

Open Access Data Research Repositories: From Data and Research Ecosystems to Artificial Intelligence and New Discovery

Raymond Uzwyshyn, Ph.D. MBA MLIS

Associate Dean, Library Collections and Strategy

Mississippi State University Libraries, USA

E-mail address: ruzwyshyn@gmail.com



Copyright © 2022 by Raymond Uzwyshyn. This work is made available under the terms of the Creative Commons Attribution 4.0 International License:

<http://creativecommons.org/licenses/by/4.0>

ABSTRACT:

Data research repositories allow sharing and archiving of research data for global research. Libraries open this sharing of data to modern metadata and interoperability for search, retrieval, and larger possibilities of global scholarly research ecosystems. Data research repositories are being leveraged to accelerate global research, promote international collaboration, and innovate on levels previously thought impossible. They link data to further content from online publications to multimedia digital communication and aggregation tools.

This article pragmatically overviews a data and content-centered ecosystem and then discusses the ecosystem's next level of possibilities. This involves questions of big data and AI infrastructures for enabling researchers towards Deep Learning (Neural Net) possibilities. These new areas show large promise in making good use of online open data repositories, digital library ecosystems and online datasets

Recent AI research also highlights the utility of several available online open-source digital library data repository and ecosystem components. An online data-centered research ecosystem accelerates open science, research and discovery on global levels. This open-source ecosystem and software infrastructure may be easily replicated by research institutions and universities globally.

Keywords: Research Libraries, Artificial Intelligence, Neural Nets, Academic Libraries, Big Data, Data Research Repositories, Open Source Research Ecosystems

Search the Texas Data Repository

 FIND

Add a Dataset



Create a Dataverse



Explore Data
Repository



Learn More



Get Help

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



Texas Data Repository, <http://data.tdl.org>

1 INTRODUCTION

This research overviews necessary infrastructures for an online research data repository and digital scholarly ecosystem. Current possibilities of online data allow discovery within Artificial Intelligence, particularly Deep Learning and Neural Nets. New potential for open science is enabled by global networks, Artificial Intelligence, online data research repositories and increasing computing processing power and storage. Online data research repositories and scholarly research ecosystem are overviewed. Examples are then utilized to show how these new infrastructures may be used to enable new potential for AI for scientific discovery in the 21st century.

2 WHAT IS AN ONLINE DATA RESEARCH REPOSITORY?

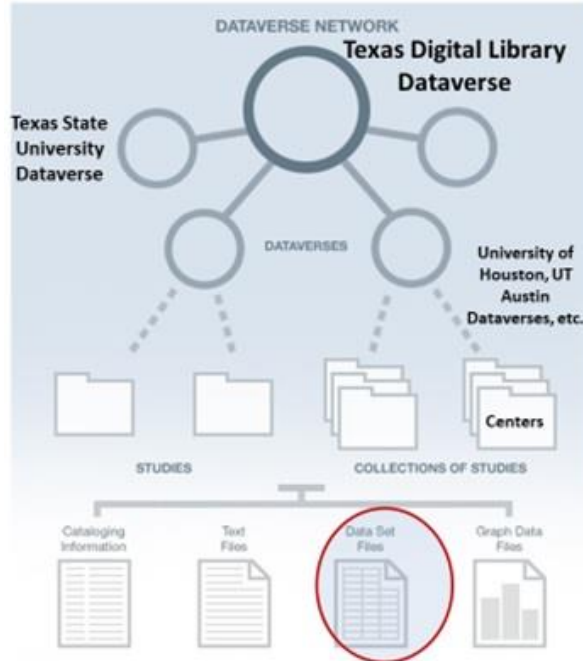
An online data research repository allows one to share, publish and archive a researcher's data. It is a platform to manage a researcher's and institution's data and metadata, a permalinking strategy for Data Citation, a way to manage grant compliance and a data archiving and sharing strategy.

The screenshot shows the homepage of the Texas Digital Library Test Dataverse. At the top left is a blue star logo with a white lightning bolt. To its right is the text "Texas Digital Library Test Dataverse" and "A statewide collaboration of higher education institutions in Texas". Below the logo is a "Metrics" bar showing "26 Downloads". On the right side, there are two buttons: an envelope icon and a share icon. The main heading reads "Share, publish, and archive your data. Find and cite data across all research fields." Below this is a welcome message: "Welcome to the Texas Digital Library Test Dataverse!" followed by an important notice: "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." A horizontal carousel of four university logos is displayed: Trinity University, UT Medical Branch, University of Texas, and Texas State University. Below the carousel is a search bar with a "Find" button and a link to "Advanced Search". On the far right is a "+ Add Data" button.

Texas Data Research Repository, <https://dataverse.tdl.org>

The Texas Data Repository is a good example of a consortial data repository. It utilizes Harvard's open source Dataverse software customized towards a consortial multi-university strategy.¹ The Texas Data Repository aggregates individual university's data for search and retrieval. It can be configured as a single instance for searching or to search across an entire group of institutions.

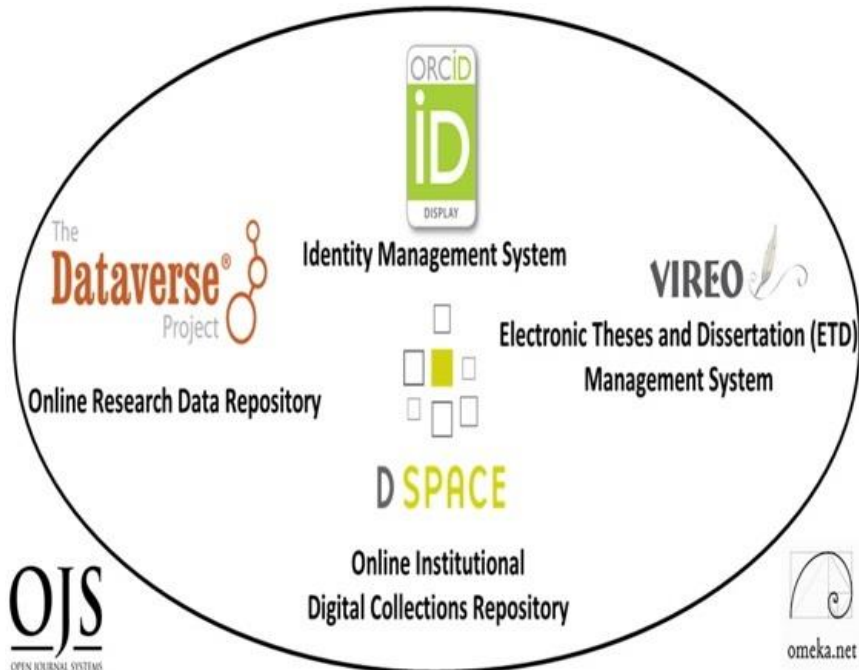
¹ See Uzwyshyn, Online Data Repositories (2016).
<https://www.researchgate.net/publication/304780954> Online Research Data Repositories the What When Why and How



Texas Data Repository Consortial Architecture

3 DIGITAL SCHOLARSHIP ECOSYSTEMS

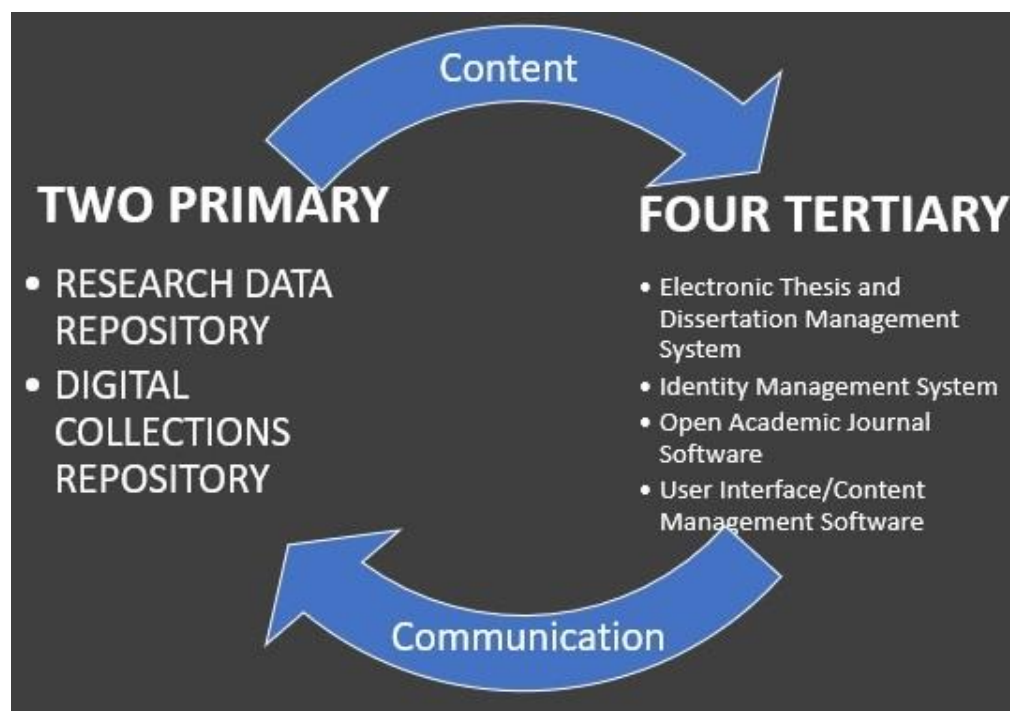
A Digital Repository may also be placed in a larger digital scholarship ecosystem which enables a wide horizon of content and global network communication.



The Texas State Digital Scholarship Research Ecosystem Consists of Six Components

The Texas State Digital Scholarship ecosystem utilizes the well-known open-source repository software, Dspace, for the university's digital collections repository. Four other tertiary components are also utilized by researchers to better enable global communication and network possibilities. These four applications are: an online electronic theses and dissertation management system, ETD System (VIREO), identity management system (ORCID), open academic journal system software (OJS3) and user interface content management software (OMEKA). Together, these function as a digital scholarship ecosystem.²

This ecosystem allows for facility in enabling data-centered methodologies. It builds strong foundations and provides foundational training data for later needed AI pathways.



A Digital Scholarship Research Ecosystem is enabled by both Content and Communication Components

The general characteristics for such a digital system are open-source software, active developer communities, communication and content repository components. The open-source software allows customizability and connection between components. Active developer communities enable a lively exchange of new possibility with regards to innovation. Open-source code allows bridges among systems. The sum of the system's capabilities exceeds separate parts. Collocating open-source digital components in a networked research ecosystem enables connections, network effects and untapped possibilities.

Together, these digital ecosystem components enable the academic research cycle. This cycle moves from original search and retrieval of data and content to gathering and analysis, to later writing, publishing and sharing online.

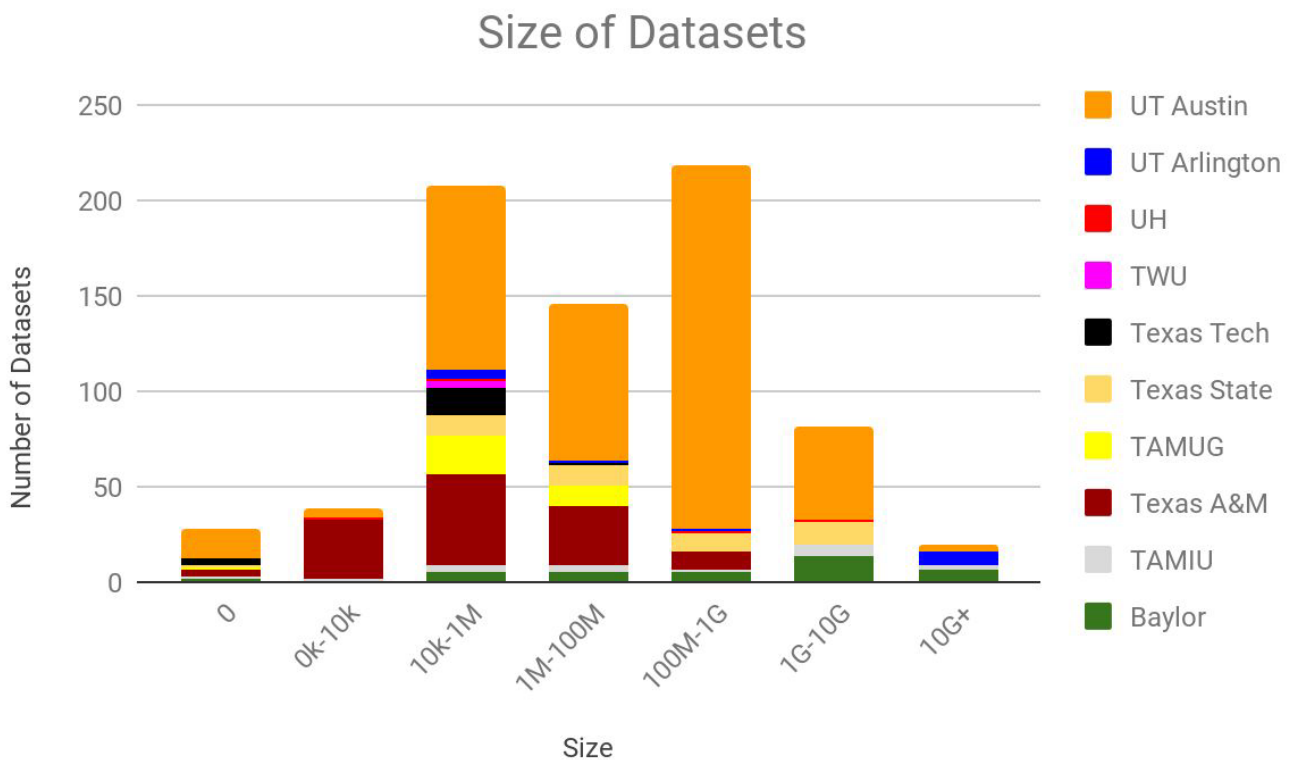
² See Uzwysyn, 2020. Available at: https://www.researchgate.net/publication/336923249_Developing_an_Open_Source_Digital_Scholarship_Ecosystem

4 DATA, DATASETS, BIG DATA

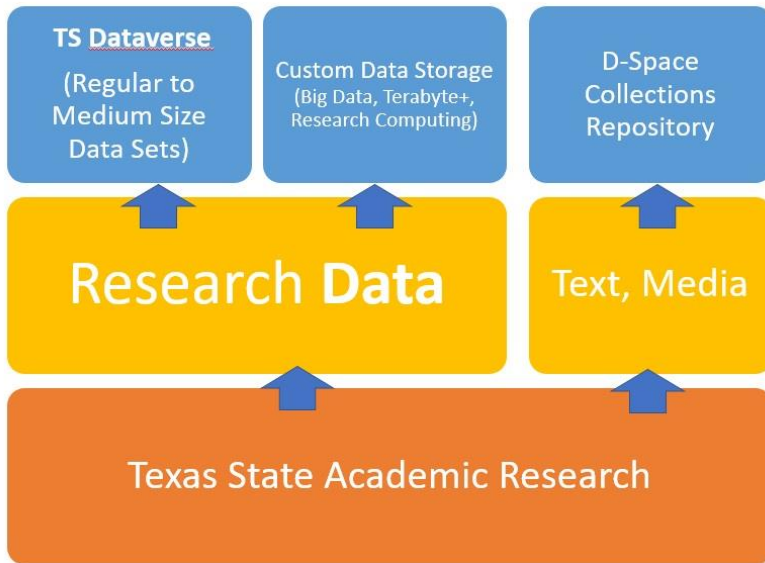
Data comes in a variety of file types, formats, media, and sizes. For AI and particularly recent Deep Learning, labelled and unlabelled datasets become important for machine training. Within information science, metadata is key. One size also does not fit all for various data research repository project needs. There are many types of sizes for data projects and repositories. The Texas Data Repository utilizing Dataverse can upload currently up to 4GB data for individual files and 10GB Datasets. This may not seem large but serve the needs of most academic researchers and have served researchers well for the last five years (2017-2022) and present but will need to expand accordingly.

Sizes of Texas Data Research Repository Datasets (See Waugh, 2020)

Most researchers' collected datasets for upload are presently in the 1 < <1000 MB range.



Currently, there is the growing recognition by researchers that 'bigger' data repositories are needed. These begin in the GB/TB ranges. For larger datasets, these may be placed with university research computing data centers or the local area supercomputing center for custom data storage should these needs arise. This type of storage is usually worked out by researchers in preliminary grant applications expecting this level of data storage needed for research work.

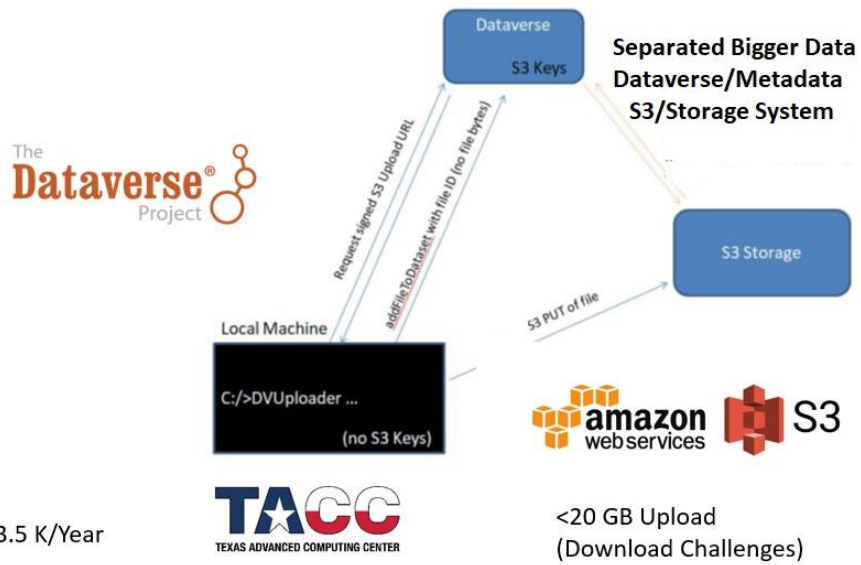


Texas State Universities Big Data Storage Model

Beyond isolated custom big data storage needs the requests for very ‘big data’ (Terabytes, Exabyte storage) are few but requests are increasing. Bigger data options and beta prototypes for these models may be currently explored. This ranges from 20GB expansions (Amazon Web Services S3 storage) and Advanced Supercomputing centers with separated metadata/storage pointer systems to more fee-based institutional models up to 300GB/dataset (Data Dryad).



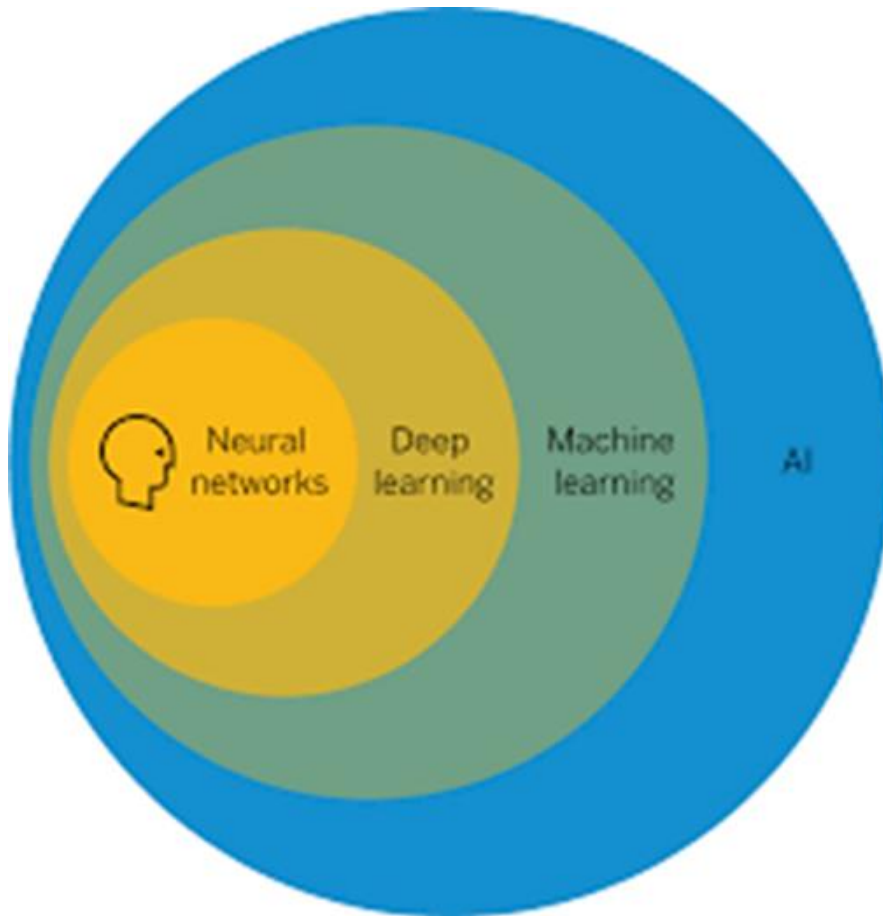
Up to 300 GB/dataset
 Fee Based Institutional Model 7.5/13.5 K/Year



<20 GB Upload
 (Download Challenges)

Beta Prototyping Big Data Texas Data Repository Architectures. 2020-2022, TACC: <https://www.tacc.utexas.edu/>, Data Dryad <https://datadryad.org/stash>

5 DATA RESEARCH REPOSITORIES, DIGITAL ECOSYSTEMS AND AI



Relationships among AI and subdomains of Machine Learning, Deep Learning and Neural Nets

The last five years (2017-2022) have shown incredible progress and gains in analytic computation tools and discovery, particularly those methodologies associated with new domains of Artificial Intelligence. Machine learning, deep learning and neural net scientific research has shown incredible potential for scientific breakthrough. This ranges from Computer Vision (Facial/Object Recognition), Natural Language Processing (speech recognition, translation), Cybersecurity (Fraud Detection) Conversational Chatbots and Strategic Reasoning (Game Theory). Breakthroughs have been enabled through a fortuitous combination of better algorithms, greater computing processing, metadata enabled online datasets and, open-source digital libraries, specifically research data repositories and ecosystems.

6 DIGITAL LIBRARIES IMAGE DATA REPOSITORIES & AI

In 2017, an innovative new cancer detection methodology was published in Nature by a Stanford University group proposing the use of Neural Nets (Esteva, Nature, 2017). The AI neural network was trained on big data and a dataset of 129,460 images of 2,032 diseases and larger dataset of AI training images (1.41 million) to classify skin cancer lesions with deep neural networks. After comparison, the neural net machine learning AI did equal to or better than 30 board certified dermatologists with decades of experience.

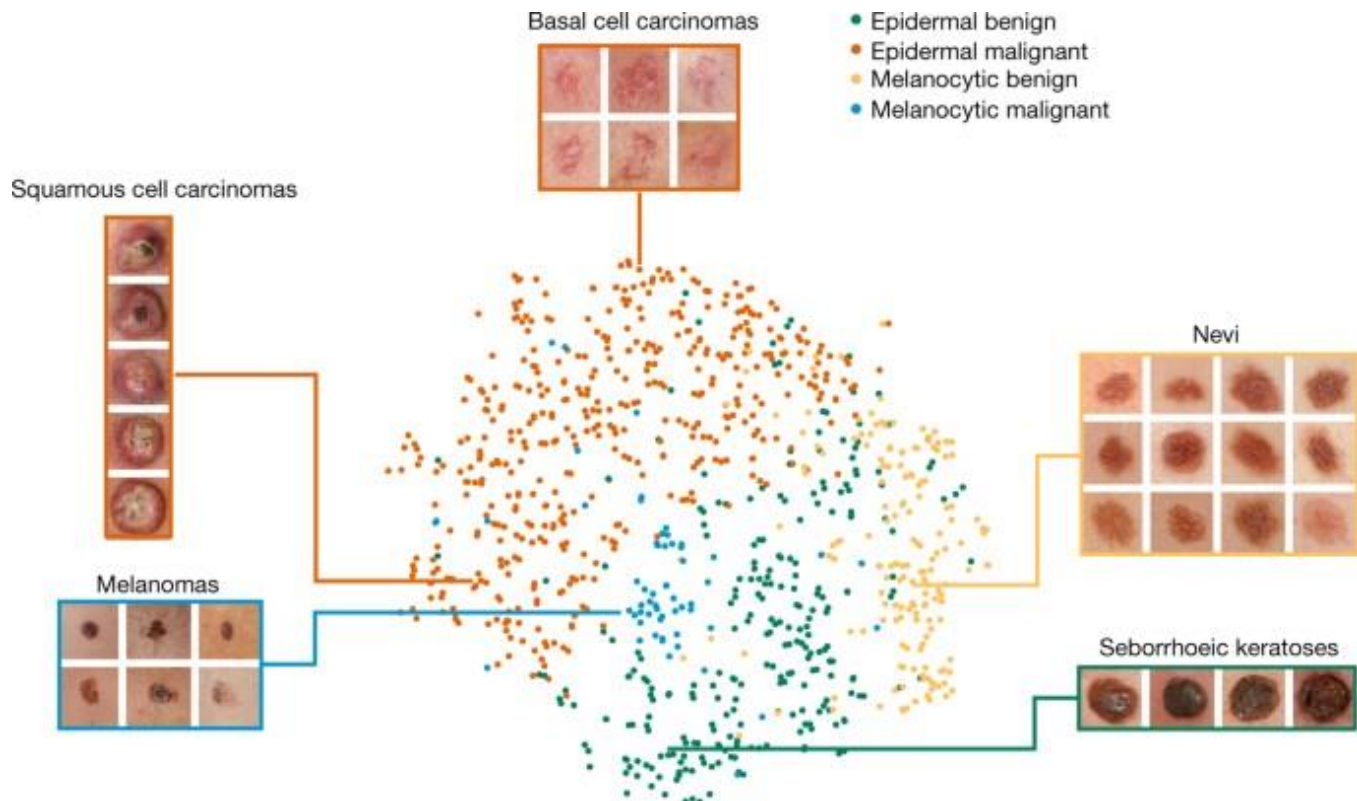


Image from Dermatologist Level Classification of Skin Cancer with Deep Neural Nets (Esteva et al, 2017) ³

The neural net here was able to classify epidermal lesions for early cancer detection into benign and cancerous (malignant) lesions better than 30-year board certified dermatologists. This method involved pixel-level differentiation and training through a multi-level neural net AI model. The large relevance of the digital image data repositories and digital libraries for initial training and metadata labelling should not be underestimated for researchers. In a recent article on Deep Learning in Cancer Pathology Surrounding a New Generation of Clinical Biomarkers (Echle, 2020), the authors emphasize the primary need for well organized digital libraries, data repositories, dataset preparation and metadata preprocessing for fundamental accuracy in training, testing, and external neural net AI validation.

³ See also, the original article from Nature. Esteva, A, Thrun, S. et al. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature, Volume 542 (February 2, 2017). pp. 115-119. doi:10.1038/nature21056 and Eschle, 2020.

7 OPEN SCIENCE, AI AND DATA-CENTERED ECOSYSTEMS

Harvard's open source Dataverse software allows for the uploading of datasets from other universities globally. Appropriate research datasets may be uploaded for sharing or use by researchers anywhere. A university can also mount its own instance of Dataverse, and make use of the software to share academic research data. As mentioned, Dataverse is open source software and any research level libraries, institution and university should be encouraged to be setting up their own instances of data repository and digital ecosystems.

To trace an example of how this is used, the HAM10,000 image dataset below is a large collection of multi-source dermatoscopic images of cancerous skin lesions. This dataset was uploaded to Dataverse by Viennese Dermatologist, Dr. Philip Tschandl, in 2018, a year after the preceding mentioned Stanford Nature Neural Net algorithmic methodology article appeared.

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
Contact Owner

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 lentiginos / seborrheic keratoses and lichen-planus like keratoses, *lck*), dermatofibroma (*df*), melanoma (*mel*), melanocytic nevi (*nv*) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, *vasc*).

HAM10000 Dataset in Dataverse Data Research Repository,
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>

As can be seen in the figures (HAM10000 Dataset and Images), the images, data and metadata can be easily downloaded, unzipped, and used by other researchers globally for neural net training purposes.







Files
Metadata
Terms
Versions

Q

Filter by

File Type: All ▾ Access: All ▾

1 to 6 of 6 Files

| | | |
|--------------------------|------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/> |  | <p>HAM10000_images_part_1.zip</p> <p>ZIP Archive - 1.3 GB</p> <p>Published Jun 4, 2018</p> <p>15,709 Downloads</p> <p>MD5: 463...e46 </p> |
| <input type="checkbox"/> |  | <p>HAM10000_images_part_2.zip</p> <p>ZIP Archive - 1.3 GB</p> <p>Published Jun 4, 2018</p> <p>12,022 Downloads</p> <p>MD5: da4...84b </p> |
| <input type="checkbox"/> |  | <p>HAM10000_metadata.tab</p> <p>Tabular Data - 810.9 KB</p> <p>Published Jan 29, 2021</p> <p>6,203 Downloads</p> <p>8 Variables, 10015 Observations UNF:6:WcXi...myQ== </p> |

HAM10000 Dermoscopic Cancer Images, Harvard Dataverse Repository,
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>

Below is a cover page from BRAC University from Dhaka Bangladesh that uses DSpace as an institutional repository to house theses and dissertations from the School of Data and Sciences, Dept. of Computer Science and Engineering. Here, the computer science and engineering students had earlier downloaded Dr. Tschandl's uploaded dermatological cancer training images, metadata and datasets. They utilized the labelled image metadata as training material to train a deep learning neural net algorithm. The model was able to recognize cancer growths with efficiency greater than, or equal to the 2017 board certified dermatologists for mobile devices. The example is very interesting for possibilities of telemedicine and the progress of open science in global populations and universities which may not have as quick access to trained specialists as those in the West.

This is a particularly good example of open science and AI possibilities operating on global institutional levels. This is occurring through the enabling power of digital scholarship ecosystems, digital libraries and data repositories. Content and specialized image data sets, with highly specialized labelled metadata that otherwise would be unavailable, are brought together with new

machine learning algorithmic techniques. New research and an exceptionally good thesis has been produced. Globally dispersed content and data, from three different continents, is aggregated instantly to advance the pursuit of knowledge and science with a speed and utility that would be unimaginable in other eras or centuries.



Institutional Repository

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

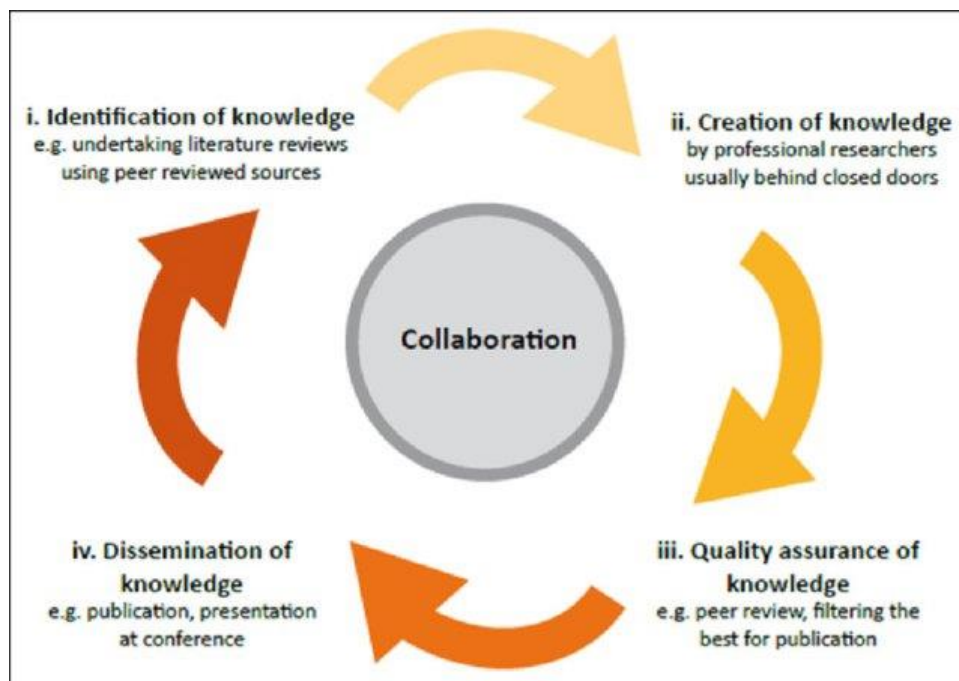
LC Subject Headings

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.

8 CONCLUSIONS – AI, DATA AND THE ACADEMIC RESEARCH CYCLE



The Academic Research Cycle, Cann, Dimitriou and Hooley, 2011.

New data repository, digital library and digital scholarly ecosystem possibilities are enabling the academic research cycle and progress of knowledge and discovery in our new millennia in amazing ways. Research libraries stand at the center of this new revolution. Open science possibilities empower a new global networked generation towards incredible new open science and knowledge discovery and creation. This can all occur through the enabling power of data, data research repositories, open access and digital scholarly ecosystems now possible for research library ecosystems.

REFERENCES

Artificial Intelligence. Machine Learning. Neural Networks. Future Technology. Bloomberg Businessweek Canada. 2022. <https://www.youtube.com/watch?v=ypVHymY715M>

Cann, A., Dimitriou, K. Hooley, T. 2011. *Social Media: A Guide for Researchers.* Research Information Network. University of Derby, UK.

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

ColdFusion (2018). *Why Deep Learning Now?* (Documentary Overview).
https://www.youtube.com/watch?v=b3IyDnB_ciI

Echle et al. Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. *British Journal of Cancer.* November 2020. <https://www.nature.com/articles/s41416-020-01122-x>

Esteva, A, Thrun, S. et al. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, Volume 542 (February 2, 2017). pp. 115-119. doi:10.1038/nature21056

Fridman, Lev. *MIT Deep Learning and Artificial Intelligence Lectures.*
<https://deeplearning.mit.edu/> 2022.

Islam, A., Khan, D. and Chowdhury, R. 2021. *An Efficient Deep Learning Approach to Detect Skin Cancer Undergraduate Thesis.* BRAC University DSpace Institutional Repository, 2021. Available: <http://dspace.bracu.ac.bd/xmlui/handle/10361/15932>

Kleinveltdt, Lynn. Smarter high education learning environments through AI: What this means for academic libraries. *Trends and Issues in Library Technology: Special Issue on AI:* June 2022. pp. 12-15. <https://repository.ifla.org/handle/123456789/1940>

Mitchell, Tom. 2022 *Where on Earth is AI Headed?* Carnegie Mellon.
<https://www.youtube.com/watch?v=ij9vqTb8Rjc>

Peters, T. and Waugh, L. Larger Data Storage Report: Research Data Management Initiatives and Planning, January 2022. Texas State University Libraries (Unpublished White Paper)

Texas Data Repository 2022. <https://dataverse.tdl.org/>

Tschandl, Phillip et al. *Human-computer Collaboration for Skin Cancer Recognition.* *Nature Medicine*, 22 June 2020, 1229-1234. See: <https://www.nature.com/articles/s41591-020-0942-0>.

Uzwyshyn, R. 2022. Steps Towards Building Library AI Infrastructures: Research Data Repositories, Scholarly Research Ecosystems and AI Scaffolding. *New Horizons in Artificial Intelligence in Libraries* (IFLA Satellite Conference), National University of Ireland, Galway, IR.

Uzwyshyn, R. 2021. Frameworks for Long Term Digital Preservation Infrastructures. *Computers in Libraries*. September 2021. pp.4-8.

Uzwyshyn, R. 2020. *Developing an Open-Source Digital Scholarship Ecosystem*. ICEIT2020. St. Anne's College Oxford, United Kingdom. February 2020. Available at: https://www.researchgate.net/publication/336923249_Developing_an_Open_Source_Digital_Scholarship_Ecosystem.

---. *Open Digital Research Ecosystems: How to Build Them and Why*. *Computers in Libraries*, (40) 8. November 2020. https://www.researchgate.net/publication/345956074_Online_Digital_Research_Ecosystems_How_to_Build_Them_and_Why

---. Online Research Data Repositories: The What, When Why and How. *Computers in Libraries*. 36:3, April 2016. pp. 18-21.
<http://rayuzwyshyn.net/TXU2016/OnlineDataResearchRepositoriesUzwyshyn.pdf>

Waugh, L. *Texas State University Annual Usage Report 2020*. TXST Dataverse Repository. Texas Conference on Digital Libraries Presentation. Texas State University.