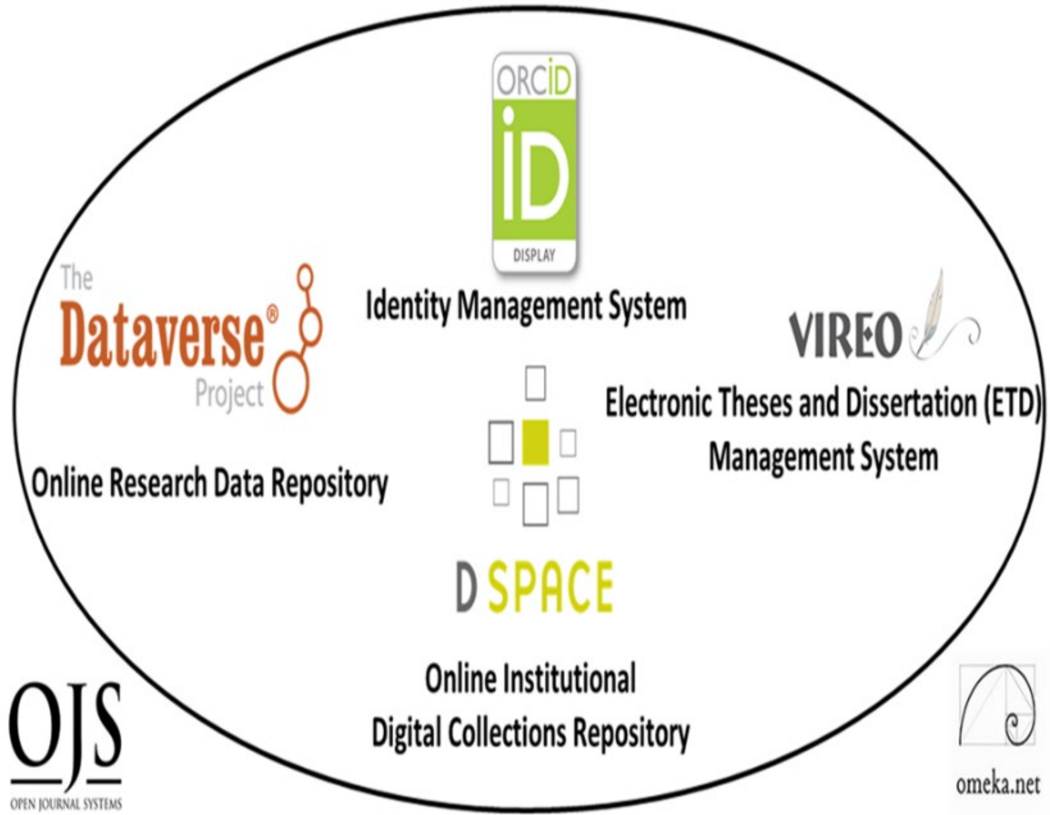
Online Data Research Repositories

From Research **Data and Datasets** to **Artificial Intelligence and Discovery**

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries
ruzwyshyn@txstate.edu
http:// ruzwyshyn.net

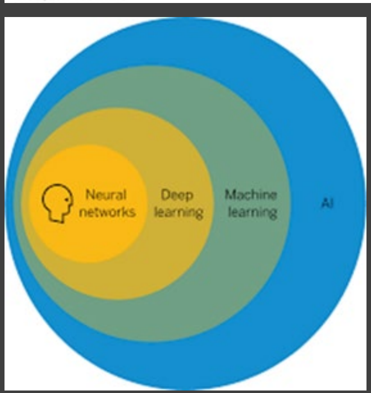


Identity Management System

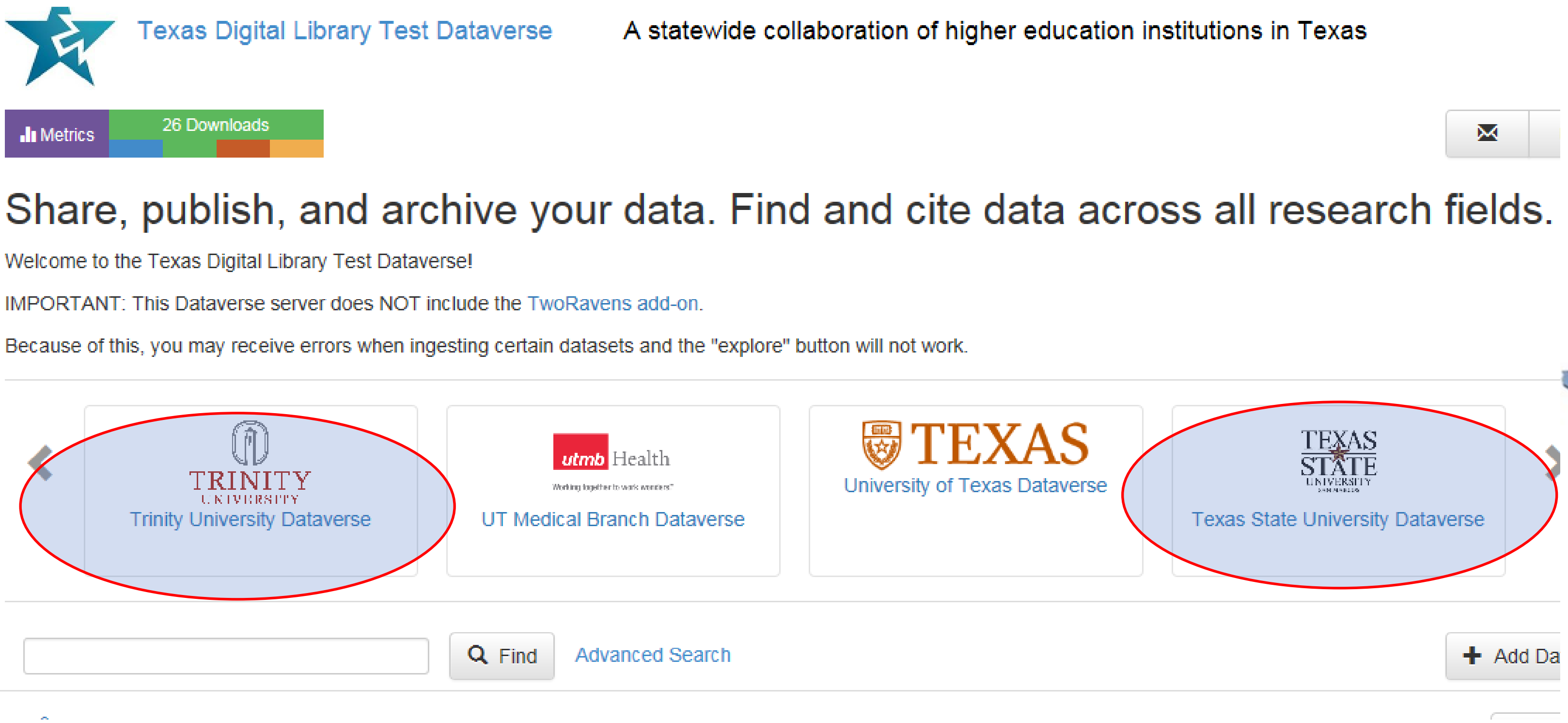
Online Research Data Repository

Electronic Theses and Dissertation (ETD) Management System

Online Institutional Digital Collections Repository



What is an Online Data Research Repository?



Texas Digital Library Test Dataverse

A statewide collaboration of higher education institutions in Texas

Share, publish, and archive your data. Find and cite data across all research fields.

Welcome to the Texas Digital Library Test Dataverse!

IMPORTANT: This Dataverse server does NOT include the TwoRavens add-on.

Because of this, you may receive errors when ingesting certain datasets and the "explore" button will not work.

Trinity University Dataverse

UT Medical Branch Dataverse

University of Texas Dataverse

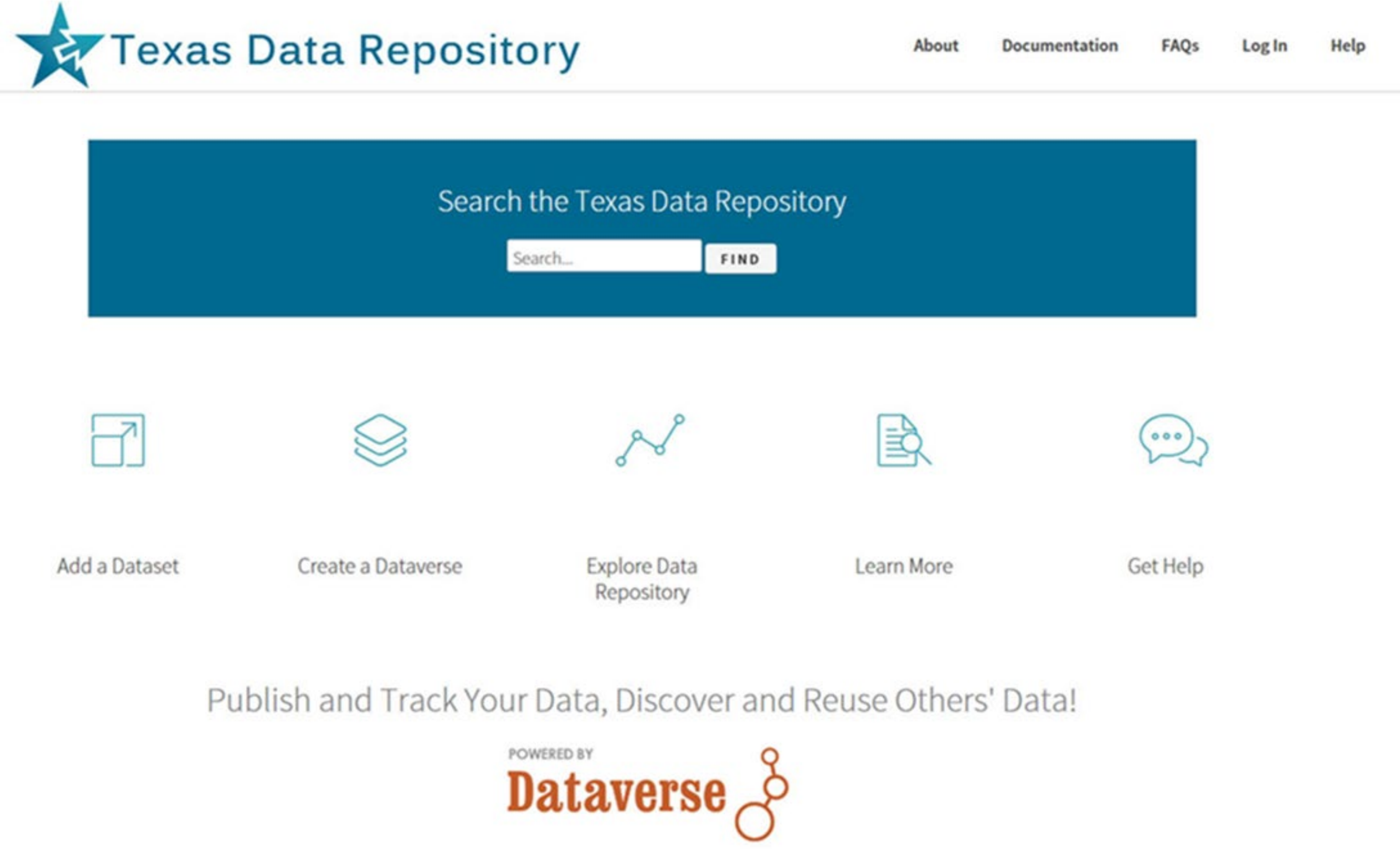
Texas State University Dataverse

TRINITY UNIVERSITY
utmb Health
TEXAS
TEXAS STATE UNIVERSITY
Find
Advanced Search
Add Da

West Texas A&M University
UNT
SFA
UT Southwestern Medical Center
UT DALLAS
ASU
BU
A&M
SH
HOUSTON
TEXAS STATE
University of Houston Clear Lake
A&M Galveston
TRINITY
utmb Health
TEXAS A&M UNIVERSITY CORPUS CHRISTI
UTRGV

Texas Data Repository which is a shared repository of several Texas Universities leveraging technological cooperation and expertise among academic research libraries libraries, https://dataverse.tdl.org

What is the Utility of An Online Research Data Repository?

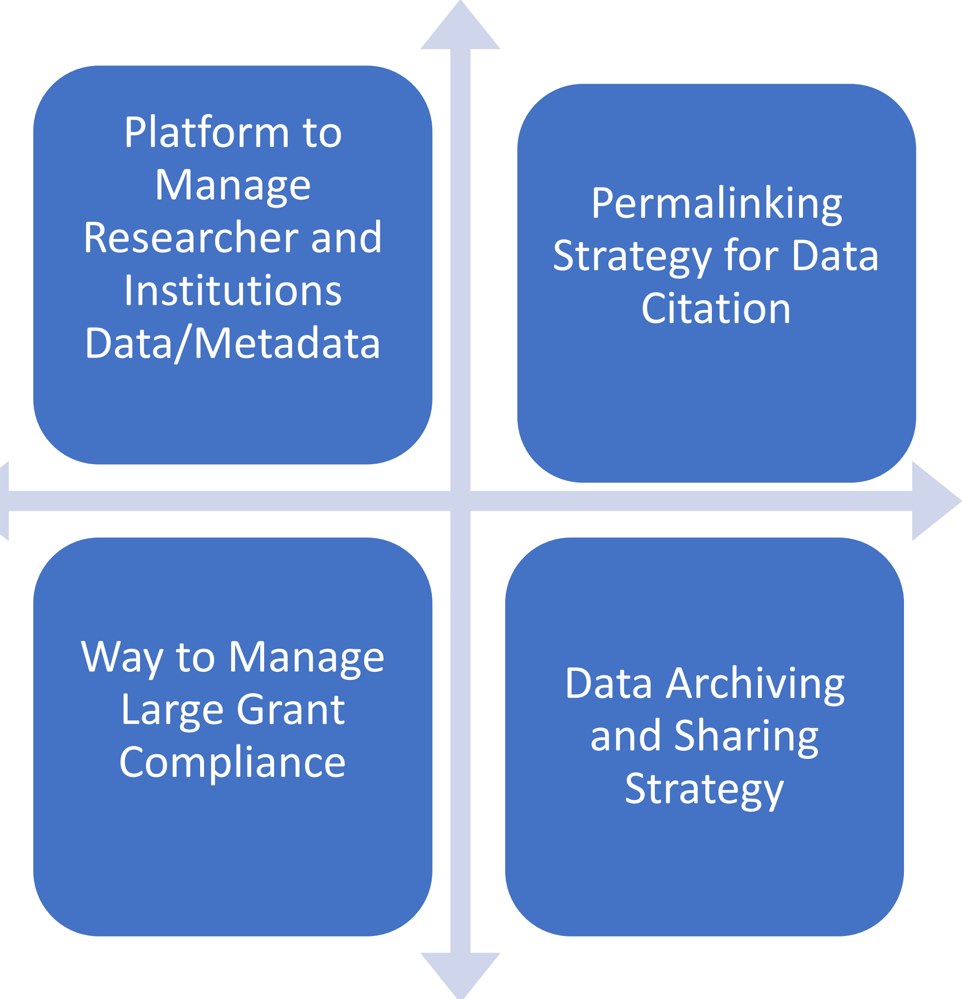


Texas Data Repository

Search the Texas Data Repository

Publish and Track Your Data, Discover and Reuse Others' Data!

https://dataverse.tdl.org/



Platform to Manage Researcher and Institutions Data/Metadata

Permalinking Strategy for Data Citation

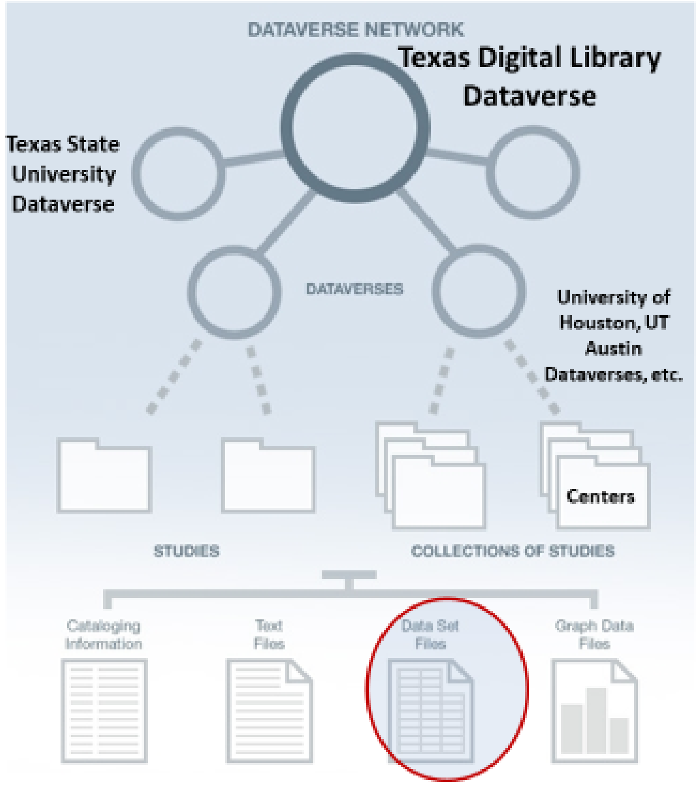
Way to Manage Large Grant Compliance

Data Archiving and Sharing Strategy

Texas State University Dataverse:
Can be configured as Single Instance or as a Consortial Model

Dataverse Architecture

(Consortial)

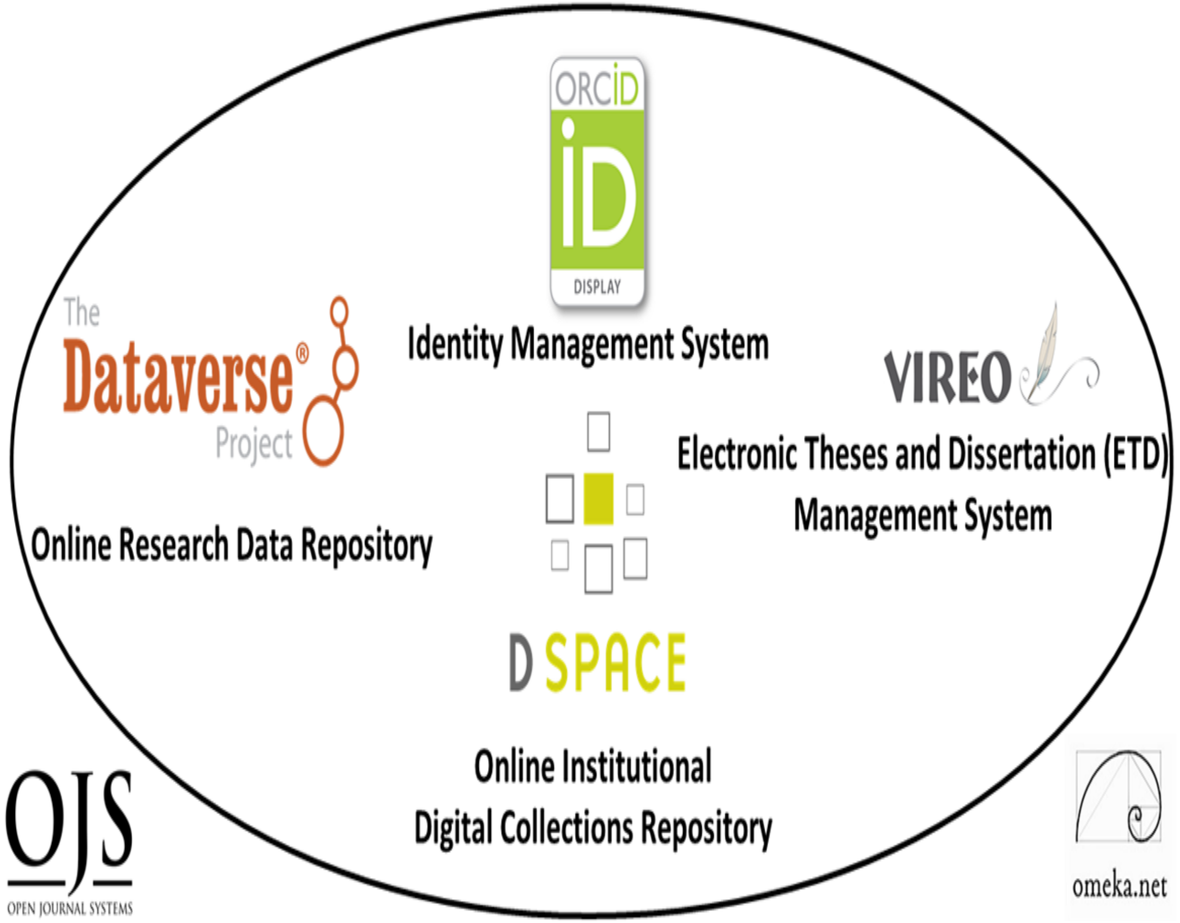


https://dataverse.tdl.org/

(Texas Aggregates Various Individual Universities through the Texas Digital Library)

Texas State University Dataverse
A platform for publishing and archiving Texas State University's research data.
Dataverse
TEXAS STATE
UNIVERSITY LIBRARIES

A Data Repository May Also Be Placed Within a Larger Digital Scholarship Research Ecosystem



Identity Management System

Online Research Data Repository

Electronic Theses and Dissertation (ETD) Management System

Online Institutional Digital Collections Repository

Digital & Web Services : University Libraries : Texas State University (txstate.edu)

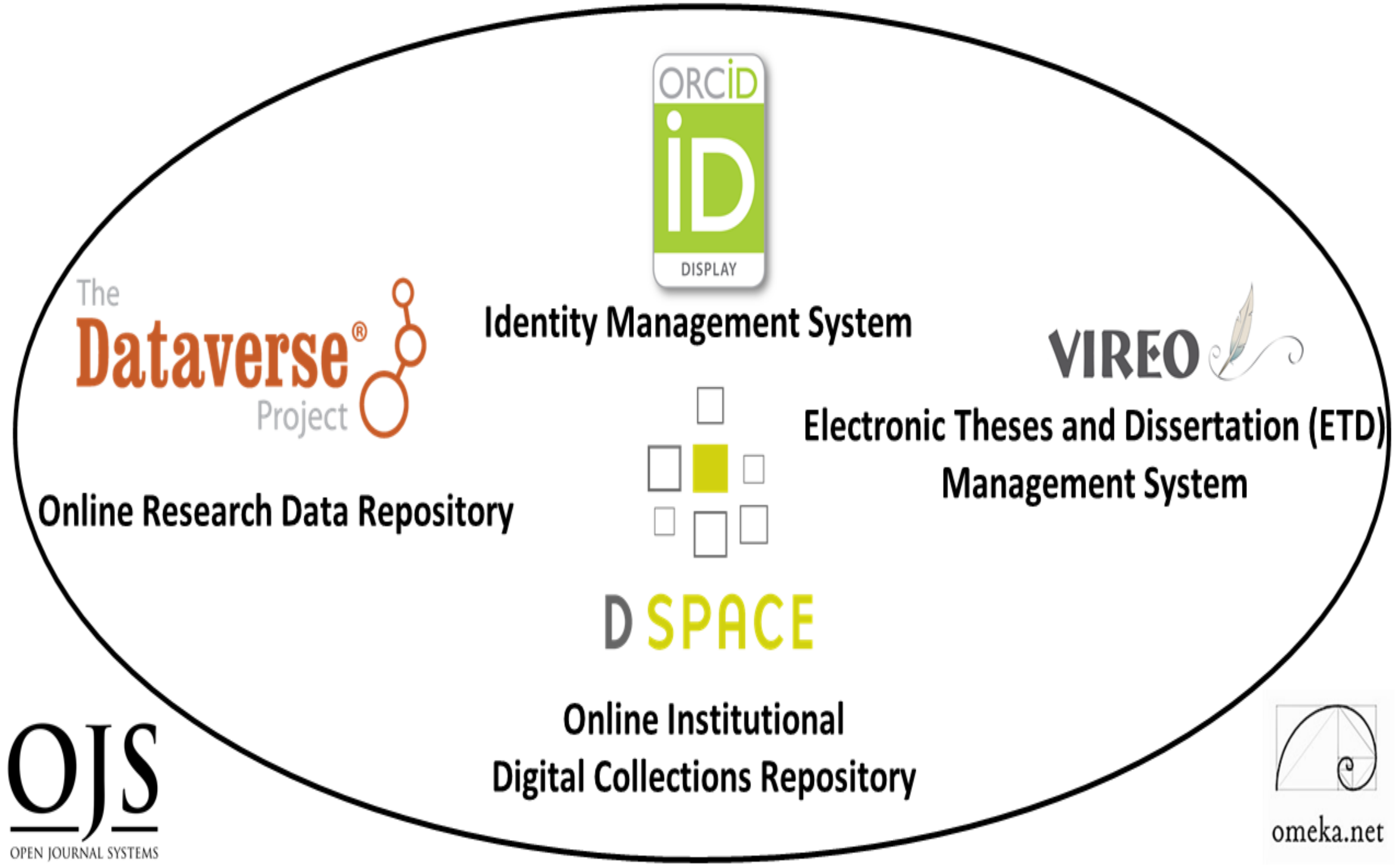
TWO PRIMARY COMPONENTS
(Content)

- **RESEARCH DATA REPOSITORY**
- DIGITAL COLLECTIONS REPOSITORY

FOUR TERTIARY COMPONENTS
(Communication)

- Electronic Thesis and Dissertation Management System
- Identity Management System
- Open Academic Journal Software
- User Interface/Content Management Software

What are the General Common Characteristics for a Data Repository and Digital Scholarship Ecosystem?



Identity Management System

Online Research Data Repository

Electronic Theses and Dissertation (ETD) Management System

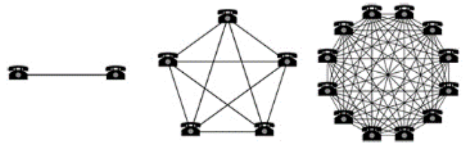
Online Institutional Digital Collections Repository

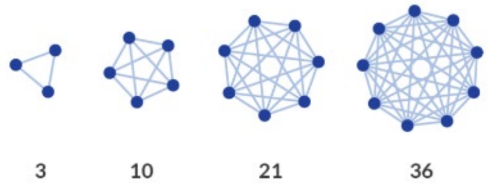
Open Source Software

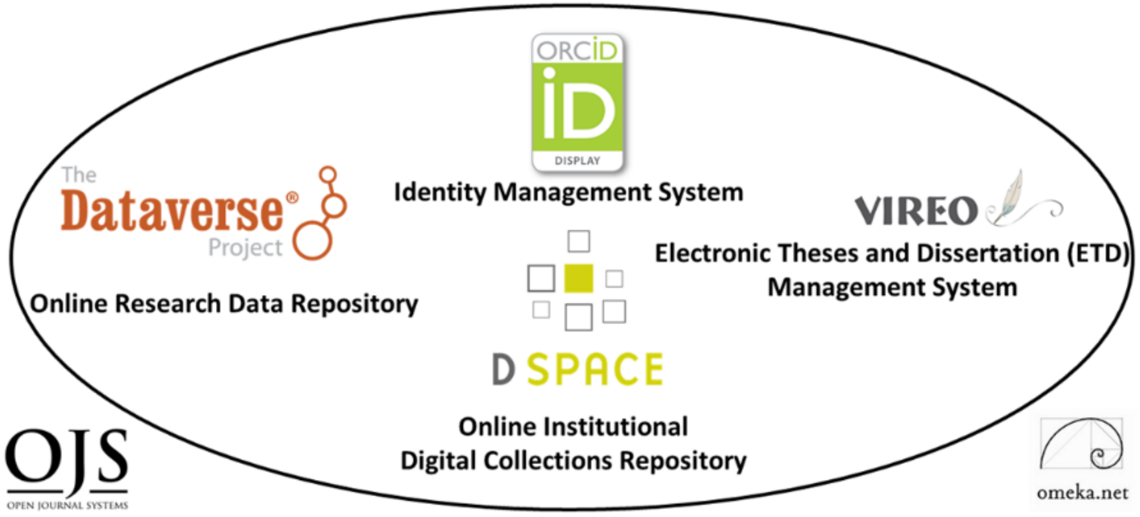
Active Developer Communities

Customizable Components

Collocating Open Source Digital Components in a Networked Research Ecosystem Enables Larger Connections and/or Network Effects







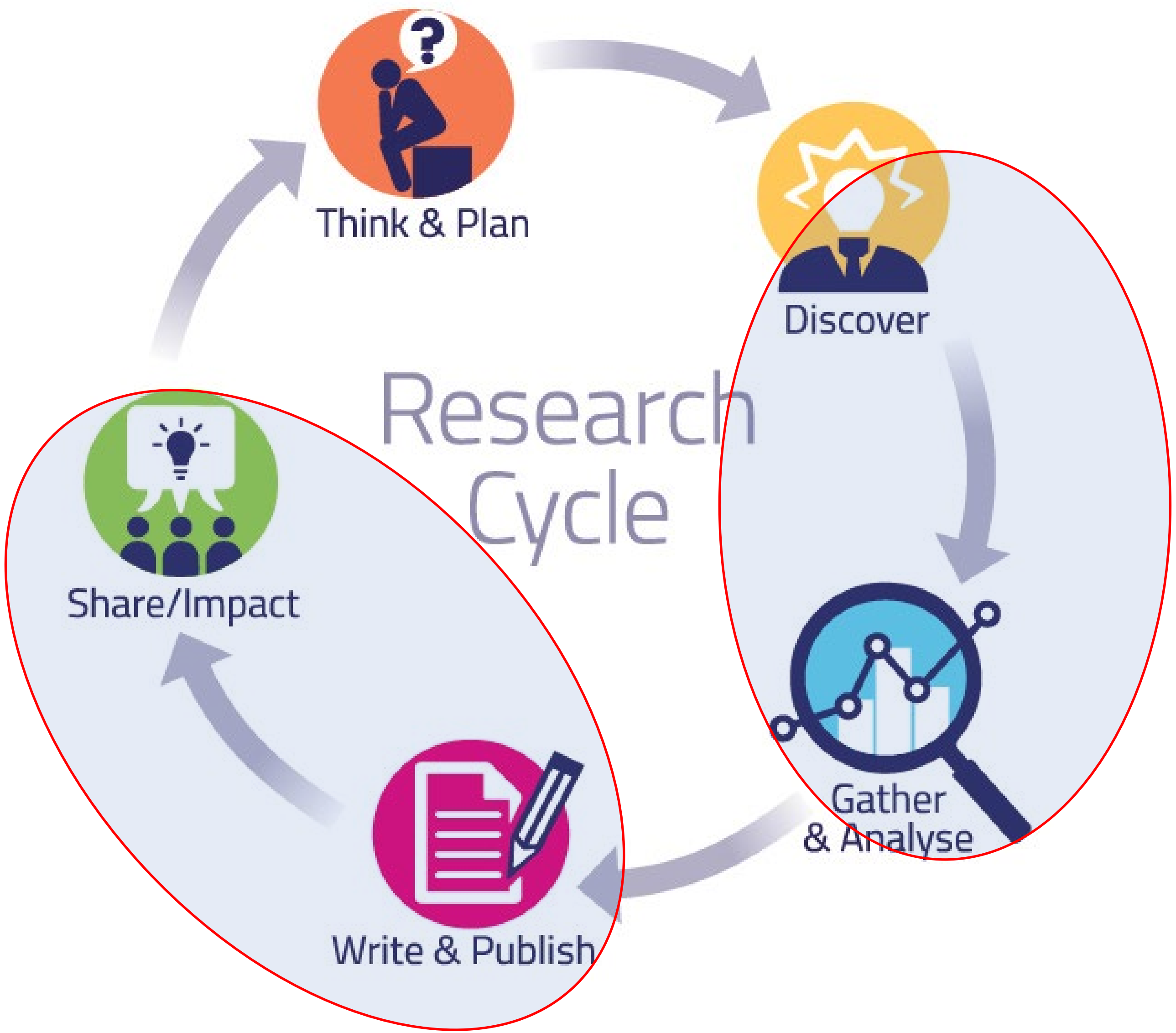
Identity Management System

Online Research Data Repository

Electronic Theses and Dissertation (ETD) Management System

Online Institutional Digital Collections Repository

Together These Digital Ecosystem Components Enable the Academic Research Cycle



Pragmatic Levels

Big, Bigger Data and Big Data

One Size Does Not Fit All for Various Data Research Repository Project Needs

Many Types of Data Projects (Sizes)

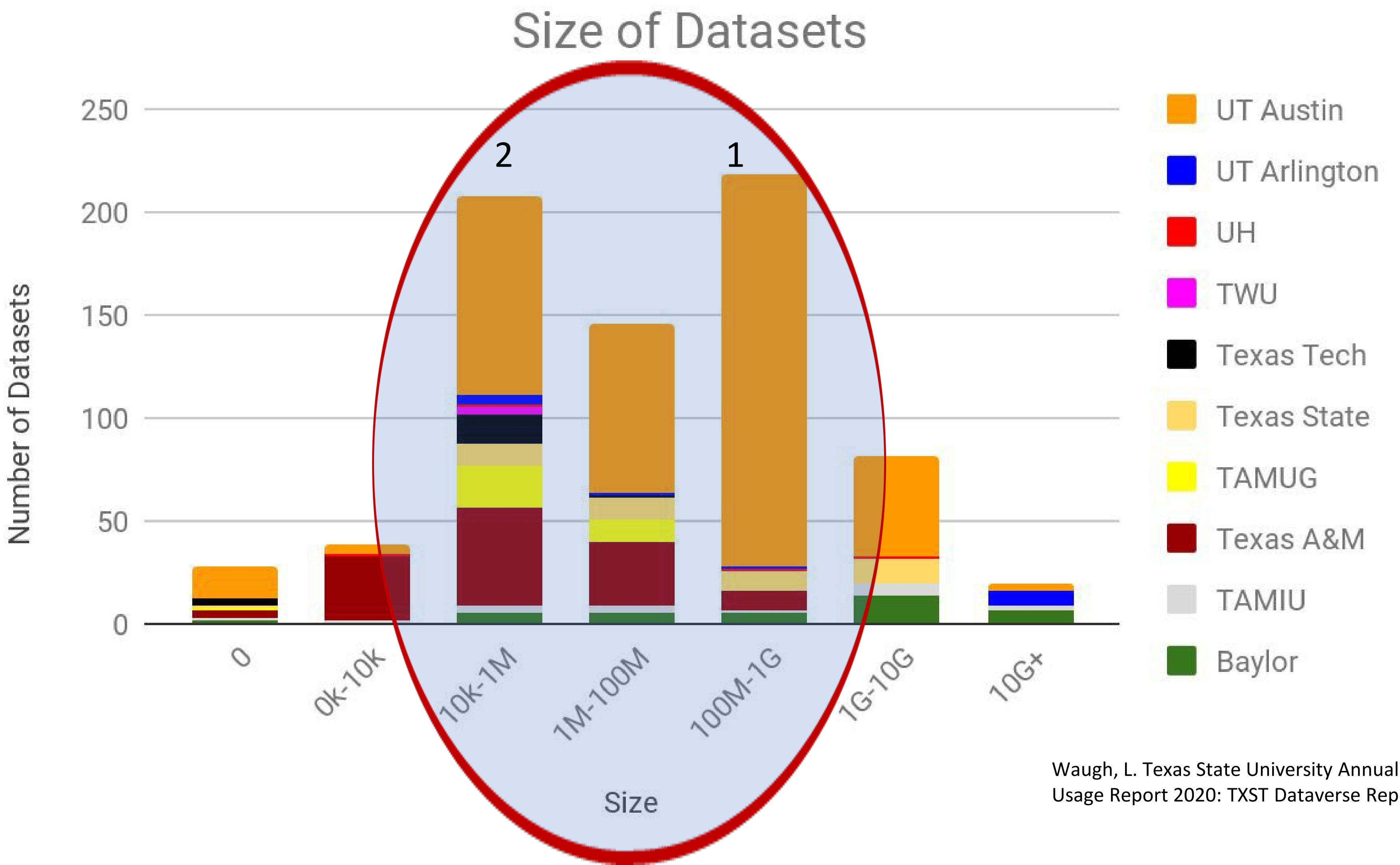
1) Normal range (<4GB Files <10GB Datasets)
Files/Data Fit on Server/Cloud, may be uploaded to the Data Repository, 4GB files, 10GB Datasets)

2) Large Projects, Bigger Data <TB
(Data may require specialized university IT Support, i.e. terabyte/petabyte tape drives, Pointers possible)

3) Huge Projects, Big Data
(Projects require consortial possibilities, national models, **Texas Advanced Computer Center TAAC**, Duracloud, AWS S3, Custom Solutions)

Present Sizes of Texas Data Repository Datasets

Most 1MB <1GB, Greater than 10 GB+ Rare



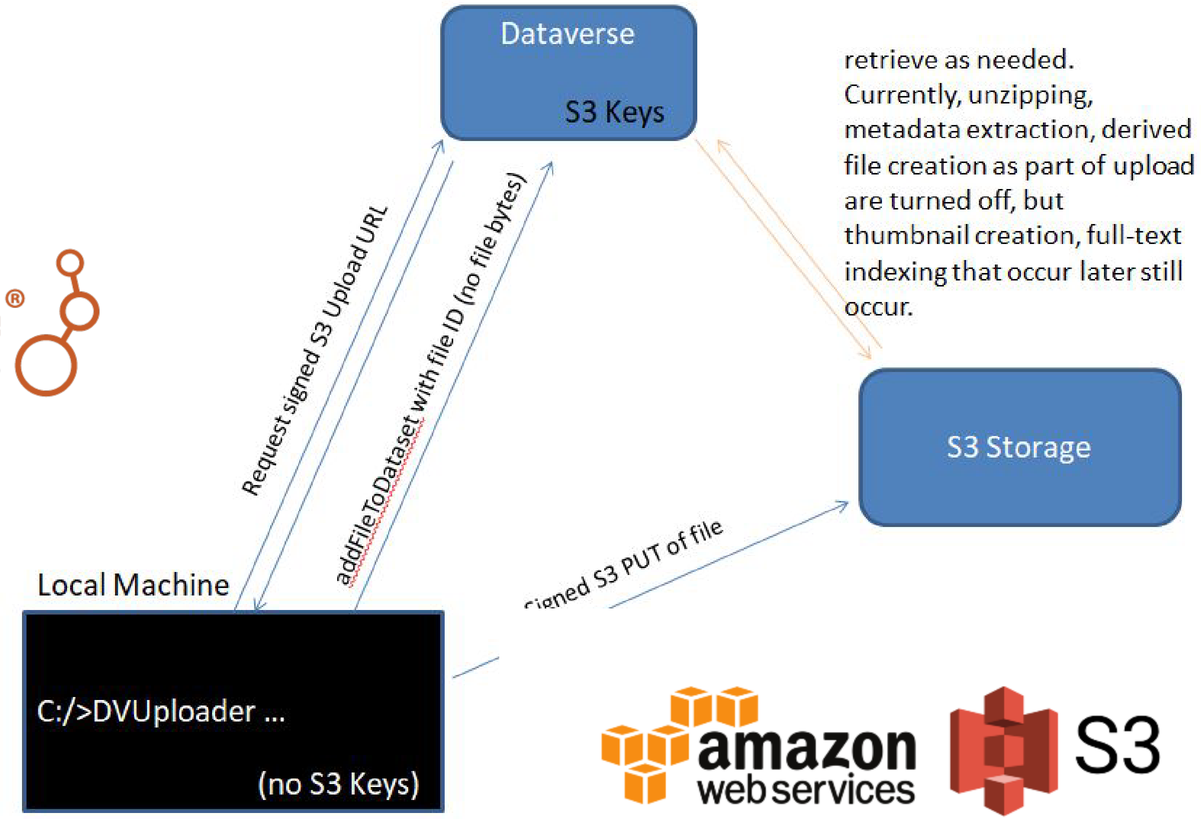
Waugh, L. Texas State University Annual
Usage Report 2020: TXST Dataverse Repository. Texas State University

Beta Prototyping Bigger Data Options

2020-2022

DRYAD

The Dataverse® Project



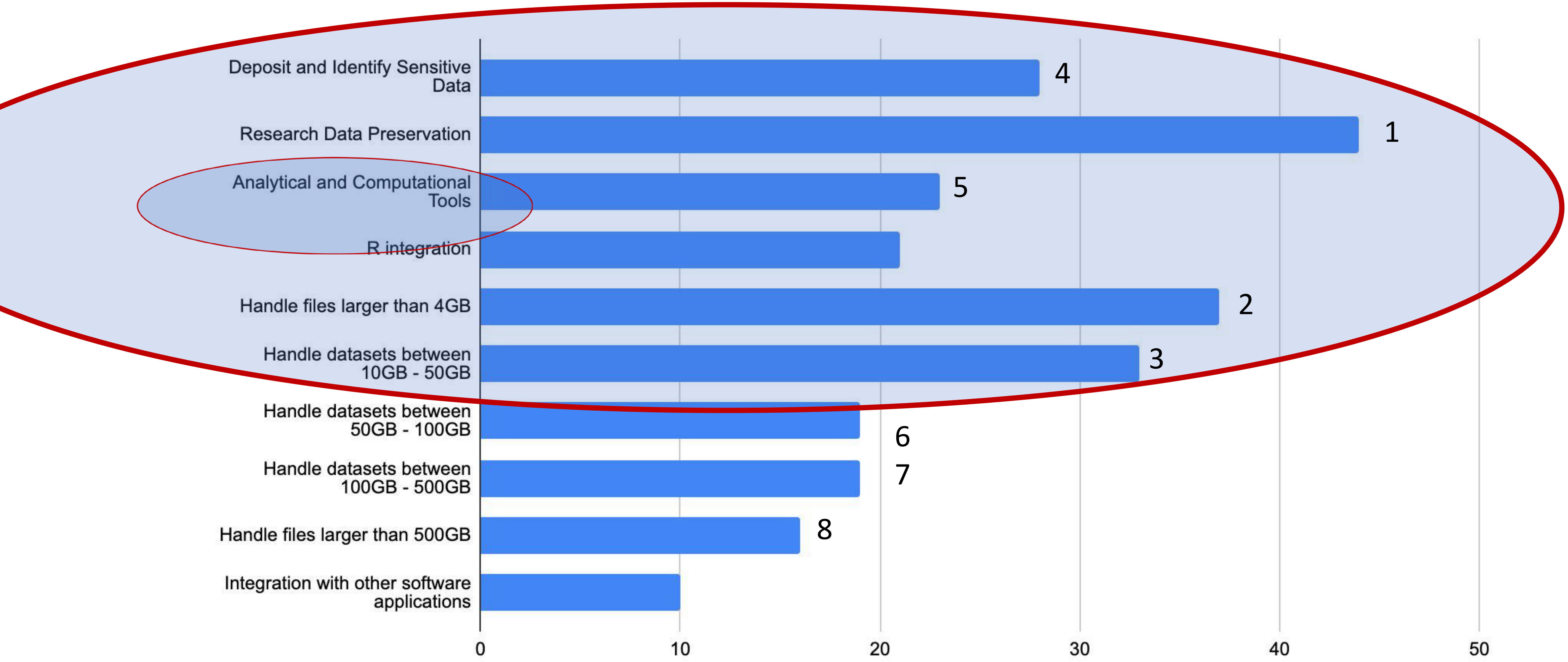
retrieve as needed. Currently, unzipping, metadata extraction, derived file creation as part of upload are turned off, but thumbnail creation, full-text indexing that occur later still occur.

Up to 300 GB/dataset
Fee Based Institutional Model 7.5/13.5 K/Year

TACC
TEXAS ADVANCED COMPUTING CENTER

<20 GB Upload
(Download Challenges)

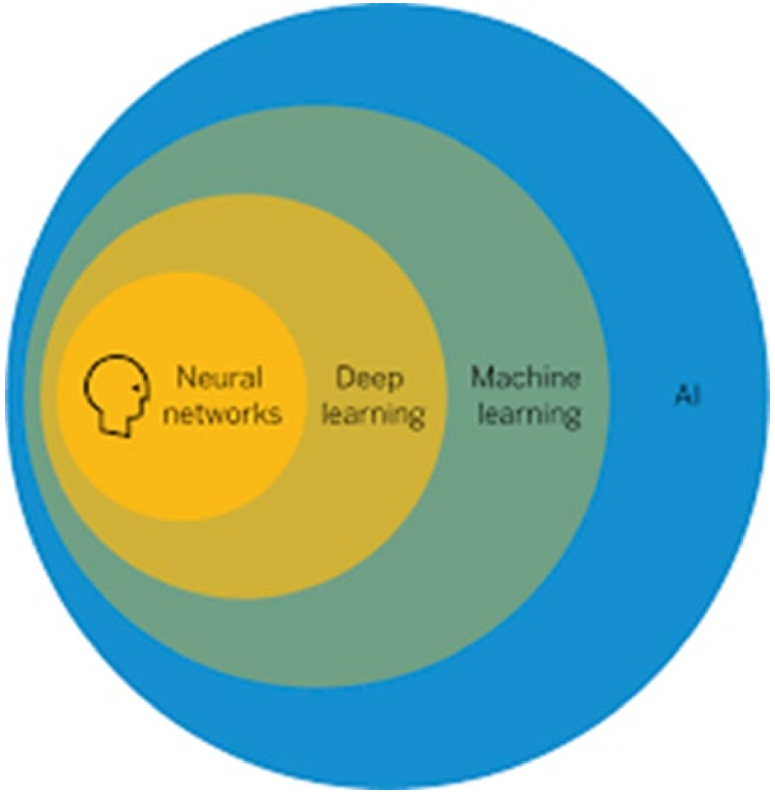
What New Data Repository Features Would Users Like to See in 2022?



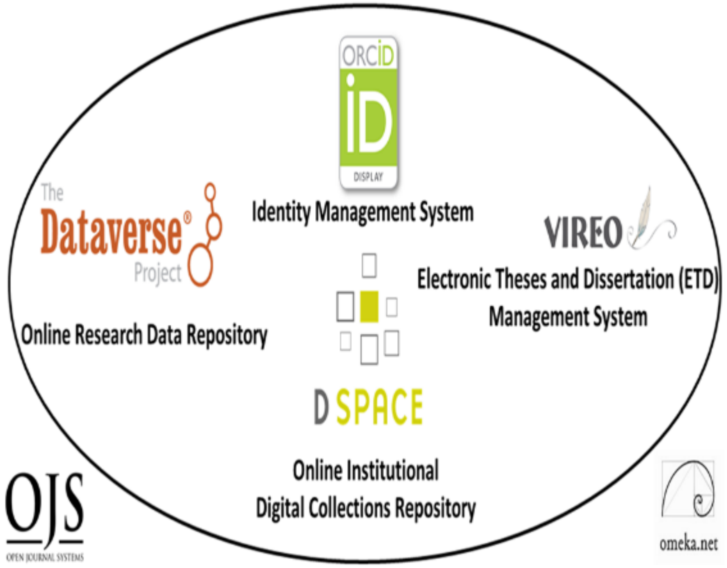
Chan- Park, C. and Sare, L. Waugh, S. Results of the Texas Data Repository User Survey, 2022. Texas Conference on Digital Libraries, 2022.

Last Five Years Has Shown Incredible Progress of, Analytical Computational Tools, Particularly, AI

Artificial Intelligence (Machine Learning (Deep Learning)) = Better Algorithms +
Greater Computing Power +
Large Data Sets

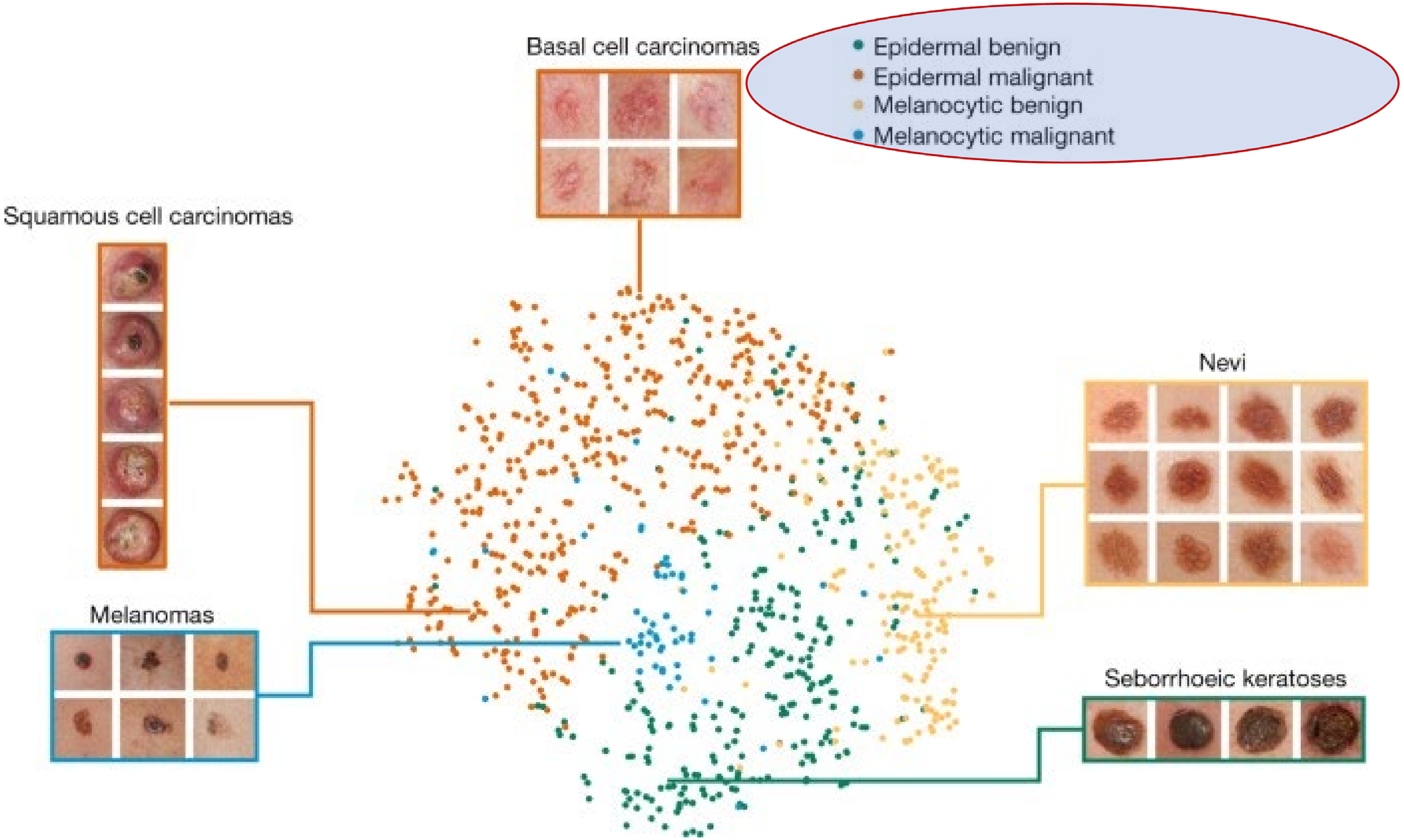


- **Computer Vision (Facial/Object Recognition Cancer Cell Detection))**
- Natural Language Processing (Speech to Text, Translation)
- Cybersecurity, Fraud Detection
- Conversational Chatbots & Robotic Agents
- Strategic Reasoning (AlphaGo)


ORCiD
iD
DISPLAY
Identity Management System
The Dataverse Project
Online Research Data Repository
VIREO
Electronic Theses and Dissertation (ETD) Management System
DSPACE
Online Institutional Digital Collections Repository
OJS
OPEN JOURNAL SYSTEMS
omeka.net

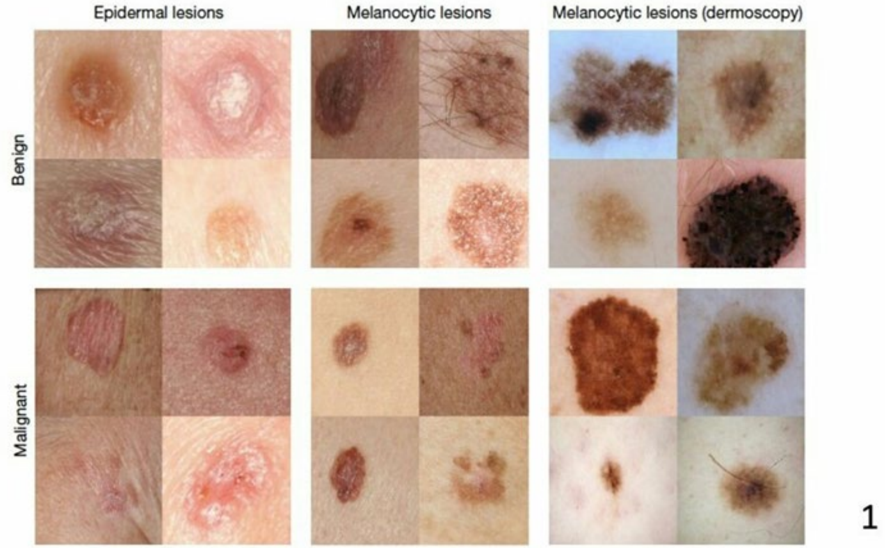
Computational Tools, Digital Ecosystems and AI Example

Major Cancer Detection Discovery Through AI Neural Nets 2017



Dermatologist-level Classification of Skin Cancer with Deep Neural Networks, Video Stanford

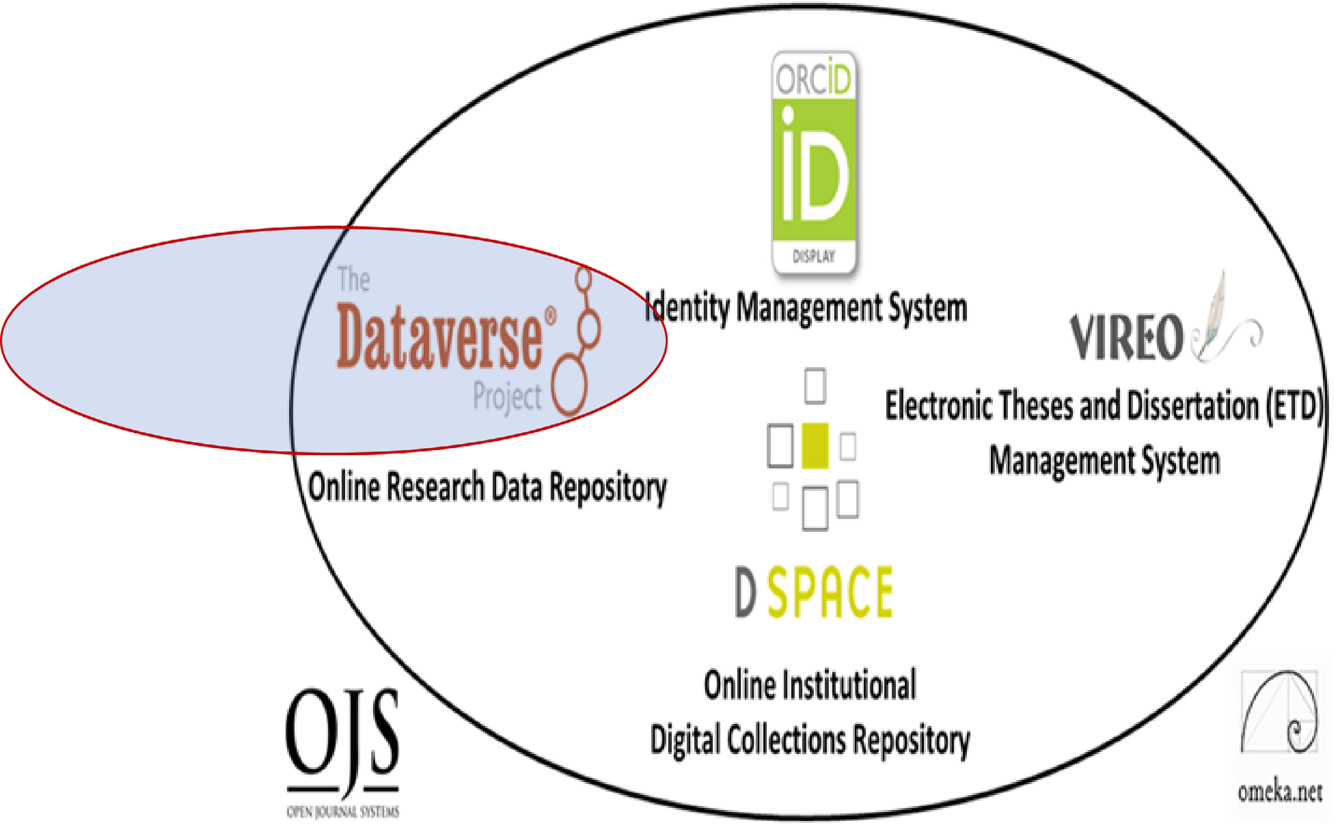
21 Board Certified Stanford Dermatologists
129,450 images of 2,032 diseases
1.41 million AI training images



Dermatologist-level Classification of Skin Cancer with Deep Neural Networks, Andre Esteva, Brett Kupress, Sebastian Thrun et al. Nature **2017**, AI Models, Deep Learning, Convolutional Neural Nets, Labeled Medical Data from Image Data Archives

Combining Data Centered Research Ecosystems + Artificial Intelligence

(Many New Possibilities for Global Open Science, New Insights and NewDiscovery)

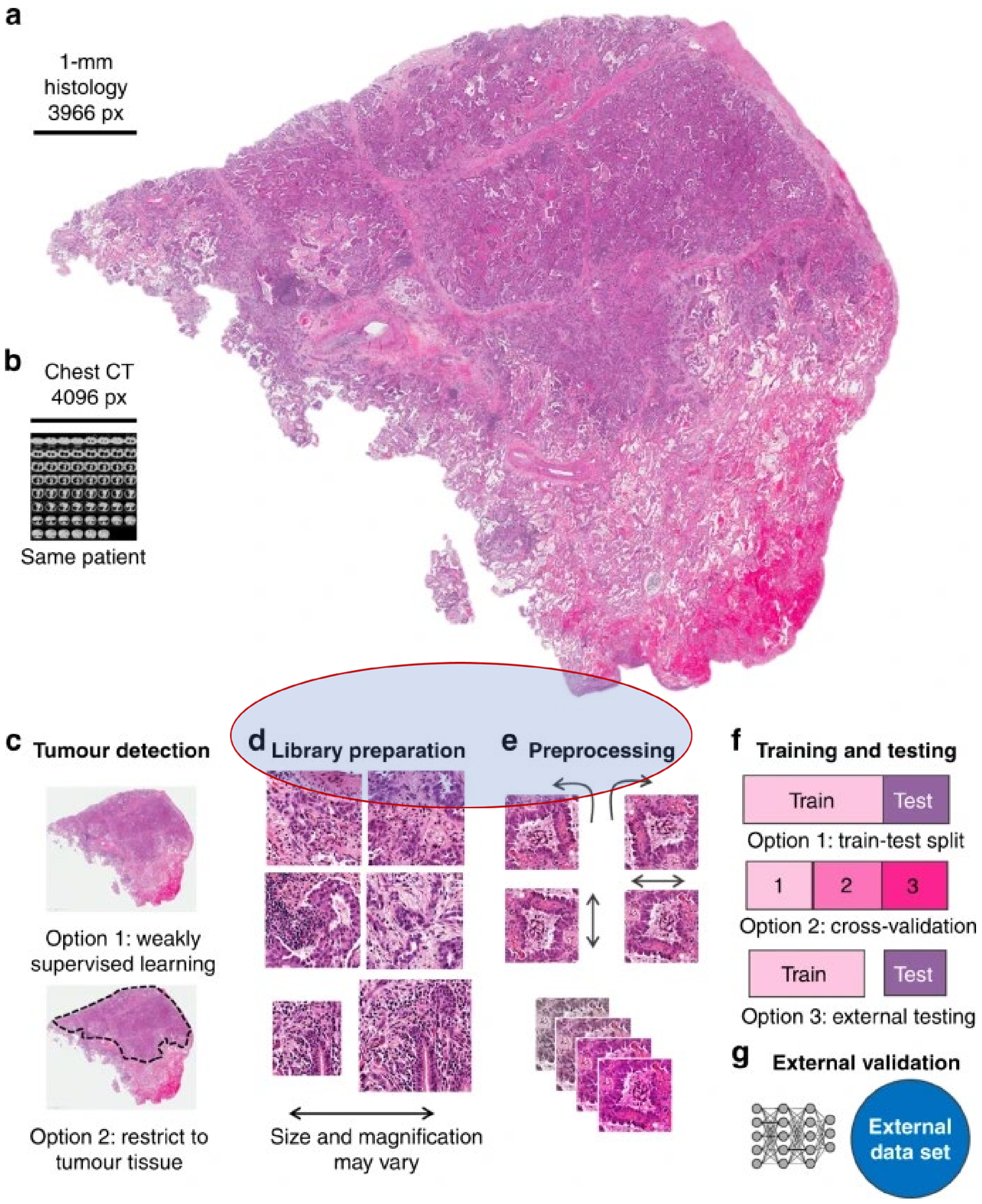


Identity Management System

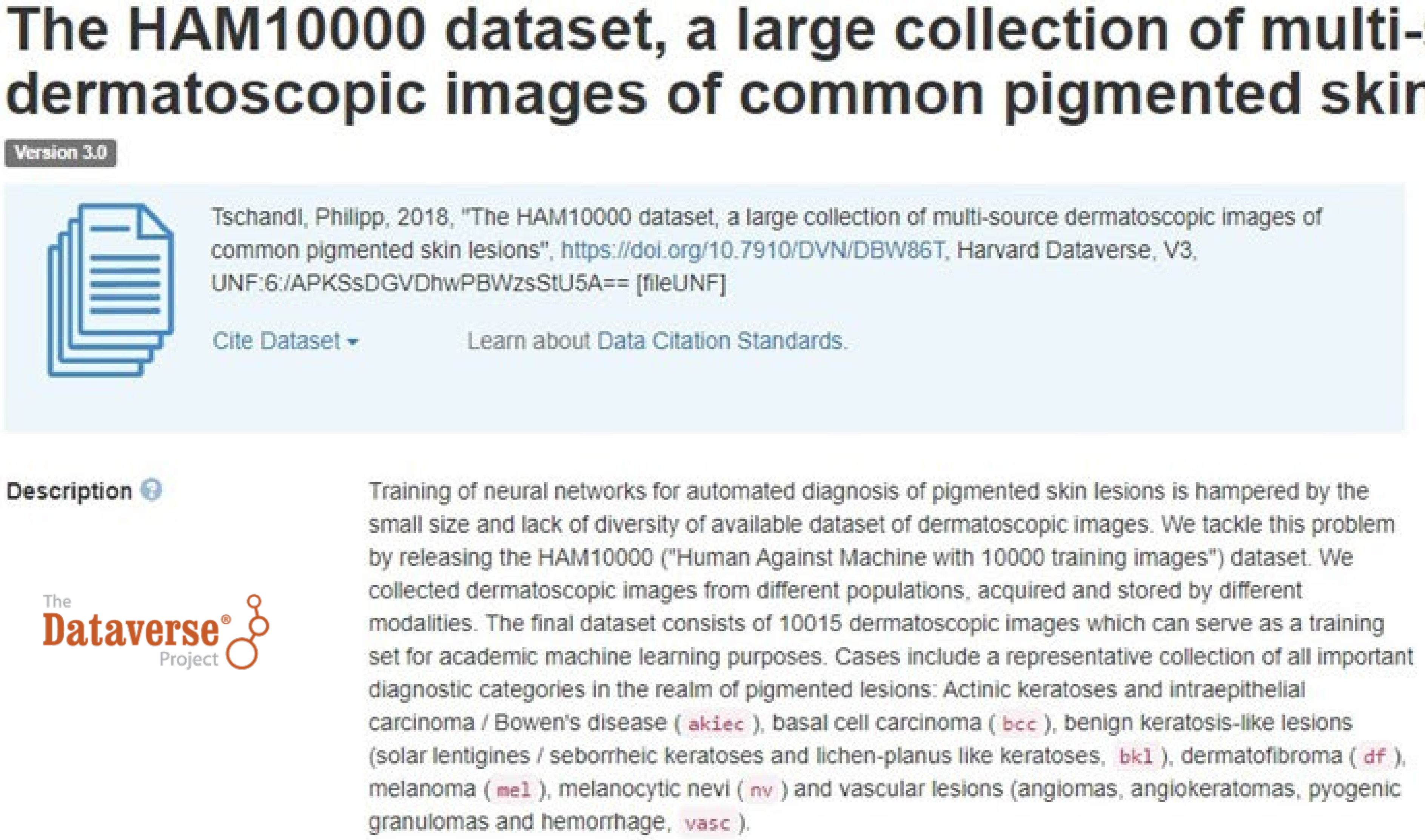
Electronic Theses and Dissertation (ETD) Management System

Online Research Data Repository

Online Institutional Digital Collections Repository



Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. British Journal of Cancer, Echle et al. November 2020



The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions

Version 3.0

Tschandl, Philipp, 2018, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", https://doi.org/10.7910/DVN/DBW86T, Harvard Dataverse, V3, UNF:6:/APKSsDGVDhwPBWzsStU5A== [fileUNF]

Cite Dataset ▾ Learn about Data Citation Standards.

Access Dataset ▾

Contact Owner | Share

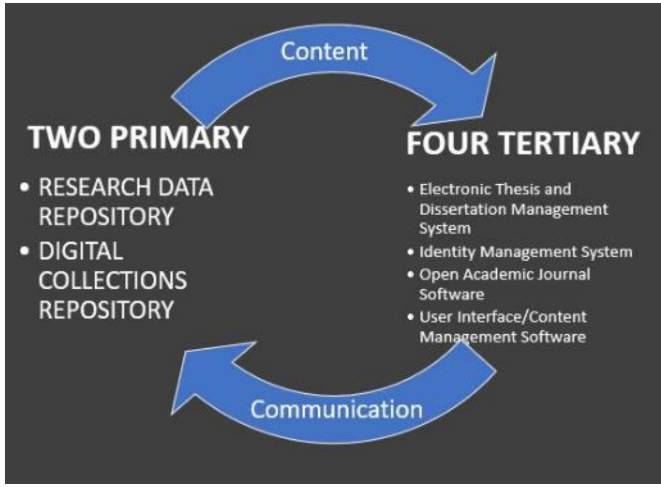
Dataset Metrics

58,334 Downloads

Description

Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available dataset of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (`akiec`), basal cell carcinoma (`bcc`), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, `bkl`), dermatofibroma (`df`), melanoma (`mel`), melanocytic nevi (`nv`) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, `vasc`).

The Dataverse® Project



TWO PRIMARY

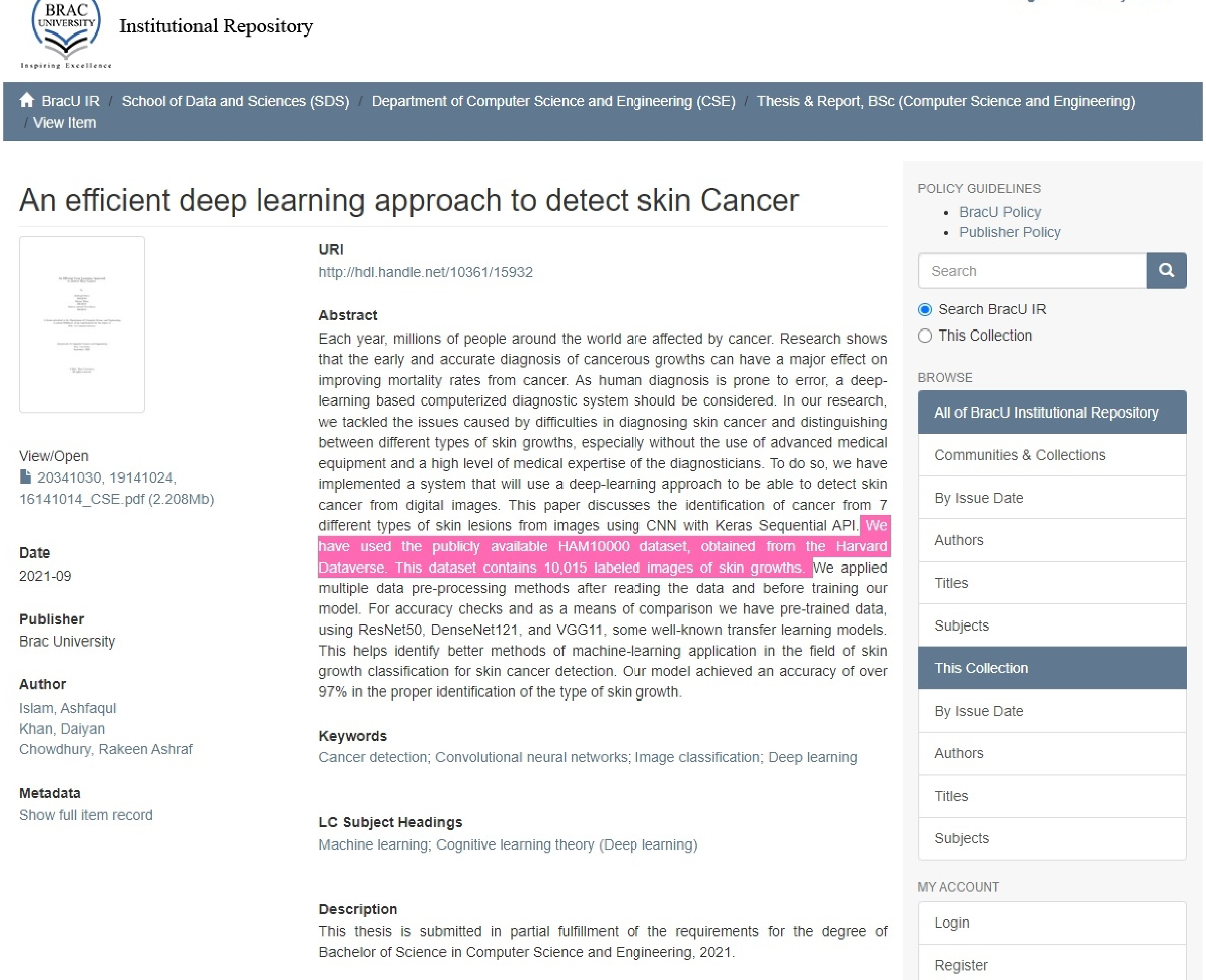
- RESEARCH DATA REPOSITORY
- DIGITAL COLLECTIONS REPOSITORY

FOUR TERTIARY

- Electronic Thesis and Dissertation Management System
- Identity Management System
- Open Academic Journal Software
- User Interface/Content Management Software

Harvard Dataverse Data Repository Open Science Dermatology Image Dataset, Philip Tschandl

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T



BracU IR / School of Data and Sciences (SDS) / Department of Computer Science and Engineering (CSE) / Thesis & Report, BSc (Computer Science and Engineering) / View Item

An efficient deep learning approach to detect skin Cancer

View/Open
20341030, 19141024, 16141014_CSE.pdf (2.208Mb)

Date
2021-09

Publisher
Brac University

Author
Islam, Ashfaqul
Khan, Daiyan
Chowdhury, Rakeen Ashraf

Metadata
Show full item record

URI
http://hdl.handle.net/10361/15932

Abstract

Each year, millions of people around the world are affected by cancer. Research shows that the early and accurate diagnosis of cancerous growths can have a major effect on improving mortality rates from cancer. As human diagnosis is prone to error, a deep-learning based computerized diagnostic system should be considered. In our research, we tackled the issues caused by difficulties in diagnosing skin cancer and distinguishing between different types of skin growths, especially without the use of advanced medical equipment and a high level of medical expertise of the diagnosticians. To do so, we have implemented a system that will use a deep-learning approach to be able to detect skin cancer from digital images. This paper discusses the identification of cancer from 7 different types of skin lesions from images using CNN with Keras Sequential API. We have used the publicly available HAM10000 dataset, obtained from the Harvard Dataverse. This dataset contains 10,015 labeled images of skin growths. We applied multiple data pre-processing methods after reading the data and before training our model. For accuracy checks and as a means of comparison we have pre-trained data, using ResNet50, DenseNet121, and VGG11, some well-known transfer learning models. This helps identify better methods of machine-learning application in the field of skin growth classification for skin cancer detection. Our model achieved an accuracy of over 97% in the proper identification of the type of skin growth.

Keywords
Cancer detection; Convolutional neural networks; Image classification; Deep learning

LC Subject Headings
Machine learning; Cognitive learning theory (Deep learning)

Description
This thesis is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering, 2021.

POLICY GUIDELINES
- BracU Policy
- Publisher Policy

Search

● Search BracU IR
○ This Collection

BROWSE

All of BracU Institutional Repository
Communities & Collections
By Issue Date
Authors
Titles
Subjects

This Collection
By Issue Date
Authors
Titles
Subjects

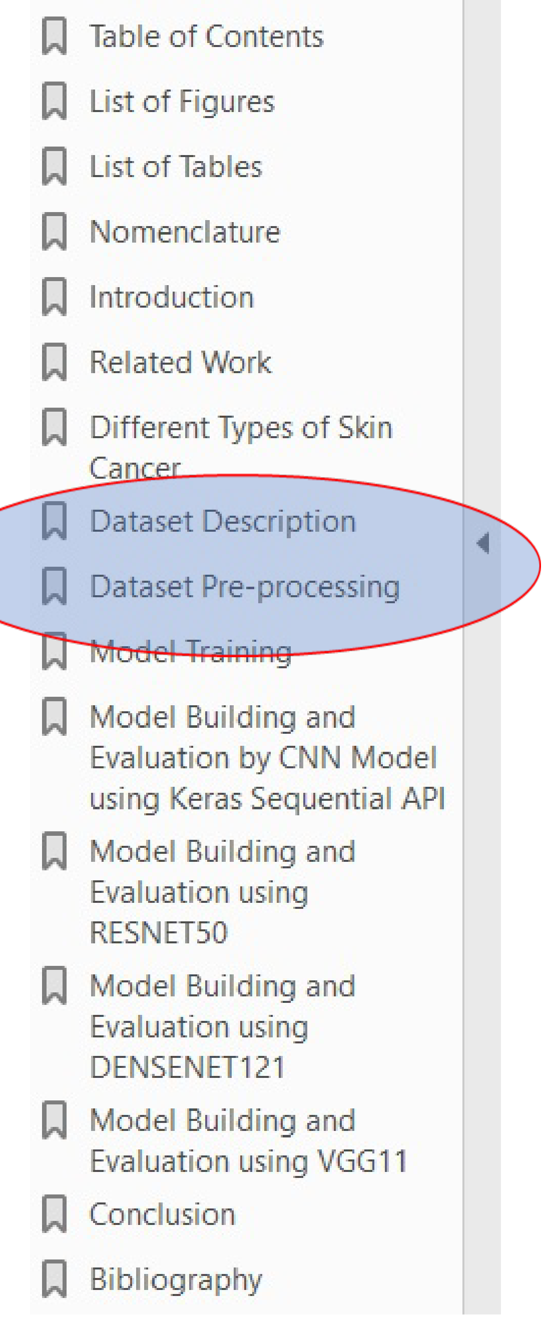
MY ACCOUNT
Login
Register

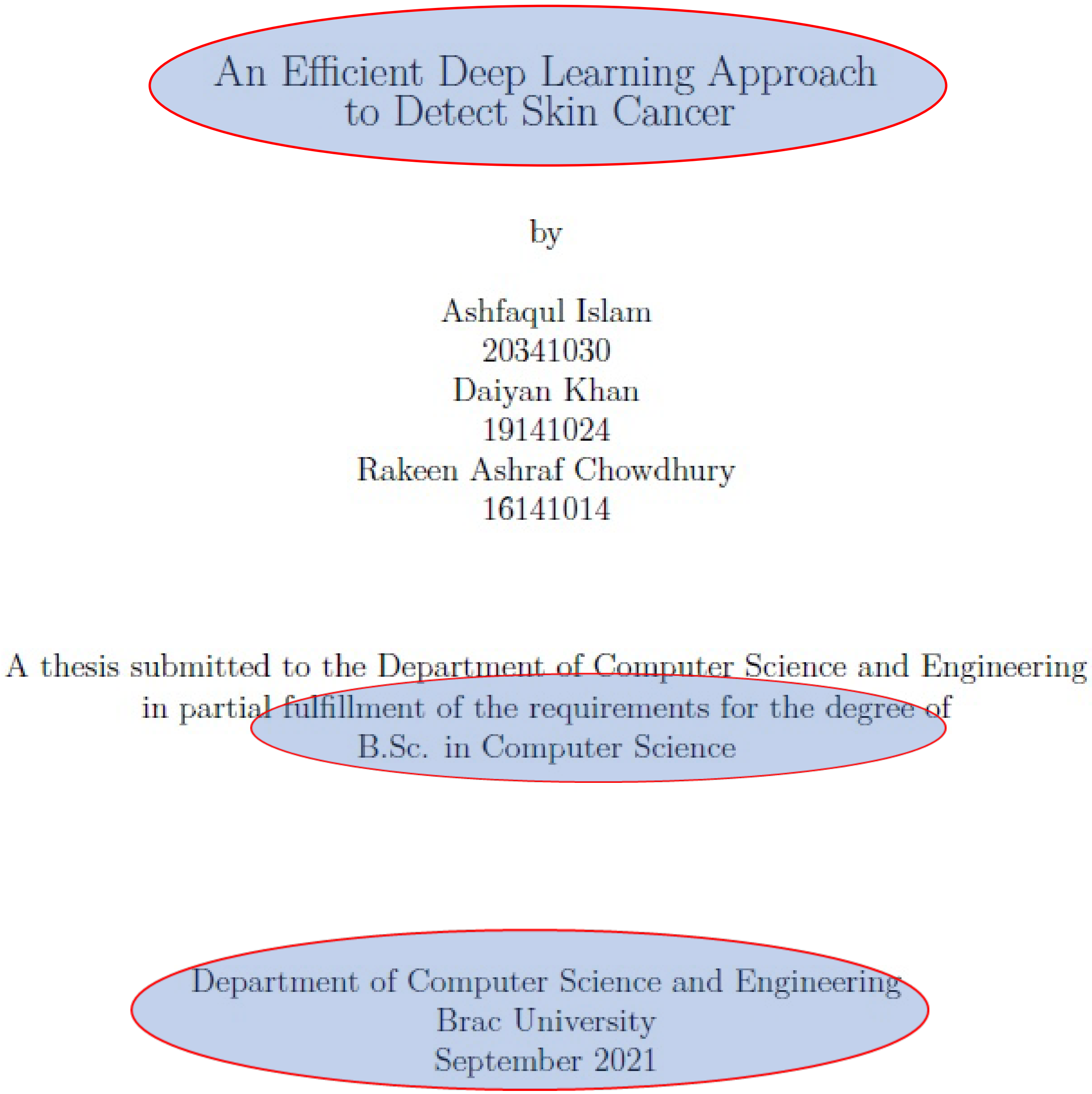
BRAC University Institutional Repository

Digital Collections Repository

Dspace
http://dspace.bracu.ac.bd/xmlui/handle/10361/15932

DSPACE





An Efficient Deep Learning Approach to Detect Skin Cancer

by

Ashfaqul Islam
20341030
Daiyan Khan
19141024
Rakeen Ashraf Chowdhury
16141014

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2021

The Progress of Knowledge

2017 Stanford Nature Deep Learning Cancer ID Article

2018 Viennesse Doctor uploaded Dermatalogical Image Library to **Harvard Dataverse Data repository**

2019-2020 Global Open Science Through Network Possibilities

2021 (November) Dspace Repository Undergraduate Thesis BRAC University, Dhaka Bangladesh, Dept. of Computer Science and Engineering

Downloaded July 2022 Texas, USA

Questions & Comments

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries
ruzwyshyn@txstate.edu
http://rayuzwyshyn.net

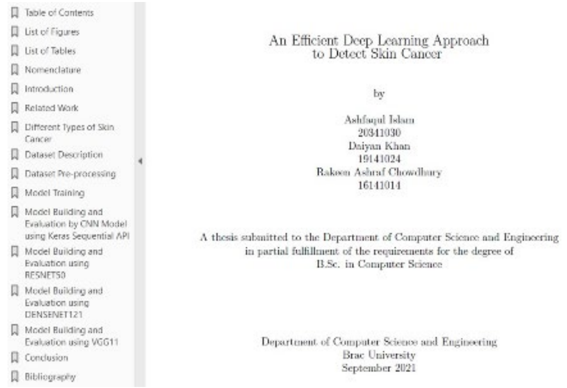
The Progress and Potential of AI, Discovery, Data and Big Data Ecosystems for Libraries and Research Institutions

nature
LESIONS LEARNT
Artificial intelligence powers detection of skin cancer from images
STAR TREK
SAFETY CATCH
MATERIAL GAINS

2017 Stanford
Nature Deep Learning Cancer ID Article

2018 Viennesse Doctor uploaded Dermatalogical Image Library to **Harvard Dataverse Data repository**

2019 Global Open Science Through Network Possiblities



An Efficient Deep Learning Approach
to Detect Skin Cancer

by

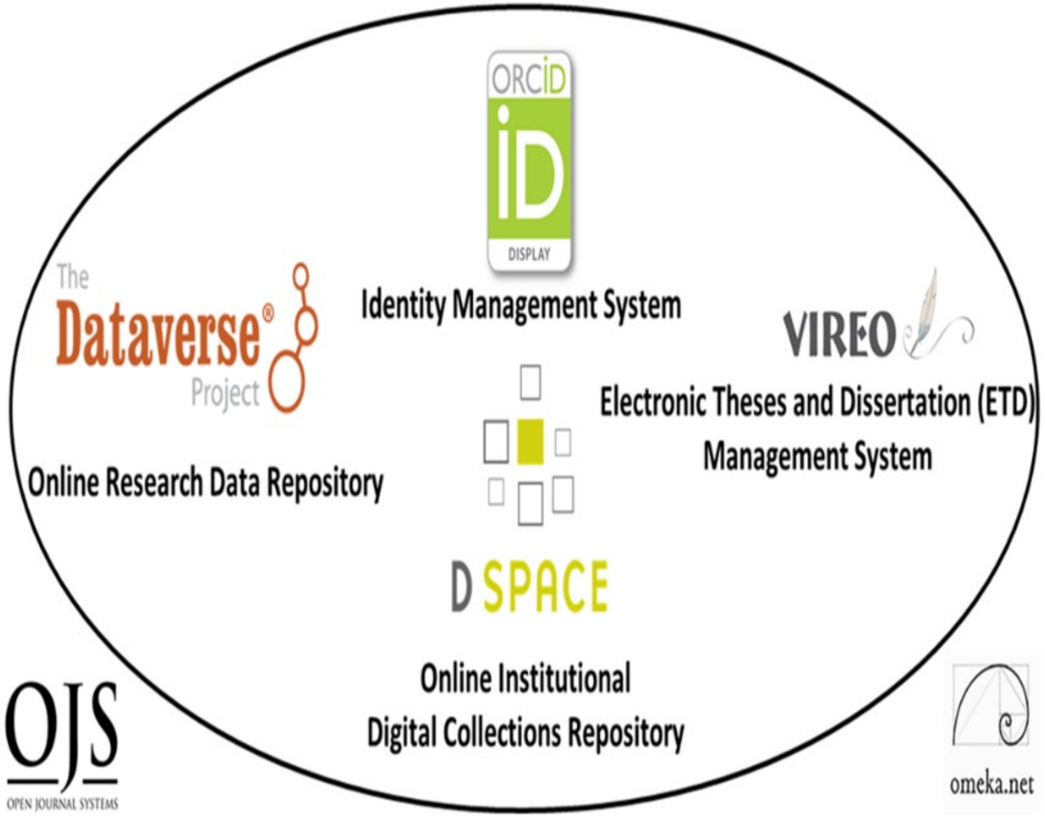
Ashfaqul Islam
20341030
Daiyan Khan
19141024
Rakeen Ashraf Chowdhury
16141014

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2021

2021 (November)
Dspace Repository
Undergraduate Thesis BRAC University, Dhaka Bangladesh, Dept. of Computer Science and Engineering

Downloaded July 2022
Texas, USA



The Dataverse® Project

Online Research Data Repository

ORCiD iD DISPLAY

Identity Management System

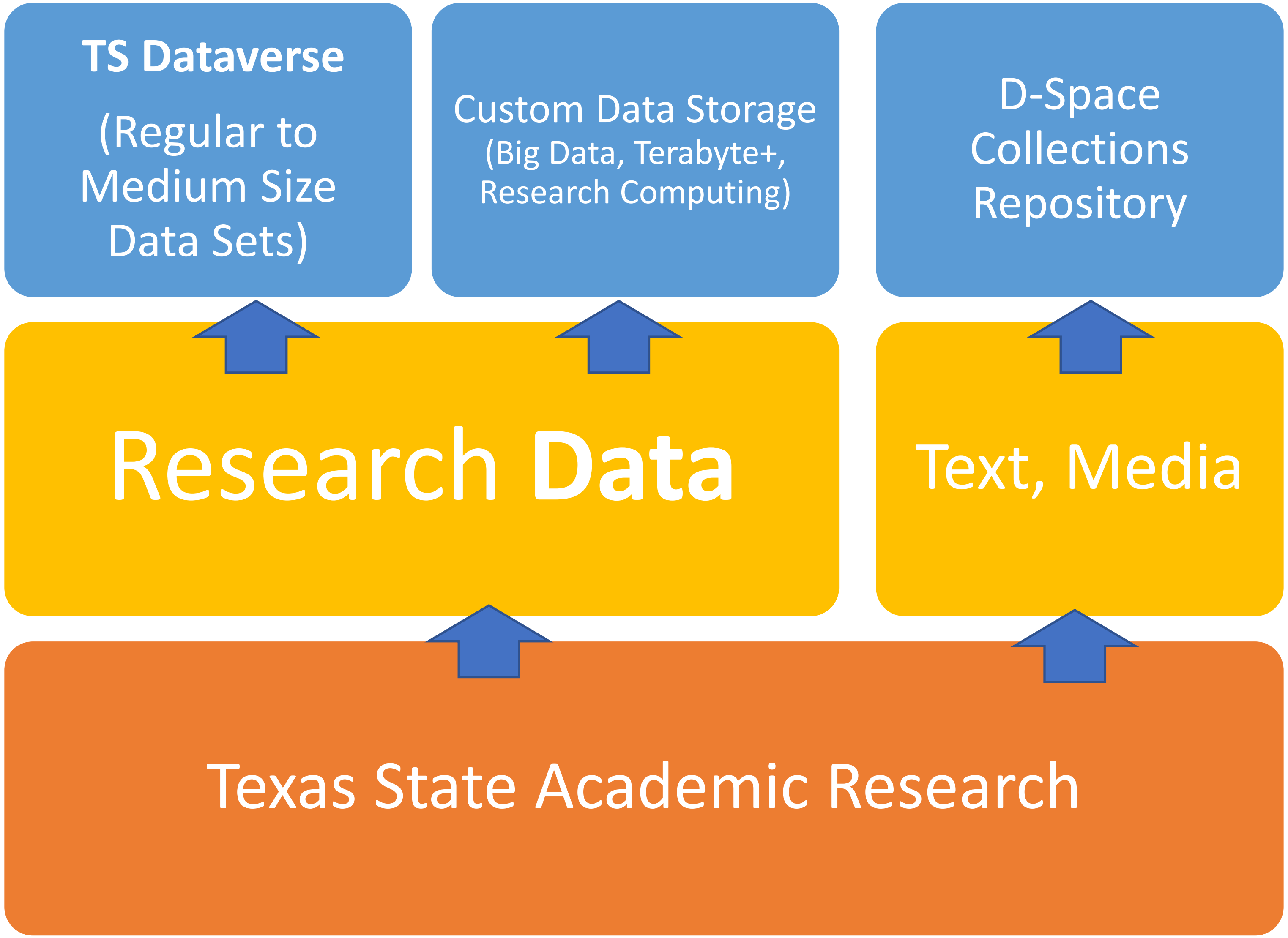
VIREO

Electronic Theses and Dissertation (ETD) Management System

D SPACE

Online Institutional Digital Collections Repository

Texas State Repositories Architecture



TS Dataverse

(Regular to Medium Size Data Sets)

Custom Data Storage (Big Data, Terabyte+, Research Computing)

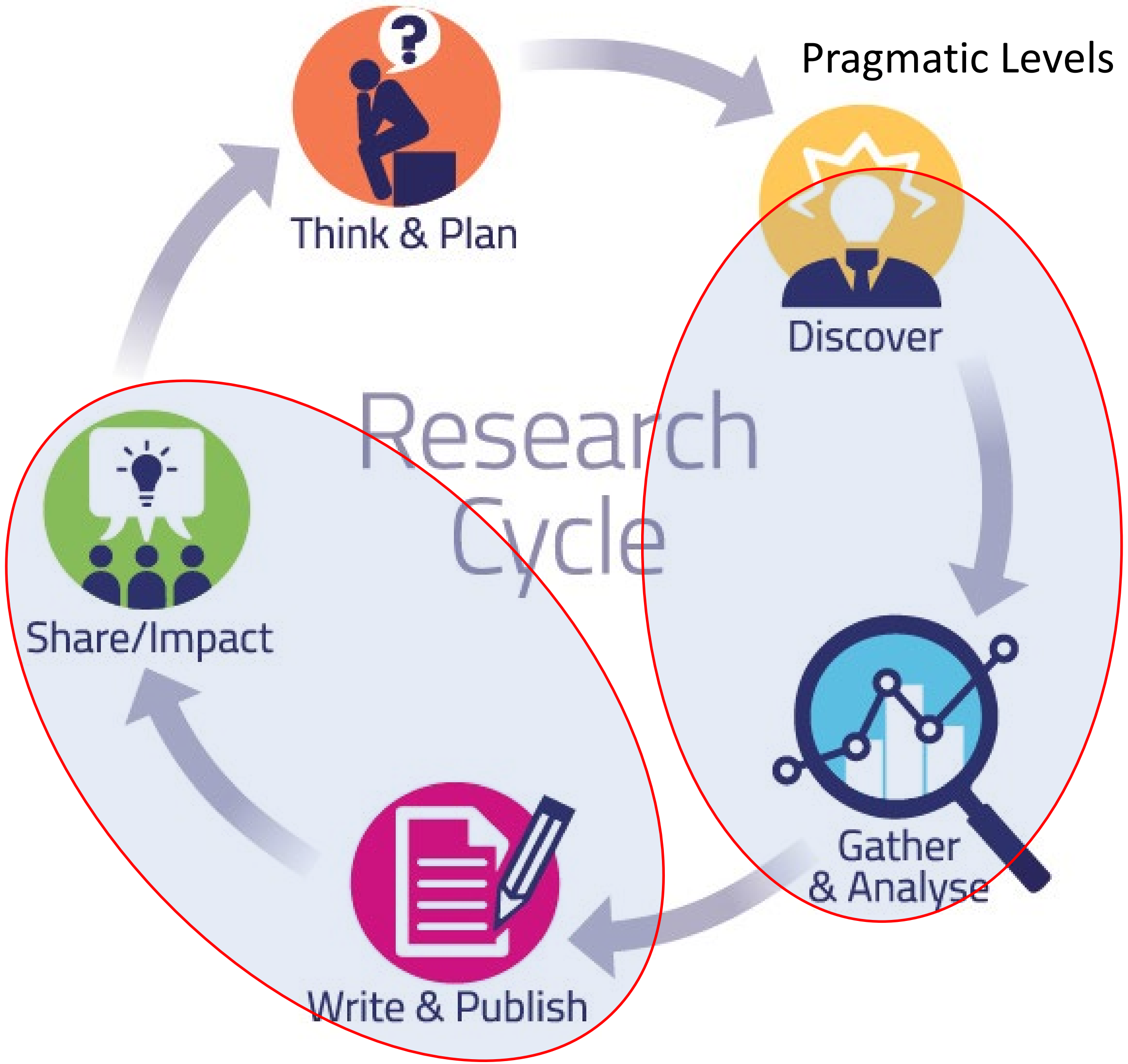
D-Space Collections Repository

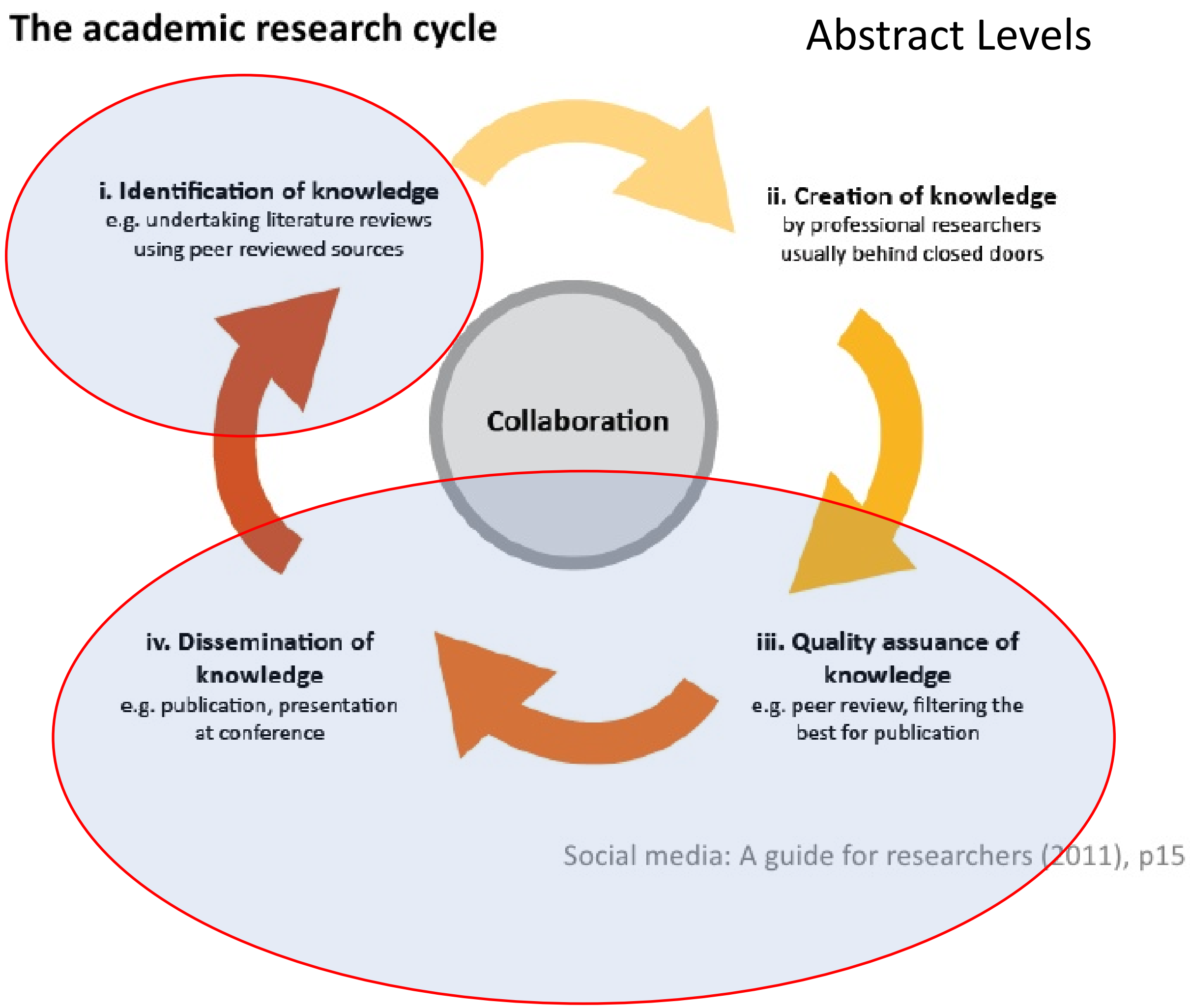
Research **Data**

Text, Media

Texas State Academic Research

Together These Digital Ecosystem Components Enable the Academic Research Cycle

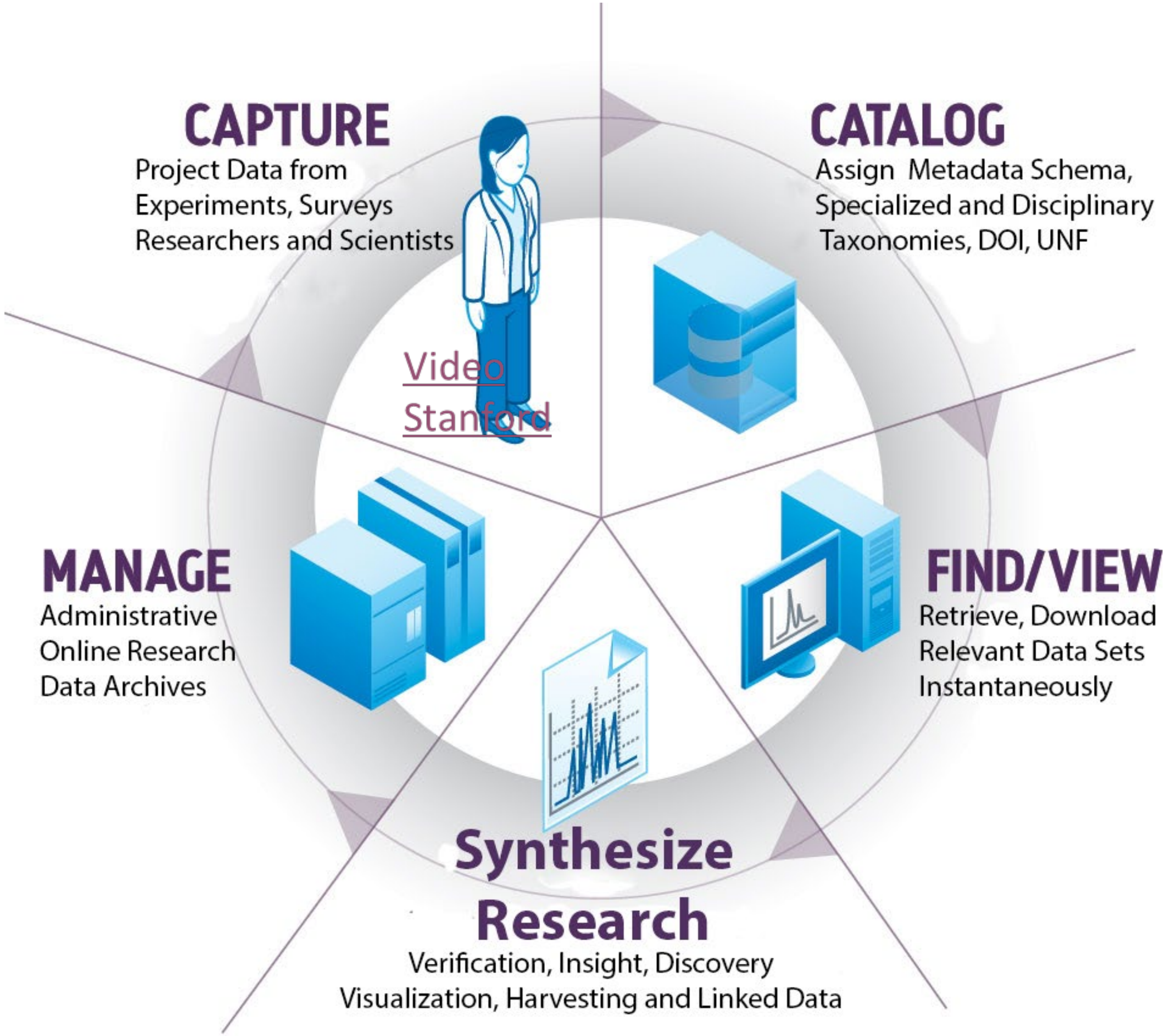



The academic research cycle
Abstract Levels
i. Identification of knowledge
e.g. undertaking literature reviews using peer reviewed sources
ii. Creation of knowledge
by professional researchers usually behind closed doors
Collaboration
iii. Quality assuance of knowledge
e.g. peer review, filtering the best for publication
iv. Dissemination of knowledge
e.g. publication, presentation at conference

Social media: A guide for researchers (2011), p15

The Research Data Repository Lifecycle

Setting Better Foundations & Organization for AI Infrastructures



CAPTURE
Project Data from
Experiments, Surveys
Researchers and Scientists

CATALOG
Assign Metadata Schema,
Specialized and Disciplinary
Taxonomies, DOI, UNF

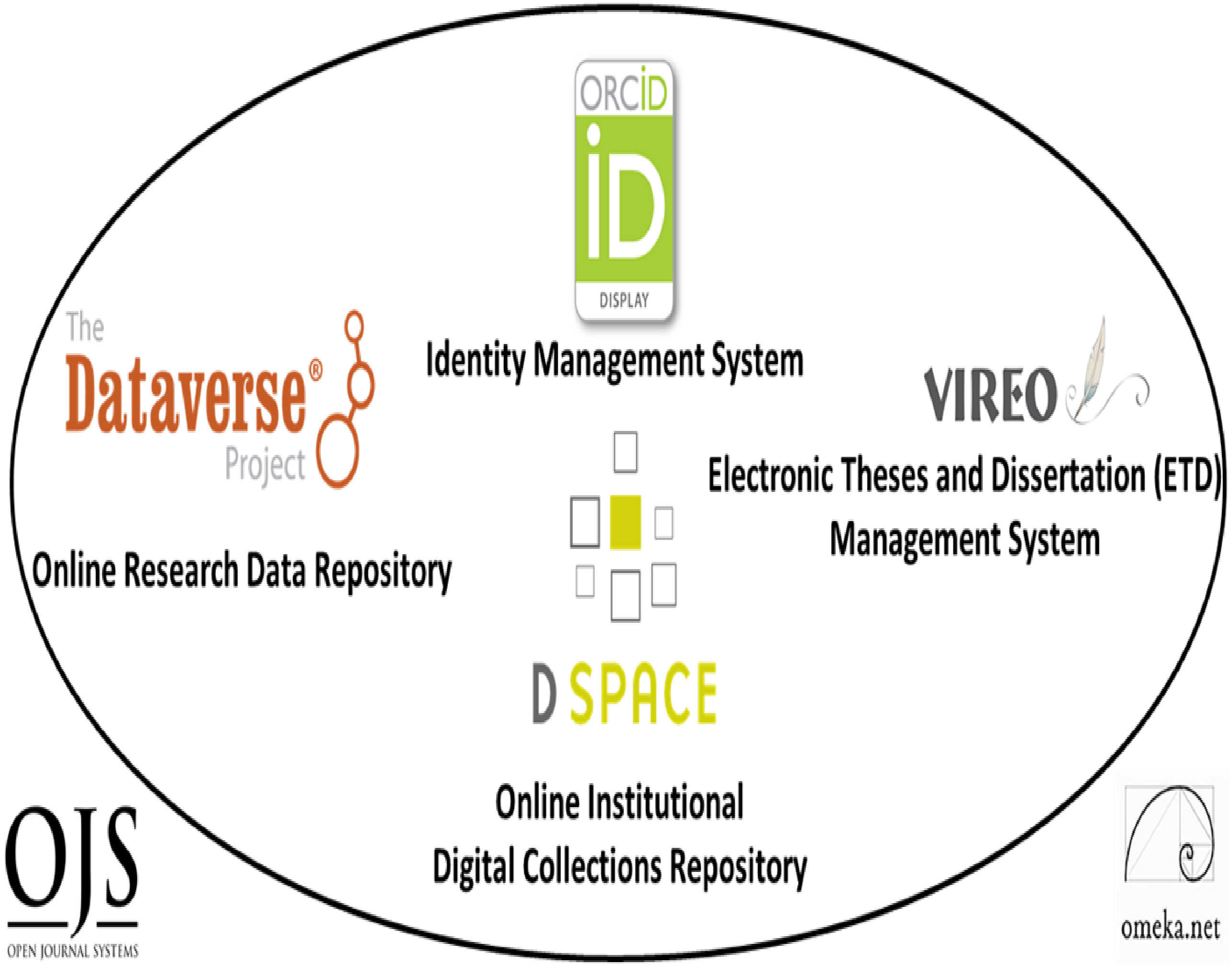
FIND/VIEW
Retrieve, Download
Relevant Data Sets
Instantaneously

Synthesize Research
Verification, Insight, Discovery
Visualization, Harvesting and Linked Data

MANAGE
Administrative
Online Research
Data Archives

Video Stanford

Digital Scholarship Ecosystem Centered on Research Data Repository and Collections Repository

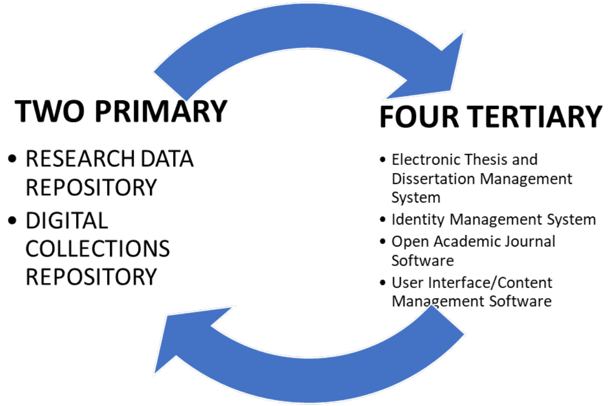


Identity Management System

Online Research Data Repository

Electronic Theses and Dissertation (ETD) Management System

Online Institutional Digital Collections Repository



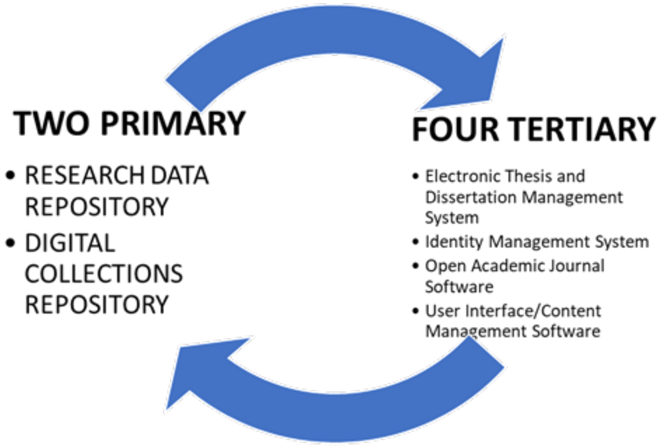
TWO PRIMARY

- RESEARCH DATA REPOSITORY
- DIGITAL COLLECTIONS REPOSITORY

FOUR TERTIARY

- Electronic Thesis and Dissertation Management System
- Identity Management System
- Open Academic Journal Software
- User Interface/Content Management Software

Questions Comments



TWO PRIMARY

- RESEARCH DATA REPOSITORY
- DIGITAL COLLECTIONS REPOSITORY

FOUR TERTIARY

- Electronic Thesis and Dissertation Management System
- Identity Management System
- Open Academic Journal Software
- User Interface/Content Management Software

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries
ruzwyshyn@txstate.edu
http://rayuzwyshyn.net