**ASU Biodiversity Data Science Workshop**

*November 28, 2018*

Decision Theater, Orchid House at the Brickyard, Tempe, AZ 85281

**Discovering biodiversity data signals from natural history collections**

*Nico M. Franz* ([nico.franz@asu.edu)](mailto:nico.franz@asu.edu))

**Abstract:** This presentation will illustrate how digitized, networked biodiversity data from natural history collections can be discovered and visualized to foster an understanding of data availability and suitability of data-driven analysis and decision making. The focus is on Arizona and on data and services produced by Arizona State University's Natural History Collections. The presentation will demonstrate how ASU's biocollections are currently discoverable through a series of public portals sustained by the Symbiota software platform - a popular, open source, ASU-developed information management system that offers a broad range of biodiversity data research and learning opportunities (<https://doi.org/10.3897/BDJ.2.e1114)>. The examples will hopefully stimulate discussions about integrating collections-based data with the overall workshop theme and its different components.

* Google: "How many mammal species are there?"
* Google: "How many insect species are there in Arizona?"
  + Refinement: "How many cicada species are there in Arizona?"
* Go instead to: <http://scan-bugs.org/portal/>
  + Regular Collections search for Cicadidae / Arizona
  + Explore Species List – according to various taxonomic perspectives
  + What about just ASUHIC?
  + What about all Insecta species for Arizona? ~ 17,000 available records..
* Another perspective: Visualize ant species diversity in the Chiricahua Mountains via Map Search
  + Download and explore the results in .csv format
* Another perspective: Dynamic Key search of members of the daisy/sunflower family in the McDowell Mountain Range at: <http://swbiodiversity.org/seinet/index.php>
  + Explore the dynamic key, then then specimens upon which the distribution map is based
* Another perspective: <http://swbiodiversity.org/seinet/checklists/checklist.php?clid=3>
* Stepping back a bit: How and when did natural history collection digitization originate?
  + 1970s herbarium databases; 1980s formation of TDWG; ~ 1995-2005: emergence of the Darwin Core Standard
    - <https://doi.org/10.1371/journal.pone.0029715>
    - <https://dwc.tdwg.org/terms/>
* Meta-aggregation of natural history collection and observation records
  + Symbiota: <http://symbiota.org/docs/>
    - <http://scan-bugs.org/portal/collections/misc/collprofiles.php?collid=121>
  + iDigBio: <https://www.idigbio.org/>
    - <https://www.idigbio.org/portal/recordsets/ab4b6a2b-a90a-44ce-95a1-2c44c911fcc6>
  + GBIF: <https://www.gbif.org/>
    - <https://www.gbif.org/dataset/750b7bfc-3577-4b26-8aaf-3e4be9f0d639>
* Issue of de-/centralization – good quality control pushes towards decentralization; global coverage pushes towards centralization – where is the right technological, sociological, and economic path to both? What role do standards play, versus extra-standard practices?
* And what are we learning?
  + <http://rstb.royalsocietypublishing.org/content/374/1763>
  + <https://doi.org/10.1371/journal.pone.0185809>
  + <https://www.nytimes.com/2018/11/27/magazine/insect-apocalypse.html>
  + <https://www.neonscience.org/>
  + <https://asunow.asu.edu/20180821-asu-receives-multimillion-grant-nsf-create-national-biorepository>
  + <https://doi.org/10.1080/14636778.2018.1546573>
* What (else) is missing?
  + Data discovery: signal properties (data quality, coverage, decisiveness, etc.) and signal message – typically require an intimate (independent) understanding of data bodies.
    - Sample question: which dataset is the most complete / suitable for a particular inference task?
  + Trend discovery (Δ) – typically requires advanced data analysis skills and *ad hoc* workflow / versioning routines.