ARK persistent identifiers come of age: 21 years, 1035 institutions

John Kunze, Dave Vieglais, Tim Clark

December 2022

ARK Alliance

arks.org

CC BY

Three ARK (Archival Resource Key) Use Cases

1. Physical Samples for Earth Sciences

Dave Vieglais
University of Kansas
vieglais@ku.edu

2. ARKs for Biomedical AI

Tim Clark
University of Virginia
twc8q@virginia.edu

3. Metadata Terms Crowdsourced

John Kunze
ARK Alliance
jakkbl@gmail.com

arks.org

Why care about ARK identifiers?

Because robust web links are rare – the average URL lifetime is 100 days

ARKs serve as persistent identifiers (PIDs) with metadata

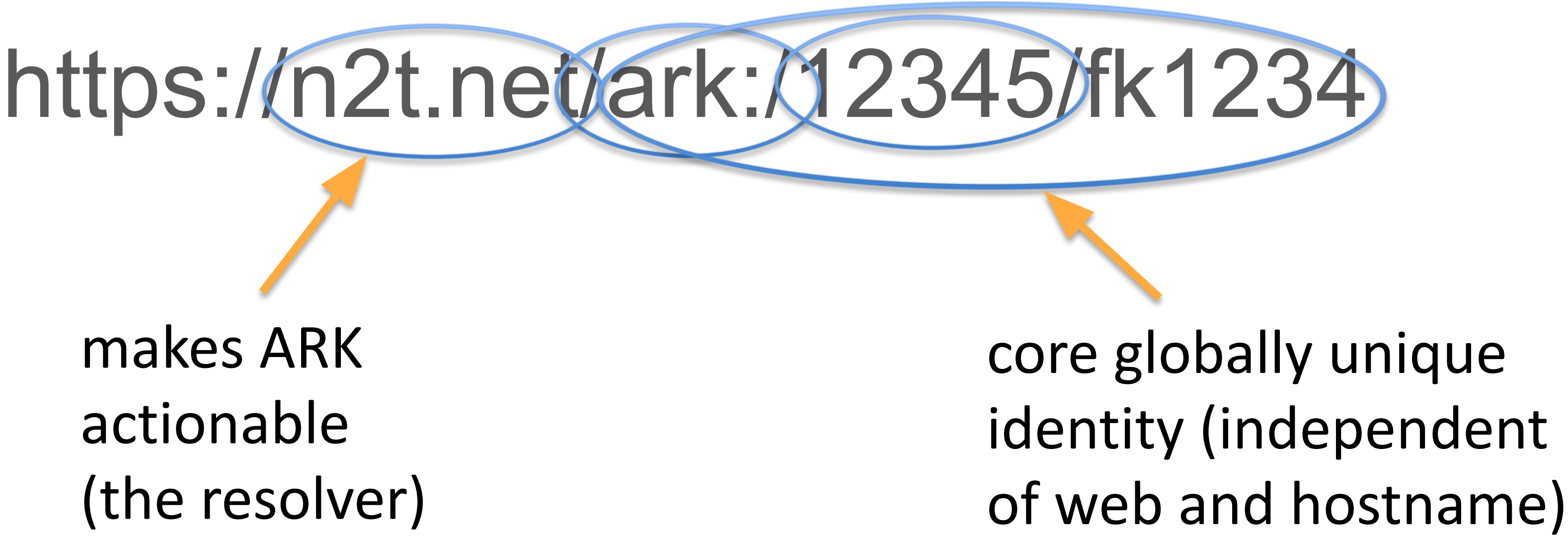
- found in the Data Citation Index, HathiTrust, Wikipedia, Wikidata, Internet Archive, ORCID profiles, etc.

In contrast to other PIDs, ARKs have

- no fees, no limits, no walled gardens (decentralized)
- very flexible metadata, including none
- can be assigned to anything digital, physical, or conceptual

ARK anatomy

A labelled URL with a globally unique identity inside it



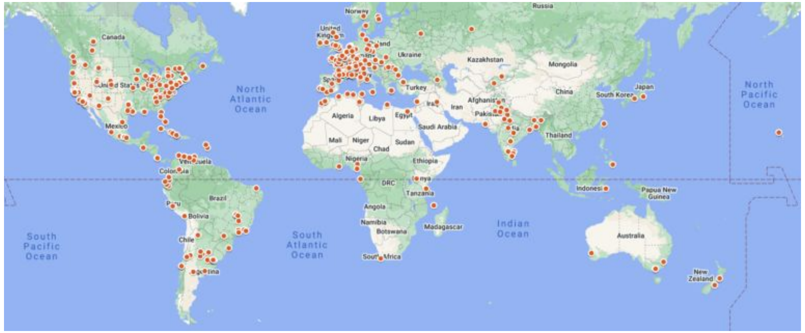
https://n2t.net/ark:/12345/fk1234

makes ARK actionable (the resolver)

core globally unique identity (independent of web and hostname)

ARK organizations

3.2 billion ARKs created by 1035 institutions – libraries, archives, museums, publishers, educators, etc. For example,



Internet Archive
Bodleian Libraries
Berkeley Law Library
Bibliothèque Mazarine
New York Public Library
French National Archives
National Library of Austria
Library and Archives Canada

University of California Berkeley
Smithsonian National Museum
National Library of France
University of Chicago
Musée du Louvre
Family Search
British Library
Google

https://n2t.net/ark:/53355/cl010066723 →

Next up: Dave Vieglais

1. Physical Samples
for Earth Sciences

Dave Vieglais
University of Kansas
vieglais@ku.edu

arks.org

iSamples: The Internet of Samples

The Internet of Samples (iSamples) is a standards-based collaboration to uniquely, consistently, and conveniently **identify material samples**, **record core metadata** about them, and **link** them to **other samples, data, and research products**.

iSamples

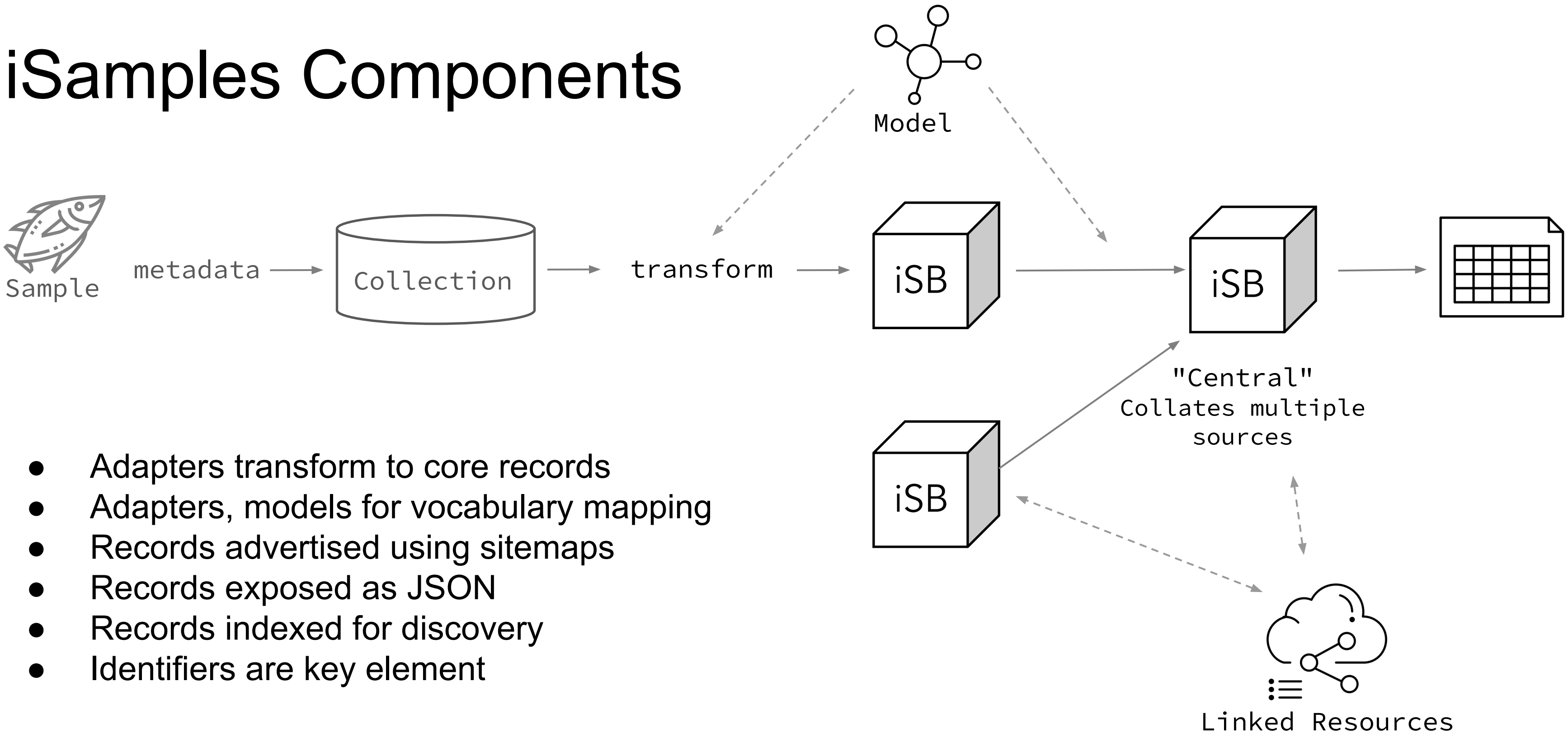
Funded by the US National Science Foundation (CSSI) projects 2004815, 2004839, 2004562, and 2004642.

arks.org

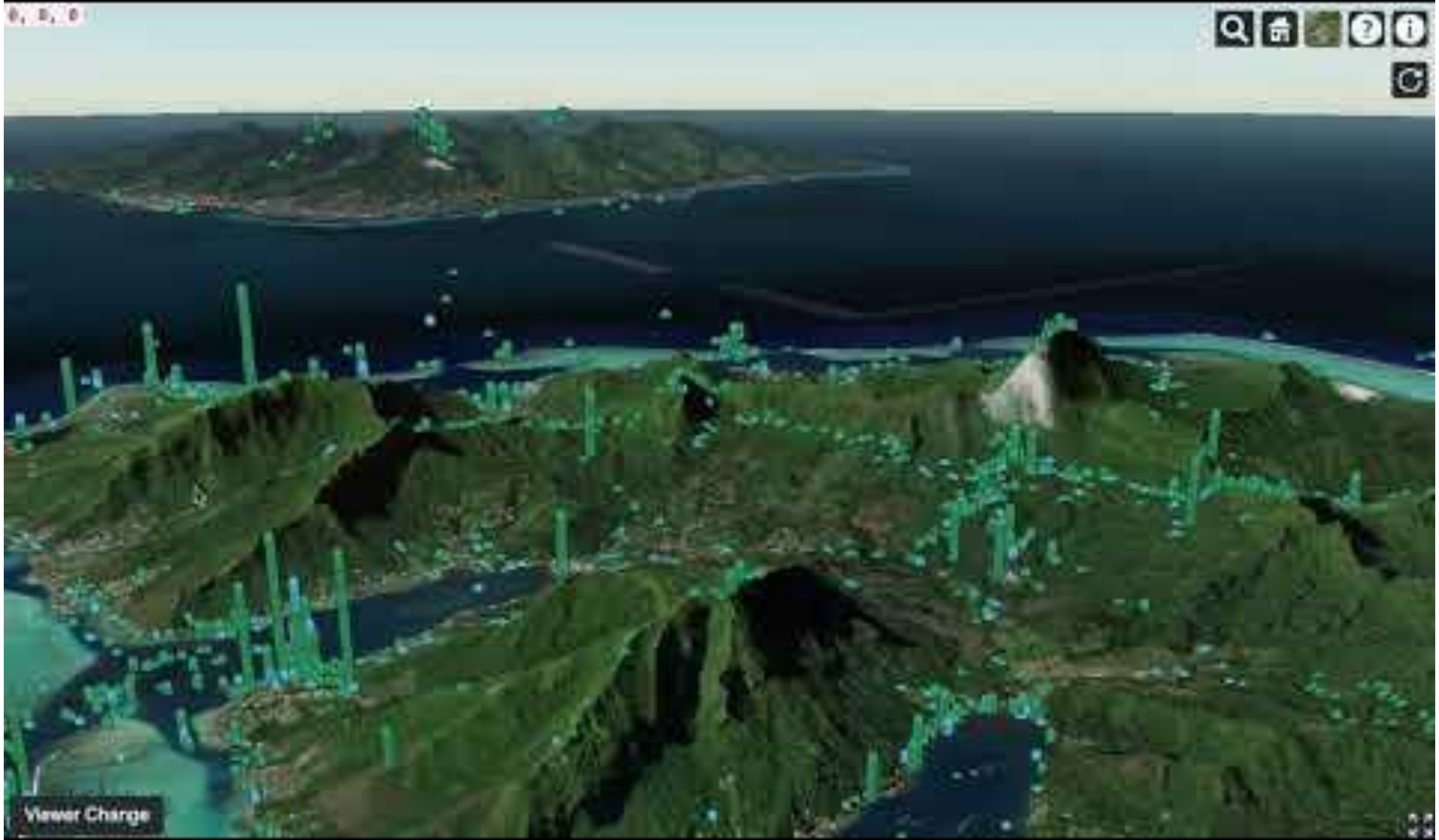
Factual Basis

- Physical samples typically gathered for analysis
- Analysis may be immediate or many years later
- Samples may be consumed in analysis or preserved in perpetuity
- Derived products such as analyses and publication form graphs of knowledge
- Often beneficial to verify analysis, traversing from publication to original samples
- Linking by identifiers is also essential for attribution, ideally transitive from publications to original sample collector

iSamples Components



- Adapters transform to core records
- Adapters, models for vocabulary mapping
- Records advertised using sitemaps
- Records exposed as JSON
- Records indexed for discovery
- Identifiers are key element



iSamples

arks.org

iSamples Summary

- Globally unique, resolvable identifiers are a critical core attribute of physical samples.
- These identifiers facilitate the rapid development of power linked data environments.
- ARK identifiers fill this role with a combination of technical characteristics, reliability, and accessibility.

iSamples

arks.org

Next up: Tim Clark

2. ARKs for
Biomedical AI

Tim Clark
University of Virginia
twc8q@virginia.edu

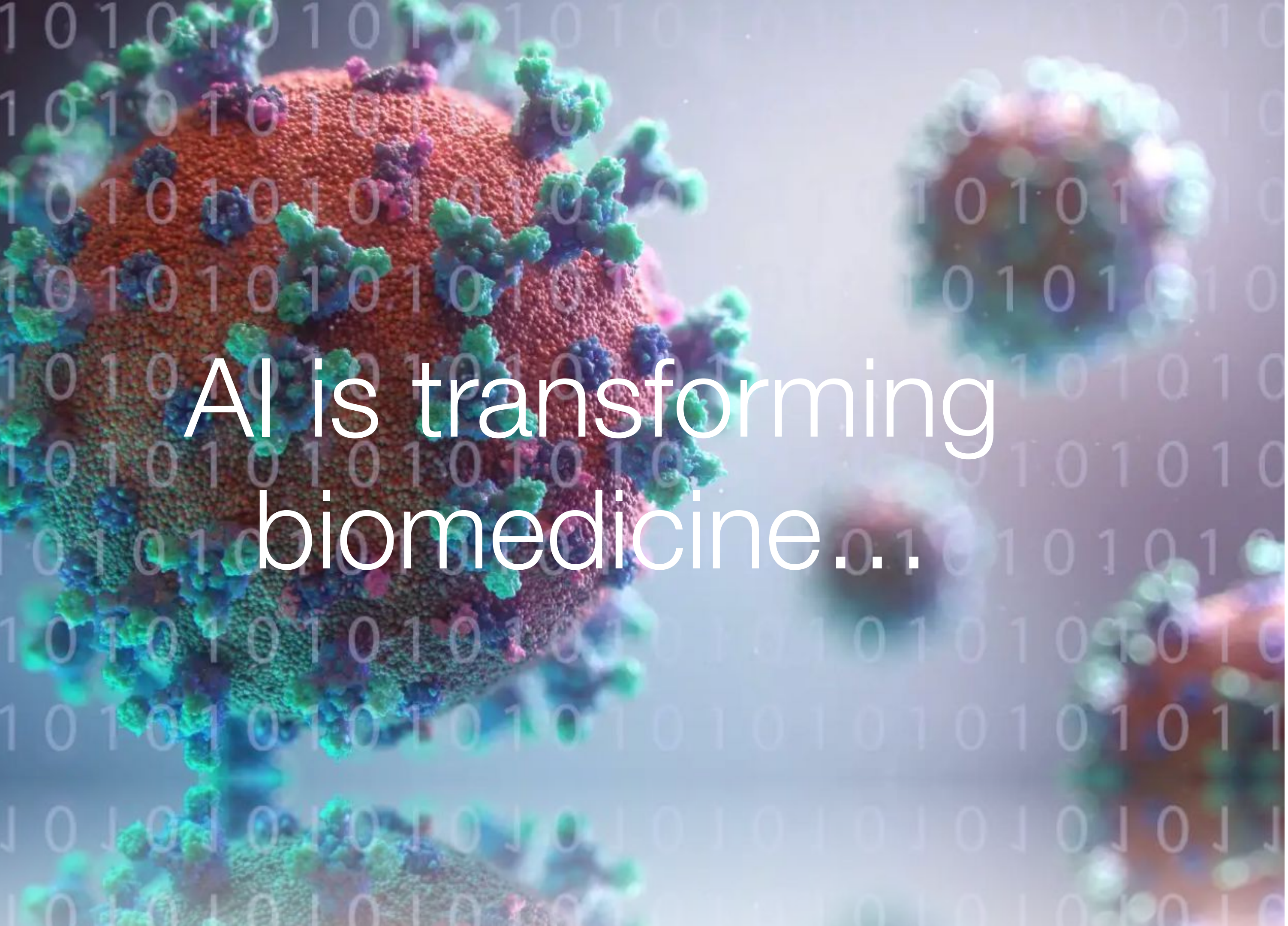
Using ARKs in AI for biomedical research

Timothy Clark, Ph.D.
School of Medicine & School of Data Science
University of Virginia

December 2022

UNIVERSITY of VIRGINIA | SCHOOL of MEDICINE | SCHOOL of DATA SCIENCE

arks.org

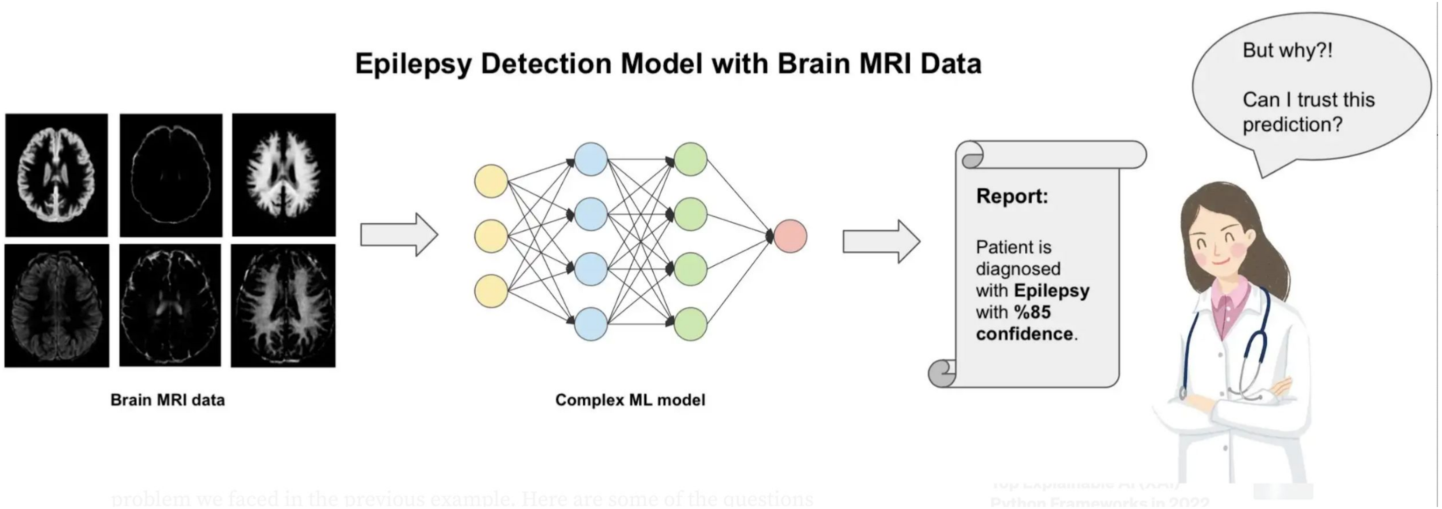


AI is transforming biomedicine...

...by making very complexpredictions...

…which require explanation and robust evidence!

arks.org



Epilepsy Detection Model with Brain MRI Data

Brain MRI data

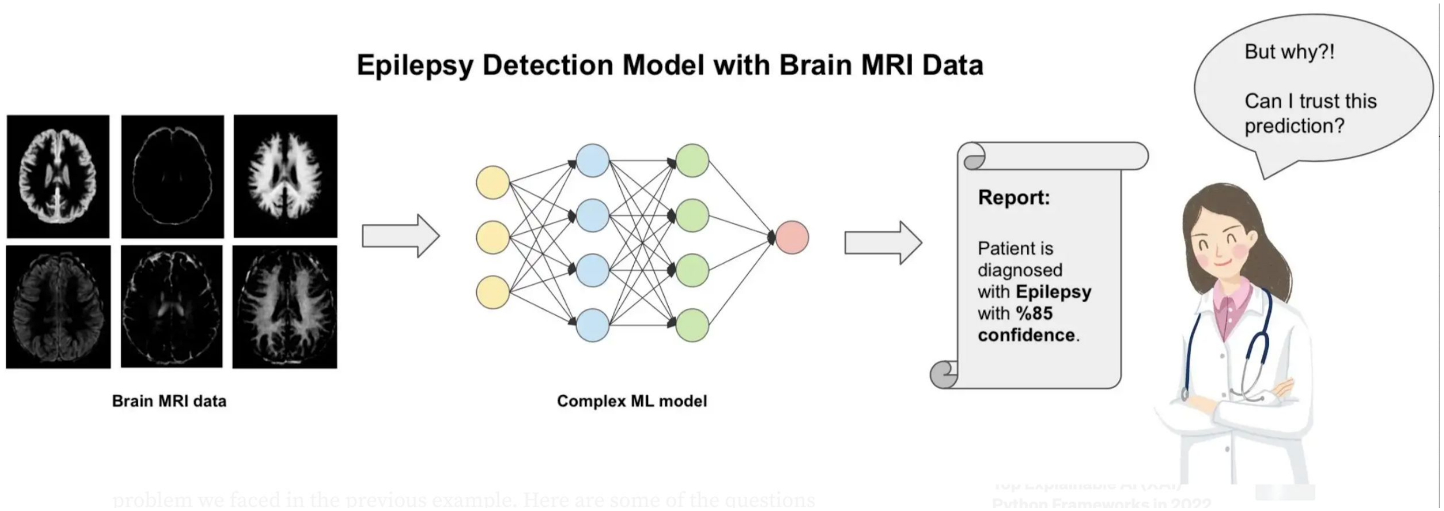
Complex ML model

Report:

Patient is diagnosed with **Epilepsy** with **%85 confidence**.

problem we faced in the previous example. Here are some of the questions

Top Explainable AI (XAI) Python Frameworks in 2022


Epilepsy Detection Model with Brain MRI Data
Brain MRI data
Complex ML model
Report:
Patient is diagnosed with Epilepsy with %85 confidence.
But why?!
Can I trust this prediction?

problem we faced in the previous example. Here are some of the questions

Top Explainable AI (XAI) Python Frameworks in 2022

Why did you make that prediction?

Can I trust the AI models?

Two Biomedical Use Cases for Deep AI/ML

1. Predictive model of cellular response to drugs based on very high-dimensional research data and deep learning.

2. Predictive model notifying physicians 7 days ahead of potential life-threatening events on infants in NICU.

Results Validation and Explainability

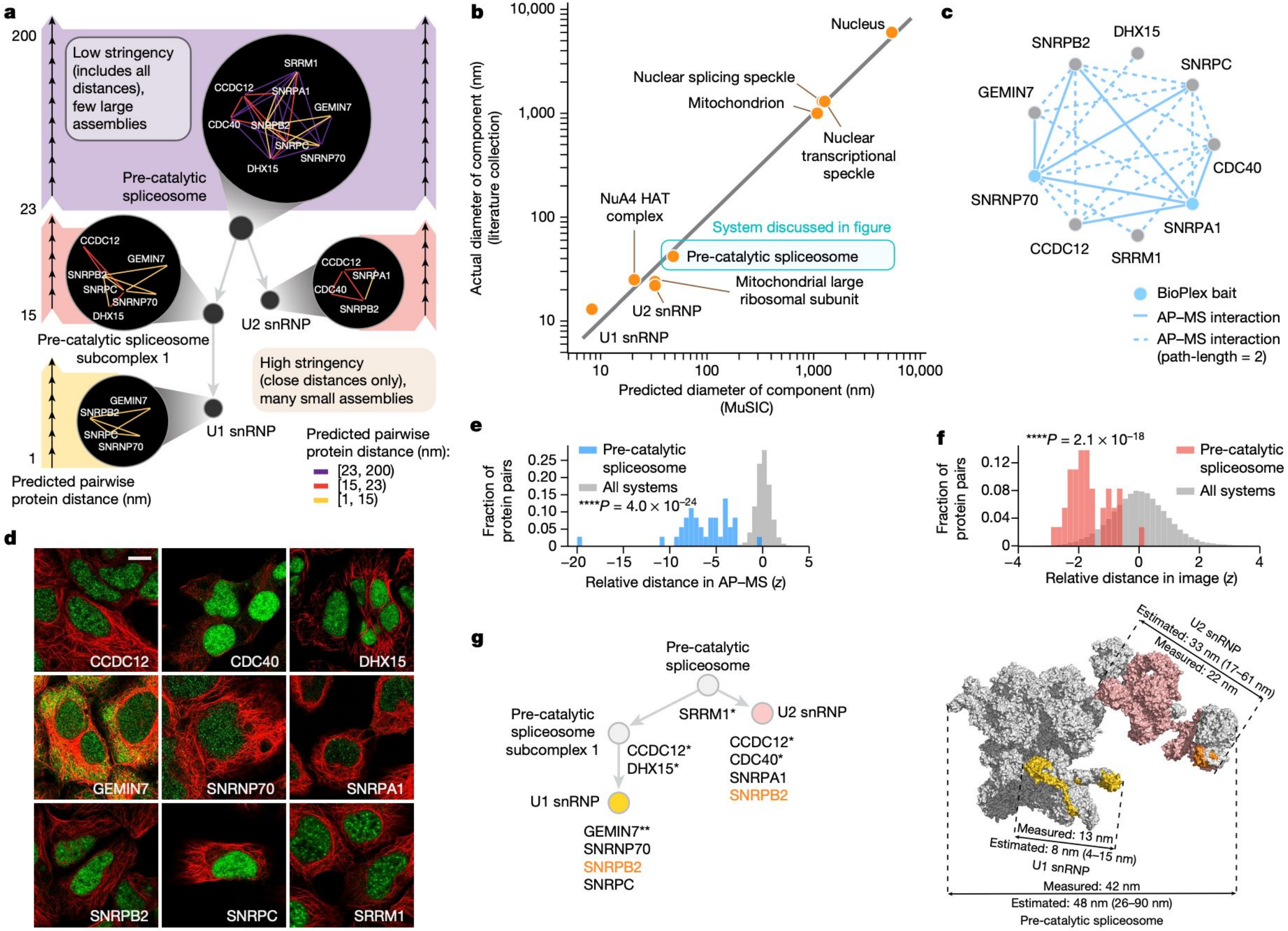
- Evidence graphs describe how results were obtained & provide supporting evidence for correctness of results.

- We give graph node an ARK resolvable to supporting data, software, and computational parameters- explainability!

- ARKs have huge advantage where thousands of chained computations are performed, and provide open metadata.

1. Predictive response modeling of normal & diseased human cells

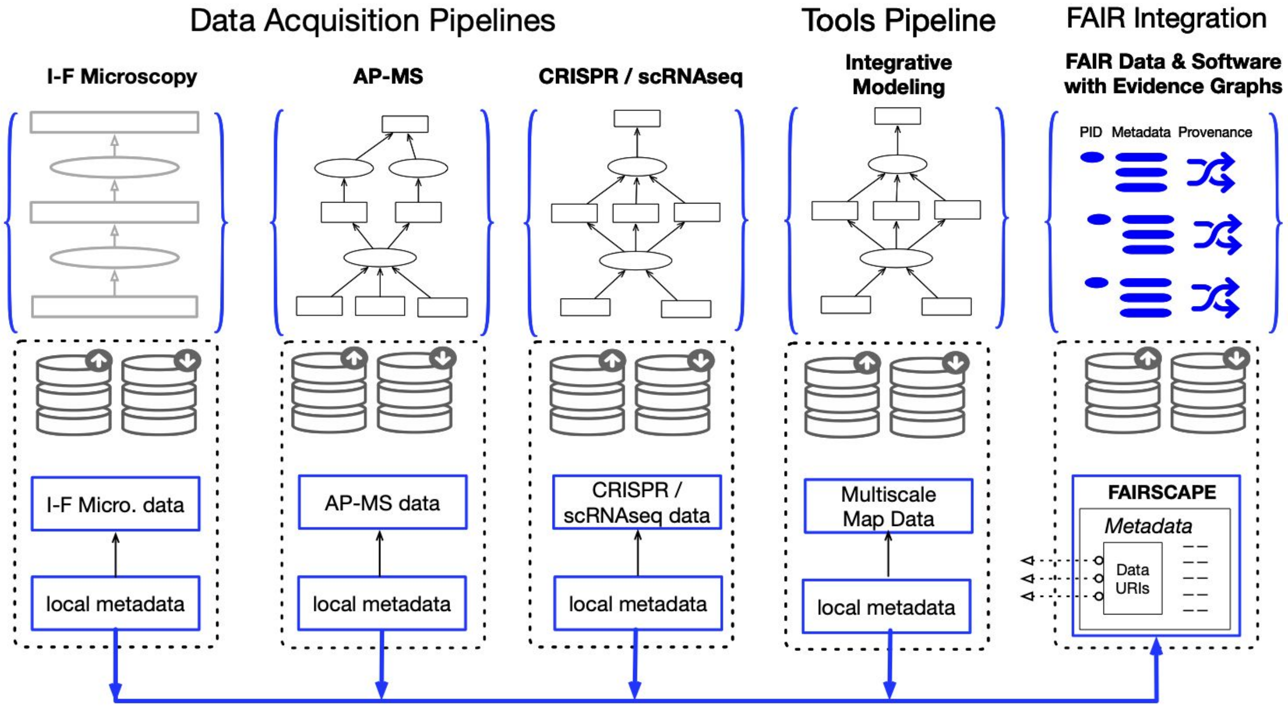
- Construct an accurate cellular component architecture based on very high-dimensional research data and AI.
- Accurately predict cellular response to biochemical perturbation (e.g. new drugs, etc.) using deep learning.
- Be able to interpret and explain model results robustly.



adapted from Qin et al. 2021 - A multi-scale map of cell structure fusing protein images and interactions. ***Nature*** 600:536-542. 16 December 2021. https://doi.org/10.1038/s41586-021-04115-9

arks.org

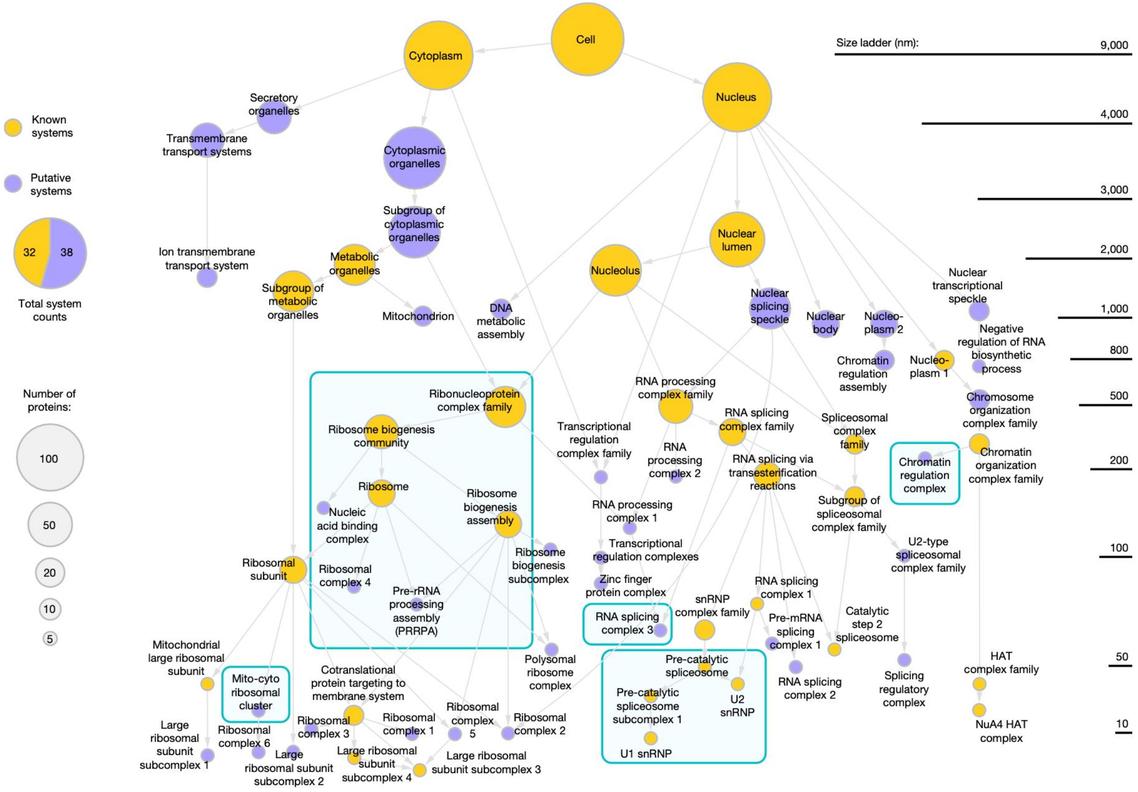
CM4AI FAIR Integration Model



FAIRSCAPE integration with CM4AI Tools and Data pipelines (adapted from CM4AI Data Dictionary Requirements Version 1.0). FAIRSCAPE holds comprehensive FAIR metadata and deep provenance graphs on all CM4AI data objects. Data is initially held locally, and accessed through contentURIs in the CM4AI metadata.

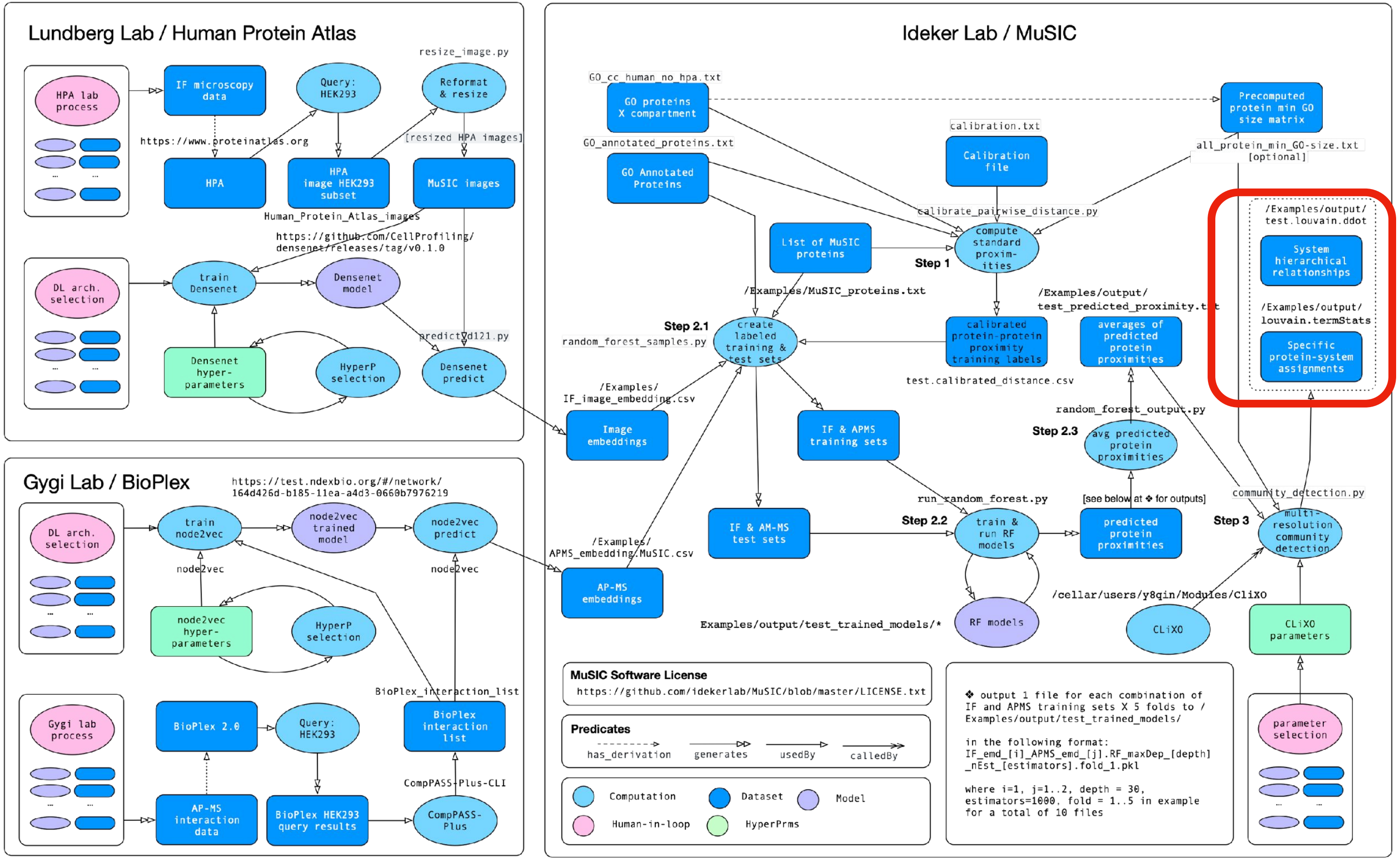
arks.org

Preliminary Map of Cell Architecture



Cell
Cytoplasm
Nucleus
Size ladder (nm):
9,000
4,000
3,000
2,000
1,000
800
500
200
100
50
10
Known systems
Putative systems
32
38
Total system counts
Number of proteins:
100
50
20
10
5
Secretory organelles
Transmembrane transport systems
Cytoplasmic organelles
Subgroup of cytoplasmic organelles
Ion transmembrane transport system
Metabolic organelles
Subgroup of metabolic organelles
Mitochondrion
DNA metabolic assembly
Nucleolus
Nuclear lumen
Nuclear splicing speckle
Nuclear body
Nucleo-plasm 2
Nuclear transcriptional speckle
Negative regulation of RNA biosynthetic process
Chromatin regulation assembly
Nucleo-plasm 1
Chromosome organization complex family
Chromatin regulation complex
Chromatin organization complex family
Ribonucleoprotein complex family
Ribosome biogenesis community
Ribosome
Nucleic acid binding complex
Ribosome biogenesis assembly
Ribosomal subunit
Ribosomal complex 4
Pre-rRNA processing assembly (PRRPA)
Ribosome biogenesis subcomplex
Transcriptional regulation complex family
RNA processing complex family
RNA processing complex 2
RNA splicing complex family
RNA splicing via transesterification reactions
Spliceosomal complex family
Subgroup of spliceosomal complex family
RNA processing complex 1
Transcriptional regulation complexes
Zinc finger protein complex
RNA splicing complex 3
snRNP complex family
RNA splicing complex 1
Pre-mRNA splicing complex 1
Catalytic step 2 spliceosome
U2-type spliceosomal complex family
Polysomal ribosome complex
Pre-catalytic spliceosome
Pre-catalytic spliceosome subcomplex 1
U2 snRNP
U1 snRNP
RNA splicing complex 2
Splicing regulatory complex
HAT complex family
NuA4 HAT complex
Mitochondrial large ribosomal subunit
Mito-cyto ribosomal cluster
Large ribosomal subunit subcomplex 1
Ribosomal complex 6
Large ribosomal subunit subcomplex 2
Cotranslational protein targeting to membrane system
Ribosomal complex 3
Large ribosomal subunit subcomplex 4
Ribosomal complex 1
Ribosomal complex 5
Large ribosomal subunit subcomplex 3
Ribosomal complex 2

adapted from Qin et al. 2021 - A multi-scale map of cell structure fusing protein images and interactions. ***Nature*** 600:536-542. 16 December 2021. https://doi.org/10.1038/s41586-021-04115-9



Lundberg Lab / Human Protein Atlas

HPA lab process

IF microscopy data

Query: HEK293

Reformat & resize

resize_image.py

https://www.proteinatlas.org

[resized HPA images]

HPA

HPA image HEK293 subset

MuSIC images

Human_Protein_Atlas_images

https://github.com/CellProfiling/densenet/releases/tag/v0.1.0

DL arch. selection

train Densenet

Densenet model

predict_d121.py

Densenet hyper-parameters

HyperP selection

Densenet predict

Gygi Lab / BioPlex

https://test.ndexbio.org/#/network/164d426d-b185-11ea-a4d3-0660b7976219

DL arch. selection

train node2vec

node2vec trained model

node2vec predict

node2vec

node2vec

node2vec hyper-parameters

HyperP selection

BioPlex_interaction_list

Gygi lab process

BioPlex 2.0

Query: HEK293

BioPlex interaction list

CompPASS-Plus-CLI

AP-MS interaction data

BioPlex HEK293 query results

CompPASS-Plus

Ideker Lab / MuSIC

GO_cc_human_no_hpa.txt

GO proteins X compartment

GO_annotated_proteins.txt

GO Annotated Proteins

calibration.txt

Calibration file

Precomputed protein min GO size matrix

all_protein_min_GO-size.txt [optional]

calibrate_pairwise_distance.py

List of MuSIC proteins

Step 1

compute standard proxim-ities

/Examples/MuSIC_proteins.txt

/Examples/output/test_predicted_proximity.txt

Step 2.1

random_forest_samples.py

create labeled training & test sets

calibrated protein-protein proximity training labels

averages of predicted protein proximities

test.calibrated_distance.csv

/Examples/IF_image_embedding.csv

Image embeddings

IF & APMS training sets

random_forest_output.py

Step 2.3

avg predicted protein proximities

/Examples/output/test.louvain.ddot

System hierarchical relationships

/Examples/output/louvain.termStats

Specific protein-system assignments

/Examples/APMS_embedding.MuSIC.csv

AP-MS embeddings

IF & AM-MS test sets

run_random_forest.py

Step 2.2

train & run RF models

[see below at ❖ for outputs]

predicted protein proximities

community_detection.py

Step 3

multi-resolution community detection

/cellar/users/y8qin/Modules/CliXO

Examples/output/test_trained_models/*

RF models

CLiXO

CLiXO parameters

MuSIC Software License

https://github.com/idekerlab/MuSIC/blob/master/LICENSE.txt

Predicates

has_derivation generates usedBy calledBy

● Computation ● Dataset ● Model

● Human-in-loop ● HyperPrms

❖ output 1 file for each combination of IF and APMS training sets X 5 folds to /Examples/output/test_trained_models/

in the following format:
IF_emd_[i]_APMS_emd_[j].RF_maxDep_[depth]_nEst_[estimators].fold_1.pkl

where i=1, j=1..2, depth = 30, estimators=1000, fold = 1..5 in example for a total of 10 files

parameter selection

Bridge2AI / CM4AI Standards Module - MuSIC 1.0 Provenance Model - Version 1.0 - 2022.12.31 - T. Clark - twc8q@virginia.edu

adapted from Clark & Al Manor et al. 2023. CM4AI MuSIC 1.0 Provenance Map. NIH Bridge2AI program, CM4AI Deliverable 1.3.

arks.org

FAIRSCAPE: initial use case

- NICU time series adverse event prediction on ~ 6,000 babies, > 100 algorithms, ~ 3,000 candidate features, 100 TB data
- Algorithmic result clustering > 17,000 computations per result

[7] J. Niestroy et al., “Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis,” (2022) *npj Digital Medicine*. preprint: https://doi.org/10.1101/2021.03.26.437138.

2. Predictive Analytics in the Neonatal ICU

- Predictive analytics: neonatal ICU @ UVA Med Center
- 6,000 NICU babies, 10 years of vital signs data
- 80 time series algorithms X dozens of parameter sets
- Goal: Predict adverse medical events 7 days in advance

Single Patient, 7-node Graphlet for HCTSA Computations

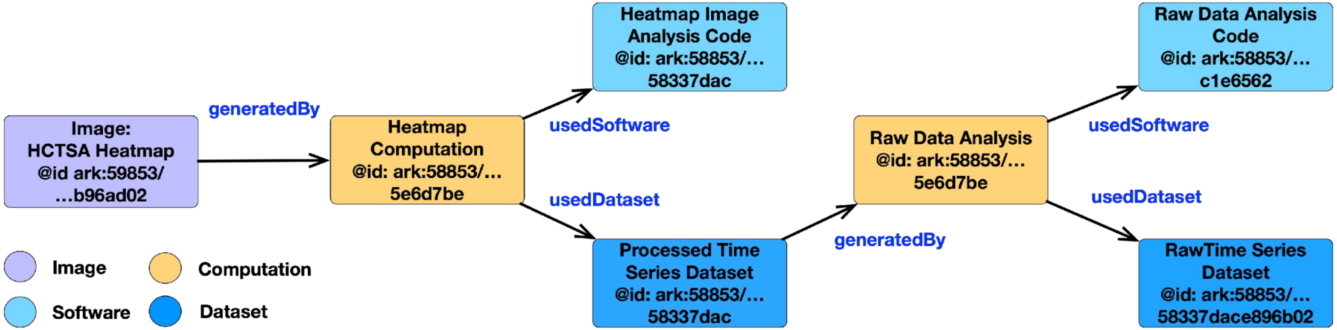


Image:
HCTSA Heatmap
@id ark:59853/
…b96ad02

Heatmap
Computation
@id: ark:58853/…
5e6d7be

Heatmap Image
Analysis Code
@id: ark:58853/…
58337dac

Processed Time
Series Dataset
@id: ark:58853/…
58337dac

Raw Data Analysis
@id: ark:58853/…
5e6d7be

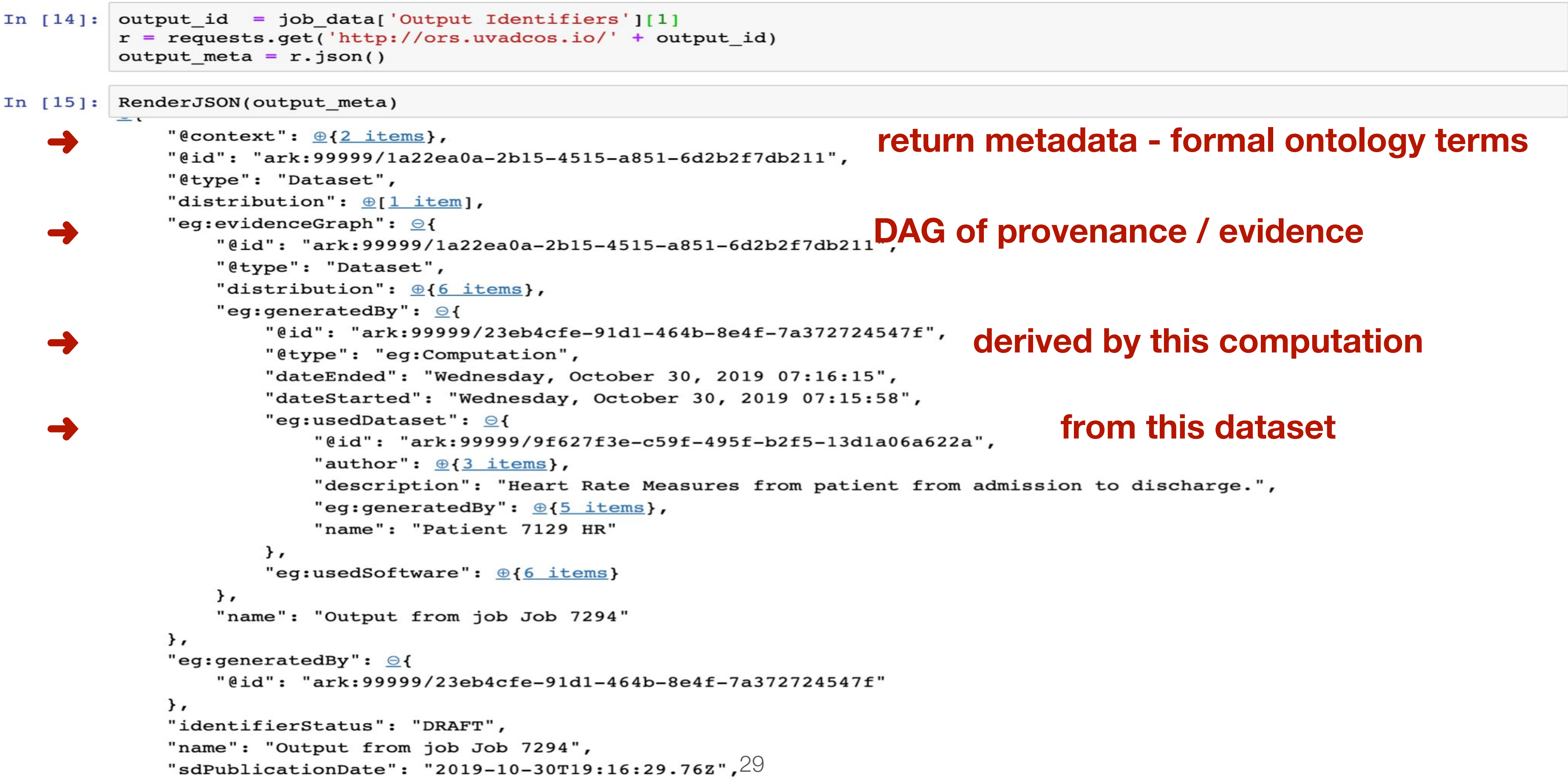
Raw Data Analysis
Code
@id: ark:58853/…
c1e6562

RawTime Series
Dataset
@id: ark:58853/…
58337dace896b02

(complete graph for results > 17,000 nodes)

arks.org

JSON-LD, schema.org & EVI Evidence Graphs



```
In [14]: output_id   = job_data['Output Identifiers'][1]
         r = requests.get('http://ors.uvadcos.io/' + output_id)
         output_meta = r.json()
```

```
In [15]: RenderJSON(output_meta)
```

```
⊖{
    "@context": ⊕{2 items},
    "@id": "ark:99999/1a22ea0a-2b15-4515-a851-6d2b2f7db211",
    "@type": "Dataset",
    "distribution": ⊕[1 item],
    "eg:evidenceGraph": ⊖{
        "@id": "ark:99999/1a22ea0a-2b15-4515-a851-6d2b2f7db211",
        "@type": "Dataset",
        "distribution": ⊕{6 items},
        "eg:generatedBy": ⊖{
            "@id": "ark:99999/23eb4cfe-91d1-464b-8e4f-7a372724547f",
            "@type": "eg:Computation",
            "dateEnded": "Wednesday, October 30, 2019 07:16:15",
            "dateStarted": "Wednesday, October 30, 2019 07:15:58",
            "eg:usedDataset": ⊖{
                "@id": "ark:99999/9f627f3e-c59f-495f-b2f5-13d1a06a622a",
                "author": ⊕{3 items},
                "description": "Heart Rate Measures from patient from admission to discharge.",
                "eg:generatedBy": ⊕{5 items},
                "name": "Patient 7129 HR"
            },
            "eg:usedSoftware": ⊕{6 items}
        },
        "name": "Output from job Job 7294"
    },
    "eg:generatedBy": ⊖{
        "@id": "ark:99999/23eb4cfe-91d1-464b-8e4f-7a372724547f"
    },
    "identifierStatus": "DRAFT",
    "name": "Output from job Job 7294",
    "sdPublicationDate": "2019-10-30T19:16:29.76Z",
```

➜ **return metadata - formal ontology terms**

➜ **DAG of provenance / evidence**

➜ **derived by this computation**

➜ **from this dataset**

Why do we use ARKs?

- Persistent IDs for data (the evidence graph nodes)
- Free to mint ARKs and their metadata is flexible
- Large ecosystem of users & developers

We used ARKs for

- Complex evidence graph on AI/ML predictions
- Every node (dataset, computation, software) resolvable
- Each node persistently identified with an ARK

Conclusion

- ARKs are a useful, flexible, scalable persistent ID model.
- Especially useful for traceable complex computations.
- We are happy to chat with prospective ARK users.

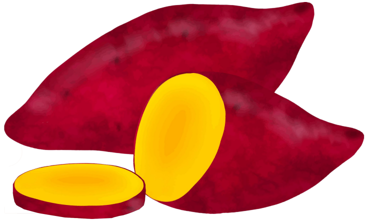
Next up: John Kunze

3. Metadata Terms
Crowdsourced

John Kunze
ARK Alliance
jakkbl@gmail.com

arks.org

Yamz.net and ARKs for metadata terms



Vocabulary creation, sharing, and standards – better, faster, cheaper

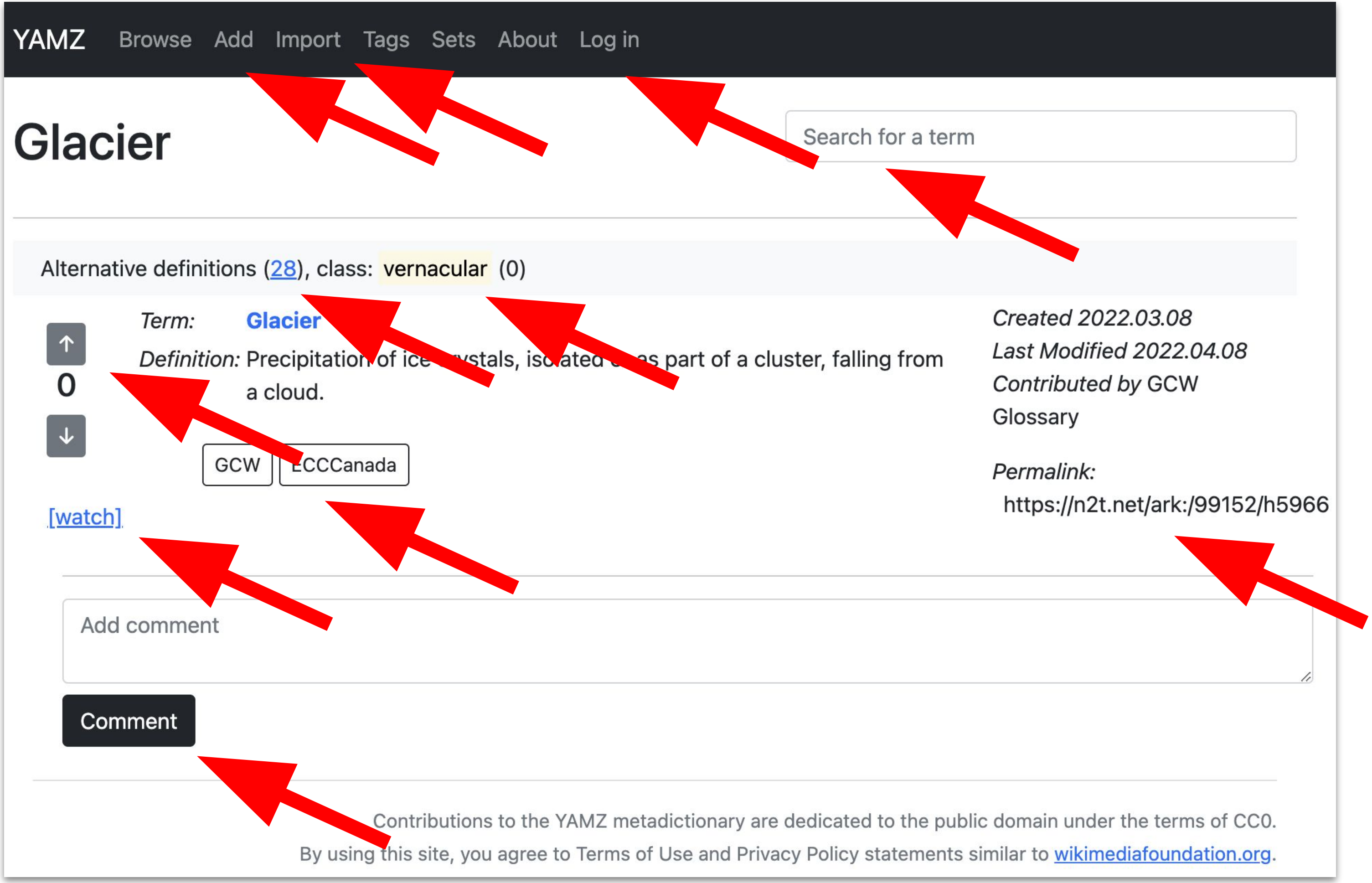
John Kunze

<MRC> Drexel UNIVERSITY

arks.org

Yamz.net

(Yet another metadata zoo)



Glacier

Alternative definitions (28), class: vernacular (0)

Term: **Glacier**

Definition: Precipitation of ice crystals, isolated or as part of a cluster, falling from a cloud.

Created 2022.03.08
Last Modified 2022.04.08
Contributed by GCW Glossary

Permalink:
https://n2t.net/ark:/99152/h5966

Contributions to the YAMZ metadictionary are dedicated to the public domain under the terms of CC0.
By using this site, you agree to Terms of Use and Privacy Policy statements similar to wikimediafoundation.org.

arks.org

Standardized metadata is rare

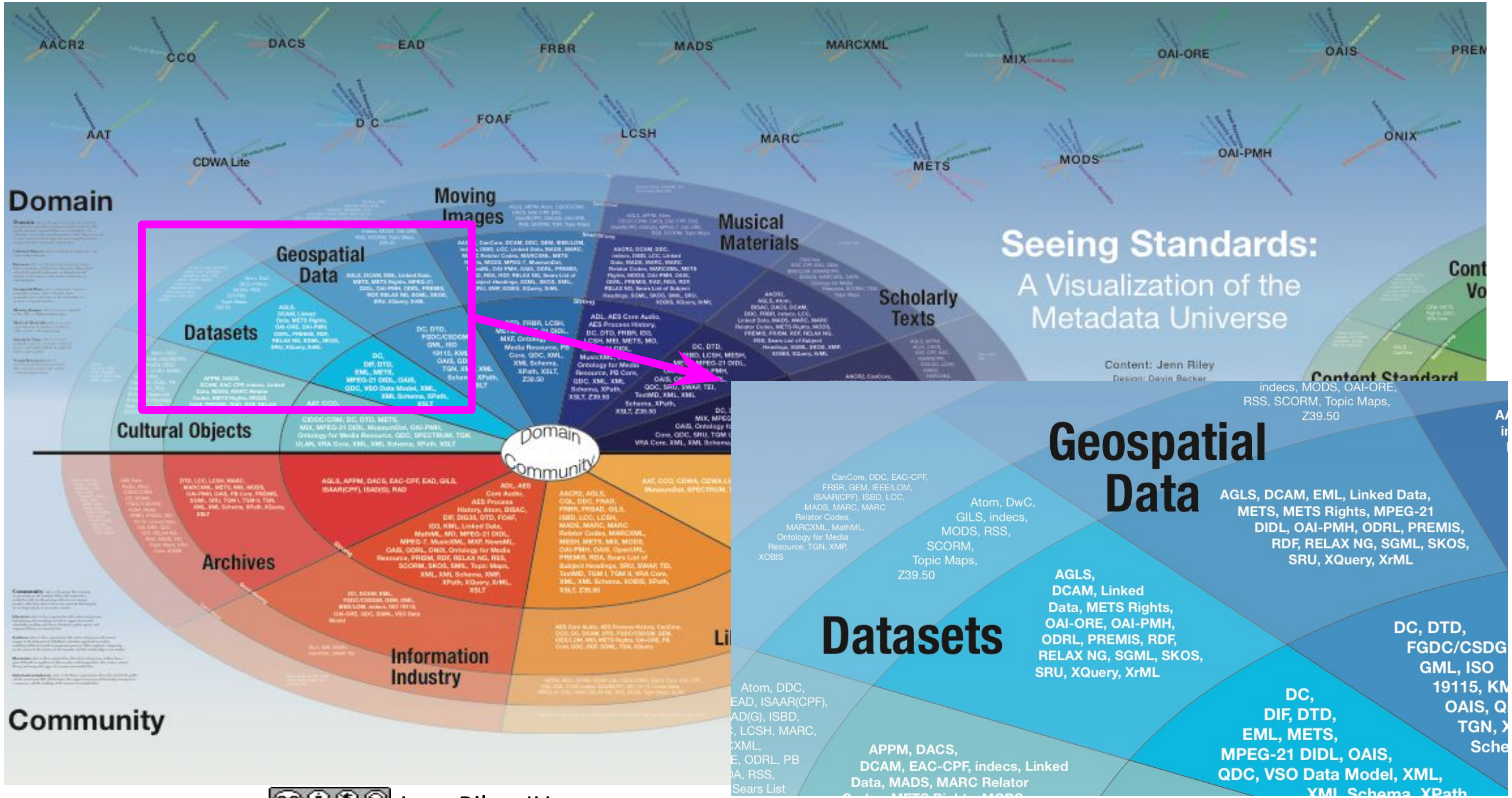
Theory

- Our boss: “*We use Dublin Core, PREMIS, X, Y, and Z, for interoperability.*”

Practice

- Cataloger/Archivist/Scientist: “*Frankly, we use those vocabularies, but with our own local modifications to make them work for our objects.*”
- Many unofficial dialects – per institution, per laboratory, per project
- Poor interoperability (see *Metadata's Bitter Harvest*, Library Journal, 2004)

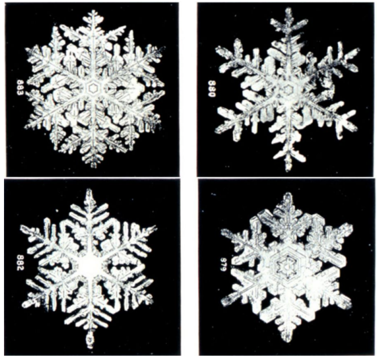
The Metadata Universe



AACR2
CCO
DACS
EAD
FRBR
MADS
MARCXML
MIX
OAI-ORE
OAIS
PREM
AAT
CDWA Lite
D C
FOAF
LCSH
MARC
METS
MODS
OAI-PMH
ONIX
Domain
Moving Images
Musical Materials
Scholarly Texts
Geospatial Data
Datasets
Cultural Objects
Archives
Information Industry
Community
Domain
Community
Seeing Standards:
A Visualization of the Metadata Universe
Content: Jenn Riley
Design: Devin Becker
Content Standard
indecs, MODS, OAI-ORE, RSS, SCORM, Topic Maps, Z39.50
Geospatial Data
AGLS, DCAM, EML, Linked Data, METS, METS Rights, MPEG-21 DIDL, OAI-PMH, ODRL, PREMIS, RDF, RELAX NG, SGML, SKOS, SRU, XQuery, XrML
CanCore, DDC, EAC-CPF, FRBR, GEM, IEEE/LOM, ISAAR(CPF), ISBD, LCC, MADS, MARC, MARC Relator Codes, MARCXML, MathML, Ontology for Media Resource, TGN, XMP, XOBIS
Atom, DwC, GILS, indecs, MODS, RSS, SCORM, Topic Maps, Z39.50
Datasets
AGLS, DCAM, Linked Data, METS Rights, OAI-ORE, OAI-PMH, ODRL, PREMIS, RDF, RELAX NG, SGML, SKOS, SRU, XQuery, XrML
DC, DTD, FGDC/CSDG GML, ISO 19115, KM OAIS, Q TGN, X Sche
DC, DIF, DTD, EML, METS, MPEG-21 DIDL, OAIS, QDC, VSO Data Model, XML, XML Schema, XPath
Atom, DDC, EAD, ISAAR(CPF), AD(G), ISBD, LCSH, MARC, XML, E, ODRL, PB A, RSS, Sears List
APPM, DACS, DCAM, EAC-CPF, indecs, Linked Data, MADS, MARC Relator

Jenn Riley, IU

Domain dialects – similar but different

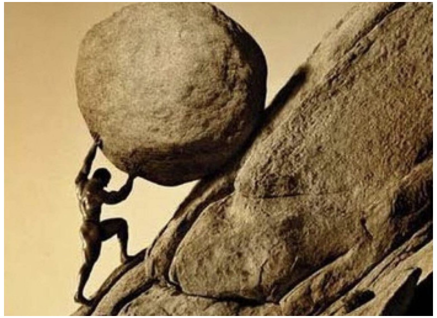


Example: Earth Science > Cryospheric (frozen water) Science

- 28 different definitions of “glacier”
- 8 different definitions of “puddle”
- 13 different definitions of “firn” (old snow)
- 10 different definitions of “frazil ice” (fine spicules of floating ice)
- 7 different definitions of “ogive” (bands of light and dark ice in a glacier)
- … and on and on

Sound familiar? What about your domain?

Which term definitions to keep or toss?

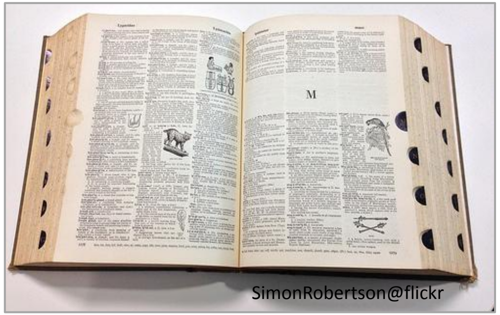


Traditional approach

- Busy experts in <*your fast-moving field*> on metadata standards committees
- … often take years to reach consensus
- … often no field testing is logistically feasible

An alternate approach is Yamz.net.

Yamz.net (Yet another metadata (zoo))


SimonRobertson@flickr

Yamz is not a standard, nor an ontology

- Yamz is a *living* dictionary of metadata terms
- Each term gets an ARK permalink, becoming a kind of
- … proposed nano-standard – some are upvoted, others not
- Reputation-based voting (like Stack Overflow) helps choosing

Yamz is a *microservice* for sharing, testing, and revising metadata dialect

- All parts of metadata “speech”, all domains, all ontologies
- Field testing via practitioners and reputation-based voting
- Ontologies, software, and predicates reference metadata terms via ARKs

Please reach out with feedback

1. Physical Samples for Earth Sciences	2. ARKs for Biomedical AI	3. Metadata Terms Crowdsourced
Dave Vieglais University of Kansas **vieglais@ku.edu**	Tim Clark University of Virginia **twc8q@virginia.edu**	John Kunze ARK Alliance **jakkbl@gmail.com**

arks.org