

Camera Trap Data Preservation: Wildwatch Kenya

Ariel Hammond, Tara Fitzgerald, Joan Hsiao, Janice Bially Mattern, Kimberly
McCauley, Estelle Reed, Millicent Ryan, Erica Saito, and Lindsey Schaefer

17:610:596:90 Data Curation

Rutgers - The State University of New Jersey

School of Communication and Information

Instructor: Michael Lesk

Fall 2018

Abstract

Although easily identifiable among Africa's famous mammals, the behaviors of the reticulated giraffe are little understood, and it is becoming increasingly evident that the species is on the brink of extinction (Wildwatch Kenya, n.d.). Wildwatch Kenya (WWK) is a collaborative initiative with the goal of gaining increased comprehension into the behaviors of giraffes to increase population sizes. The initiative collects images from camera traps, aerial data, and geological data to study the population size of groups of reticulated giraffes in Kenya's Namunyak Community and Loisaba Conservancies. The data is sent to San Diego Zoo Global (SDZG) where it is uploaded to Zooniverse, a citizen science website that crowdsources the task of tagging the images. Currently SDZG does not have a digital curation and preservation plan in place for the data. A team of nine graduate students enrolled in Rutgers University's Master of Information program worked with SDZG employees to create a set of customized guidelines for curating and preserving the WWK camera trap data. This paper outlines those guidelines focusing on content selection, rights and open access, data cleaning, description and metadata, file formatting, data organizing, and data compressing, with attention throughout to sustainability and risk management. The paper concludes by recommending specific archiving and access repositories most appropriate for WWK data and SDZG objectives.

Keywords: Wildwatch Kenya, reticulated giraffe, camera trap, data curation, San Diego Zoo Global, Rutgers University

Table of Contents

Introduction	3
Content Selection	5
Rights and Open Access	7
Data Cleaning	10
Section 1: Data Cleansing Guidance	10
Section 2: Four levels of Data Documentation	11
Section 3: Four Levels of Wildwatch Kenya Data Documentation	12
Section 4: Suggestions for the Wildwatch Kenya Team	14
Description and Metadata	15
Organizing Data	19
Glacier	21
Quality Control	22
Creating Vaults	23
Tagging Vaults	23
Uploading Archives to Vaults	23
Retrieving Archives	24
Recommendations for Wildwatch Kenya	24
File Formatting	26
Data Compressing	30
eMammal	31
Other Repositories Considered	34
Conclusion	38
References	41
Appendix 1: Manual for Submitting Legacy Data to eMammal Data Repository	48
Appendix 2: Data Cleanse for Camera Trap Images Inventory	53
Appendix 3: Codebook for Inventory of Camera Trap Images	55
Appendix 4: Sample Dublin Core Application Profile	57
Appendix 5: XML Template and Example	68
Appendix 6: Readme Example	70
Appendix 7: Glacier Storage Index	71
Appendix 8: Legacy Metadata Template	72

Camera Trap Data Preservation: Wildwatch Kenya

Launched in 2016, Wildwatch Kenya (WWK) is a “collaborative conservation initiative to study reticulated giraffe” populations in “two sites in northern Kenya: the 56,000-acre Loisaba Conservancy, and the 800,000-acre Namunyak Community Conservancy” (Wildwatch Kenya, About: Research). With the goal of reversing the current declining population trend of the reticulated giraffe, the WWK team uses movement-activated camera traps, aerial data from spotters in planes, and geolocated data from observers on the ground to develop a better understanding of how giraffes live in their natural habitat.

The images from the camera traps are uploaded from memory cards onto an external drive and then shipped to a team at San Diego Zoo Global (SDZG). Copies are then uploaded to the citizen science website Zooniverse, in order to crowdsource the large task of tagging the images. The on-the-ground observational data uses a laser and compass to calculate the coordinates of an animal. This data is then entered into a GIS software called Survey123, and sent via CSV to San Diego.

Currently, SDZG does not have a comprehensive preservation plan for camera trap datasets. Additionally, although the WWK team hopes to make the data open and accessible to the public in the long term, SDZG does not have a comprehensive plan to ensure that the WWK data can be made accessible to others. Since the long term value of data to other researchers depends upon proper curation and preservation, our objective in this project was to recommend a course of preservation and data curation that SDZG could implement to support WWKs goals (Ruggles, 2018). Working with SDZG employees, Ariel Hammond, Tara Fitzgerald, Joan Hsiao, Janice Bially Mattern, Kimberly McCauley, Estelle Reed, Millicent Ryan, Erica Saito, and

Lindsey Schaefer, a group of nine students from Rutgers University's Master of Information program, created the following set of guidelines for SDZG on how to best curate and preserve WWK data as their final project for Professor Michael Lesk's Fall 2018 Data Curation course (17:610:596).

After carefully reviewing several repository options, we recommend a two-pronged strategy. For immediate preservation needs, we identified Amazon Glacier, an archival cloud storage service offered by Amazon Web Services (AWS) in which the SDZG staff can store data with relatively little object-level data curation. For the longer term goal of making the data available to other researchers, we identified eMammal as a compelling possibility. eMammal is a Smithsonian-run project that both preserves and makes openly accessible camera trap images of wildlife (eMammal, 2017, About eMammal).

This paper is divided into seven main sections. The "Content and Selection" section of this paper outlines the types of content the WWK uses. A subsection of "Content and Selection," "Rights and Open Access" discusses ownership of the data rights and open access. The "Data Cleaning" section is divided into four subsections that outline general guidance on data cleaning, the four levels of data documentation and how they apply to the WWK project, as well as suggestions and questions for the WWK team about data cleaning. The "Description and Metadata" section outlines our reasoning for recommending WWK use the Dublin Core Metadata schema. Information about how to create a Dublin Core Application Profile is also discussed in that section. Focusing on image organization, the "Organizing Data" section describes WWK's current organization system, current inconsistencies (folder organization, size, and naming conventions), and lists our recommendations for improvement. The "Glacier"

section of our paper encompasses both a detailed description of the Glacier storage system, file formatting, and data compression. The “File Formatting” subsection of “Glacier” explains the reasoning behind our suggestion that WWK should convert everything for the archive to TIFF (Tagged Image File Format) for compression and to keep the JPEGs (Joint Photographic Experts Group) on SDZG’s hard drives as they are. Since Glacier charges downloads by the GB, the “Data Compressing” subsection notes the importance of WWK reducing the size of the files to upload them to Glacier to optimize the project resources. Similarly to the “Glacier” section, the “eMammal” section of this paper provides a detailed description about why we chose to propose eMammal as a suggestion for this project. In addition, this section also lists all other repositories that were considered for this project and our reasoning behind why we did not choose them. The sustainability and risk management of our suggestions for WWK are outlined in the Conclusion of this paper. This paper is reinforced by eight appendices located after the references list. We believe in WWK’s goal of trying to ensure a future for the Kenyan reticulated giraffe population, thus every suggestion outlined in this paper was selected with that goal in mind.

Content and Selection

The content collected by WWK is made up of images, metadata from citizen tagging in the Zooniverse site, and GIS data from on the ground observations, which is sent to SDZG and derived data. Referencing literature and using similar conversation projects as a guide (Swanson, Kosmala, Lintott, Simpson, Smith, & Packer, 2015; LaFleur, Sauther, Cuzzo, Yamashita, Youssouf, & Bender, 2014), our research has demonstrated that the decision to include supplemental documents will be heavily influenced by the digital storage options available. As we are recommending two simultaneous pathways, comprehensive project

preservation in Amazon's Glacier, and selected data preservation accompanied by a forward facing portal in eMammal, two different content selection criteria must be followed.

The goal for Glacier is to have a comprehensive copy of the data, so that, in the event of emergency, the data can be completely understandable by those unfamiliar with the project. Thus, for Glacier, we recommend preserving substantial supplementary documentation, including a ReadMe file, maps of the camera trap distribution grids, organizational image finding aids, a codebook to decipher the camera trap image inventory and related WWK data, and official reports and documents.

eMammal requires a data package that adheres to the Camera Trap Metadata Standard, "which captures information about study design, the type of camera used, and the location and species names for all detections in a standardized way"(Forrester, O'Brien, Fegraus, Jansen, Palmer, Kays, Ahumada, & McShea, 2016). The eMammal team does not accept supplemental documentation such as maps, official reports, or additional notes on the data. As a forward-facing archive, which is focused on making published data open access, they take the position that supplementary documents should be maintained by the original researcher or included in the dataset's corresponding published paper (J. Zhao, personal communication, November 14, 2018). eMammal must consider available resources when making content selection decisions, and thus cannot accept supplemental documentation for all of the datasets that they house.

In determining content to be made publically available, WWK must also consider issues of privacy and security. Both eMammal and Zooniverse ensure that location information

concerning endangered and threatened species is masked and images of humans are not made publically available.

Rights and Open Access

Copyright and Data

Copyright “is a form of legal protection for intellectual and creative works” (Rutgers, 2018). In the case of data, there are “layers of copyrights” (Carroll, 2015, p.4). Copyright is “founded on certain science-friendly policies” that allow for commonly known facts and ideas, including “raw observational and experimental data” to be shared with no restrictions (Carroll, 2015, p.3).

Databases, however, and any process that interprets the raw data can have copyright protection. In particular, “annotations, visualizations, and other forms of metadata can receive separate copyright protection if they are sufficiently original” (Carroll, 2015, p. 4). When “there is a copyright layer to a dataset or database, the owner(s) of the copyright(s) associated with this layer would be the one(s) who chose how to organize, arrange, annotate, or visualize the data” (Carroll, 2015, p.7). It is important to clarify ownership in order to promote data reuse and “it is incumbent on the owner(s) of these rights to mark the data with the associated permissions.” (Carroll, 2015, p. 9). When the images are published to eMammal, eMammal then becomes the copyright holder. However, eMammal uses a Creative Commons licenses which allows WWK to publish the photos elsewhere provided they follow the terms of eMammal’s Creative Commons license (eMammal, 2017, Frequently Asked Questions).

Open Access

There are degrees of Open Access depending on the licensing associated with the data. While data may be “open access” in the sense that anybody can access the data for free, the creators and owners of the data need to stipulate how others can use the data. Otherwise confusion can arise over whether or not the data falls under the restrictions of copyright. While facts are open domain, research data can encompass both facts and authorship where “the amount of creativity or editorial discretion to create a work of authorship is quite minimal. As a result, some aspects of a dataset are likely to have a copyright attached to them. Copying the whole dataset will involve copying the copyrighted layer”(Carroll, 2015, p.3). In order to maximize the reuse and sharing of data, licenses are an effective tool to communicate to potential users “what they are allowed to do with the data” (Ball, 2012, p.3). There are two organizations that offer licenses that address Open Access Data:

1. **Creative Commons.** This is the best known licensing provider and offers “simple but robust licenses for creative works” (Ball, 2012, p.6). It offers a Public Domain Dedication (CC0 1.0) license which gives “creators a way to waive all their copyright and related rights in their works to the fullest extent allowed by law” (Creative Commons, n.d.). This license is “suitable for use with data” (Ball, 2012, p.10). It also offers an Attribution 4.0 International (CC BY 4.0) license which “lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation” (Creative Commons, n.d.). This license may be suitable for figures and papers that are created from the data.

2. **Open Database Commons.** This project was created to provide licenses designed particularly for use with databases. It offers a Public Domain Dedication and License (PDDL) that “places the data(base) in the public domain (waiving all rights)” (Open Domain Commons, n.d.). It also offers the Open Data Commons Attribution License (ODC-BY) that allows copying, distributing, and modifying the database as long as the original database is attributed in the manner specified by the license.

There is no definitive answer as to which licenses between Creative Commons and Open Database Commons are preferable. Many advocates of open data recommend using either the CC0 or PDDL license in order to make the data as open as possible and to avoid the problem of “attribution stacking” where “a derivative work must acknowledge all contributors to each work from which it is derived, no matter how distantly. If a dataset is at the end of a long chain of derivations, or if large teams of contributors were involved, the list of credits might well be considered too unwieldy.” (Ball, 2012, p.4). However, WWK may prefer to take a more conservative approach. eMammal, for example, uses the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license which allows the sharing and adaptation of content but requires attribution and does not allow use of the material for commercial purposes. eMammal posts this license information clearly on its webpage for accessing its photos. Also on its website, eMammal notes “any publication, sharing, or use of Data accessed through the Web Site must clearly and unambiguously acknowledge eMammal as a reference source, acknowledge the projects from which Data are obtained, and acknowledge the owners of the specific project data” (eMammal, 2017, Data Provider Terms and

Conditions). It may be helpful for the WWK team to consult with the SDZG's legal department that handles copyright issues for its input.

Data Cleaning

Data cleaning presupposes thorough dataset documentation because documentation defines the variables and values, which allows for clearer identification of errors (Cadman & Gonzalez, 2014, p.40). This discussion of data cleansing has been divided into four subsections: Section 1: Data Cleaning Guidance, which compiles general guidance on cleaning data; Section 2: Four Levels of Data Documentation, which defines and sketches the four levels of data documentation outlined in both the “Manual for Submitting Legacy Data to eMammal Data Repository”, included in Appendix 1 of this paper, and “An Open Standard for Camera Trap Data” (Forrester et al., 2016, p.3); Section 3: Four Levels of Wildwatch Kenya Data Documentation, which describes the content needed at each level of the WWK project; and lastly, Section 4: Suggestions for the Wildwatch Kenya Team, which covers errors we noticed when reviewing WWK's current data, and suggestions for future entry and cleaning policies.

Section 1: Data Cleansing Guidance

Arthur Chapman, author of *Principles and Methods of Data Cleaning*, defines data cleansing as the “process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions” (Chapman, 2005, p.1). This process is important because it prevents errors from occurring, which “is cheaper and more efficient” than trying to identify and correct errors after the fact (Chapman, 2005, p.1). It has been noted that “primary species data encompass a whole range of data – from museum and herbarium data, through observational data (point-based, regional or area-based,

and systematic or grid-based), to survey data, both systematic and other” (Chapman, 2005, p.1).

The data compiled by the WWK is composed of all the data forms listed above excluding survey data. Data cleansing needs to be done for all data gathered including camera trap data, observational data and Zooniverse data, file names, documentation for all data, and all metadata created for WWK.

We have identified the best basic practices for data cleansing as checks on format, completeness, reasonableness, and limits, as well as a review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and an assessment of data by subject area experts (e.g. taxonomic specialists). Best practices in workflows of data cleansing include defining and determining error types, searching, identifying, documenting, and correcting error instances, as well as modifying data entry procedures to reduce future errors.

Section 2: Four Levels of Data Documentation

This section defines the four levels of data documentation: Project, Deployment, Sequence, and Image. The project level of documentation include metadata of the sort used to catalog a book, article, or other work. A project is “a scientific study that has a certain objective, defined methods, and a defined boundary in space and time” (Forrester et. al, 2016, p.3). The Project level “contains information about the project name, design, and objectives. Projects can either be of limited duration or be long-term monitoring” (Forrester et. al., 2016, p.3).

Deployments are a camera in a specific location for a specific time period. Documentation of deployment “contains all information related to specific locations where cameras are placed, including separate identifiers for a deployment of a camera and for deployment locations”

(Forrester et al., 2016, p.3). Relevant data at the deployment level may include information about the camera equipment, placement, settings, and number of deployments.

The sequence level of data documentation “is a group of images that are all captured by a single detection event, defined as all pictures taken within 60 seconds of the previous picture or another time period defined by the Project. A sequence can either be a burst of photographs or a video clip” (Forrester et al., 2016, p.3). In addition, “the Sequence section contains metadata for groups, or sequences, of images that are captured as part of a single detection event of an animal or group of animals” (Forrester et al., 2016, p.3). The image level describes “an individual image captured by a camera trap, which may be part of a multi-image sequence” (Forrester et al., 2016, p.3). Just as the sequence level contains metadata for groups of images, the Image level contains metadata for individual images (Forrester et. al., 2016, p.3).

Section 3: Four Levels of Wildwatch Kenya Data Documentation

Using the definitions outlined above, this section provides more detail on the content of the Project, Deployment, Sequence, and Image levels of the WWK data.

Project level. This level requires information on the project name, objectives, owner, principal investigator, and contact information, as well as the project latitude and longitude for the central project point. This level of documentation is complete.

Deployment level. This level requires information on the deployment of each camera, at a specific location and for a specific period of time. Deployment beginning and ending dates, actual latitude and longitude, and camera failure details are all needed. Much of this information already existed in the WWK inventory, and simply needed proper cleansing and organization.

The cleaning of this level of documentation is also complete, examples of which appear in Appendix 2.

For this level, as well as the Sequence and Image levels, eMammal also creates and includes a Camera deployment ID. This is a key that links the data across each level. At this level eMammal also requires information on the camera(s), such as Camera ID and settings, as well as whether bait was used. These additional levels of data will be completed in advance of the trial crosswalk.

Sequence level. For this level, data is needed about each sequence or burst of photos, such as the Image Sequence ID, the date and time of the first and last images in the sequence. In addition data about the captured image is also needed, such as animal count, species genus name, species common name, and IUCN threatened species status. Information regarding threatened species statuses can be accessed via the IUCN website (IUCN, 2017). This level will not be included in Glacier, and will be addressed in advance of the eMammal crosswalk.

Image level. The information outlined in this subsection is needed for each image within a sequence and deployment. Although this level of description is not required for using Glacier, this section will serve as advice for the WWK team on whether there is more data that should be gathering at the image level and/or how to record it. Information is needed on Image Sequence ID, Image ID, Location, Image File Name, Photo Type, Photo Type Identified by, and Date/Time Captured. Further information regarding the Genus species, Species Common Name, Integrated Taxonomic Information, and the IUCN ID of subjects in the image is necessary for this level of data documentation. Information on Embargo Period, Restrictions on Access, and Image Use Restrictions is also required.

Some image level metadata may be possible to retrieve from Zooniverse. To increase the potential usability of Zooniverse data, we recommend the WWK team ensure that the data downloaded from Zooniverse is clean (e.g., uniform names; no random spaces etc.; what to do about blanks/nulls). Additionally, to ensure that the Zooniverse data is intelligible to those who have not directly administered that system, we recommend that a SDGZ Zooniverse administrator create a codebook for the Zooniverse data. This codebook can be similar to that which we have created for the WWK image inventory spreadsheet (Appendix 3) but should specify the meaning of the Zooniverse column variables and cell values. As with the WWK image inventory spreadsheet, the Zooniverse codebook should be integrated into each Zooniverse CSV file for easy reference. Matching the Zooniverse data to individual images should not be a problem, as it includes the original folder and image name. WWK Zooniverse tagging is also very controlled, and species common names are fairly granular, offering specific tags such as “eland” instead of the more generic “antelope.” However, the fact that the WWK Zooniverse site does not include taxonomic information may be a challenge. For some tags, such as aardvark, the taxonomic name is easy to determine. Yet other tags, such as livestock, could represent any number of species. Parsing out taxonomic information for eMammal will take time, and multiple scenarios are being discussed to determine how to address this

Section 4: Suggestions for the Wildwatch Kenya Team

With regard to data cleansing, we have cleaned the data currently available by standardizing table formats, converting degree/time latitude and longitude coordinates into decimal coordinates, specifying deployment beginning and ending dates, and identifying missing variables (further information in Appendix 2). We have also created a codebook that describes

the meaning of the column variables and cell values used the WWK camera trap images inventory spreadsheet. This codebook, which can be found in Appendix 3, should be integrated into the inventory file as a readme sheet.

Going forward, we think that it is in WWK's best interests to follow our standardized variables and coding rules. Although "correcting errors in data and eliminating bad records can be a time consuming and tedious process" (Chapman, 2005, p.2), improving the quality of the data so that it is able to be used by other researchers will assist the initiative in its goal of helping giraffes and reversing the negative population trend they are currently experiencing .

Description and Metadata

Metadata, that is data that provides information about other data, fuels effective search and retrieval. The metadata used to best describe a collection of data is dependent upon the nature of the data itself, the intended users of said data, and the amount of time/resources available to devote to metadata creation. Dublin Core, "a vocabulary of fifteen properties for use in resource description," was selected as the suggested primary metadata schema for WWK to use when describing the camera trap images. This particular schema was selected because "its elements are broad and generic, usable for describing a wide range of resources" (Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2012). The broad nature and wide use of the schema mean it will be easier to implement on an already existing project, such as WWK. Additionally, Dublin Core has a high degree of interoperability which makes the schema an appropriate choice to begin describing a large quantity of images. "The adoption of a common metadata format (such as Dublin Core) opens the way to partnerships between libraries that either have adopted this format, or who are willing to crosswalk - that is, to map their existing

metadata - to that standard format” (Khoo & Hall, 2010, p.361). Properly describing the data using a standard metadata schema is useful for both Glacier and other repositories. Since the purpose of metadata is to describe an item, using the DC metadata schema to describe files will allow both users and the WWK team to know more about the contents of the file. The Dublin Core schema is comprised of the following fifteen core elements:

- *Contributor*: An entity responsible for making contributions to the resource.
- *Coverage*: The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. Best practice suggests the use of the Getty Thesaurus of Geographic Names Controlled Vocabulary (The Getty Institute, 2017).
- *Creator*: An entity primarily responsible for making the resource.
- *Date*: A point or period of time associated with an event in the lifecycle of the resource. Best practices suggests the use of the W3 Consortium Date and Time Formats encoding scheme (W3 Consortium, 1997).
- *Description*: An account of the resource.
- *Format*: The file format, physical medium, or dimensions of the resource.
- *Identifier*: An unambiguous reference to the resource within a given context.
- *Language*: A language of the resource.
- *Publisher*: An entity responsible for making the resource available.
- *Relation*: A related resource.
- *Rights*: Information about rights held in and over the resource.
- *Source*: A related resource from which the described resource is derived.

- *Subject*: The topic of the resource. Best practice suggest the use of the Library of Congress Subject Headings Controlled Vocabulary (Library of Congress, n.d.).
- *Title*: A name given to the resource.
- *Type*: The nature or genre of the resource. Best practices suggest using the Dublin Core Type Controlled Vocabulary (DCMI Type Vocabulary DCMI Metadata terms, 2012).

In addition to the fifteen core elements used for Dublin Core description, cataloguers are able to further describe, or refine, items with Dublin Core Metadata terms. The following six DCMI terms were identified as those most likely to be used by WWK (DCMI, 2012, DCMI Type Vocabulary DCMI Metadata Terms):

- *accessRights*: Information about who can access the resource or an indication of its security status. Access Rights may include information regarding access or restrictions based on privacy, security, or other policies.
- *Extent*: The size or duration of the resource.
- *Medium*: The material or physical carrier of the resource.
- *rightsHolder*: A person or organization owning or managing rights over the resource.
- *Spatial*: Spatial characteristics of the resource.
- *Temporal*: Temporal characteristics of the resource.

We believe that the WWK team should create a Dublin Core Application Profile (DCAP) to specify and describe the metadata used for this project. A DCAP allows organizations “to document which metadata schemas they they have decided to use and how they have decided to use them,” (Corrado & Moulaison, 2017, p.77). Application profiles are declarations on how standard schemas are used during description (Heery & Patel, 2000). According to the Dublin

Core Metadata Initiative, a successful application profile will communicate five essential components. The profile will describe what a community wants to accomplish with its application (functional requirements), characterize the types of things described by the metadata and their relationships (Domain Model), enumerate the metadata terms to be used and the rules for their use (Description Set Profile and Usage Guidelines), and define the machine syntax that will be used to encode the data (Syntax Guidelines and Data Formats)(Coyle & Baker, 2009).

To best communicate how and why the metadata schema is used, a DCAP should include seven clarifications. These include all required elements, any permitted Dublin Core elements, any permitted Dublin Core qualifiers, any permitted schemes and values (e.g. use of a specific controlled vocabulary or encoding scheme), any library domain elements used from another namespace, any additional elements/qualifiers from other application profiles that may be used (e.g. DC-Education: Audience), and finally, any refinements of standard definitions. To help illustrate these suggestions, we have put together a sample DCAP (modeled after the Utah Academic Library Consortium Mountain West Digital Library Dublin Core Application Profile) which can be updated and modified to meet the specifics of the WWK data. This sample can be found in Appendix 4 of this paper.

In digital library systems, metadata files are increasingly using Extensible Markup Language (XML) as an encoding language (Corrado & Moulaison, 2017, p.70). This encoding language is used essentially to annotate text with additional information in order to be machine readable. Though not visible to the end user, the XML defines what the data does. “XML’s ability to exchange data on the Web and store that data in digital library and digital preservation systems makes XML-based languages an appealing choice for metadata formats in the LAM

communities” (Corrado & Moulaison, 2017, p.70). For reference purposes, we have created an XML Template and sample XML record, located in Appendix 5 of this paper.

Organizing Data

In order to improve clarity regarding the organization of data, we recommend that there should be consistency in the file/folder naming conventions and folder organization across all platforms, so that it is easier to track files across all data sources. This section will focus on image organization since the camera trap images comprise the majority of WWK data. A description of WWK’s current organization system, current inconsistencies (folder organization, size, and naming conventions), and recommendations to bring the folders into a uniform organization scheme will also be discussed in this section.

WWK camera trap images are currently stored on external hard drives and the organizational scheme consists of main folders and subfolders. The main folders are organized by hard drive and date range. Each subfolder contains images from one camera trap from a specific date range. When a camera trap produced more than 1000 images, the memory card stored additional images in a second sequential folder. Once that folder reached 1000 images, another folder was created by the memory card, and so on.

For example, the following folders contain images from camera trap L1, for date ranges February-May 2016, and the last component represents the sequential folder names that were created by the memory card. During this time period, camera trap L1 captured 15,516 images. The first 15 folders contain approximately 1000 images, while the last folder contains approximately 516 images. These folders are stored in the “L2_Feb-May_2016” subfolder because the images contained within are stored on a hard drive that the WWK team named “L2.”

Table 1	
Sub Folder Name:	Main Folder Name:
L1_February-May_2016-100EK113	L2_Feb-May_2016
L1_February-May_2016-101EK113	L2_Feb-May_2016
L1_February-May_2016-102EK113	L2_Feb-May_2016
L1_February-May_2016-103EK113	L2_Feb-May_2016
L1_February-May_2016-104EK113	L2_Feb-May_2016
L1_February-May_2016-105EK113	L2_Feb-May_2016
L1_February-May_2016-106EK113	L2_Feb-May_2016
L1_February-May_2016-107EK113	L2_Feb-May_2016
L1_February-May_2016-108EK113	L2_Feb-May_2016
L1_February-May_2016-109EK113	L2_Feb-May_2016
L1_February-May_2016-110EK113	L2_Feb-May_2016
L1_February-May_2016-111EK113	L2_Feb-May_2016
L1_February-May_2016-112EK113	L2_Feb-May_2016
L1_February-May_2016-113EK113	L2_Feb-May_2016
L1_February-May_2016-114EK113	L2_Feb-May_2016
L1_February-May_2016-115EK113	L2_Feb-May_2016

As seen in Table 1, the prevailing folder naming conventions are HardDrive_Month-Range_Year for main folder and CameraTrap_Month-Range_Year-IdFromMemoryCard for sub folders. However, there are a significant amount of sub folders that do not match this naming convention. We recommend that the sub folder names be updated to match the prevailing naming conventions as seen above, and that large folders be broken up into smaller folders (approximately 1000 images per folder) to

make the folder names and sizes uniform. Appendix 7, the Glacier Storage Index, documents folder organization inconsistencies and recommendations discussed here. Organization and Appendix 7 are also discussed in more detail with regards to Glacier storage in the next section. We also recommend that WWK follow the prevailing naming conventions and folder structure/size in the future as new images are obtained from the camera traps.

Glacier

There are two goals for WWK and Amazon Glacier archival storage. First, Glacier serves as a temporary storage place for the images until a front facing repository is selected. Second, WWK needs a secure backup for the images in the event of catastrophic loss, even after a front facing repository is selected. In this section, we outline the benefits of Amazon Glacier and pertinent information about working with Glacier storage. Finally, this section will end with recommendations on a storage and organization plan for the archival of WWK images.

Amazon Glacier is an archival cloud storage service offered by Amazon Web Services that provides low cost cloud storage that secure, durable, and flexible. Glacier leverages AWS' security platform to provide strict access control and regular security audits. Its storage achieves 99.999999999% durability by storing data on multiple devices in multiple locations (Amazon Web Services, 2018, Amazon S3 Glacier). As for flexibility, Glacier provides on demand but long term storage that does not require any upfront cost commitment or need to maintain a local physical storage infrastructure. While data storage in Glacier is low cost, there are also costs associated with retrieving files that are assessed by the number and size of files that are retrieved. Therefore Glacier storage best used to store “archival data,” as in data that is infrequently accessed but will or may be needed in the future (Lazier & Zhang, 2013). Thus, we believe that

\$1000 should be budgeted towards Glacier, to cover costs for ingest as well as to allow for emergency file retrieval in the future.

As with most AWS products, Glacier uses unique terminology to refer to the objects and storage structure. First, it stores “archives,” which are essentially files. Archives can be single files or a group of single files aggregated in a ZIP or TAR file. Second, it uses containers to organize archives called “vaults.” Vaults can be organized by data type, age, or any criteria a user would like. Vaults allow for container level access policies and are extremely durable, as they cannot be deleted while any files reside inside. Vaults do not automatically index the contents, so a master index must be kept locally and updated regularly. However in the event of index loss, vaults do offer an inventory of contents that can be used to recreate an index (Lazier & Zhang, 2013).

Quality Control

Glacier requires that a SHA-256 checksum, a type of Cryptographic Hash Algorithm, be generated for each archive and included with the archive when it is uploaded. The checksum is returned when retrieving and downloading archives, as well as used by Glacier when archives are moved internally. This ensures that archives remain unchanged throughout storage.

The WWK team will need to generate checksums prior to uploading files to Glacier. AWS provides instructions and sample Java and .NET scripts in the Glacier documentation (Amazon Web Services, 2018, Computing Checksums). Alternatively, the WWK can upload files through a client that generates checksums prior to uploading. Three options are FastGlacier, Cloudberry Explorer, and CrossFTP. These programs also offer serve as a graphical interface to

interact with Glacier without having to use the command line or the limited-functionality of the Glacier console.

Creating Vaults

The first step in migrating archives into Glacier is to create a Vault. Vaults are used to store and organize archives. Vaults can be created in the console, via command line, or through a client (FastGlacier, Cloudberry, etc.). When vaults are created, the user must specify a unique name for each vault along with a Glacier storage region. Data cannot be stored in Glacier without being stored inside a vault (Amazon Web Services, 2018, Creating Vaults in Amazon S3 Glacier).

Tagging Vaults

Vaults may be tagged with user-defined Key-Value pairs. These tags can be used to offer additional information about the contents of a vault, track costs, set policies, or manage access control (Amazon Web Services, 2018, Tagging Your Amazon S3 Vaults).

Uploading Archives into Vaults

While vaults may be created, deleted, and managed in the Glacier Console, files cannot be uploaded via the console. Files can be uploaded using the command line interface, REST API, Amazon's software development kits in various languages, or third party client. Archives under 4GB in size may be uploaded in a single transaction. Archives over 4GB (but not over 4TB) can be uploaded in parts using Glacier's multipart upload, where Glacier assigns a unique identifier to each part that is used to put the file back together once all parts are uploaded. When using the multipart upload, a checksum must be generated for each archive part. It is important to note that Glacier updates vault inventory once daily. Therefore, uploaded archives cannot be immediately

seen inside vaults and can only be seen once the inventory has been updated (Amazon Web Services, 2018, Uploading an Archive in Amazon S3 Glacier).

Retrieving Archives

Retrieving files is a two step process. First, a user must initiate a request for specific archives. Once the request is completed, the archives are moved into a landing zone where the archives can then be downloaded by the user. Glacier then sends the user a notification that the job is done and the files can be downloaded. Users have 24 hours to download the requested archives before the request expires. The user must make sure there is enough bandwidth to download the data within the 24 hour period. If the data is not downloaded within the 24 hour period, the archives will be deleted from the landing zone to make room for other requests and the request will have to be initiated and charged again (Amazon Web Services, 2018, Retrieving Glacier Archives).

Similar to uploads, requests and downloads are initiated via REST API, command line, Amazon's software development kits, or third party client. Checksums are returned with downloaded files to ensure fixity. Because of the programming knowledge needed to work with the REST API, command line interface, and software development kits, we strongly recommend using a third party Glacier client for generating checksums, uploading archives, downloading archives, and managing vaults.

Recommendations for Wildwatch Kenya

For Glacier storage, it is best to aggregate small files in order to reduce uploading and request costs. For the camera trap images, we recommend creating ZIP files based on current folder organization and naming conventions. A vault should be created for each main folder, and

follow the same naming conventions as WWK main folders. The ZIP files should then be uploaded into the appropriate vault, based on their original main folder. The reason for keeping a similar organizational scheme for Glacier storage is to help the WWK team to understand how their archives are stored and create consistency across platforms. In addition, in the event of catastrophic loss, the team will know exactly which folders are damaged or lost and which archives can replace those files.

However, as discussed in the Organizing Data section, there are some current WWK folders that do not conform to the prevailing organization and naming conventions. There are sometimes one folder for thousands of images, as they are not broken into 1000 image. Others are broken into 1000 image groups, but do not follow the prevailing naming conventions. It is strongly recommended that the WWK team edit these non-uniform folder names to match the naming conventions of the rest of the folders. It is also recommended that the large subfolders be divided into smaller folders containing a maximum of 1000 images each. This not only ensures uniformity, but also imposes a limit on the size of the ZIP files. Controlling ZIP file size lessens the risk of errors or failure when zipping or unzipping files.

In order to visualize organization and provide name-change recommendations, an organizational index will be created for each conservation location (Loisaba and Namunyak) detailing the following data elements: Camera Trap Name, Current WWK Folder Name, Approx Number of Images contained in each ZIP file/folder, Current and Proposed WWK Main Folder Name, Proposed Glacier ZIP File Name, Proposed Glacier Vault Name. This index can serve as a guide for ingest and a finding aid for retrieval. Where the Current WWK Folder Names do not match the prevailing naming conventions, recommendations are provided on how to change the

folder names to match the naming conventions. The Glacier Storage Index is located in Appendix 7 at the end of this report.

Although the camera trap images comprise the bulk of WWK data, the goal for Glacier storage is to have a comprehensive copy of all WWK data. Thus, all supplementary documentation, such as the ReadMe file, codebook, finding aids, and camera trap grid maps should also be uploaded into Glacier. A separate vault, named ‘Documentation,’ should be created for the storage of these documents. These files do not need to be aggregated into ZIP files, and can be uploaded as single files.

In order to make the ZIP files as user friendly as possible, the metadata schema described in the “Description and Metadata” section of this paper will need to be added to them. The suggested best practice is to add a TXT document entitled “ReadMe-metadata.” A TXT file (an example of which can be found in Appendix 6 of this document) does not require a license to be opened, as with Word. The basic metadata must be represented in this file, for example: title, publisher, date, and rights. This simple action will allow users to understand what is inside each ZIP file.

File Formatting

The file format used by WWK is JPEG. With images, JPEG can be problematic due to it having relatively small file sizes and the fact that it is a default on most modern cameras. An additional issue with JPEG is that every time the image is modified and saved, it compresses the file causing data to be overwritten. This can result in digital images becoming damaged overtime, which affects their quality when printed and viewed. It is for that reason that the majority of institutions that engage in digital preservation prefer uncompressed TIFF files

(Corrado & Moulaison-Sandy, 2014, p.146). TIFF also allows for lossless and lossy compression based on the algorithm that is selected at the time the file is saved. The problem that can arise with using uncompressed or lossless compression with TIFF is that the size of the files can be several times larger. Figure 1 below contains two screenshots that compare the same sixteen images taken from the WWK project. The screenshot on the left has the original images in JPEG and the screenshot on the right shows the same screenshots being converted into uncompressed TIFF files.

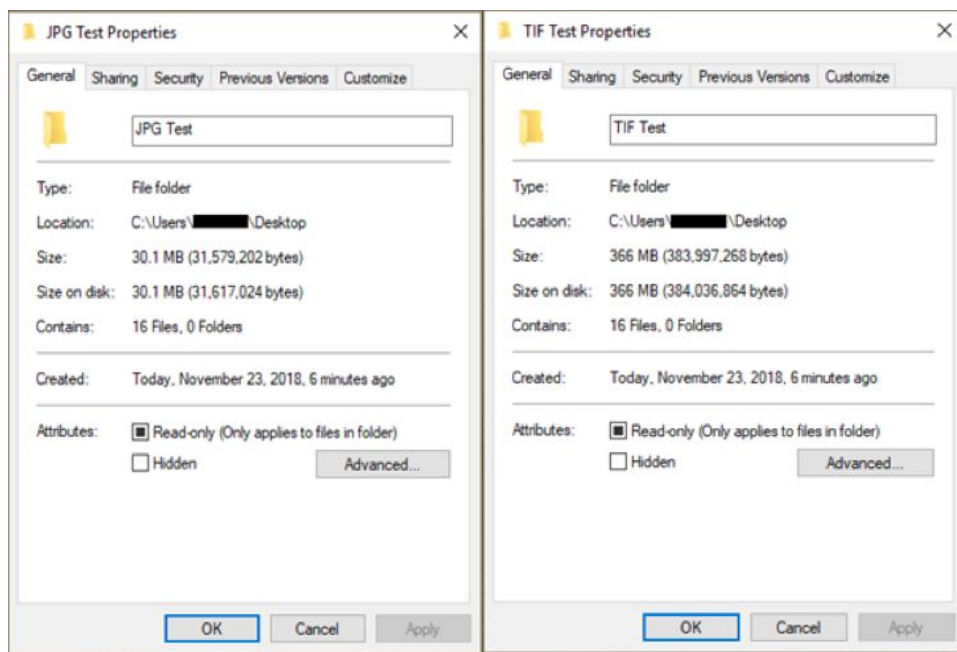


Figure 1. Screenshots comparing the size of the same sixteen images in both JPEG and TIFF file format.

Figure 1 shows that the TIFF file is about twelve times larger than the one with the original JPEG images. Table 2 below describes a test of resizing TIFF images and looking at how much detail was lost in making the image smaller then resizing it. This particular test

showed that when resizing down to an incredibly small file that the resizing filters in Adobe Photoshop did not help preserve the image.

	File Size	Pixels (width x height)	Resolution
Original image that has been converted to TIFF	34.9 M	3094 x 4032 px	72
TIFF image that has been reduced	2.18 M	756x1008 px	72
TIFF image rescaled with Adobe's bicubic smoother enlargement setting	35.7 M	3094 x 4032 px	72
TIFF image rescaled with Adobe's preserved detail enlargement setting	35.7 M	3094 x 4032 px	72
TIFF image rescaled with Adobe's automatic relarging setting	35.7 M	3094 x 4032 px	72

Many institutions view the JPEG 2000 file format as an option for consideration because it allows for lossless compression just like TIFF, but the file size is significantly smaller (Corrado & Moulaison-Sandy, 2014, p.146). JPEG 2000, however, does not have backward compatibilities. It can therefore, be very resource intensive when encoding, and there are very few programs that will open the format. The lack of backward compatibilities is just one of the reasons that JPEG 2000 has not replaced TIFF as the preferred file format for preservation. When JPEG 2000 was created, it was written from an entirely new code, not built from the old JPEG code. The intent was to allow for the possibility of software developers to code their programs to accept both JPEG and JPEG 2000. Since JPEG 2000 is not very common, there is no demand for developers to program software to open both formats. JPEG has priority since it is the older format and more commonly used.

Our advice for WWK is to convert everything for the archive to TIFF for compression and to keep the JPEGs on SDZG's hard drives as they are. It is a common practice for many preservation repositories to contain "multiple versions, or representations, of the same intellectual entity" (Corrado & Moulaison-Sandy, 2014, p.133). There is a practice in photography that many digital preservation specialists use called the 3-2-1 Rule (Witmer, 2017). This rule describes a situation in which three copies of a file are made: two should be stored on different types of storage media and can also be saved in different formats, and the final copy should be stored in a different location. This guarantees that if one copy of the file become corrupted or destroyed there are copies that can be retrieved. By converting at least one copy to TIFF, WWK can hopefully guarantee that there would be at least one high-quality copy of the images if anything would happen to the JPEGs on the Zoo's hard drives. JPEGs can be converted to TIFF through a myriad of ways.

JPEGs can be converted to TIFF individually with GIMP or Adobe Photoshop. This method is best for when there are a small number of files that need to be converted. They can also be converted using a Bulk Image Converter. This is a free program that markets itself as an easy and simple converter for multiple images (Sourceforge, 2018). XnConvert, a highly rated free program, that converts JPEGs to TIFF is appealing because it converts between lots of image file formats and can convert images on Windows, Linux, and macOS; it can convert multiple images at once; it has lots of advanced settings you can customize; and there is a portable option available (XnConvert, 2018). Another JPEG and TIFF converting program is Adapter. This free program allows users to "adjust sizing, compression and add text or image overlays as desired." It will also watermark as it converts (Macroplant LLC, 2018). In addition to

the programs discussed above, there are countless other browser converters but those would require having to upload and download everything, which would be a long and arduous process.

Data Compressing

Data compression “is defined as the process of encoding information using fewer bits than the original representation of the information would use (Sharma and Naaz Mir, 2018, p.39).” It is relevant to reduce the size of the files to upload them to Glacier to optimize the project resources in case the team needs to download them later since Glacier charges downloads per GB.

According to Mark and Gailly (n.d.), there are two types of compression: Lossy and Lossless. The last one allows the data compression without compromising the quality of the files and it is the most recommended for this project because the team wants to preserve high quality images for long-term preservation. TAR and ZIP files are two examples of lossless tools used to compress data, and they combine multiple files in a single one. TAR is used for Linux and ZIP files can be used for Windows or Mac. Amazon Glacier accepts TAR and ZIP files.

According to the inventory, the project has now 1,203,324 images. One sample folder was selected to test data compression with the following formats: ZIP, TAR, and TAR GZ. The folder had 1,004 images and its size on disk was 1893MB. After compressing the folder, it has the same size because all the images are in JPEG, a format already compressed using a lossy algorithm.

1893MB L1_February-May_2016-100EK113

1889MB L1_February-May_2016-100EK113.tar

1878MB L1_February-May_2016-100EK113.tar.bz2

1877MB L1_February-May_2016-100EK113.tar.gz

1877MB L1_February-May_2016-100EK113.zip

The WWK will not save space by compressing the JPEG files in the hard disks, it is only worth to do so to download the files easily since it is possible to combine multiple files in a single folder as previously mentioned.

For TIFF images to be uploaded in Glacier, we also tested the compression to check if there was any storage saving. An image converted to TIFF sizing 23 MB was also compressed using different formats and it was able to detect a file reduction of estimated 18%. We selected ZIP files to compress the WWK images because it is more popular and most of the computers do not require a special software to extract the files like others such as TAR.

eMammal

Whereas the Amazon Glacier archival storage system provides catastrophe protection, eMammal offers a promising option for relatively low cost forward-facing open access to the camera trap images. The eMammal project works with citizen scientists and camera traps to conduct large scale mammal surveys. Coordinated by the Smithsonian, eMammal supports citizen and professional science by providing camera trappers with software that allows them to capture images of wildlife, identify animals and upload identified images for review. This project also archives those identified images in the Smithsonian's repository, and makes them publicly viewable on an open access basis.

It is in this latter capacity that eMammal is of interest. Specifically, our research on repositories (discussed in the next section) yielded two key insights. First, for SDZG to provide

open access to camera trap images by supporting its own digital library¹ would require technological and human resources that represent a considerable and sustained commitment. As a team, we felt we could not assume that level support would be forthcoming. Therefore, we opted to explore more ‘turnkey’ or already-existing options. As an established data management system, archive, and forward-facing, open-access database with its own dedicated staff and equipment, eMammal surfaced as an attractive, sustainable alternative. Second, eMammal camera traps are far more discoverable to end users than would likely be the case for a zoo-hosted solution. This is because eMammal is more than an archive or digital library. It is also an ongoing project that engages citizens and professionals, providing them with equipment and software for camera trapping, and building a community around the effort. As such the eMammal database is well known among interested parties (eMammal, 2017, Browse Data). By merging WWK images into the eMammal platform it is more likely that they will be found than if it is siloed by itself in a locally hosted Zoo repository.

With this in mind, and with the support of San Diego Zoo Global, we have begun a pilot to evaluate the viability of migrating WWK camera trap data to eMammal. Most of the camera trap images in eMammal’s system are generated by researchers that used eMammal software and equipment from the start. As a result, they are already organized and formatted following Open Standard for Camera Trap Data that is understood by the automated eMammal ingest process. This is not so for WWK data. Generated using different equipment and organized and described specifically for the WWK project, the data must be crosswalked to conform to eMammal’s data management structure and metadata requirements. Following a manual for legacy data provided

¹ Meaning a forward facing database with or without an underlying preservation solution/repository

to us by eMammal staff, located in Appendix 1, we have begun the crosswalk on a subset of 2000 images. Links to the crosswalk template and our partially-completed population of that template with WWK data can be accessed from Appendix 8. Although some issues have arisen--for instance, the need to sequence the WWK images and correct some camera-generated date stamps--most of the challenges appear manageable.

One challenge that poses a somewhat more significant problem is the quality of WWK image level metadata identifying animals. WWK's crowdsourced Zooniverse data for Common Species Name does not rise to the level of the expert-approved identifiers required for eMammal. Additionally, WWK images are not currently tagged (or vetted by experts) for Genus species, Integrated Taxonomic Information, and the IUCN ID. In the case of the pilot, the relatively few number of images means that it is possible for us to address these matters by hand coding. However in the event of a successful pilot and continued interest merging WWK camera traps the will need to dedicate some resources and expertise to systematically addressing these gaps. If these issues are resolved, and the crosswalk proves successful, we recommend that WWK budget \$4000 to eMammal. Since the eMammal system requires everything to be sent through Amazon's S3 storage before it goes into the repository, they require all participants to pay \$4 per camera month of images captured. WWK currently has 920 camera months, which amounts to a cost of \$3,680. However, as many of these camera months may only contain "blank" images, and would not need to be sent to eMammal, the current camera month count is likely much lower. Since WWK is still actively collecting images, though, it is in their best interest to follow the original budget.

Other Repositories Considered

While we eventually chose Glacier and eMammal Repository, the following repositories outlined in this section were considered for this project. We review them here in order to clarify why Glacier and eMammal were the best options as well as to share what we have learned about the ecosystem of repository options with others facing similar choices for digital preservation projects.

The ArchivesSpace repository is “designed to support core functions in archives administration such as accessioning; description and arrangement of processed materials including analog, hybrid, and born-digital content; management of authorities (agents and subjects) and rights; and reference service” (ArchivesSpace, 2018, About: Mission & Guiding Principles). Even so, this repository does not offer any file management or workflow for digital assets and there is an annual fee of \$15,000. For those reasons, this repository was not encouraged for use by the WWK.

The Databank Repository is a free UK based repository that is one of the two components that make up DataFlow, the other being DataStage (DataBank, n.d., About DataBank). It is a scalable research data repository, that also allows for permanent storage. This repository was not selected because it does involve command-line usage and there is a lot of coding that might go into this program.

Dataverse Repository is really two things: an individual or institutional dataverse that is hosted remotely by Harvard; and an OS repository software you can install, configure, and administer from your own server. This repository is “open to all researchers worldwide in all disciplines” (Dataverse Project, n.d., Home: Researchers). There is currently no mention made of

space limitations or price for extensive space on the repository's website, however queries made to Harvard were denied on account of WWK's large data size.

Dryad Repository is a free "platform that members to share different types of data and encourage its reuse to generate new knowledge. It is maintained by Dryad, a non-profit organization. The repository was created because there was a necessity to share publications data, also known as Joint Data Archiving Policy (JDAP), among a group of leading journals and scientific societies in evolutionary biology and ecology" (Dryad, 2018, The Organization: Overview). This repository is "flexible about data format, while encouraging the use and further development of community standards" and "fits into the manuscript submission workflow of its partner journals, making data submission easy" (Dryad, 2018, The Repository: Key Features). Even so, this repository is built upon the open-source DSpace and is very limited as a result.

Environmental Data Initiative "is an NSF-funded project accelerating curation and archive of environmental data, emphasizing data from projects funded by the NSF DEB" (Environmental Data Initiative, n.d.). This hosted repository allows private and embargoed data, provides DOI, and is a DataONE member node that makes data more discovered and a helpful resources section for users (DataONE, n.d.). The main negative aspects of this repository that kept it from being suggested for use was the emphasis on data funded by the NSF and the size limitations of the repository. The fact that EDI only allows Ecological Metadata Language also affected the consideration of this repository, for while it may prove more limiting than Dublin Core, the fact that it was developed specifically for the ecological sciences (DCC, n.d.) means that it could potentially connect with similar datasets better, or offer more discipline-specific granularity.

Fedora Repository “is a robust, modular, open source repository system for the management and dissemination of digital content. It is especially suited for digital libraries and archives, both for access and preservation. It is also used to provide specialized access to very large and complex digital collections of historic and cultural materials as well as scientific data” (Fedora, 2018, About). Fedora was not selected as the repository for this project because it is not an immediate solution. It might be a possibility for the long-term combined access and preservation solution for WWK, but for the short-term it was not a feasible option. Simply creating the repository would take months and we noticed the need for WWK to have an immediate solution. Islandora, “an open-source software framework designed to help institutions and organizations and their audiences collaboratively manage, and discover digital assets using a best-practices framework,” is used to connect Fedora repository storage with a Drupal website (Islandora, n.d., About). This software framework was only considered with Fedora, and since Fedora was not selected for this project, neither was Islandora.

According to its website, “figshare allows users to upload any file format to be previewed in the browser so that any research output, from posters and presentations to datasets and code, can be disseminated in a way that the current scholarly publishing model does not allow” (figshare, n.d., About). This repository allows users to “upload files up to 5 GB” and “20 GB of free private space that allows you to share research privately until you choose to make it public” for free (figshare, n.d., figshare features). These features were found to be too small for the WWK, which is why this repository was not selected.

Open Science Framework is a free cloud based data management repository for collaborating and sharing research project data. OSF is supported and maintained by the Center

for Open Science” (Open Source Framework, 2018). Research showed that text based documents, such as reports and datasets, are best suited for this repository. WWK has a lot of images, and it is our notion that Open Science Framework is not the best place to store them due to the amount of images and file structure.

RUresearch’s data portal “provides a place for Rutgers researchers to share their research data with the global scholarly community. RUresearch leverages all the capabilities of RUcore and adds additional tools and services specific to research data” (Rutgers University Libraries, 2018, RUresearch Data Portal). This hosted on-campus repository allows open access, restricted and embargoed data, uses DOIs for data, encourages supplementary materials, and hides rights metadata from users (Rutgers University Libraries, 2018, About RUresearch). The main disadvantage to using this repository is that it would have been used for storage only, and would not have hosted a portal. As the immediate preservation needs of WWK required a storage solution, this was a strong option, however knowing that a forward-facing portal would be required in the future made eMammal a more practical option.

Finally, the last repository we considered for this project, but ultimately did not use is Zenodo. Zenodo “was launched in 2013 as a partnership between CERN and OpenAIRE” (Zenodo, n.d., About). Zenodo welcomes “research outputs from across all fields... assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citable.. [and] allows you to create your own collection and accept or reject uploads submitted to it” (Zenodo, n.d., features). The Zenodo website is very vague regarding size limitations however. Their website states that they “currently accept up to 50GB per dataset,” but that those with larger datasets must contact them directly. (Zenodo, 2018, Frequently Asked

Questions). As the current WWK data size is 110% larger than their commonly accepted dataset size, Zenodo's ability to handle the data seemed questionable. In addition, there is no information on the Zenodo website about the cost for implementing the repository. Due to this lack of information, Zenodo was not selected as the repository for this project.

Conclusion

Corrado and Moulaison-Sandy, authors of *Digital Preservation for Libraries, Archives, and Museums*, state that "risk management is an essential aspect of digital preservation," and as such have identified the following key strategies for ensuring access long-term access: starting early in the creation process, using open and well-documented standards and systems, documenting decisions, using accepted metadata schemas, understanding user expectations, having a backup strategy and succession plan (Corrado & Moulaison-Sandy, 2014, p.69). Since WWK and SDZG currently do not have a digital curation plan in place, it was determined that they did not follow digital preservation best practices early in the creation process.

As this paper makes clear, implementing a set of guidelines for SDZG on how to best curate and preserve WWK data is a multi step complex endeavor. In order to test that our recommendations to SDZG and WWK were feasible, we cleaned a sample of the data, created a codebook for the WWK images inventory, and a sample DCAP, reviewed the naming conventions used for folders, ran tests on a sample of data to determine the differences in both TIFF and JPEG format file size, and evaluated the compression of a sample of data files to ZIP format. To evaluate the viability of eMammal as an access repository, we have begun a pilot on a subset of the WWK camera trap images. We are currently preparing this subset for ingest through a crosswalk. The work that we have done highlights the need for SDGZ and WWK to

implement the suggestions proposed in this paper. The recommendations outlined in this paper follow digital preservation best practices since they are comprehensive, follow open and well-documented standards, and suggest the use accepted metadata schemas. The systems suggested for implementation in this paper were also chosen after carefully considering user expectations and the sustainability of the technology.

Since even the best archival digital preservation system in use today is unlikely to remain as such well into the future, ensuring that the digital preservation solution is sustainable “will reduce risk in the long run” (Corrado & Moulaison-Sandy, 2014, p.71). As Harvey notes in *Digital Curation: A How-To-Do-It Manual*, “the active management of risks over time is essential to preserve digital objects and to ensure that they remain usable in the future” (Harvey, 2010, p.83-84). In order to plan for succession and future usability, the recommendations proposed in this paper focused on considering the sustainability of the project.

Sustainability can be equated to “long-term accessibility, authenticity, and integrity, [to ensure their] digital objects need to be managed over the whole of their lifecycle...To achieve the active management of data throughout their lifecycle, constant maintenance and elaborate life support systems must be planned” (Harvey, 2010, p.83). According to Corrado and Moulaison-Sandy, there are five key factors that influence the sustainability of digital projects: leadership, a clear value proposition, the problem of direct costs, developing revenue streams, and goals and assessments (Corrado & Moulaison-Sandy, 2014, p.72-73). The WWK’s website describes its leadership hierarchy in detail and outlines the value of the project. The value of the work can be easily identified through the strides being made in understanding the behavioral patterns of the reticulated giraffe. Their goal of trying to keep the species from extinction is

incentive enough for public interest to remain for this project. Through outlining the costs involved with different software systems and technologies, this paper addresses the problem of direct costs by discussing the costs related to the project, analyses of how to minimize costs, and a relative budget for the two storage solutions. It is our opinion that the “developing revenue streams” factor listed in the Corrado and Moulaison-Sandy text does not directly apply to the topics covered in this paper.

Through presenting best practice suggestions on rights and open access, content selection, data cleaning, description and metadata, file formatting, data organizing, and data compressing, this paper set targets for WWK to reach, which addresses the goals and assessments factor. In that way, this paper is helping WWK follow best practices for sustainability, ensuring the project’s long-term success. It is our hope that SDZG can use this paper as a guide for similar projects.

References

- Amazon Web Services. (2018). *Amazon S3 Glacier*. Retrieved from <https://aws.amazon.com/glacier/>
- Amazon Web Services.(2018). *Computing Checksums*. Retrieved from <https://docs.aws.amazon.com/amazonglacier/latest/dev/checksum-calculations.html>.
- Amazon Web Services. (2018).*Creating a Vault in Amazon S3 Glacier*. Retrieved from <https://docs.aws.amazon.com/amazonglacier/latest/dev/creating-vaults.html>.
- Amazon Web Services.(2018). *Retrieving Glacier Archives*. Retrieved from <https://docs.aws.amazon.com/amazonglacier/latest/dev/downloading-an-archive-two-steps.html>
- Amazon Web Services. (2018). *Tagging Your Amazon S3 Glacier Vaults*. Retrieved from <https://docs.aws.amazon.com/amazonglacier/latest/dev/tagging-vaults.html>.
- Amazon Web Services. (2018). *Uploading an Archive in Amazon S3 Glacier*. Retrieved from <https://docs.aws.amazon.com/amazonglacier/latest/dev/uploading-an-archive.html>.
- ArchivesSpace. (2018). *About: Mission & Guiding Principles*. Retrieved from <https://archivesspace.org/about/mission>
- Ball, A. (2011). *How to license research data*. Digital Curation Centre. Retrieved from <http://www.dcc.ac.uk/resources/how-guides/license-research-data>
- Cadman, M., & González-Talaván, A. (Eds.). (2014). *Publishing Camera Trap Data, a Best Practice Guide*. Contributed by Athreya, V., Chavan, V., Ghosh, M., Hanssen, F., Harihar, A., Hirsch, T., Lindgaard, A., Mathur, V.B., Mehlum, F., Pandav, B., Talukdar,

- G. & Vang, R. Copenhagen: Global Biodiversity Information Facility. Retrieved from http://www.gbif.org/orc/?doc_id=6045.
- Carroll M. W. (2015). Sharing Research Data and Intellectual Property Law: A Primer. *PLoS biology*, 13(8), e1002235. [doi:10.1371/journal.pbio.1002235](https://doi.org/10.1371/journal.pbio.1002235)
- Chapman, A. D. (2005). *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0*. Report for the Global Biodiversity Information Facility, Copenhagen. Retrieved from https://assets.ctfassets.net/uo17ejk9rkwj/46SfGRfOesU0IagMMAOIkk/1c03ea3e21fcd9025cc800d786890e72/Principles_20and_20Methods_20of_20Data_20Cleaning_20-_20ENGLISH.pdf
- Cloudberry Labs. (n.d.). *Cloudberry Explorer for Amazon Glacier*. Retrieved from <https://www.cloudberrylab.com/explorer/amazon-glacier.aspx>.
- Corrado, Edward M., and Sandy, Heather Moulaison. (2014). *Digital Preservation for Libraries, Archives, and Museums*. Rowman & Littlefield Publishers, ProQuest Ebook Central. Retrieved from <http://ebookcentral.proquest.com/lib/rutgers-ebooks/detail.action?docID=1676304>
- Coyle, K. and Baker, T. (2009). *Guidelines for Dublin Core application profiles (working draft)*. Retrieved from <http://dublincore.org/documents/profile-guidelines/>
- Creative Commons (n.d.). *About the licences*. Retrieved from <https://creativecommons.org/licenses/>
- CrossFTP. (n.d.) *CrossFTP- Amazon Glacier Client*. Retrieved from <http://www.crossftp.com/amazon-glacier-client.htm>.

DataBank. (n.d.). *About DataBank*. Retrieved from <https://databank.ora.ox.ac.uk/about>

DataONE. (n.d.). *Current Member Nodes*. Retrieved from

<https://www.dataone.org/current-member-nodes#uploads>

Dataverse project. (n.d.). *Home: Researchers*. Retrieved from <https://dataverse.org/researchers>

DCC. (n.d.). *EML - Ecological Metadata Language*. Retrieved from

<http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language>

DCMI Usage Board. (2012). *DCMI Type Vocabulary DCMI Metadata Terms*. Retrieved from

<http://dublincore.org/documents/dcmi-type-vocabulary/>

DCMI Usage Board. (2012). *DCMI Metadata Terms*. Retrieved from

<http://dublincore.org/documents/dcmi-terms/>

DCMI. (2012). *Dublin Core Metadata Element Set, Version 1.1: Reference Description*.

Retrieved from <http://www.dublincore.org/documents/dces/>

Dryad. (2018). *The organization: Overview*. Retrieved from

<https://datadryad.org/pages/organization>

Dryad. (2018). *The Repository: Key Features*. Retrieved from

<https://datadryad.org/pages/repository>

eMammal. (2017). *About eMammal*. Retrieved from <https://emammal.si.edu/about>

eMammal. (2017). *Browse Data*. Retrieved from <https://emammal.si.edu/analysis/data-download>

eMammal. (2017). *Data Provider Terms and Conditions*. Retrieved from

<https://emammal.si.edu/data-provider-terms>

eMammal.(2017). *Frequently Asked Questions*. Retrieved from

<https://emammal.si.edu/about/FAQ>

eMammal. (2017). *View Photos*. Retrieved from <https://emammal.si.edu/favorite-photos>

Environmental Data Initiative. (n.d.). *Welcome*. Retrieved from

<https://environmentaldatainitiative.org/>

FastGlacier. (n.d.). *What is FastGlacier*. Retrieved from <https://fastglacier.com/>.

Federal Agencies Digital Guidelines Initiative. (2016). *Technical Guidelines for Digitizing*

Cultural Heritage Materials: Creation of Raster Image Files [Brochure]. Author.

Retrieved from http://www.digitizationguidelines.gov/guidelines/FADGI_Federal

[Agencies Digital Guidelines Initiative-2016 Final_rev1.pdf](http://www.digitizationguidelines.gov/guidelines/FADGI_Federal_Agencies_Digital_Guidelines_Initiative-2016_Final_rev1.pdf)

Fedora. (2018). *About*. Retrieved from <https://duraspace.org/fedora/about/>

Figshare. (n.d.). *About*. Retrieved from <https://figshare.com/about>

Figshare. (n.d.). *Figshare features*. Retrieved from: <https://figshare.com/features>

Forrester, T., O'Brien, T., Fegraus, E., Jansen, P. A., Palmer, J., Kays, R., Ahumada, J., McShea,

W. (2016). "An Open Standard for Camera Trap Data." *Biodiversity Data Journal* 4,

p.1-8. Retrieved from <https://doi.org/10.3897/BDJ.4.e10197>

Harvey, R. (2010). *Digital Curation: A How-To-Do-It Manual*. New York: Neal-Shulman

Publishers.

Heery, R., & Patel, M. (2000). "Application Profiles: Mixing and Matching Metadata Schemas."

Ariadne: Web Magazine for Information Professionals (25). Retrieved from

<http://www.ariadne.ac.uk/issue25/app-profiles/>

Islandora. (n.d.). *About*. Retrieved from <https://islandora.ca/about>

ITIS. (2017). *Integrated Taxonomic Information System*. Retrieved from <https://www.itis.gov/>

- IUCN. (2018). *IUCN Red list of Threatened Species*. Retrieved from <https://www.iucn.org/theme/species/our-work/iucn-red-list-threatened-species>
- Khoo, M., and Hall, C. (2010). Merging metadata: a socio-technical study of crosswalking and interoperability. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries* (pp. 361-364). Retrieved from <https://dl.acm.org/citation.cfm?doid=1816123.1816180>
- Kozlowski, W. (2014, May 30). *Guidelines for writing “readme” style metadata*. Retrieved from http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf
- LaFleur, M., Sauther, M., Cuozzo, F., Yamashita, N., Youssouf, I. A.J., & Bender, R. (2014). Cathemerality in wild ring-tailed lemurs (*Lemur catta*) in the spiny forest of Tsimanampetsotsa National Park: camera trap data and preliminary behavioral observations. *Primates*, 55(2), 207–217. Retrieved from <https://doi.org/10.1007/s10329-013-0391-1>
- Lazier, C., & Zhang, H. (2013, October 21). *AWS Webcast - Archiving in the Cloud - Amazon Glacier Best Practices*. Retrieved from <https://www.youtube.com/watch?v=SnmMJJAD2mM&t=2505s>.
- Library of Congress. (n.d.). *Library of Congress Subject Headings*. Retrieved from <http://id.loc.gov/authorities/subjects.html>
- Liss, J. (2018, November 28). *Principles for Evaluating Metadata Standards*. Retrieved from <http://metaware.buzz/2016/08/04/principles-for-evaluating-metadata-standards/>
- Macroplant LLC. (2018). *Adaptor*. Retrieved from <https://macroplant.com/adaptor/image-converter>

Mark, N., & Gailly, J. (n.d.). *The Data Compression Book*. Retrieved from

http://staff.uob.edu.bh/files/781231507_files/The-Data-Compression-Book-2nd-edition.pdf

Miller, S. J. (2014). *Metadata for Digital Collections: A How-To-Do-It Manual*. New York: Neal-Schuman.

Open Domain Commons. (n.d.) *Open Data Commons Public Domain Dedication and License (PDDL)*. Retrieved from <https://opendatacommons.org/licenses/pddl/index.html>

Open Source Framework. (2018). *Open Source Framework: A scholarly commons to connect the entire research cycle*. Retrieved from <https://osf.io/>

Ruggles, S. (2018). The Importance of Data Curation. In Vannette D., Krosnick J. (eds) *The Palgrave Handbook of Survey Research* (pp.303-308). Palgrave Macmillan, Cham.

Rutgers University. (2018). *Copyright Basics*. Retrieved from

<https://www.libraries.rutgers.edu/copyright/copyright-basics>

Rutgers University Libraries. (2018). *RUresearch Data Portal*. Retrieved from

<https://rucore.libraries.rutgers.edu/research/>

Rutgers University Libraries. (2018). *RUresearch: About*. Retrieved from

<https://rucore.libraries.rutgers.edu/research/about/>

Sharma, V., & Naaz Mir, R. (2018). "Digital Preservation and Data Compression." *International Journal of Computer Science And Technology*,37-43. Retrieved from

<http://www.ijcst.com/vol9/issue2/8-vipul-sharma.pdf>

Sourceforge. (2018). *Bulk Image Converter*. Retrieved from

<https://sourceforge.net/projects/bulkimageconver/>

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, 2, 150026. Retrieved from

<https://doi.org/10.1038/sdata.2015.26>

The Getty Research Institute. (2017). *Getty Thesaurus of Geographic Names® Online*. Retrieved from <http://www.getty.edu/research/tools/vocabularies/tgn/index.html>

Utah Academic Library Consortium. (2011, July 20). *Mountain West Digital Library Dublin Core Application Profile*. Retrieved from

https://mwdl.org/docs/MWDL_DC_Profile_Version_2.0.pdf

W3 Consortium. (1997). *Date and Time Formats*. Retrieved from

<https://www.w3.org/TR/NOTE-datetime>

Wildwatch Kenya. (n.d.). *About: Research*. Retrieved from

<https://www.zooniverse.org/projects/sandiegozooglobal/wildwatch-kenya/about/research>

Witmer, S. D. (2017). *Personal Digital Archiving Guide Part 1: Preservation Planning*.

Retrieved from

<https://www.lib.umich.edu/blogs/bits-and-pieces/personal-digital-archiving-guide-part-1-preservation-planning>

XnConvert. (2018). *XnConvert: Best Batch Image Processing, Image Convertor, Image Resizer*.

Retrieved from <https://www.xnview.com/en/xnconvert/>

Zenodo. (n.d.). *About*. Retrieved from <http://about.zenodo.org/>

Zenodo. (n.d.). *Features*. Retrieved from <http://help.zenodo.org/features/>

Zenodo. (n.d.). *Frequently Asked Questions*. Retrieved from <http://help.zenodo.org/>

Appendix 1: Manual for Submitting Legacy Data to eMammal Data Repository

Manual for submitting Legacy Data to eMammal Data Repository Data Package

Each project should be a folder named with the Project Name that contains 4 csvs, one for each worksheet below (Project, Deployment, Sequence, Image), and the project photos. The project photos may be in folders as long as the folder structure is accurately documented in the location field in the image worksheet and all folder names are unique. If you have file names that are not unique in the project (i.e. img_001.jpg in multiple deployments) you **MUST** place the images in folders with unique names.

All cells in the spreadsheet must contain valid input (defined for each column). Some cells may be left blank, but others will throw errors if not filled in.

Bolded fields indicate they **cannot be left blank**.

****Starred** fields indicate that they will be filled in for you.

See attached Excel sheet for template. Columns in green indicate required fields for **you, the researcher, if appropriate**.

PROJECT Worksheet

The project page provides information about the dataset to be submitted.

****Project ID** – Text Format. This can be a dummy field unless you have an existing project in eMammal. If the latter is true, it must be in the format p[number], ex: p153.

Publish Date – Date Format MM/DD/YYYY. Must be valid date, or else it will be rejected. In the repository this represents the date to publish the data, not the date that a paper or report was published in the past using this data. This date should reflect the embargo in Project Data Use and Constraints

Project Name – Text Format. The name of your project.

Project Objectives – Text Format. A paragraph of text describing your project.

Project Owner – Text Format. The person or organization that owns the project. Must be an eMammal user.

Project Owner email – Text Format, valid email address only. Must be associated with eMammal user.

Principal Investigator – Text Format. Person responsible for project. Must be an eMammal user.

Principle Investigator Email – Text Format, valid email address only. Must be associated with eMammal user.

Project Contact – Text Format. Text Format. Person responsible for project other than PI. Must be an eMammal user.

Project Contact email – Text Format, valid email address only. Must be associated with eMammal user.

Project Latitude – Numeric Format. Latitude of the central project point in decimal degrees using WGS84 datum. Accuracy to 5 decimal places.

Project Longitude – Numeric Format. Longitude of the central project point in decimal degrees using WGS84 datum. Accuracy to 5 decimal places.

Country Code – Text Format. 3-character ISO code for your country.

Project Data Use and Constraints – Text Format. Possibilities: 1. No constraints; 2. 1-year embargo; 3. 3-year embargo

DEPLOYMENT Worksheet

****Camera Deployment ID** – Text Format. This can be a dummy field unless you have an existing deployment created in eMammal. If the latter is true, it must be in the format d[number], ex: d23504.

Camera Site Name – Text Format. Must be a name for the camera site, but also unique in time. Deployments are a camera in a specific location for a specific time period. Ex: HTL12-H01B is a project in the HTL region of China; 12 indicates 2012; H is the specific location; 01 is the month of the year; B is the camera ID.

Camera Deployment Begin Date – Date Format MM/DD/YYYY. Must be valid date, or else it will be rejected. Date of camera trap setup, aka first picture in SD card.

Camera Deployment End Date – Date Format MM/DD/YYYY. Must be valid date, or else it will be rejected. Date of camera trap pickup, aka last picture in SD card.

Actual Latitude – Numeric Format. Latitude of a deployment point in decimal degrees using WGS84 datum. Accuracy to 5 decimal places.

Actual Longitude – Numeric Format. Longitude of a deployment point in decimal degrees using WGS84 datum. Accuracy to 5 decimal places.

Camera Failure Details – Text Format. Describes condition of the camera during deployment. Only the following phrases can be entered, **with these exact words. Separation by semicolon means a different phrase**: Camera Functioning; Wildlife Damage; Camera Hardware Failure; Memory Card/Film Failure; Vandalism/Theft; Unknown Failure.

Bait – Text Format. Only the following phrases can be entered, **with these exact words.**

Separation by semicolon means a different phrase: No Bait; Other Bait; Acoustic; Visual; Meat; Scent.

Bait Description – Text Format. Describe baiting method and Other Bait, if necessary. Example Bait Type = Scent, Bait Description = Calvin Klein Perfume.

Feature – Text Format. Describes a feature that the camera trap was set on or near. Only the following phrases can be entered, **with these exact words. Separation by semicolon means a different phrase**: Not Entered; Road, paved; Road, dirt; Trail, hiking/people; Trail, game; Road underpass/overpass/bridge; Culvert; Burrow; Nest site; Carcass; Water source/Spring; Fruiting tree; Other.

Feature Methodology – Text Format. Describes 'Other' Feature Type.

Camera ID – Text Format. Layout depends on project PI.

Quiet Period Setting – Numeric Format. The length of time, in seconds, before a camera will take another picture. Example: Camera is set to take 3 pictures per trigger at 1 second intervals, then be quiet for 2 minutes before the next trigger is possible. Quiet Period Setting = 120.

Sensitivity Setting – Text Format, only takes values: high; medium; low. Setting of camera, we usually do “high.”

**SEQUENCE Worksheet

If you do not have your data sorted into sequences, eMammal will handle this for you, as long as you accurately fill out the Image Worksheet. Consider all of the following to be starred.

Observation Type – Text Format, only Researcher can be entered here.

Deployment ID – Text Format, See above. Same as Camera Deployment ID in DEPLOYMENT Worksheet.

Image Sequence ID – Numeric Format. Assign a number to each sequence, these numbers must be consecutive and not skip a number or else the ingest will fail. Ex: 1, 2, 4, 5 will fail, but 1, 2, 3, 4 will not.

Date_Time Begin – Custom Date Time Format MM/DD/YYYY HH:MM:SS. Must be valid date-time, or else it will be rejected. Date and time of first picture in sequence.

Date_Time End – Custom Date Time Format MM/DD/YYYY HH:MM:SS. Must be valid date-time, or else it will be rejected. Date and time of last picture in sequence.

Genus species – Text Format. Latin Binomial for the species. IUCN Red list should be used as the taxonomic authority.

Species Common Name – Text Format. Most widely accepted common name for the species. IUCN Red list should be used as the authority.

Age – Text Format, only takes values: Adult; Juvenile; Unknown. Can be blank.

Sex – Text Format, only takes values: Male; Female; Unknown. Can be blank.

Individual ID – Text Format. If you want to attach an individual identifiable code, you may do so.

Count – Numeric Format. Number of animals in sequence. A count must be a number and greater than 0, but can be left as 1 if only presence/absence data was recorded.

Animal recognizable – Text Format, values are Y; N. Indicates if an animal is recognizable – has visible tags, or has a unique pattern (spots, stripes, etc).

Individual.Animal.Notes – Text Format, used to note anything unusual/interesting. Should be used if you have classified birds because we recommend changing those to “Other Bird species.”

~TSN.ID – Numeric Format, [Integrated Taxonomic Information System](#) must be used.

~IUCN.ID – Numeric Format, [IUCN Red list](#) must be used.

~IUCN Status – Text Format. [IUCN Red list](#) must be used. Helps determine embargo status of deployment (if even one EN or CR animal appears in a deployment, the entire deployment

location is fuzzed). Only the following phrases can be entered, **with these exact words**.

Separation by semicolon means a different phrase: LC; NT; VU; EN; CR; EW; EX; DD; NE; US.

~We have a comprehensive list of all animals in eMammal that we have added to our database. If you would like a copy, please email emammal@si.edu.

IMAGE Worksheet

****Deployment ID** – Text Format. See above. Same as Camera Deployment ID in DEPLOYMENT Worksheet.

****Image Sequence ID** – Text Format, See above. Same as Image Sequence ID in SEQUENCE Worksheet.

Image ID – Text Format. **Unique** identification for image within deployment, usually generated by AGOUTI or another processing program, but can also be self-generated numbers.

Location – Text Format. The full directory path to the image location within the project folder. Example if the project folder is GTM-001: GTM-001/Tikal_2005/EscanNo.30/00.JPG.

Image File Name – Text Format. Must be exact name of image, ex: IMG_001.JPG, and the extension (JPG) **must be capital letters**.

Photo Type – Text Format. Class of image. Only the following phrases can be entered, **with these exact words**. **Separation by semicolon means a different phrase:** end; start; set up; blank; animal; staff; unknown; unidentifiable; time lapse.

Photo Type Identified by – Text Format, name of the person who classified the images. If more than one person, names must be separated by a semi-colon. **Genus Species** – Text Format. Latin Binomial for the species. IUCN Red list should be used as the taxonomic authority.

Genus species – Text Format. Latin Binomial for the species. IUCN Red list should be used as the taxonomic authority.

Species Common Name – Text Format. Most widely accepted common name for the species. IUCN Red list should be used as the authority.

~**TSN.ID** – Numeric Format, [Integrated Taxonomic Information System](#) must be used.

~**IUCN.ID** – Numeric Format, [IUCN Red list](#) must be used.

~**IUCN Status** – Text Format. [IUCN Red list](#) must be used. Helps determine embargo status of deployment (if even one EN or CR animal appears in a deployment, the entire deployment location is fuzzed). Only the following phrases can be entered, **with these exact words**.

Separation by semicolon means a different phrase: LC; NT; VU; EN; CR; EW; EX; DD; NE; US.

~We have a comprehensive list of all animals in eMammal that we have added to our database. If you would like a copy, please email emammal@si.edu.

Date_Time Captured – Custom Time Format. Date and time that image was recorded in YYYY/MM/DD HH:MM:SS format.

****Interest Ranking** – Text Format, only takes values: None; Favorite. Indicates if a photo is presentation/outreach-worthy.

Age – Text Format, only takes values: Adult; Juvenile; Unknown. Can be blank.

Sex – Text Format, only takes values: Male; Female; Unknown. Can be blank.

Individual ID – Text Format. If you want to attach an individual identifiable code, you may do so.

Count – Numeric Format. Number of animals in sequence. A count must be a number and greater than 0, but can be left as 1 if only presence/absence data was recorded.

Animal recognizable – Text Format, values are Y; N. Indicates if an animal is recognizable – has visible tags, or has a unique pattern (spots, stripes, etc).

Individual Animal Notes – Text Format

Digital Origin – Text Format. Only the following phrases can be entered, **with these exact words. Separation by semicolon means a different phrase**: born digital; reformatted digital; digitized microfilm; digitized other analog.

Embargo Period – Text Format

Restrictions on Access – Text Format

Image Use Restrictions – Text Format

Appendix 2: Data Cleanse of Camera Trap Images Inventory

ORIGINAL_camera_trap_images_inventory_7114

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

75% \$ % .0 .00 123- Calibri 12 B I U A

1	Camera Trap	Dates	Round	Actual Latitude	Actual Longitude	Approximate # of images	Hardrive	Given to Marketing	Batch Uploaded to Zooniverse
2	L1	February - May 2016	Pilot	00 32.892		15,516	L2	YES	
3	L2	February - May 2016	Pilot			2,092	L2	YES	
4	L3	February - May 2016	Pilot			15,574	L2	YES	
5	L4	February - May 2016	Pilot			14,665	L2	YES	
6	L5	February - May 2016	Pilot			14,278	L2	YES	
7	L8	February - May 2016	Pilot			15,951	L2	YES	
8	L10	February - May 2016	Pilot			15,669	L2	YES	
9	L11	February - May 2016	Pilot			15,863	L2	YES	
10	L12	February - May 2016	Pilot			15,772	L2	YES	
11	L13	February - May 2016	Pilot			15,813	L2	YES	
12	L14	February - May 2016	Pilot			15,639	L2	YES	
13	L15	February - May 2016	Pilot			16,082	L2	YES	
14	L16	February - May 2016	Pilot			15,655	L2	YES	
15	L17	February - May 2016	Pilot			15,846	L2	YES	
16	L18	February - May 2016	Pilot			15,469	L2	YES	
17	L19	February - May 2016	Pilot			15,654	L2	YES	
18	L20	February - May 2016	Pilot			15,819	L2	YES	
19	L21	February - May 2016	Pilot			15,472	L2	YES	
20	L22	February - May 2016	Pilot			15,576	L2	YES	
21	L23	February - May 2016	Pilot			15,700	L2	YES	
22	L24	February - May 2016	Pilot			15,819	L2	YES	
23	L28	February - May 2016	Pilot			13,423	L2	YES	
24	L29	February - May 2016	Pilot			16,026	L2	YES	
25	23								
26					Total # of Images =	327,857			
27	L2	May - October 2016	1	N00 37.609		11,564	L3	YES	
28	L26	May - October 2016	1			17,886	L3	YES	

Loisaba Loisaba Camera Traps Namunyak Namunyak Camera Traps Explore

The above image is of the original Camera Trap Image Inventory (coordinates hidden to protect endangered species).

camera_trap_images_inventory_7114

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

75% \$ % .0 .00 123- Calibri 12 B I U A

1	Grid Index Location	Dates	Deployment Begin Date	Deployment End Date	Round	Actual Latitude	Actual Longitude	Approximate # of images
2	L1	February - May 2016	02/01/2016	05/31/2016	Pilot	0 78111		15,516
3	L2	February - May 2016	02/01/2016	05/31/2016	Pilot			2,092
4	L3	February - May 2016	02/01/2016	05/31/2016	Pilot			15,574
5	L4	February - May 2016	02/01/2016	05/31/2016	Pilot			14,665
6	L5	February - May 2016	02/01/2016	05/31/2016	Pilot			14,278
7	L8	February - May 2016	02/01/2016	05/31/2016	Pilot			15,951
8	L10	February - May 2016	02/01/2016	05/31/2016	Pilot			15,669
9	L11	February - May 2016	02/01/2016	05/31/2016	Pilot			15,863
10	L12	February - May 2016	02/01/2016	05/31/2016	Pilot			15,772
11	L13	February - May 2016	02/01/2016	05/31/2016	Pilot			15,813
12	L14	February - May 2016	02/01/2016	05/31/2016	Pilot			15,639
13	L15	February - May 2016	02/01/2016	05/31/2016	Pilot			16,082
14	L16	February - May 2016	02/01/2016	05/31/2016	Pilot			15,655
15	L17	February - May 2016	02/01/2016	05/31/2016	Pilot			15,846
16	L18	February - May 2016	02/01/2016	05/31/2016	Pilot			15,469
17	L19	February - May 2016	02/01/2016	05/31/2016	Pilot			15,654
18	L20	February - May 2016	02/01/2016	05/31/2016	Pilot			15,819
19	L21	February - May 2016	02/01/2016	05/31/2016	Pilot			15,472
20	L22	February - May 2016	02/01/2016	05/31/2016	Pilot			15,576
21	L23	February - May 2016	02/01/2016	05/31/2016	Pilot			15,700
22	L24	February - May 2016	02/01/2016	05/31/2016	Pilot			15,819
23	L28	February - May 2016	02/01/2016	05/31/2016	Pilot			13,423
24	L29	February - May 2016	02/01/2016	05/31/2016	Pilot			16,026
25	23							
26							Total # of Images =	327,857
27	L2	May - October 2016	05/01/2016	10/31/2016	1	0.78583		11,564
28	L26	May - October 2016	05/01/2016	10/31/2016	1			17,886
29	L28	May - October 2016	05/01/2016	10/31/2016	1			10,000

Project Loisaba Loisaba Camera Traps Namunyak Namunyak Camera Traps

The image above is of the cleansed Camera Trap Image Inventory (coordinates hidden to protect endangered species).

Specific changes done to clean the Camera Trap Image Inventory are as follows:

- Removed working notes upon approval (*All pages*)
- Identified null entries and inconsistencies (*All pages*)
- Created Project page based on eMammal standards, and filled required variables (*Project page*)
- Standardized Grid Index variables and entries (*All pages*)
- Standardized comprehensive deployment Dates (*Loisaba and Namunyak pages*)
- Introduced Deployment Begin Date and Deployment End Date columns (*Loisaba and Namunyak pages*)
- Introduced standard MM/DD/YYYY variables to Deployment Begin Date and Deployment End Date columns (*Loisaba and Namunyak pages*)
- Standardized Actual Latitude and Actual Longitude columns, from degree/time format to decimal format (*Loisaba and Namunyak pages*)
- Standardized Camera Failure Details column (*Loisaba and Namunyak pages*)

Appendix 3: Codebook for Inventory of Camera Trap Images

Sheets

Loisaba

Inventory of deployments at Loisaba conservancy site

Loisaba Camera Traps

Inventory of geographic coverage of deployments at Loisaba conservancy site

Namunyak

Inventory of deployments at Loisaba conservancy site

Namunyak Camera Traps

Inventory of geographic coverage of deployments at the Namunyak conservancy site

Variables

Round (currently column A)

Period of time for which camera trap devices were deployed in the field; used to sample across time periods; length of rounds varies due to social and/or mechanical contingencies.

Grid Index Location (currently Camera Trap)

Identifier for geographic location of camera trap device on a grid map of the Loisaba site or the Namunyak site; used to ensure geographic coverage of the sites; use of grid index locations varies due social contingencies.

Dates

Beginning and end date of a particular round

Actual Latitude

Actual latitude recorded at the location where the camera was placed in the field.

Actual Longitude

Actual longitude recorded at the location where the camera was placed in the field.

Approximate # of images

Estimated total number of images collected during each round per round camera trap location.

Hard drive

Identifier for the hard drive on which images are saved

Notes

Camera failure details or other deployment problems

Grid Index Number

Identifier assigned to a grid index location (currently Camera Trap)

Coordinates of Grid Location

Latitude and longitude of Grid Index Number as determined through GIS

Deployed/Round

Whether a grid location was used in any given round (Yes or No)

Comments

Anomalies in deployment (similar to Notes)

JPG Directory

Each folder includes all the images gathered from a particular grid index (i.e., camera trap location) during a particular round.

Appendix 4: Sample Dublin Core Application Profile

Table Key:

DC Element Name	The unique name given to the element within the list of Dublin Core terms or elements namespace.
Label	Commonly assigned name for the field in the metadata record.
DC Definition	Definition as stated in the <i>DCMI DC Metadata Element Set</i> vocabulary: http://www.dublincore.org/documents/dces/
Is Field Required?	Indicates the extent to which this field is required in a record: Required: Field must be included in record. Mandatory if applicable: If the data is known, field must be included in record. Optional: Inclusion of field is up to the discretion of the collection manager.
Is Field Repeatable?	Yes or No to indicate if the field is repeatable within a single documentary unit.
How to Use	Guidelines for what to put into a field and how to enter the data.
Refines/Refinement	Refines the parent element by the element listed or notes refinements that may be made.
Schemas	Identifies controlled vocabulary or format convention used to structure the data contained in the field; these are called schemes. For more information about the listed scheme, click the scheme name.
DC Mapping	Lists the Dublin Core element to map to for harvesting purposes.

Elements Tables (In alphabetical order)

DC Element Name	<i>contributor</i>
Label	Contributor
DC Definition	An entity responsible for making contributions to the resource.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Use for other people or entities who contributed to making the intellectual content of the resource, but who are not covered in the creator field.
Refines/Refinement	None
Schemas	Library of Congress Name authority File (LOCNAF)
DC Mapping	dcterms:contributor

DC Element Name	<i>coverage</i>
Label	Coverage
DC Definition	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Use specific refinements <i>spatial</i> and/or <i>temporal</i> .
Refines/Refinement	<i>Spatial</i> or <i>Temporal</i>
Schemas	None
DC Mapping	dcterms:coverage

DC Element Name	<i>creator</i>
Label	Creator
DC Definition	An entity primarily responsible for making the resource.
Is Field Required?	Mandatory if applicable
Is Field Repeatable?	Yes
How to Use	Person or entity responsible for creating intellectual content of resource such as a person, organization or service. It is suggested that the form of the name is verified in the Library of Congress Name Authority File (LCNAF). If the name is not listed in the LCNAF, give name in the following format: Last name, First name, Middle initial and period, year of birth and/or death if known, separated by a hyphen.
Refines/Refinement	None
Schemas	Library of Congress Name authority File (LOCNAF)
DC Mapping	dcterms:creator

DC Element Name	<i>date</i>
Label	Date
DC Definition	An entity primarily responsible for making the resource.
Is Field Required?	Required

Is Field Repeatable?	No
How to Use	A resource may have several dates associated with it. The date covered by this table refers to creation of the original resource, that is, when the resource was first created, before undergoing any conversion.
Refines/Refinement	None
Schemas	W3C Date Time Format profile of ISO 8601 (W3CDTF)
DC Mapping	dcterms:date

DC Element Name	<i>description</i>
Label	Description
DC Definition	An account of the resource.
Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	Anything significant about the digital resource not covered elsewhere. Use standard punctuation and grammar.
Refines/Refinement	None
Schemas	None
DC Mapping	dcterms:description

DC Element Name	<i>extent</i>
Label	Extent
DC Definition	The size or duration of the resource.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Describe the file size and, if applicable, duration of the digital object.
Refines/Refinement	Refines <i>Format</i>
Schemas	None
DC Mapping	dcterms:extent

DC Element Name	<i>format</i>
Label	Format
DC Definition	The file format, physical medium, or dimensions of the resource.
Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	Describe the file format of the resource using the Internet Media Type (IMT) scheme. Use of the scheme will imply the software needed to display or operate the resource.
Refines/Refinement	<i>Extent</i> or <i>Medium</i>
Schemas	Internet Media Type (IMT). A commonly used subset of the entire IMT set of acceptable formats is available for download on the MWDL website.
DC Mapping	dcterms:format

DC Element Name	<i>identifier</i>
Label	Identifier
DC Definition	An unambiguous reference to the resource within a given context.
Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	<p>Many digital asset management systems, such as CONTENTdm and bepress Digital Commons, automatically create, populate, and map the main identifier field. This main identifier is a URL for the resource.</p> <p>Additional identifier fields may be created as needed for local use, such as call number, filename, etc. However, only the identifier that contains the URI leading back to the resource should be mapped to dcterms:identifier because of the possibility of confusion for harvesters. Handle these additional identifiers one of three ways: map the field to “None”; map it to some other field besides dcterms:identifier; or begin the value of the field with something other than a URI scheme.</p>
Refines/Refinement	None
Schemas	None
DC Mapping	dcterms:identifier

DC Element Name	<i>language</i>
Label	Language
DC Definition	A language of the resource.
Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	Use ISO 639-3 three letter codes. For multiple languages, use separate language fields or list all in a single field, separating each with a semicolon and a space.
Refines/Refinement	None
Schemas	ISO 639-3
DC Mapping	dcterms:language

DC Element Name	<i>medium</i>
Label	Medium
DC Definition	The material or physical carrier of the resource.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	
Refines/Refinement	Refines <i>Format</i>
Schemas	Strongly recommended the use of established controlled vocabularies such as: Library of Congress Subject Headings (LCSH)
DC Mapping	dcterms:medium

DC Element Name	<i>publisher</i>
Label	Publisher
DC Definition	An entity responsible for making the resource available.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Name of the entity that created or is providing access to the resource.

Refines/Refinement	None
Schemas	None
DC Mapping	dcterms:publisher

DC Element Name	<i>relation</i>
Label	Relation
DC Definition	A related resource.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Use to indicate a if the resource is related to another resource.
Refines/Refinement	Possible refinements include: <i>isPartOf; hasFormat; hasPart; hasVersion; isFormatOf; isReplacedBy; isReferencedBy; isRequiredBy; isVersionOf; replaces; requires; and references.</i>
Schemas	None
DC Mapping	dcterms:relation

DC Element Name	<i>rights</i>
Label	Rights
DC Definition	Information about rights held in and over the resource.
Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	Use Rights to describe the copyright status of the access file, its copyright holder and contact information (if applicable), and physical ownership rights (if applicable).
Refines/Refinement	<i>accessRights</i> and <i>rightsHolder</i>
Schemas	None
DC Mapping	dcterms:rights

DC Element Name	<i>source</i>
Label	Source
DC Definition	A related resource from which the described resource is derived.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Use only when the resource is the result of digitization of non-digital originals. Provide sufficient information to identify and find the original resource. If the resource was born-digital or was converted from a born-digital resource, no <i>source</i> field is needed.
Refines/Refinement	None
Schemas	None
DC Mapping	dcterms:source

DC Element Name	<i>spatial</i>
Label	Spatial
DC Definition	Spatial characteristics of the resource.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Use to state the place or area that is described or represented by the resource, not the place where the resource was published. May name a place or specify geographic coordinates. A jurisdiction may be a named or a geographic place to which the resource applies.
Refines/Refinement	Refines <i>coverage</i>
Schemas	Strongly recommended the use of established controlled vocabularies such as: Library of Congress Subject Headings (LCSH)
DC Mapping	dcterms:spatial

DC Element Name	<i>subject</i>
Label	Subject
DC Definition	The topic of the resource.

Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	Describe what the resource content is about, expressed in keywords, phrases, names, subject headings, or classification codes. Use a separate subject field for each different vocabulary and indicate the vocabulary in the label name.
Refines/Refinement	None
Schemas	Strongly recommend the use of established controlled vocabularies such as: Library of Congress Subject Headings (LCSH) or Getty Thesaurus of Geographic Names (TGN)
DC Mapping	dcterms:subject

DC Element Name	<i>temporal</i>
Label	Temporal
DC Definition	Temporal characteristics of the resource.
Is Field Required?	Optional
Is Field Repeatable?	Yes
How to Use	Use to describe the time period covered or represented by the resource, not the date when the resource was published. Temporal topic may be a named period, date, or date range.
Refines/Refinement	Refines <i>coverage</i>
Schemas	Library of Congress Subject Headings (LCSH)
DC Mapping	dcterms:temporal

DC Element Name	<i>title</i>
Label	Title
DC Definition	A name given to the resource.
Is Field Required?	Required
Is Field Repeatable?	No
How to Use	Take the title from the digital item or file name.

Refines/Refinement	None
Schemas	None
DC Mapping	dcterms:title

DC Element Name	<i>type</i>
Label	Type
DC Definition	The nature or genre of the resource.
Is Field Required?	Required
Is Field Repeatable?	Yes
How to Use	<p>Must have at least one type field containing relevant type(s) from DCMI Type vocabulary.</p> <p>For images, refine with a second term, either <i>StillImage</i> or <i>MovingImage</i> (i.e., use <i>Image;Still Image</i> or <i>Image;MovingImage</i>).</p> <p><i>*Note that these combined terms contain no space after the semicolon and that both words in the second term are capitalized.</i></p>
Refines/Refinement	None
Schemas	DCMI Type Vocabulary
DC Mapping	dcterms:type

Vocabulary Encoding Schemas

Term	<i>IMT</i>
Full Name	Internet Media Type
DC Definition	Media types specified by the Internet Assigned Numbers Authority.
List of Terms	https://www.iana.org/assignments/media-types/media-types.xhtml

Term	<i>LCSH</i>
Full Name	Library of Congress Subject Headings
DC Definition	The set of labeled concepts specified by the Library of Congress Subject Headings
List of Terms	https://authorities.loc.gov/webvoy.htm

Term	<i>LCNAF</i>
Full Name	Library of Congress Name Authority File
DC Definition	The set of authorized personal and corporate names specified by the Library of Congress.
List of Terms	https://authorities.loc.gov/

Term	<i>DCMI Type</i>
Full Name	DCMI Type Vocabulary
DC Definition	The set of classes specified by the DCMI Type Vocabulary, used to categorize the nature or genre of the resource.
List of Terms	http://dublincore.org/documents/dcmi-type-vocabulary/

Syntax Encoding Schemas

Term	<i>ISO639-3</i>
Full Name	ISO639-3
DC Definition	639-3, Codes for the representation of names of languages--Part 3: Alpha-3 code, ISO 639-3 is a code that aims to define three-letter identifiers for all known human languages. At the core of ISO 639-3 are the individual languages already accounted for in ISO 639-2.
List of Terms	https://iso639-3.sil.org/code_tables/639/data/all

Term	<i>W3CDTF</i>
Full Name	W3CDTF
DC Definition	Defines a profile of ISO 8601, the International Standard for the representation of dates and times.
List of Terms	https://www.w3.org/TR/NOTE-datetime

Appendix 5: XML Template and Example

XML Template

```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/documents/2000/11/dcmes-xml/dcmes-xml-dtd.dtd">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description about="http://.../">

    <dc:title>Title</dc:title>
    <dc:creator>Creator</dc:creator>
    <dc:contributor>Contributor</dc:contributor>
    <dc:subject>Subject</dc:subject>
    <dc:description>Description</dc:description>
    <dc:publisher>Publisher</dc:publisher>
    <dc:date>Date</dc:date>
    <dc:type>Type</dc:type>
    <dc:format>Format</dc:format>
    <dcterms:extent>extent</dcterms:extent>
    <dcterms:medium>medium</dcterms:medium>
    <dc:identifier>Identifier</dc:identifier>
    <dc:source>Source</dc:source>
    <dc:language>Language</dc:language>
    <dc:coverage>Coverage</dc:coverage>
    <dcterms:spatial>Spatial</dcterms:spatial>
    <dcterms:temporal>temporal</dcterms:temporal>
    <dc:rights>Rights</dc:rights>
    <dcterms:accessRights>accessRights</dcterms:accessRights>
    <dcterms:rightsHolder>rightsHolder</dcterms:rightsHolder>
    <dc:relation>Relation</dc:relation>

  </rdf:Description>
</rdf:RDF>

```

XML Example

The following example XML template describes this image:



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/documents/2000/11/dcmes-xml/dcmes-xml-dtd.dtd">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description about="http://.../">

<dc:contributor>Zooniverse</dc:contributor>
<dc:contributor>San Diego Zoo Global</dc:contributor>
<dc:subject>Cows</dc:subject>
<dc:subject>Cattle</dc:subject>
<dc:subject>Livestock</dc:subject>
<dc:subject>Kenya</dc:subject>
<dc:description>Camera trap image. Photograph captures four cows </dc:description>
<dc:publisher>Wildwatch Kenya</dc:publisher>
<dc:date>2015-02-05T01:10:08+3:00+</dc:date>
<dc:type>Image</dc:type>
<dc:format>Photograph</dc:format>
<dc:identifier>02050138</dc:identifier>
<dc:source>Camera Trap L1</dc:source>
<dc:language>eng</dc:language>
<dc:coverage>00 32.892 036 51.151</dc:coverage>
<dc:rights>Copyright Protected </dc:rights>

</rdf:Description>
</rdf:RDF>
```

Appendix 6: ReadMe-Metadata Example

Contributor: Zooniverse

Contributor: San Diego Zoo Global

Description: The goal of Wildwatch Kenya is to determine giraffe habitat range. This is accomplished using camera traps (which are activated by movement), aerial data from spotters in planes, and geolocated data from observers on the ground. The images are uploaded from memory cards onto an external drive, and then shipped to the team in San Diego. Copies are then uploaded to the citizen science website Zooniverse, in order to crowdsource the large task of tagging the image. For more information, contact XXXX @email.com

Publisher: Wildwatch Kenya

Date: 2015-02-05

Type: Image

Format: Photograph

Source: Camera Trap L1

Language: eng

Coverage: 00 32.892 036 51.151

Rights: Copyright Protected

Appendix 7: Glacier Storage Index

Camera Trap	WWK Sub Folder Name - Change	WWK Sub Folder Name - Current/Suggested	Approx # of images	WWK Main Folder Name	Zip File Name	Glacier Vault Name
L1		L1_February-May_2016-100EK113	1000	L2_Feb-May_2016	L1_February-May_2016-100EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-101EK113	1000	L2_Feb-May_2016	L1_February-May_2016-101EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-102EK113	1000	L2_Feb-May_2016	L1_February-May_2016-102EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-103EK113	1000	L2_Feb-May_2016	L1_February-May_2016-103EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-104EK113	1000	L2_Feb-May_2016	L1_February-May_2016-104EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-105EK113	1000	L2_Feb-May_2016	L1_February-May_2016-105EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-106EK113	1000	L2_Feb-May_2016	L1_February-May_2016-106EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-107EK113	1000	L2_Feb-May_2016	L1_February-May_2016-107EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-108EK113	1000	L2_Feb-May_2016	L1_February-May_2016-108EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-109EK113	1000	L2_Feb-May_2016	L1_February-May_2016-109EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-110EK113	1000	L2_Feb-May_2016	L1_February-May_2016-110EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-111EK113	1000	L2_Feb-May_2016	L1_February-May_2016-111EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-112EK113	1000	L2_Feb-May_2016	L1_February-May_2016-112EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-113EK113	1000	L2_Feb-May_2016	L1_February-May_2016-113EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-114EK113	1000	L2_Feb-May_2016	L1_February-May_2016-114EK113.zip	L2_Feb-May_2016
L1		L1_February-May_2016-115EK113	516	L2_Feb-May_2016	L1_February-May_2016-115EK113.zip	L2_Feb-May_2016
L2		L2_February-May_2016-100EK113	1000	L2_Feb-May_2016	L2_February-May_2016-100EK113.zip	L2_Feb-May_2016
L2		L2_February-May_2016-101EK113	1000	L2_Feb-May_2016	L2_February-May_2016-101EK113.zip	L2_Feb-May_2016
L2		L2_February-May_2016-102EK113	92	L2_Feb-May_2016	L2_February-May_2016-102EK113.zip	L2_Feb-May_2016
L3		L3_February-May_2016-100EK113	1000	L2_Feb-May_2016	L3_February-May_2016-100EK113.zip	L2_Feb-May_2016
L3		L3_February-May_2016-101EK113	1000	L2_Feb-May_2016	L3_February-May_2016-101EK113.zip	L2_Feb-May_2016
L3		L3_February-May_2016-102EK113	1000	L2_Feb-May_2016	L3_February-May_2016-102EK113.zip	L2_Feb-May_2016
L3		L3_February-May_2016-103EK113	1000	L2_Feb-May_2016	L3_February-May_2016-103EK113.zip	L2_Feb-May_2016
L3		L3_February-May_2016-104EK113	1000	L2_Feb-May_2016	L3_February-May_2016-104EK113.zip	L2_Feb-May_2016

The above example shows the prevailing file naming conventions as the basis for organization in Glacier storage.

Camera Trap	WWK Sub Folder Name - Change	WWK Sub Folder Name - Current/Suggested	Approx # of images	WWK Main Folder Name	Zip File Name	Glacier Vault Name
N52 Low	n52 low namunyak july september 2016 100ek113	N52_Low_July-September_2016-100EK113	805	N2_July-September_2016	N52_Low_July-September_2016_100EK113	N2_July-September_2016
S1 Low	s1 low namunyak july september 2016 100ek113	S1_Low_July-September_2016-100EK113	1000	N2_July-September_2016	S1_Low_July-September_2016_100EK113	N2_July-September_2016
S1 Low	s1 low namunyak july september 2016 101ek113	S1_Low_July-September_2016-101EK113	1000	N2_July-September_2016	S1_Low_July-September_2016_101EK113	N2_July-September_2016
S1 Low	s1 low namunyak july september 2016 102ek113	S1_Low_July-September_2016-102EK113	1000	N2_July-September_2016	S1_Low_July-September_2016_102EK113	N2_July-September_2016
S1 Low	s1 low namunyak july september 2016 103ek113	S1_Low_July-September_2016-103EK113	157	N2_July-September_2016	S1_Low_July-September_2016_103EK113	N2_July-September_2016
S2 Low	s2 low namunyak july september 2016	S2_Low_July-September_2016-100EK113	579	N2_July-September_2016	S2_Low_July-September_2016_100EK113	N2_July-September_2016
S3 High	S3 HIGH - unfolded	S3_High_July-September_2016-100EK113	1000	N2_July-September_2016	S3_High_July-September_2016_100EK113	N2_July-September_2016
S3 High		S3_High_July-September_2016-101EK113	1000	N2_July-September_2016	S3_High_July-September_2016_101EK113	N2_July-September_2016
S3 High		S3_High_July-September_2016-102EK113	539	N2_July-September_2016	S3_High_July-September_2016_102EK113	N2_July-September_2016
S6 Low	s6 low namunyak july september 2016 100ek113	S6_Low_July-September_2016-100EK113	1000	N2_July-September_2016	S6_Low_July-September_2016_100EK113	N2_July-September_2016
S6 Low	s6 low namunyak july september 2016 101ek113	S6_Low_July-September_2016-101EK113	1000	N2_July-September_2016	S6_Low_July-September_2016_101EK113	N2_July-September_2016
S6 Low	s6 low namunyak july september 2016 103ek113	S6_Low_July-September_2016-102EK113	1000	N2_July-September_2016	S6_Low_July-September_2016_102EK113	N2_July-September_2016
S6 Low	s6 low namunyak july september 2016 104ek113	S6_Low_July-September_2016-103EK113	1001	N2_July-September_2016	S6_Low_July-September_2016_103EK113	N2_July-September_2016
S7 Low	s7 low namunyak july september 2016 100ek113	S7_Low_July-September_2016-100EK113	1000	N2_July-September_2016	S7_Low_July-September_2016_100EK113	N2_July-September_2016
S7 Low	s7 low namunyak july september 2016 101ek113	S7_Low_July-September_2016-101EK113	1000	N2_July-September_2016	S7_Low_July-September_2016_101EK113	N2_July-September_2016
S7 Low	s7 low namunyak july september 2016 102ek113	S7_Low_July-September_2016-102EK113	1000	N2_July-September_2016	S7_Low_July-September_2016_102EK113	N2_July-September_2016
S7 Low	s7 low namunyak july september 2016 103ek113	S7_Low_July-September_2016-103EK113	1000	N2_July-September_2016	S7_Low_July-September_2016_103EK113	N2_July-September_2016
S7 Low	s7 low namunyak july september 2016 104ek113	S7_Low_July-September_2016-104EK113	1000	N2_July-September_2016	S7_Low_July-September_2016_104EK113	N2_July-September_2016
S7 Low	s7 low namunyak july september 2016 105ek113	S7_Low_July-September_2016-105EK113	977	N2_July-September_2016	S7_Low_July-September_2016_105EK113	N2_July-September_2016
S8 High	S8 HIGH - unfolded	S8_High_July-September_2016-100EK113	1000	N2_July-September_2016	S8_High_July-September_2016_100EK113	N2_July-September_2016
S8 High		S8_High_July-September_2016-101EK113	1000	N2_July-September_2016	S8_High_July-September_2016_101EK113	N2_July-September_2016
S8 High		S8_High_July-September_2016-102EK113	1000	N2_July-September_2016	S8_High_July-September_2016_102EK113	N2_July-September_2016
S8 High		S8_High_July-September_2016-103EK113	175	N2_July-September_2016	S8_High_July-September_2016_103EK113	N2_July-September_2016
N27 High	n27 high namunyak july september 2016	N27_High_July-September_2016-100EK113	602	N2_July-September_2016	N27_High_July-September_2016-100EK113	N2_July-September_2016

The above example shows current WWK sub folders which have non-uniform naming conventions in column 2 in red. Our recommendations on how to change sub folder names, ZIP file names, and folder/ZIP file sizes are outlined in red in columns 3,4,and 6.

Full indices can be found for each conservation at the following links.

Loisaba: https://docs.google.com/spreadsheets/d/1PRWh4gSM67pjQmp-AMXs67BFzagHP851_En9daQPFIM/

Namunyak: <https://docs.google.com/spreadsheets/d/1KjcmBi2JTalwCzgAIOrjal6E7WIBfYte7h388VWrAGg/>

Appendix 8: Legacy Metadata Template

You can access the Legacy Metadata Template through the following link:

<https://docs.google.com/spreadsheets/d/1at222yVAW2DAGhyY0C0Pc4B5IfZLSz7VRhkD187J5d0/edit?usp=sharing>

You can access the partially completed pilot crosswalk of the WWK data to the Legacy Metadata Template through the following link:

<https://docs.google.com/spreadsheets/d/1571Y6rMK4y2k29nwzVoCIJrfCIImRGfInoDNIBbLFvM/edit#gid=801762637>