Information extraction, dataset referencing and linked-data research at ACIS/U. Florida

José Fortes
Advanced Computing and Information Systems (ACIS) Laboratory
University of Florida

Outline

- Sample research projects @ ACIS
  1. Integrated Digitized Biological Collections (iDigBio)
  2. Self-aware Information Retrieval
  3. Data references and citations
  4. Linked data in the biological collections domain
- Conclusions

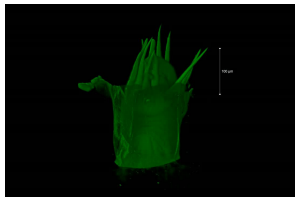
Digitization of Biocollections

- Information in biocollections can be used to understand environmental change, contaminants, biological invasions, disease transmission, and agriculture, among many other important areas.
- There are about 1 Billion specimens in Biocollections in the USA and about 3 Billion in the whole World (Estimates).
- NSF's Advancing Digitization of Biodiversity Collections (ADBC) program.

Photo by Chip Clark. U.S. National Herbarium at the Smithsonian Institution's National Museum of Natural History. Featured researchers: Dr. James Norris (right, front), research assistant Bob Sims (left, front), and associate researcher, Katie Norris (left, back).

Specimen Data: what are they?

- Records
- Images
  - Specimen
  - Labels
  - Thumbnails
  - 2D, 3D, …
- Sounds
- Video
- Other
  - Metadata
  - Providers
  - …

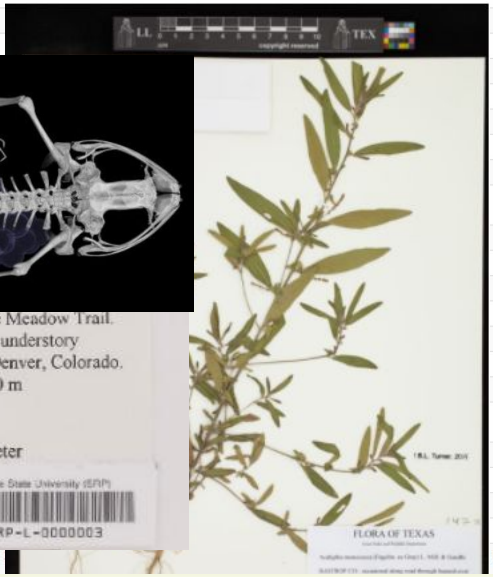


Acalypha monococca (Engelm. ex Gray) L. Mill. & Gandhi

From University of Texas Herbarium

Continent	North America	Institution Code	Prc
Country	United States	Collection Code	Tex
State/Province	Texas	Catalog Number	Tex00466196
County/Parish	Bastrop	Collected By	W.r. Carr & Mik
Locality	Both Sides Of Main Road On Hoppy Spring Tract Ca. 0.1 Mi Se Of Gate On St. Rt. 21, Ca. 0.3-0.4 Airmiles Sw Of Jct. St. Rt. 21 And F. M. 1441 Ne Of Bastrop. Bastrop State Park. Smithville Nw Quadrangle. Elev. 480 Ft.	Date Collected	2012-05-15
Latitude	30.147		
Longitude	-97.229283		

Taxonomy	
Scientific Name	Acalypha monococca
Kingdom	Plantae
Phylum	Magnoliophyta
Class	Magnoliopsida
Order	Euphorbiales
Family	Euphorbiaceae
Genus	Acalypha
Specific Epithet	monococca
Scientific Name Au	(Engelm. ex Gray) L.
Specimen	
Identified By	B. L. Turner (TEX), 2014
Catalog Number	TEX00466196
Reproductive Cond	flowering=early
Institution Code	PRC
Collection Code	TEX
Occurrence ID	urn:uuid:292610ad-5287-4a88-b
Basis of Record	PreservedSpecimen
Collection Event	
Collected By	W.R. Carr & Mike Powers
Collector Number	30593
Date Collected	2012-05-15
Year	2012
Month	5
Day	15
Start Day of Year	136
Locality	
Country	United States
State / Province	TEXAS
County	Bastrop
Continent	north america
Locality	Both sides of main road on Hoppy Spring Tract ca. 0.1 mi SE of gate on St. Rt. Bastrop. Bastrop State Park. Smithville NW Quadrangle. Elev. 480 ft
Habitat	Along road through burned-over drained, neutral to slightly acid fi Associates included Acalypha m Commelina erecta, Coreopsis sp Eragrostis 30592, Gamochaeta hookeriana, Panicum brachyantl
Decimal Latitude	30.147
Decimal Longitude	-97.229283
Verbatim Coordinat	30d 8.82m N; 97d 13.757m
Other	
Modified	2015-08-10 22:40:30
ID	1320606
dwc:datasetID	046bbc50-cae2-47ff-aa43-729fbf53f7c5
dwc:collectionID	688bf697-a37c-4d62-8fb2-2fb90259ee4a
dcterms:references	http://prc-symbiota.tacc.utexas.edu/collections/individual/index.php?occid=1320606

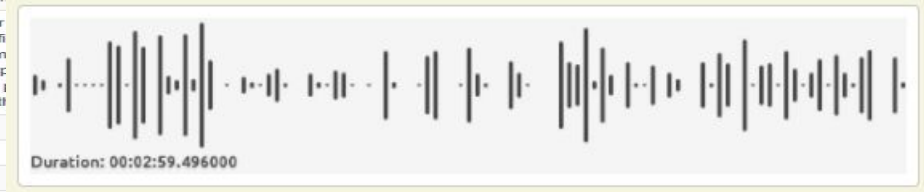


Pseudotsuga - Pinus ponderosa forest above Meadow Trail.
Moderate north slope, mixed age with open understory
Mt. Falcon Open Space Park, just west of Denver, Colorado.
39°37.85'N, 105°13.84'W elevation: 2400 m
FHM off-frame plot CO-7 13 June 1995

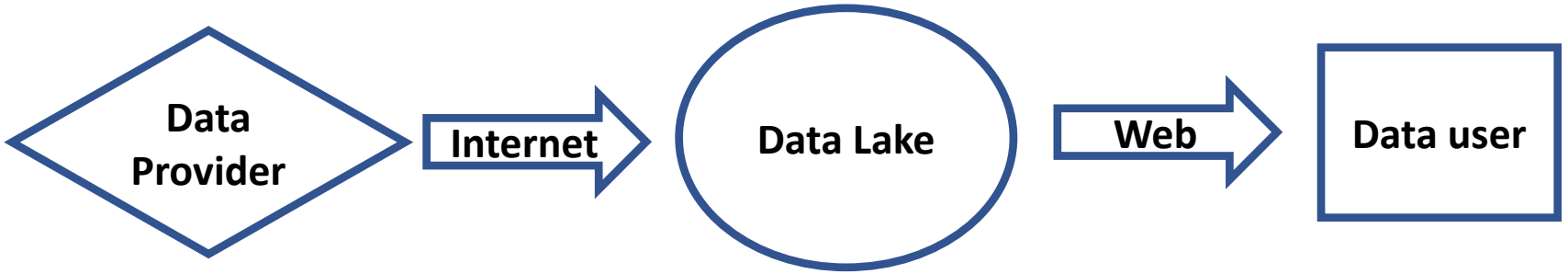
Coll: A. DeBolt, B. McCune, & R. Rosentreter
Rosentreter # 9305 -(b)

39.6308, -105.2306

SRP-L-0000003



Basic Aggregator Cyberinfrastructure

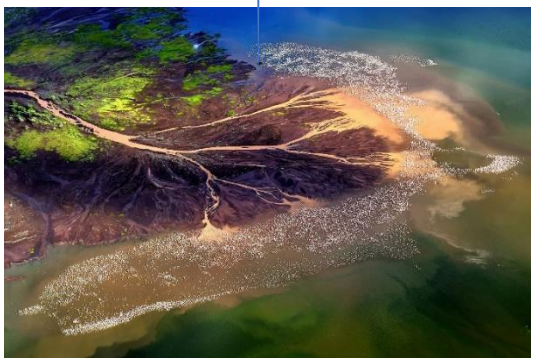


Specimens → Digitization → Records and media → Aggregated datasets → Queried datasets → Research → Scientific results

Information extraction

Running code on resources

Efficient data analytics



PHOTOGRAPH BY FRANCO CAPPELLARI,
https://www.reddit.com/r/Kenya/comments/1aubre/stunning_aerial_view_of_lake_bogoria_photo_by/
https://commons.wikimedia.org/w/index.php?curid=9880216

Information Extraction (IE) Challenge

Automated IE: Optical Character Recognition (**OCR**) + Natural Language Processing (**NLP**)

- Biocollections' images are problematic for OCR engines
- OCR results are not perfect. Handwritten text is especially problematic.

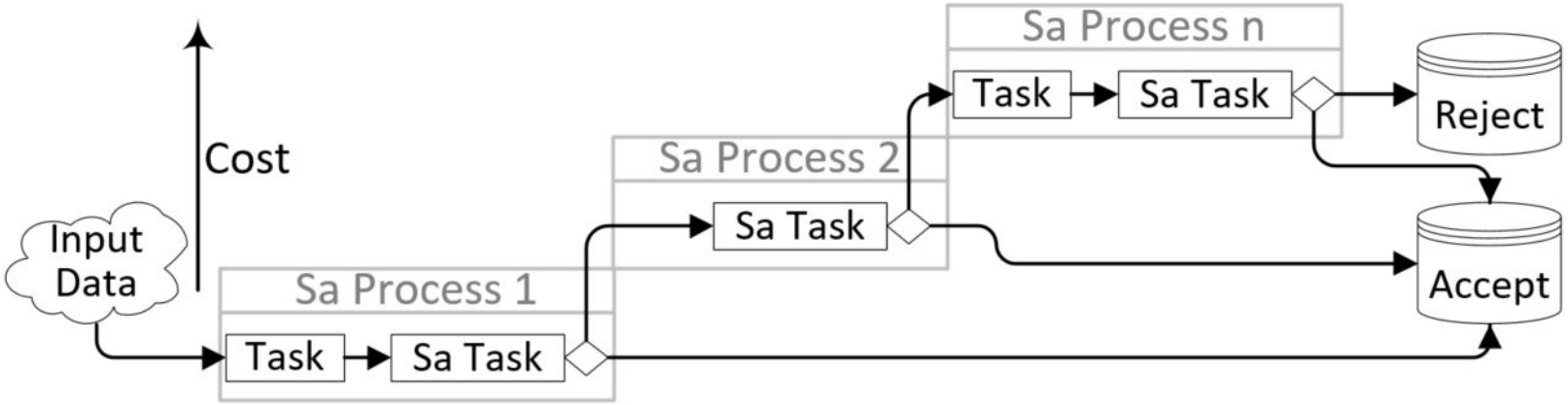
Genus: Carbrunneria
Specific Epithet: marci
Scientific Name: Carbrunneria marci
Country: Australia
Verbatim Locality: Mount Baldy, Loop Road, near Atherton, Herberton Range
Latitude: -17.266
Longitude: 145.416
Institution Code: AM
Scientific Name Authors: (Roth, 1991)
Catalog Number: K.482255
Date Collected: 2011-05-29

Self-aware computational task (ChatGPT)

- A type of task performed by a computer system or artificial intelligence (AI) that involves the system's ability to monitor its own performance, analyze its own behavior, and adapt its own algorithms or strategies to achieve better performance or optimize its performance based on changing conditions or feedback.
- In other words, a self-aware computational task is one in which the system is not only able to perform a specific task or set of tasks, but also able to reflect on its own performance, identify potential areas for improvement, and modify its own behavior or processes accordingly. This level of self-awareness is typically achieved through the use of advanced machine learning and artificial intelligence algorithms, which enable the system to learn from experience and adjust its own behavior based on that learning.

SELFIE: Self-aware Information Extraction

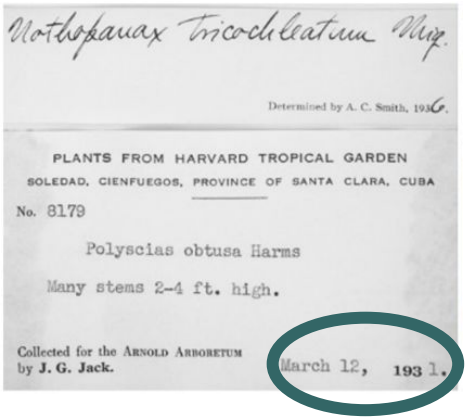
- Workflow of Self-aware Processes (SaP) consisting of Self-aware Tasks (SaT) and possibly other tasks



Self-aware Task (SaT)

Part	Input	Adaptable Script/program	Adaptable Acceptance Method	Outputs
Example	*Image x*	*/path/script1.py*	*[0,b) -> Task y* *[b,1] -> Accept*	*Image x* *Value, Confidence*

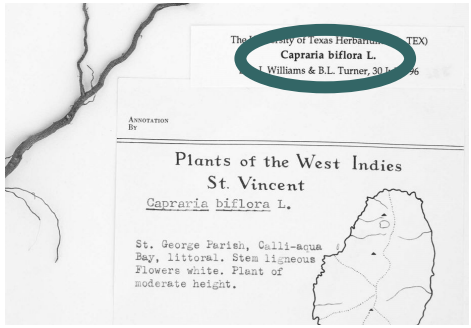
Information Extraction Examples



Dates

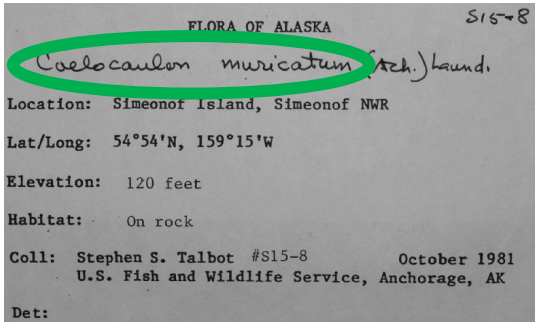
March 12, 1931

VI-4-60

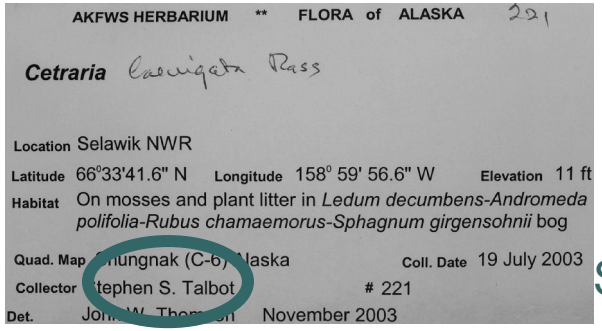


Capra**ria** biflo**ra**

Scientific names

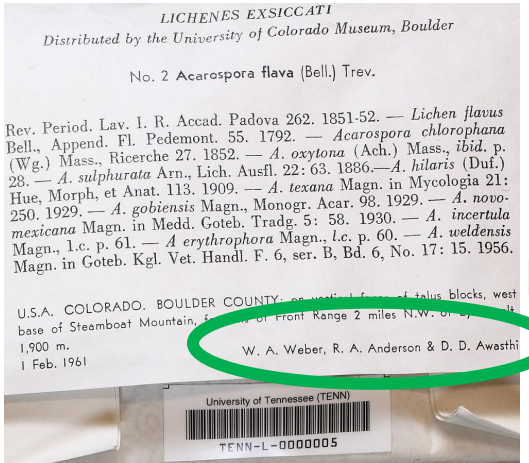


Caelocaulen muricat**um**



Collector names

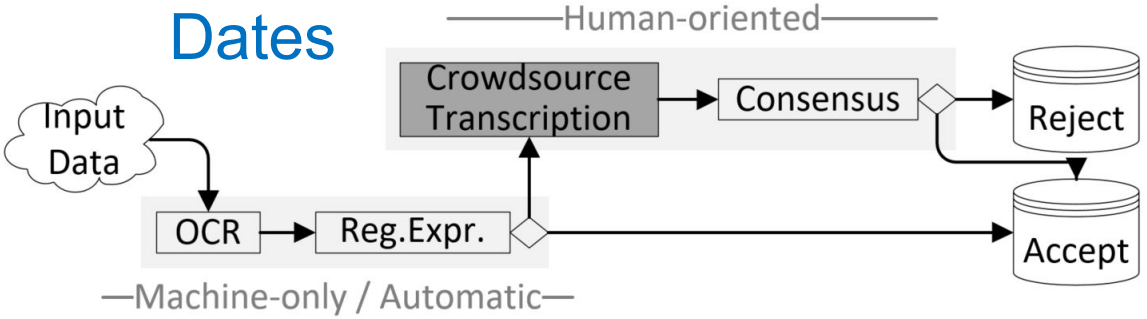
Stephen S. Talbot



W. A. Weber, R. A. Anderson & D. D. Awasthi

SELFIE Workflow Examples

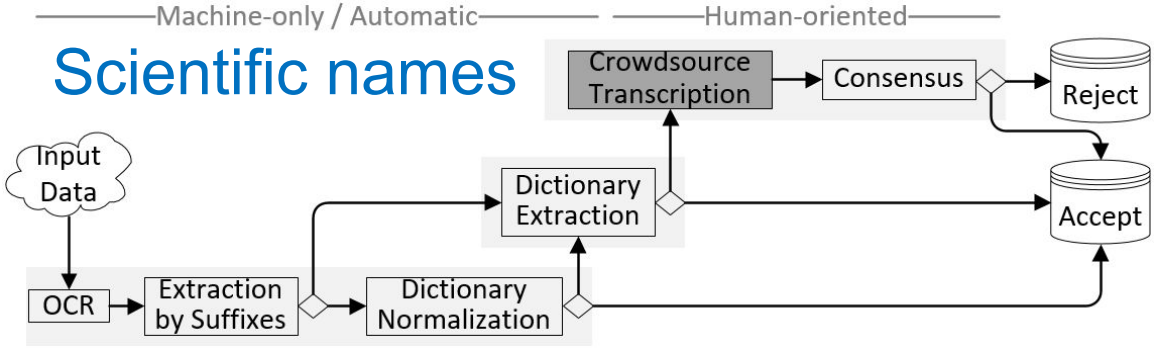
Dates



OCR: OCR software generates text with all information in the image.

Reg. Expr.: analyzer returns earliest date amor the "long" dates.

Scientific names

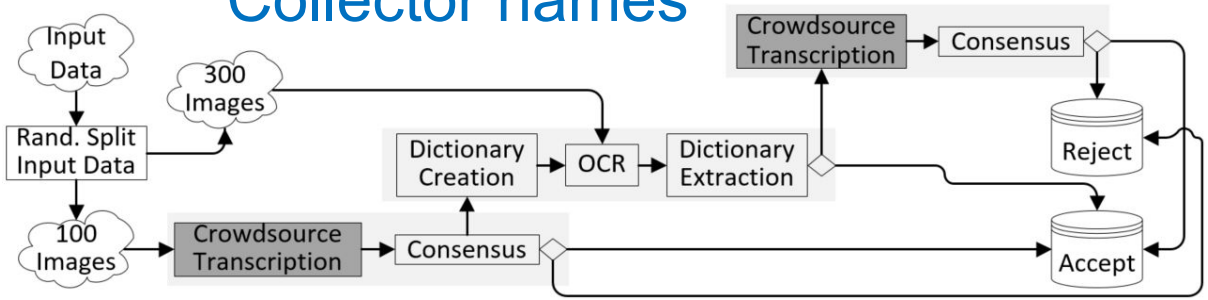


Extraction by Suffixes + Dictionary Normalization

Or/followed by

All text scan + Dictionary Extraction

Collector names



OCR + Consensus + Dictionary Creation

followed by

OCR+ Dictionary Extraction + Crowdsource + Consensus

Similarity to expert transcription

Dates

SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
Machine-only	48	0.934	0.024	0.167
Human-only	51	0.971	0.022	0.155
SELFIE	99	0.953	0.016	0.162

Scientific names

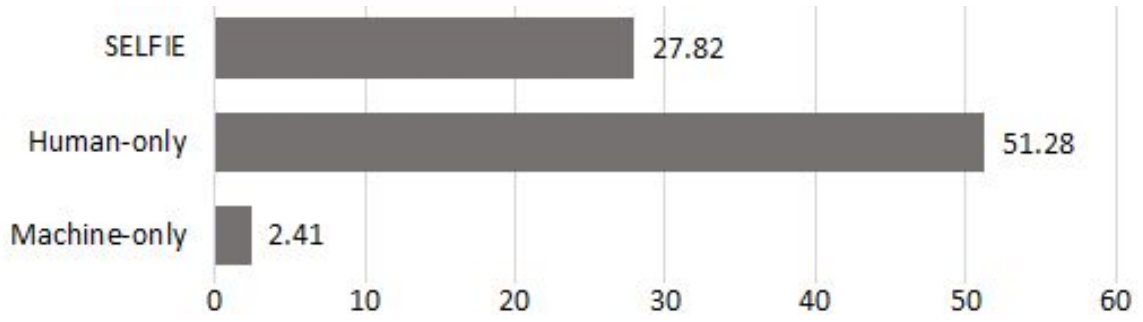
SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
1. Suffix+Dict.Ex.	15	1.0	0.00	0.00
2. Dict. Ex.	10	1.0	0.00	0.00
3. Crowd	66	0.944	0.026	0.214
SELFIE	91	0.959	0.019	0.183

Collector names

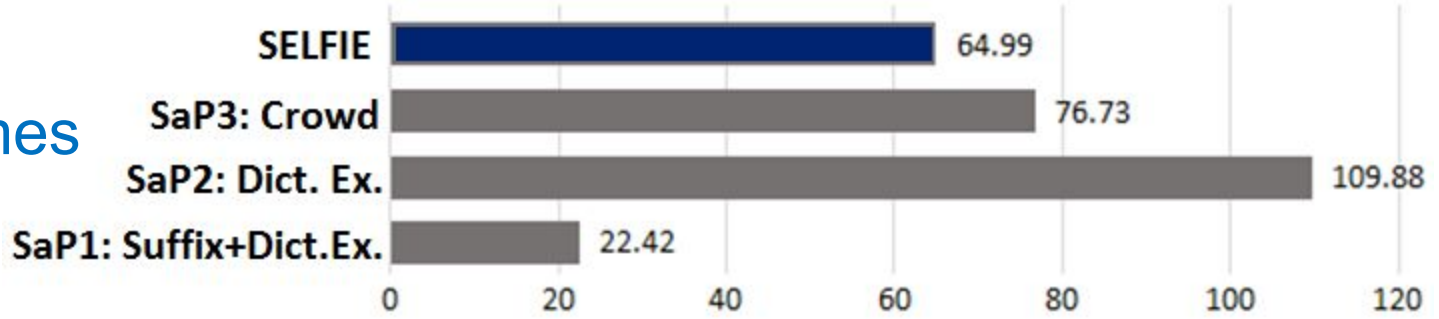
SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
1. Human 100i	92/100	0.900	0.030	0.288
2. Machine-only	94/300	0.862	0.027	0.262
3. Human 300i	191/206	0.900		
SELFIE	375/400	0.895		

Seconds per accepted item

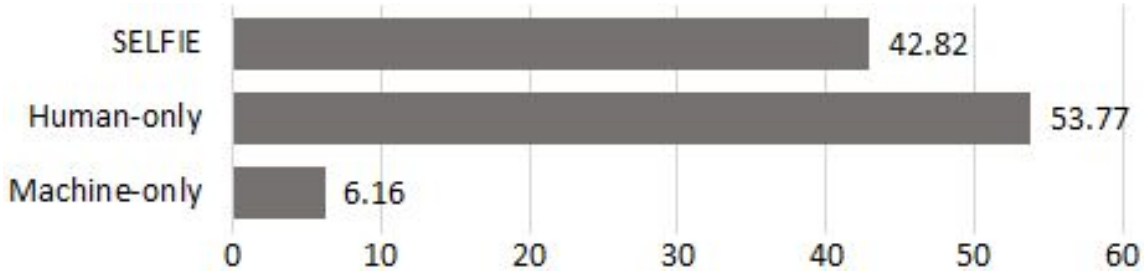
Dates



Scientific names



Collector names

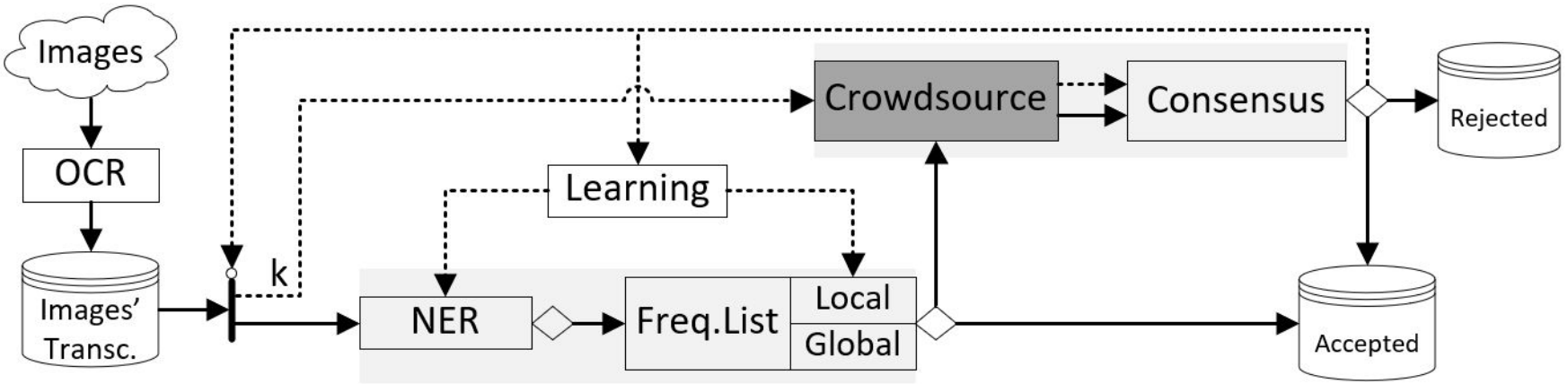


General Extraction of Candidate Values

- Optical Character Recognition
  1. The entire text of the images is extracted (Google Cloud OCR engine was used).
- Named Entity Recognition (NER)
  - Challenge: **Train** the recognition model (**training values**)
  - Algorithm for Automated Labeling of DC Terms from Crowdsourced Data:
    2. Crowdsourced values are searched in the sentences of the OCR-ed text of their correspondent images.
    3. Training sentences are prepared (start – end positions of every term in the sentence).
  - The NER model is trained with the training sentences.
    4. Use the training set of sentences to customize NER model in the NLP library (spaCy was used).

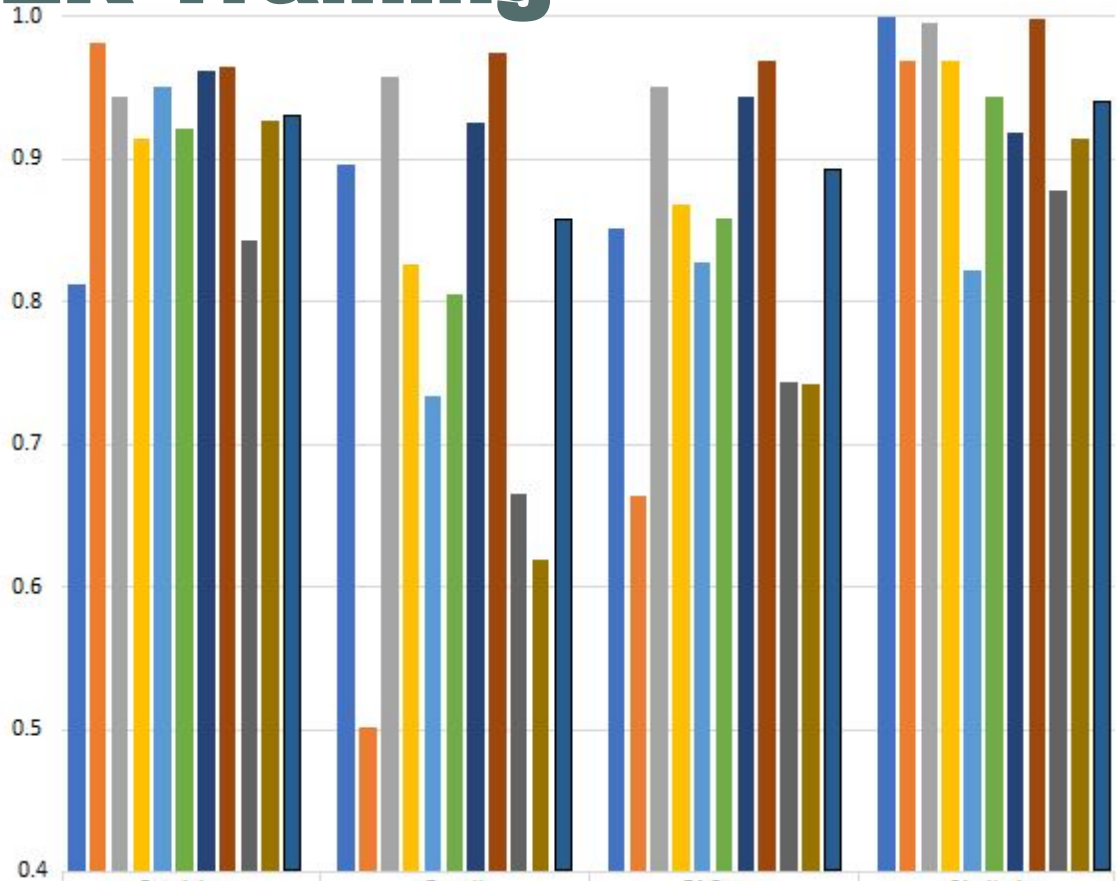
Human-in-the-Loop Workflow for DC Terms Extraction

- Iterative training of the NER model.
- When no previous data is available or useful for a specific term.



Experiments – NER Training

- The trained NER model was able to extract ~ 86% of **ALL** 10 DC values with a similarity to the ground-truth data of 0.931



	Precision	Recall	F1 Score	Similarity
cN	0.812	0.896	0.852	0.999
cry	0.981	0.502	0.664	0.968
cty	0.944	0.957	0.951	0.995
eD	0.914	0.826	0.868	0.969
ha	0.950	0.734	0.828	0.822
rB	0.921	0.805	0.859	0.944
sN	0.962	0.925	0.943	0.918
sP	0.964	0.974	0.969	0.998
vLa	0.842	0.665	0.743	0.879
vLo	0.927	0.620	0.743	0.914
all_10	0.931	0.859	0.893	0.941

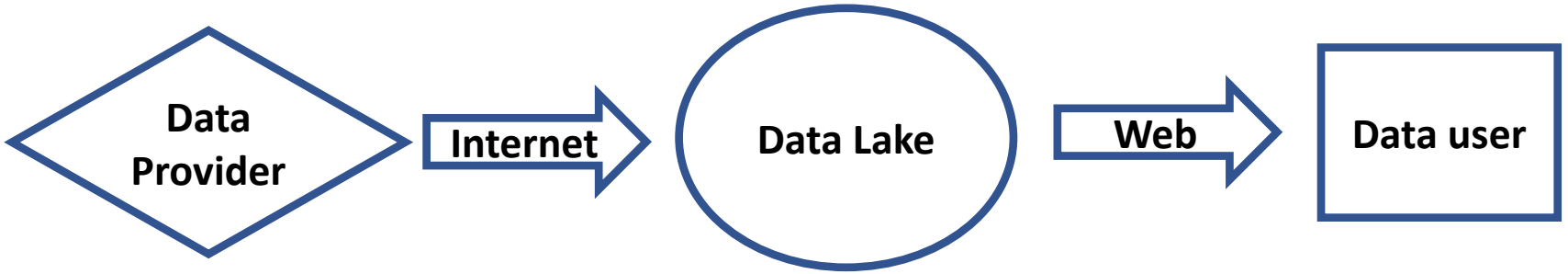
Darwin Core terms:
cN: catalogNumber
cry: country
cty: county
eD: eventDate
ha: habitat
rB: recordedBy
sN: scientificName
sP: stateProvince
vLa: verbatimLatitude
vLo: verbatimLongitude
all_10: weighted average for all the terms

Confidence for Candidate Values

- Global Frequency List (~static)
  - Per-term number of times every value has repeated.
  - Created with the data of iDigBio: more than 120 Million records.
- Local Frequency List (~dynamic)
  - Created from specimens already processed in the biocollection under study

- Algorithm (candidate_value, local_list, global_list):
  - If repetitions(candidate_value, local_list) >= 3:
    - return(Accept)
  - If repetitions(candidate_value, global_list) >= 20:
    - return(Accept)
  - return(Reject)

Biodiversity CY 1.0: Biocollections data lake

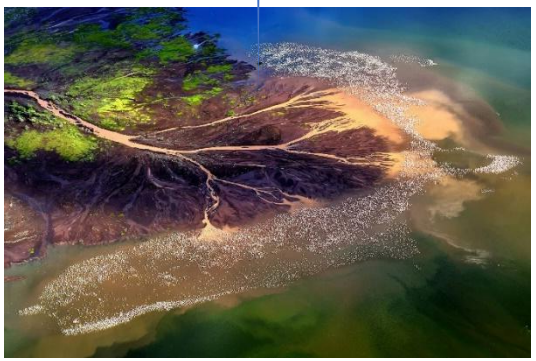


Specimens → Digitization → Records and media → Aggregated datasets → Queried datasets → Research → Scientific results

Information extraction

Running code on resources

Efficient data analytics



PHOTOGRAPH BY FRANCO CAPPELLARI,
https://www.reddit.com/r/Kenya/comments/1aubre/stunning_aerial_view_of_lake_bogoria_photo_by/
https://commons.wikimedia.org/w/index.php?curid=9880216

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

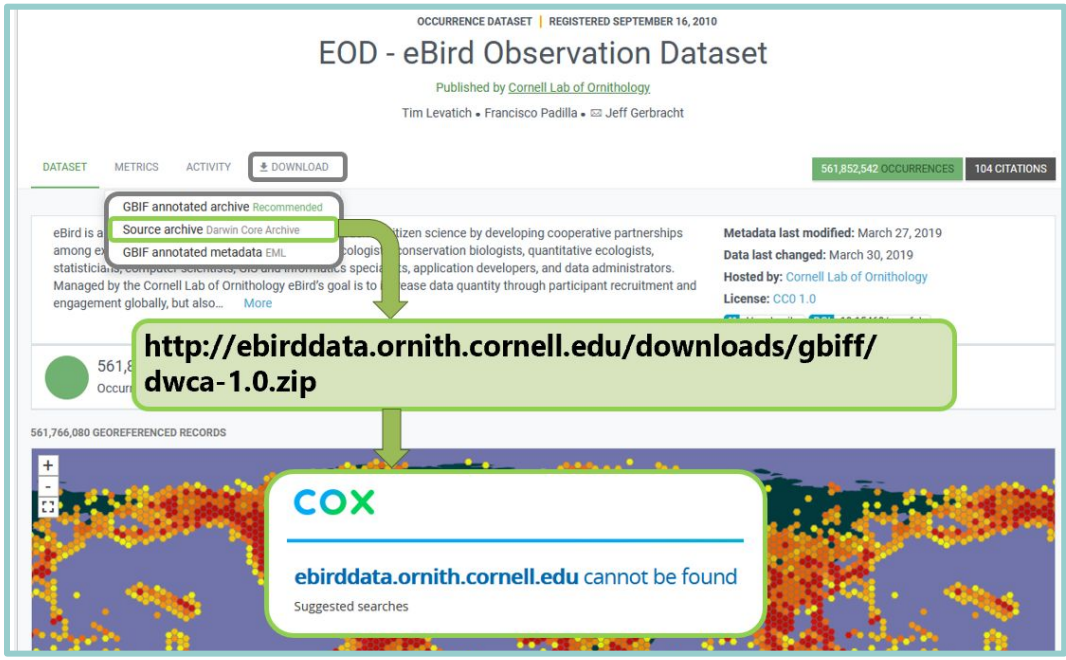
- An example reference using a URL

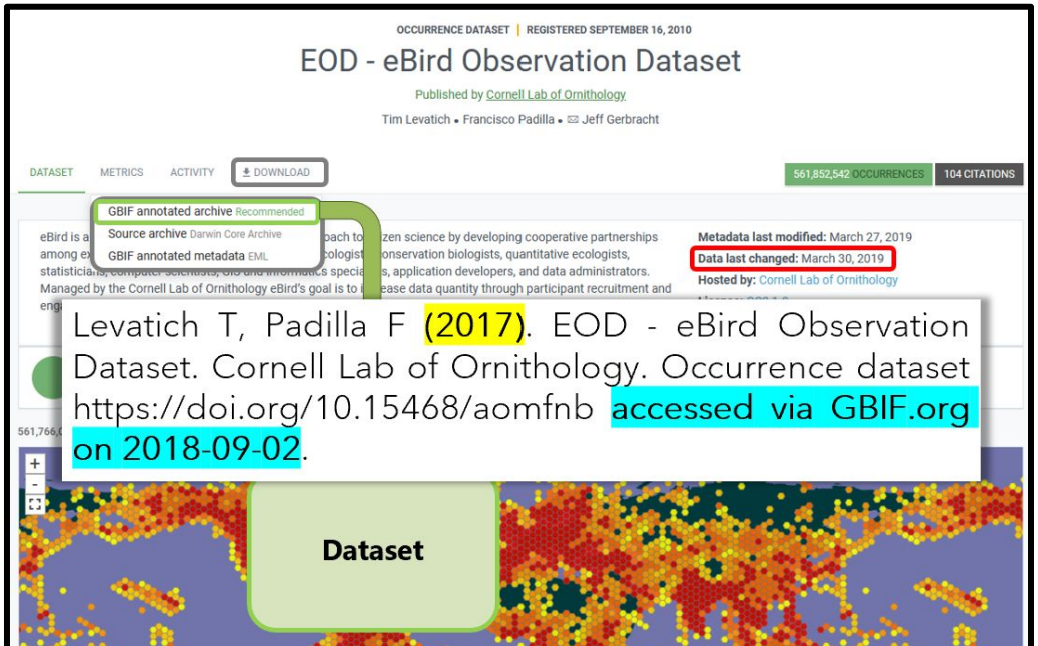
Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

Reliable references and citations

A reference is **reliable** if both

- Allows continued access to what is referenced
  - No link rot
- Only identifies what is referenced
  - No content drift/re-use
- Problem: How to reliably reference datasets served at



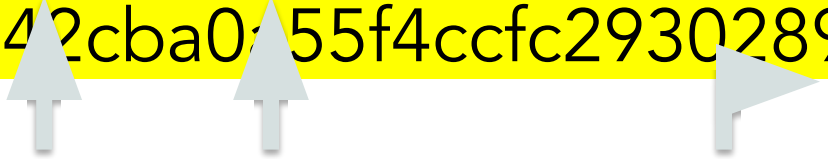


Content-based identifiers

- Cryptographic hash functions can produce unique content-based identifiers for digital datasets

Content-based identifier for the 2017 eBird dataset:

hash://sha256/29d30b566f924355a383b13cd48c3aa2
39d42cba0e55f4ccfc2930289b88b43c

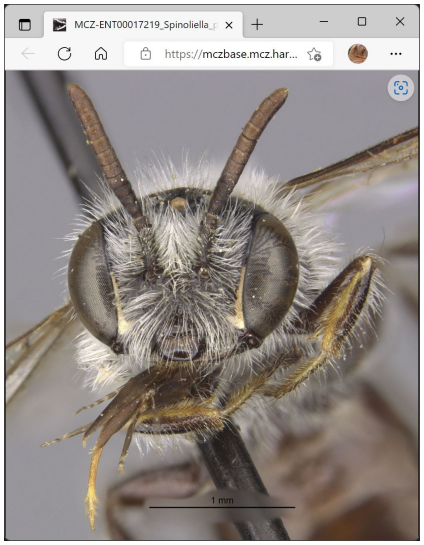


Content-based identifier for the 2019 eBird dataset:

hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40
f598a900e8be56e7645e5a24013dffc4

Signed citation

- a customary citation extended to include a content signature of the cited digital content



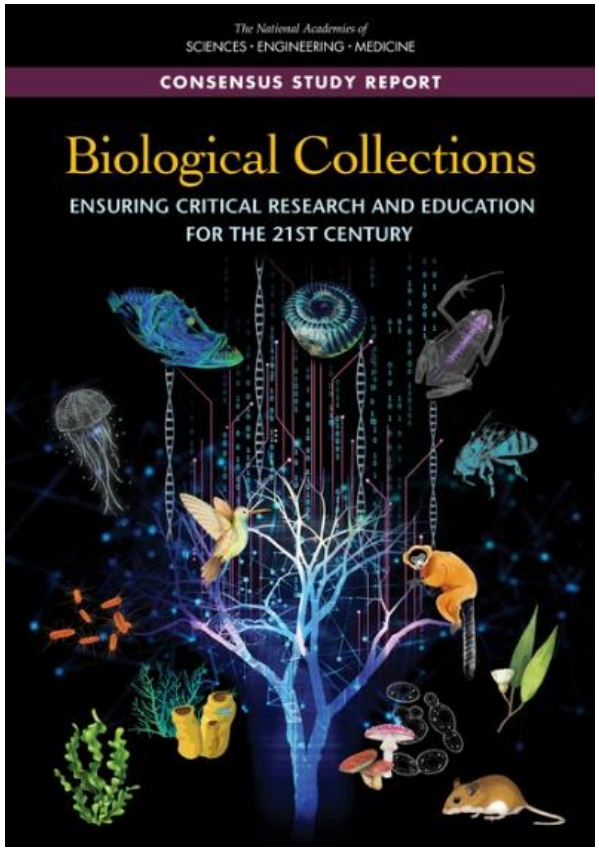
Museum of Comparative Zoology, Harvard University. 2021. Head Frontal View of MCZ:ENT:17219 Nomadopsis puellae (Cockerell, 1933) Accessed at http://mczbase.mcz.harvard.edu/specimen_images/entomology/large/MCZ-ENT00017219_Spinoliella_puellae_hef.jpg on 2021-12-07.

hash://sha256/edde5b2b45961e356f27b81a3aa51584de4761ad9fa678c4b9fa3230808ea356

Benefits of signed citations

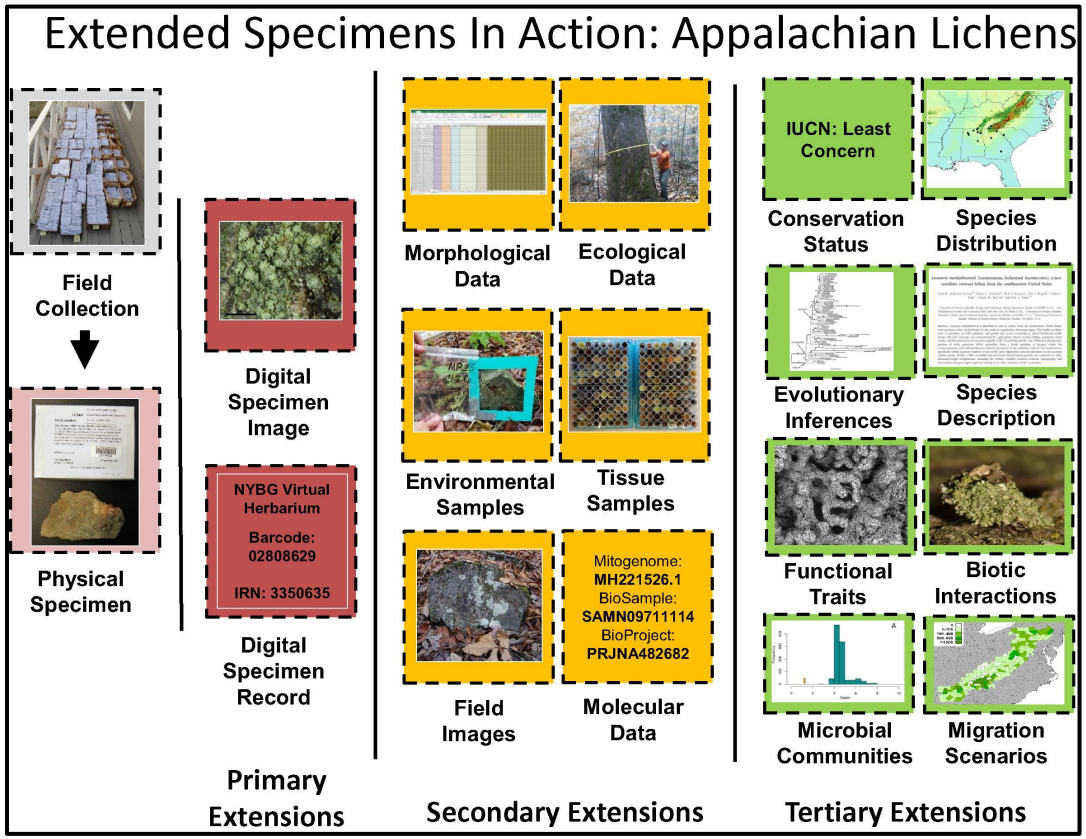
1. Verification: signatures can be regenerated
2. Unique identification: hashes are statistically guaranteed to be unique
3. Content-based: vs. location/publication based
4. Decentralized resolution: registries & repositories can be created by anyone anywhere
5. Robust resolution: rot detected, recoverable from (other) existing registries and/or copies
6. Recursive citations: citations of collections of citations *… and more*

Extended Specimen Concept



Lendemer et al. 2020, https://doi.org/10.1093/biosci/biz140.

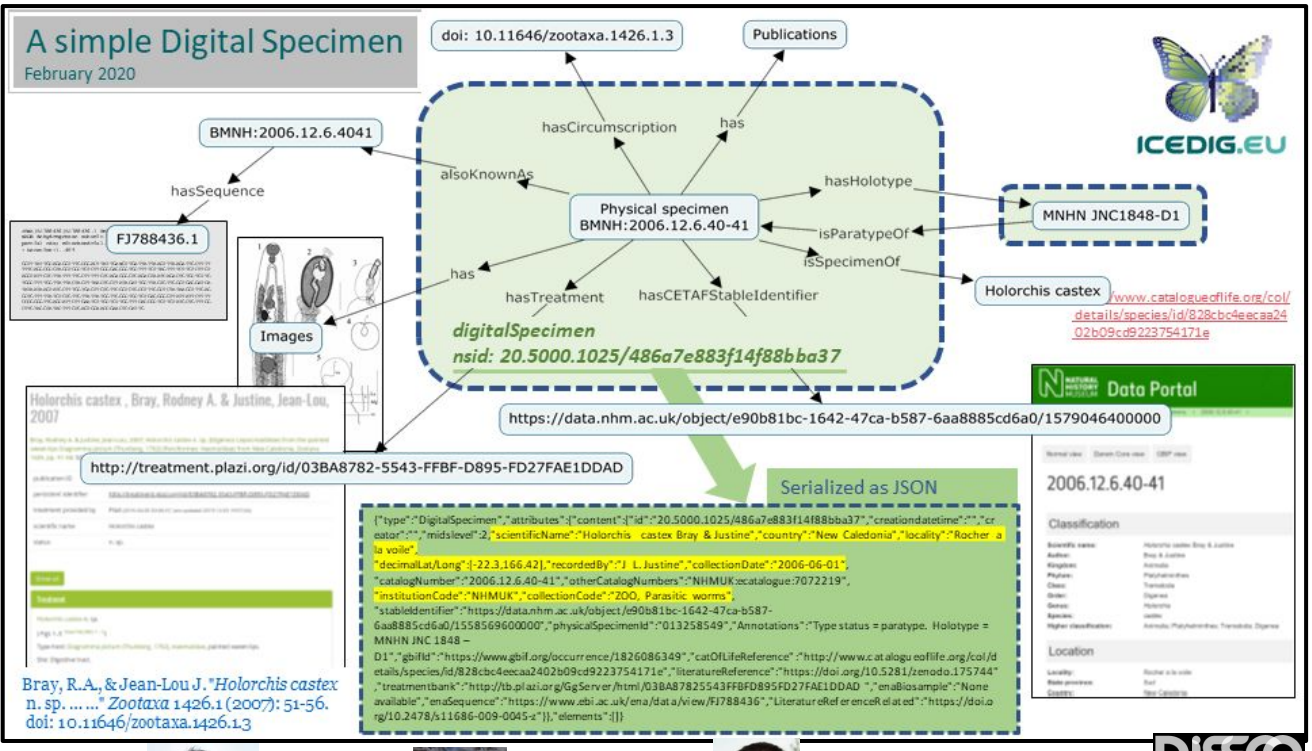
Webster et al. 2017, Chapter 1



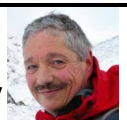
recommends building a network of extended specimens to facilitate research across taxonomic, temporal, and geospatial scales. …the "holistic" (Cook et al., 2016) or "extended specimen" concept (Webster, 2017)

Extended specimen "a constellation of specimen and data types that, in combination, capture the multidimensional phenotype (and genotype) of an individual."

Opportunity: Richer IT Abstractions



Digital specimen "a surrogate in cyberspace for a specific physical specimen, identifying its actual location and authoritatively saying something about its collection event (who, when, where) and taxonomy (what), as well as providing links to high-resolution images. A digital specimen exposes supplementary information about related literature, traits, tissue samples and DNA sequences, chemical analyses, environmental analyses, information, and much more, stored elsewhere than in the natural science collection itself."

Alex Hardisty 

Wouter Addink 

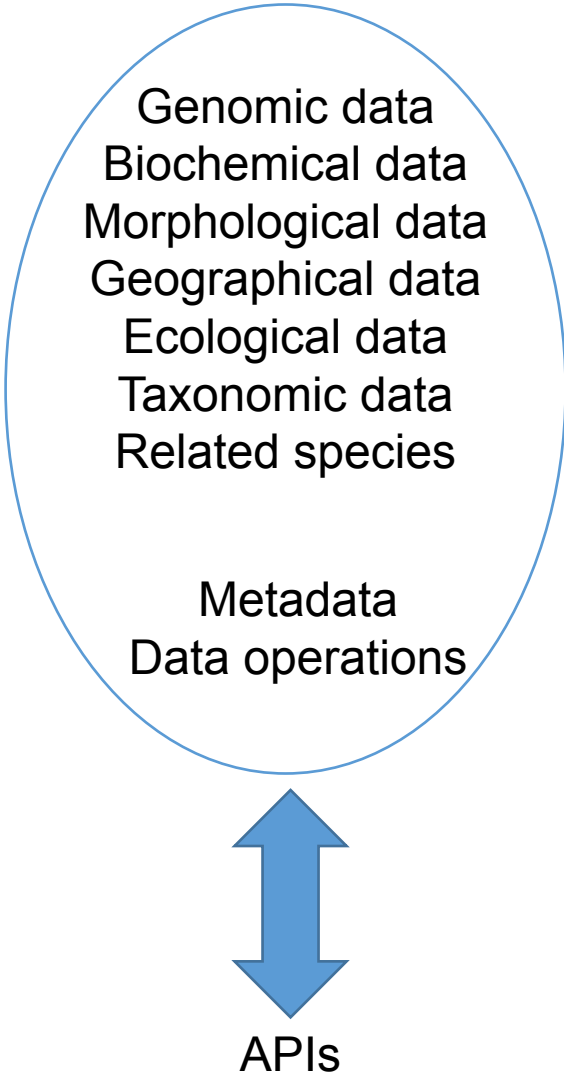
Dimitrios Koureas 

What is a Digital Specimen?

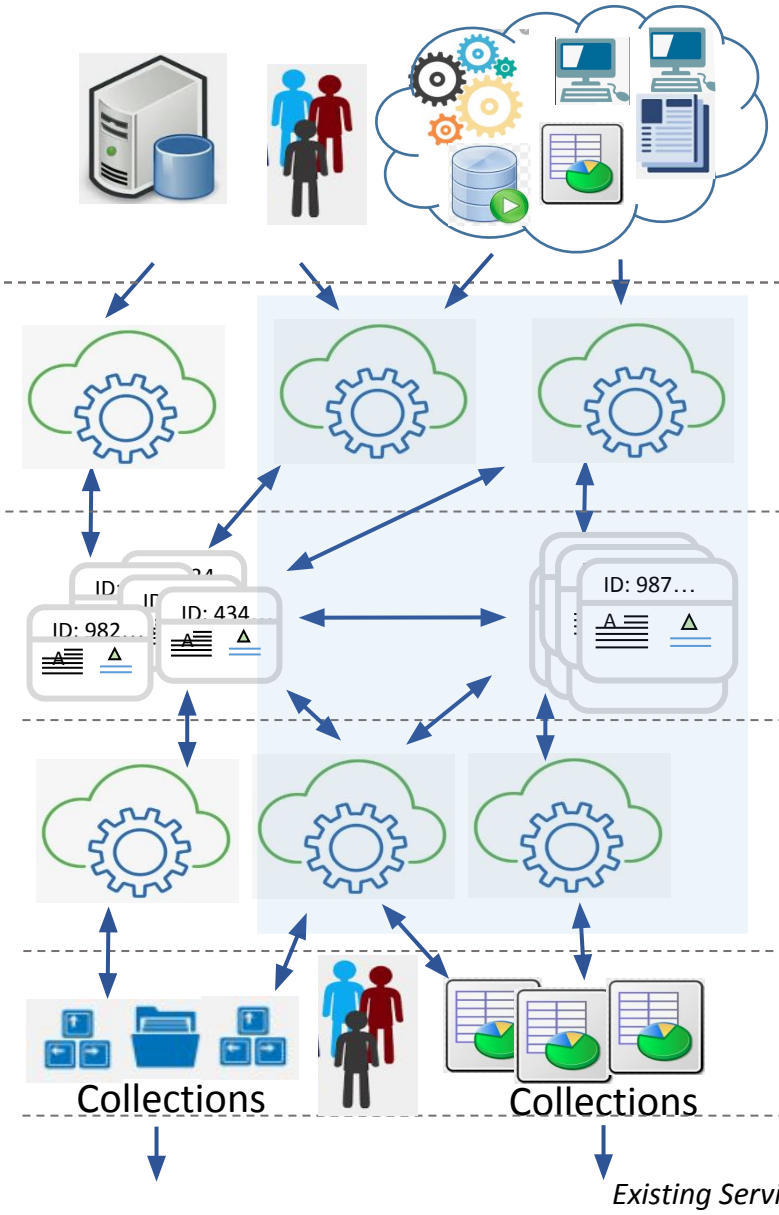
https://bit.ly/DigitalSpecimen

FAIR Digital Objects (FDOs)

- **F**indable **A**ccessible **I**nteroperable and **R**eusable
- FDO: a FAIR self-contained, typed, machine-actionable data package
  - Has a unique identifier
  - Can be discovered, accessed, moved, replicated and managed individually and programmatically



FDO Cyberinfrastructure Architecture (high-level)



FAIR objects and services

- Hosted locally/regionally by biocollections IT organizations (e.g. the unshaded mid-three layers) **OR**
- Hosted by global/community-level serving entities like today's "aggregators" (e.g the shaded area).

Acknowledgments

- National Science Foundation's Advancing Digitization of Biodiversity Collections Program [DBI-1115210 (2011-2018) and DBI-1547229 (2016-2021)
- iDigBio/ADBC Community and Collaborators
- Current ACIS IT team


Renato Figueiredo (Professor)


Chris Wilson (Software Developer)


Maureen Kelly (Developer)

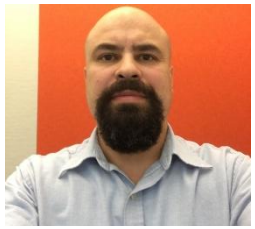

Randy Fischer (Developer)


Jesse Bennett (Developer)

Dan Stoner (Software Contractor)


Michael Elliott (PhD Student)

- SELFIE work done with


Icaro Alzuru (UF)


Andrea Matsunaga (UF)


Mauricio Tsugawa (UF)