Data integration meeting/conversation -- 18 July 2019

Phil Miller notes

Meeting Participants:
Phil Miller, CPSG
Dalia Conde, Species360
Jim Guenter, Species360
Doug Verduzco, Species360
Taylor Callicrate, SCTI (remote)
Bob Lacy, SCTI (remote)

PROBLEMS TO SOLVE
1. Dalia -- Need to plan for more species -- want to see how to serve species conservation. And we don't know for a given species, what levels of data -- biological, legislative, genetic, threats, managerial -- are available.
2. Doug -- want to serve the community, don't now what the problmes are, just want to help.
3. Phil -- two levels of organization: do more work on more species, and do better work on single species (e.g., chimps in Liberia)
3A. Taylor -- Bob and I have talked to Apex-RMS (Canada) about how they integrate all sorts of data for their predictive work. We should talk to people like that...
3B. Dalia -- yes, we don't want to re-invent things
4. Dalia -- we talk to TAGs...our members ask us, how do we prioritize species conservation work? how do we prioritize our collections? We're beginning to bring in-situ people -- Arnaud and Pati -- into ZIMS to implement a One Plan Approach for specific projects
5. Taylor -- really focused on the implementation perspective...how do we bring in tech like machine learning to improve our tool development, and how do we access knowledge around existing tools to do our job better? How do we access tools that assess global trends, emerging information, etc. We don't have these partnerships in place that will be key to our evolution.
6. Jim -- So is the problem: We don't have the data? You can't find the data? The data are not standardized?

Lots of overlap in our issues/problems/challenges!!! We gotta start working together...

Dalia will send papers on case studies of data integration that have come out of her lab...NOAA shipwrecks, etc.

Jim -- update on Conservation Science Alliance... Dalia's team in Copenhagen...also have three sponsors: WRS Singapore, Copehagen Zoo, WAZA

Bob -- who is our audience for this product? And based on this, what kind of interface do we develop to satisfy these audiences?
Jim -- Our audience is very broad: governments, specialists, academicians, zoo people, etc. We should probably start small and work our way outwards to bite off something we can chew reasonably effective...
Phil -- hmmm...do we start simple for everyone, or do we start complex for a professional and then make it simpler?

Bob -- We really need to scale up our activities...more data on more species...need to scale up our ability to collect, assemble, integrate and analyze more data to make a bigger impact.
Phil -- what about just assembling PVA model inputs into a single facility so people can look at this?! Low-hanging fruit for lots of information on lots of speces. Perhaps this is a simple, short-term, effective way to gather information. And when doing RedListing, we need to have a facility for accessing very basic information on what's known about, for example, amphibians to facilitate that knowledge.
Dalia -- Maybe the Species Knowledge Index has some of that information, but probably not for amphibians. My idea is to expand that Index to include many more types of information

UPDATE ON SPECIES360 HUB -- Doug
Possible to get his presentation?

SOLUTIONS
Jim -- our goal is to develop something like a Wikipedia page for each species, linking all the "data" that is pulled in to get a better understanding of the species, its threats, its in-situ and ex-situ status, and its conservation opportunity. Species 360 is working with Red List people to get ex-situ data into RL assessments.

First big database conference: Biodiversity Next -- Netherlands, October 2019. Doug and Johanna will be there...

Bob -- Low-hanging fruit: Develop the Species Knowledge Index into a facility that facilitates Red List assessments, providing information specifically for the questions needed to make a RL assessment. Would need demographic data like generation length, but would also need info on AOO, EOO, threats, etc.
Jim -- that seems too easy...why hasn't this happened already? Do the data not exist, or they're too hard to find? Need to find this out, across multiple taxa...
Phil -- a tool like an expanded Species Knowlede Index could be really useful in our A2P process to facilitate improved Red Listing and assessing for the purposes of planning.

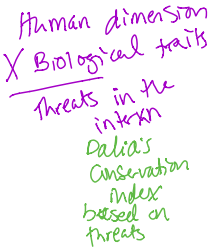
PROPOSED PHASE 1
Think about developing Dalia's Species Knowledge Index into a tool to access/provide data that are appropriate for implementing a Red List assessment across multiple species. This includes data on basic species biology/demography, distribution, and threats (ideally spatially explicit and perhaps prioritized (although that could be a post-processing task), per an A2P process).
Jim -- Phil, would this kind of tool be valuable to you?
Phil -- absolutely...we're pushing the A2P process, and this would be a really valuable way to facilitate expanded Red Listing across a larger number of taxa.
Bob -- Need to think about bringing Rest Akcakaya into this process, as he's been involved in building RAMAS Red List.
Dalia -- we don't want to step on the Red List people's toes...gotta be careful about this...this can be a more generic assessment (A2P-type) tool that we can use to facilitate later planning, and others can use to ultimately facilitate more specific Red Listing.
Jim/Phil -- this is sounding alot like the existing Red List database...are we doing someting new?
Dalia -- yes, as there's lots of data that are available that are NOT already in the Red List database...
Phil -- temperature check...where are we?
Bob -- I think we're going down a good road...we need to develop a system that allows users to access ALL the data that are helpful for status assessment, so this is a good thing. The issue I see is around resources...how do we get started? Where do we start? What's the prototype that can demonstrate proof of concept, that we can show as a successful first step?

So...we need to figure out (for, perhaps, putting together a funding proposal):
- What exactly is this thing?
- Who will be the audience?
- What kinds of data will be in it?
- How will data ownership work?
- Who has access to the data?
- Who owns the tool itself?
- Who will be our collaborators to make it happen?

How much does this thing cost?
- Depends on: how big it is, and how fast you want it...$200,000 would go pretty quick...

Notes from Taylor Callicrate follow on the next page...

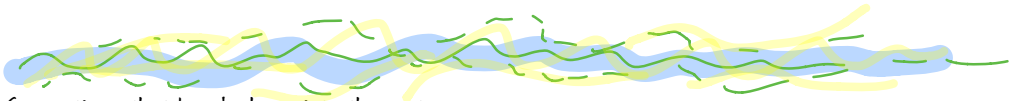
- Data needs & what problems are we trying to solve
  - Different projects have different types of data needs, but there may be some patterns
  - First need to establish what are the problems that we're running into with data access & availability
    - Dalia: conservation planning needs to be done for thousands of species; how do we work together to provide this information and also develop the tools to use it for massive-scale planning
    - Species360 has data that will be useful for species conservation, but also welfare of captive animals (health trends
      - Will make data available for research
  - From the end user standpoint, specialist groups will be coming to CPSG to do planning for all the species they've red-listed
    - Need access to a wider breadth of species information for higher-level groups ('grasshoppers' or 'freshwater fish')
    - Also need some species-specific information (I.e., chimpanzees in Liberia)
      - In addition to typical PVA info, need info about the government, human population, human age structure, distribution of oil palm plantations... and projections for all of these things
      - Species knowledge index may not be as useful when we need all this contextual information
  - We'll make a Google doc to share a list of other groups that have similar problems that we might want to reach out to
    - Like Apex RMS- they may need/use this contextual data for landscape modeling
  - Really need to be able to rapidly gather large amounts of data for assessments
    - These data are in various states: spread amongst existing databases, found in the literature, multiple formats...
    - Phil would like some kind of data portal where you can search on a taxa, and then be able to see all the information
      - It would link to other databases that have info on this species
      - Biology of related taxa
      - Projections for sea level rise or other relevant contextual/ancillary information like feral cats
  - With ancillary information, it's a two-level question
    - Is that data stream relevant (how many feral cats, and are they even a threat?)
    - If it is relevant, then how does it impact our model input?
    - Dalia's center is working to fill these knowledge gaps
    - Scale of this data is also important - both temporal and spatial



- Talking about the hub and what it would look like
  - Links between databases (nodes) that meet at key nodes where the data are aggregated and displayed
  - Working only with open data due to the MOU with Red List(?), working with Open Air and Prayce(?) and science data alliance, things that the EU is developing to support research and policy makers (Dalia will send more info on this)
  - Need to develop a working model first before having a hub that dynamically pulls info from each node and displays current data in real-time according to user request
- Doug: presentation about the data hub design

    - Who do we want to serve?
        - Especially targeting how we engage governments more effectively
    - Types of data
        - Aggregated, shared data (ZIMS)
            - Belongs to Species360
        - Local shared data (shared across ZIMS database)
        - Local shareable data (possible to share but maybe no mechanism in place)
        - Local-only data
            - All of these belong to the ZIMS members
    - What should 'open' data look like for Species360, considering organizational sustainability
        - If all data is opened up, may lose members
        - Some data will only be open to membership (medical data, for example)
        - Free access to the public (general trends, aggregate summary stats, etc)
        - A medium tier with more info than free that would be monetized
        - Kind of a mall of America concept with different hubs with access to different bits of data targeted to different audiences
        - Links with CITES, Red List, but also human population and urbanization databases, possibly via Dalia's center a U Denmark
    - Species360 making links with IUCN
        - Data standardization has been an issue
            - Vocabulary (I.e., what do you call the age of first reproduction?)
            - Taxonomy
        - Have worked with an R programmer to standardize data as you pull it
            - R open science group
        - Linkages with Oceanarium (Joal? In Portugal) and with Jon Paul to work on linkages with IUCN
            - How can we put data from oceans together to advise policy processes?
            - This is maybe a case study of a single institution that wants to work on bringing data together, and could be considered another data hub
    - BiodiversityNEXT conference about biodiversity databases in the Netherlands in October
        - Lots of discussion on operational parts of it (like standardization)
        - Implementation of machine learning to facilitate this communication
    - Linkages with GBIF which is working on linking different types of databases
- Conservation Science Alliance
    - WRS, WAZA, and Copenhagen are sponsors supporting Dalia's research team
    - The goal is to do some strategic research to illustrate the value of the data for conservation, and inspire others to do research with the data and then facilitate that work

Afternoon Session



- Question categories - 6 questions that break down into three streams

    - What data do we want to improve our conservation planning?
        - And how much of those data are actually available/accessible
    - Who's already working in this space of data accessibility/integration?
        - How do we most effectively work with them and others?
    - What are the gaps in existing work on data accessibility/integration?
        - How do we act to begin filling the gaps
- Based on these questions, how can we work together as a group of groups, given that there is quite a bit of overlap between our needs and expertise
    - We already have some conversation summaries from previous CPSG meetings regarding what we want
    - Dalia (and others?) will distribute those summaries
- Who is this for?
    - Have to carefully consider who it's for in deciding the design and how user-friendly it is
        - To consider the balance between technical and powerful, and being user-friendly
- Need to be careful that we're not trying to create a mega database, but rather a meta database
- What can we do relatively quickly/simply that will have a big impact?
    - With red listing (amphibian example), right now a bunch of experts sit around a table and think about who might know something
        - Hard to know if the data exist, or if they just exist in hard to find formats, or something else
        - And the challenges of assembling the data will be different for different taxa
    - It would be easier to have a database or index that would just tell you if that info is available and where it is
    - Demographic knowledge of species index hosts some data, but for some users have to click to visit external database
    - Would be difficult to query databases using a tool that just pulls from various databases because they each have their own rules for who can run queries; there are legal complications
- Using the species knowledge index to facilitate red list assessments
    - Would need distribution and threats information
    - What's been done so far
        - Using simple text mining to get the threats
- Might be possible to make a prototype using data sources that are open, to develop an aggregator tool that will pull data needed for red list assessments
    - First phase: red list assessments via providing or directing to data using the species index
        - Possibly letting users select which variables/data streams are important to them, for example region so they could retrieve data related to threats in that region only
    - Second phase: with more partners, data
    -