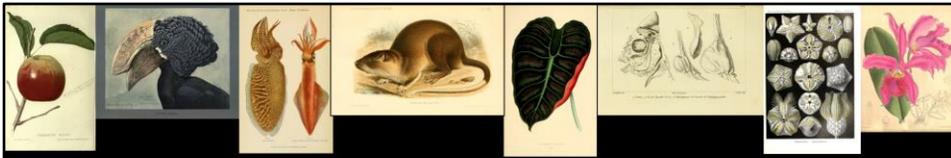
Featured Collection
 Savants of Napoleon's Egyptian Campaign

VRA Conference Milwaukee WI 2014 Trish Rose-Sandler, Missouri Botanical Garden VRA Core Unbound session

The consortium has been digitizing literature since 2007 and has nearly 43 million pages of content that we serve both through the Internet Archive and through a specialized portal at biodiversitylibrary.org



Here is a screen shot of our book viewer



[BHL](#) About BHL Help [Facebook](#) [Twitter](#)

Cassell's book of birds : From the text of Dr. Brehm /

Pages: Page 202, Page 203, Page 204, Page 205, Page 206, Page 207, Page 208/Pg. 42, Page 209/Pg. 42, Page 210, Page 211

URL for Current Page: <http://biodiversitylibrary.org/1006/13770060>

Scientific Names on this Page: [Page 204/Pg. 42](#), [Jacana](#)

Powered by [iLib](#)

Contributed by [Harvard University, MCZ, Digitized by \[Biodiversity Library\]\(#\)](#)

VRA Conference Milwaukee WI 2014 Trish Rose-Sandler, Missouri Botanical Garden VRA Core Unbound session

What many people do not know about the BHL is that it also contains millions of visual resources found within its pages. Unfortunately these are mostly hidden because there is no identifying information about them.



BHL Problem statement

- users want access to images, access to images is limited
- How to broaden the audiences for BHL content?

VRA Conference Milwaukee WI 2014

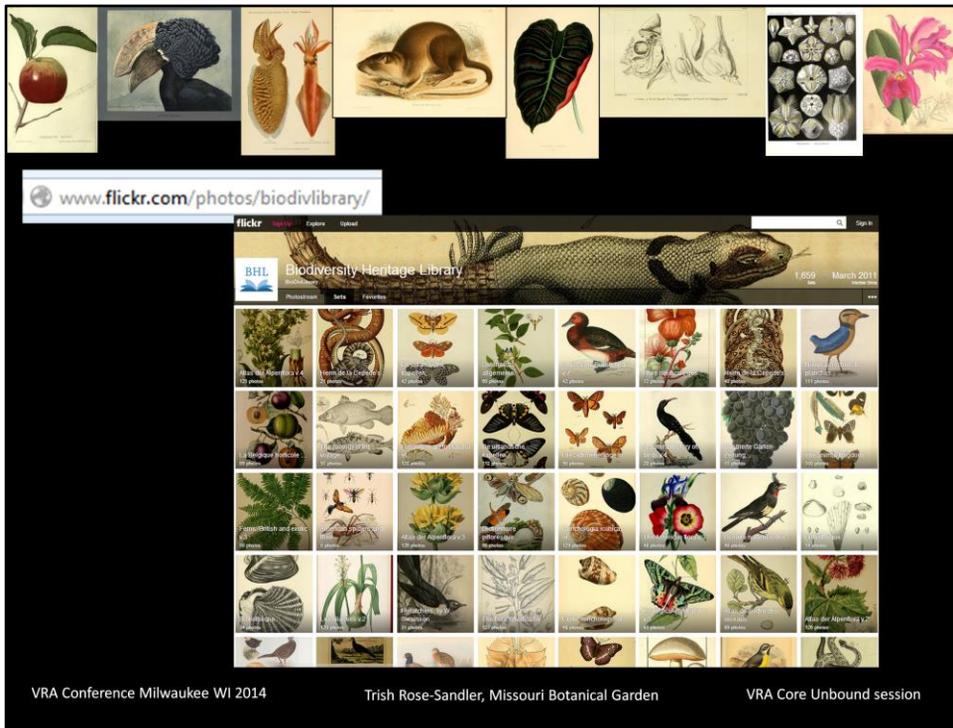
Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

The Art of Life project really grew out of a BHL problem statement

We had a critical mass of textual content online, BHL users knew there were amazing images within the BHL pages but there was no easy way to find them other than opening up a BHL book or volume and scrolling through page by page to find illustrations. There is no descriptive metadata attached to the illustration that would tell you the content of the image, date when they were created or who was involved in their creation.

We also wanted to expand BHL to new audiences and domains and felt the illustrations were the pathway for doing that. Knew these illustrations would be of interest not only to biologists, but also to artists, and historians in both the arts and science, educators; librarians/curators.



One way we've tried to address the need for image discovery is by pushing selected images to Flickr. We have created a BHL account in Flickr and pushed over nearly 90,000 images so far but this is all a very manual process that takes considerable staff time. We estimate that we have millions of illustrations within BHL so this manual process does not scale well. The address is flickr.com/photos/biodivlibrary/ and I would encourage you to explore and have your patrons explore this incredible collection of natural history images. These are all public domain images so can be re-used for any purpose without permission.



5 Primary Objectives of Art of Life

Objective 1: Define an appropriate metadata schema for natural history illustrations

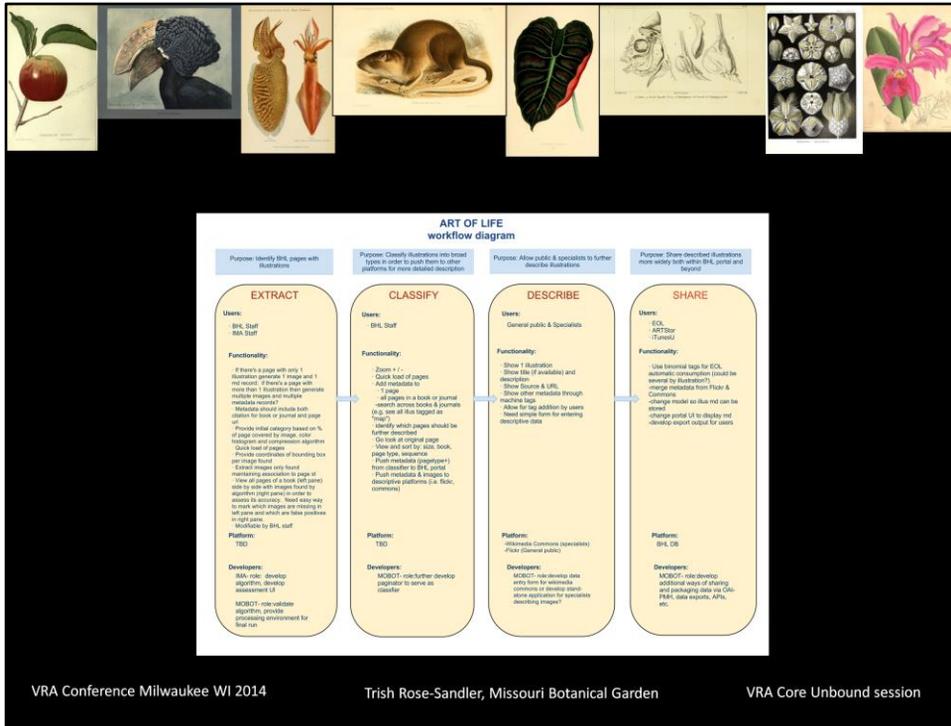
Objective 2: Build software tools to automatically identify illustrations in the BHL corpus

Objective 3: Enhance existing tools to enable the initial sorting, viewing, and editing of these identified visual resources.

Objective 4: Integrate tagging applications to enable a community of users to edit descriptive metadata for the illustrations

Objective 5: Integrate the descriptive metadata generated by users back into BHL portal both for access and preservation

Object 1 is what I will focus on in this talk.



This is the Art of Life workflow diagram which identifies the 4 processes the illustrations will go through as they move through each stage of the workflow. They include: Extract, Classify, Describe, and Share.

The Extract stage is where BHL pages will be run through the algorithms to identify which pages contain illustrations, whether they be full plates or only a section of the page. At the Classify stage, the pages with illustrations will be tagged by Art of Life staff as being one or several broad types such as drawing/painting, photograph, diagram, or map. For the Describe stage, the illustrations will be pushed into platforms such as Flickr and Wikimedia Commons where both the general public and specialists can describe them in much greater detail such as adding a title, creator, date (if different from date of publication), and subjects. Wikimedia Commons is where the schema can play a role. Because Wikimedia allows you to create templates we can provide guidance to more expert taggers on what information to record and how to record it. In the Share stage, the metadata contributed in Flickr and Wikimedia Commons will be ingested into the BHL portal both for preservation and discovery. Because many of these new audiences don't know about BHL and wouldn't go to the BHL platform to discover the illustrations we also want to push the illustrations out to environments where those audiences are familiar with: Encyclopedia of Life, ARTstor, and even iTunesU where we already have some themed collections at the book level.



Art of Life Schema

Needs to support three purposes:

- 1) to enable the discovery, description and use of the identified images by artists, biologists, humanities scholars, librarians, and educators
- 2) to make BHL's metadata and images available to other platforms
- 3) to import crowdsourced metadata generated in other platforms back into BHL.



Schema landscape review

- VRA Core 4.0 (art image community)
- LIDO (museum community)
- Dublin Core (Web community)
- Darwin Core (biodiversity community)
- Audubon Core (biodiversity community)

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

Our initial step was to do a landscape review of what metadata standards already existed that might suite the needs of natural history illustrations. We focused in on 5 element sets:

VRA Core - which needs no introduction in this community.

LIDO was designed for museum objects and has begun to supercede CDWA.

Dublin Core of course is the default standard to consider for any online digital repository

Darwin Core - is a standard to describe biological specimens and their digital surrogates

Audubon core – is a standard to describe multimedia resources about biodiversity such as images, audio or video

We determined that VRA Core really was found to be the best fit for the natural history illustrations. Its elements and attributes were mostly closely aligned with the types of information we wanted users to record. But also because its relationship of works to one or more images fit nicely with the book structure which often contain one or more illustrations on a single page. The only thing the VRA Core lacked was a way to record an acceptedName and CommonName for a species. VRA Core has a subject attribute type of scientificName but Taxonomists need more specificity.

Darwin Core was able to fulfill this need and so we borrowed 2 elements from that schema.



<http://tinyurl.com/9hm7nsb>



...Notes & News from the BHL Staff

THURSDAY, AUGUST 30, 2012

Interested in improving access to millions of digital images?

The Biodiversity Heritage Library (BHL) has made significant contributions to the research community over the past five years. One of the largest has been to successfully digitize a significant mass of biodiversity literature (nearly 40 million pages) and make that literature available for open access and responsible use as a part of a global "biodiversity commons."

Yet despite this success, BHL continues to have several challenges with access to and distribution of its digitized content. One of which is the ability for users to easily find the millions of natural history illustrations hidden within the pages of the BHL corpus. Only a small percentage of pages have been tagged as having illustrations because this is currently a labor-intensive manual task (a small selection of the diversity of BHL images can be viewed in its Flickr stream at <http://www.flickr.com/photos/biodivlibrary/sets/>). Once tagged, users still cannot search on the illustration's content using criteria such as species names, dates, and creators because images have not been described at that level of detail.

The NEH-funded Art of Life project has set out to solve this problem both by developing an algorithm to automatically identify which pages contain illustrations and by creating a schema to further classify and guide the description of the illustrations so as to increase their accessibility to users. Once the algorithm tags pages containing illustrations, they will be pushed out to image-sharing platforms such as Flickr and Wikimedia Commons for crowdsourcing of the descriptions. The schema will provide guidance on the recording of fields and their values. [View file](#)

ABOUT BHL

A consortium of major natural history museum libraries, botanical libraries, and research institutions have joined to form the Biodiversity Heritage Library. The participating libraries have over two million volumes of biodiversity literature collected over 200 years to support the work of scientists, researchers, and students in their home institutions and throughout the world.

TWITTER UPDATES

UPTWEET

 [View more](#)

FACEBOOK SHARE



VRA Conference Milwaukee WI 2014 Trish Rose-Sandler, Missouri Botanical Garden VRA Core Unbound session

We have a draft of the schema which is out for public review . The easiest way to get to it is from a blog post <http://tinyurl.com/9hm7nsb> we did about the schema which links to the draft and has a brief survey for feedback. We would love to get feedback from the VRA community.



ART OF LIFE SCHEMA ELEMENTS <small>red =required</small>
Title
Type
Date
Copyright
Source
Agent
Subjects
Description
Inscription

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

We ended up with 9 elements total, 7 of which came from VRA Core 4.0 and 2 which came from Darwin Core. The elements in red are required

The Darwin Core fields are not shown in this slide but they subsumed under the subject element. While VRA core 4.0 does have a type of “scientific name” in the restricted list of types for Subject it wasn’t quite granular enough for the needs of biologists. Therefore we used the Darwin Core elements of vernacularName, and acceptedNameUsage to record the various taxonomic name variations under which a species can be known.



Example of illustration described using Art of Life schema

	Title	Stictospiza formosa
	Type	Paintings
	Date	Publication: 1898
	Agent	Author: Arthur G. Butler (1844-1925) Illustrator: F.W. Frohawk (1861-1946)
	Description	A pair of finches with green and yellow bodies resting on reeds
	Subjects	Birds, finches Scientific name: Amandava formosa Vernacular Name: Green Avadavat or Green Munia Accepted Name: Amandava formosa (Latham, 1790)
	Inscriptions	bottom center: Green Amaduvade Waxbill (Stictospiza formosa)
	Rights	Public domain

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

Here is an illustration described using the schema

If you look at Subjects its clearer here how that information would get recorded where you might have a common name of birds and finches but then be able to record the appropriate taxonomic terms for those species as well

Current status of the schema is that we are developing an application profile for it. I don't have a finished profile to show you today but I thought it would be helpful to walk you through steps of creating one.



Application Profiles

Definition

“application profiles as schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application”

“The Resource Discovery Framework (RDF) syntax has provided the enabling technology for the combination of individual elements from a variety of differing schemas, thus allowing implementors to choose which elements are best fit for their purpose.”

Sept 2000 article in Ariadne “Application Profiles: Mixing and Matching Metadata Schemas”
<http://www.ariadne.ac.uk/issue25/app-profiles>

API is basically a way to specify your location implementation of a schema or schemas.



Why create an Application Profile?

- Metadata schemas are rarely adopted “as is” – local needs related to our users, cataloging systems, or display systems require adaptations
- specify what elements/attributes from a schema we will use and how
- identify elements/attributes that don't exist in any schemas but that we need

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

Metadata element sets, schemas are great for guiding us on what to consider when needing to describe resources or collections that will be accessed not only within our individual institutional environments but also with a wider audience on the Web. They help to ensure a certain level of interoperability when we begin sharing those records in aggregated environments and help with consistency across different catalogers. But it is rare for anyone to adopt them “as is” in full or in part. We all have practical needs imposed either by our local users needs, cataloging systems, or display systems that have to be met. Application profiles allow us to specify what elements/attributes from a schema will be used and how and create new elements/attributes that may not exist in any current schemas but are needed for our particular use.



What should be included in an Application Profile?

From *Framework for Dublin Core Application Profiles*

- Functional Requirements
- Domain Model
- Description Set Profile and Usage Guidelines
- Syntax Guidelines and Data Formats

Found this list from the *Framework for Dublin Core Application Profiles*. While written specifically for DC the principles are universal across metadata sets



What should be included in an Application Profile?

Functional Requirements – example from Art of Life

- specify descriptive information needed by the science and art communities in searching and using natural history illustrations
- Guide the creation of a template in Wikimedia Commons on how to record information about the natural history illustrations found within BHL
- Export those descriptions from Wikimedia Commons back into BHL portal for discovery.
- Within BHL portal want to view illustrations across all books related to: a) subject area b) geographic region Will want to view all illustrations for a specific species
- Exchange those descriptions and images more widely beyond the BHL repository e.g. ARTstor, Encyclopedia of Life. Need records to interoperate well in aggregated environments, therefore its critical that Art of Life chooses elements and attributes based on internationally accepted standards

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

Functional Requirements - Describes what a community wants to accomplish with its application



What should be included in an Application Profile?

Functional Requirements

The screenshot shows the Wikimedia Commons page for the template 'Information Art of Life'. The page is titled 'Template:Information Art of Life' and includes a 'Documentation' section. The documentation text states: 'This template is used to structure the metadata used by the Art of Life project of the Biodiversity Heritage Library. We (the Art of Life developers) plan to eventually use this template to automatically upload images from BHL into the Wikimedia Commons directly, but for now we hope that it can be used to fully describe natural history illustrations and images in the Commons. We also plan to develop other templates which will closely integrate with existing natural history organization schemes on the Commons, such as Category:Phylogenetic tree of life. Examples of images using this template are available on this template's gallery page; to see a complete list of images using this template, use this link.' Below the documentation is a 'Contents' table of contents and a 'Usage' section with code examples for using the template.

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

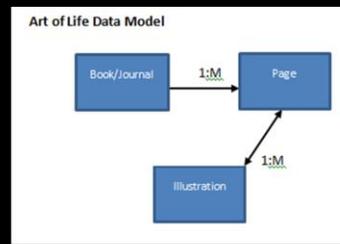
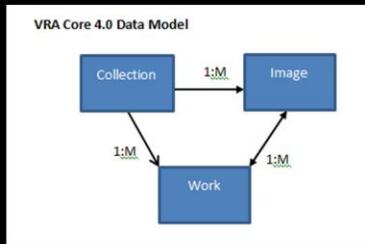
VRA Core Unbound session

This is a screen shot of a template we created in Wikimedia Commons for guiding users how we want them to record the information



What should be included in an Application Profile?

Domain Model



VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

Date Models - characterizes the types of things described by the metadata and their relationships

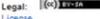
The Art of Life data model is very close to VRA data model with some variations. With VRA we have the primary entities of Collections, Works and Images. Collections contain works which contain images. In Art of Life the entities are Books/Journals, Pages, and Illustrations. Books contain Pages and pages contain illustrations. Pages really serve as the surrogates or images in this case because that is the level at which BHL digitizes. Pages then contain one or more Illustrations. The focus of description in VRA Core is primarily the Work entity and in Art of Life it is the Illustration entity.



What should be included in an Application Profile?

Description Set Profile and Usage Guidelines

ART OF LIFE SCHEMA [DRAFT]

Title: Art of Life schema
Date Issued: Aug 31, 2012
Date Modified:
Abstract: This schema is designed for the cataloging and description of natural history illustrations from the [Biodiversity Heritage Library](#) (BHL), as part of the NEH funded [Art of Life project](#). The schema needs to support the needs of a wide variety of audiences, including 1) Artists 2) Biologists 3) Humanities Scholars 4) Librarians 5) Educators
Creators: Gaurav Vaidya, University of Colorado Boulder; Trish Rose-Sandler, Missouri Botanical Garden
Contributors: William Ullate, Missouri Botanical Garden; Robert Guralnick, University of Colorado Boulder; Andrea Thomer, University of Illinois at Urbana-Champaign
Legal:  Art of Life schema is licensed under a [Creative Commons Attribution-ShareAlike 3.0 Unported License](#).
Current Version: 1.0 [draft]
Replaces: n/a

Below is an overview of the schema elements, their definitions and use. The values for each element are shown in both their displayed and indexed form. Display is meant for human viewing while indexing is for search, cataloging and retrieval of the illustrations via computer processing.

At the end of the document is a crosswalk designed to demonstrate how this schema maps to elements found in national and international schema standards. We also provide a set of visual resources that have been [made-up according to the schema](#) so you can see how it performs in practice. We intentionally chose a wide variety of illustrations from the BHL corpus to test the robustness of schema. Suggestions on how both metadata terms and their values are expressed in the Art of Life schema are appreciated.

Alpha List of elements
Red=required

Agent	Copyright	Date	Description	Inscriptions
Source	Subjects	Title	Type	

VRA Conference Milwaukee WI 2014

Trish Rose-Sandler, Missouri Botanical Garden

VRA Core Unbound session

The Description set profiles and usage guidelines specify the terms to be used and rules for use – in a sense we’ve already done this in schema draft but would need to get more proscriptive.

E.g. date or personal name formats, controlled vocabs, whether we want them to follow the VRA Core 4 restricted or unrestricted schema, whether to record display or encoded data or both.



What should be included in an Application Profile?

Syntax Guidelines and Data Formats

Art of Life? TBD

- XML
- RDF/XML

- Darwin Core is in RDF and VRA Core 4 in XML
- Task force looking into RDF ontology for Core 4

To turn an application profile into something that will function in a software environment you need to specify how the metadata will be encoded. You do this by laying out your syntax guidelines and data formats.

There are lots of options – Art of Life will probably use XML, RDF or a combination Darwin Core is currently in RDF but VRA Core 4 is only available in XML. There is a task force right now to look into an RDF version for Core 4.



Art of Life team

PI

Trish Rose-Sandler, Missouri Botanical Garden

Algorithm development

Ed Bachtá, Charlie Moad, Kyle Jaebker, Indianapolis Museum of Art

Classification User Interface

Joel Richard, Smithsonian Institution Libraries

Schema design

Gaurav Vaidya and Robert Guralnick, University of Colorado, Boulder

William Ulate, Missouri Botanical Garden

Programming

Mike Lichtenberg, Missouri Botanical Garden

Consultants

Doug Holland, Chuck Miller, and Mike Blomberg, Missouri Botanical Garden; Chris Freeland, Washington University (former PI for Art of Life)