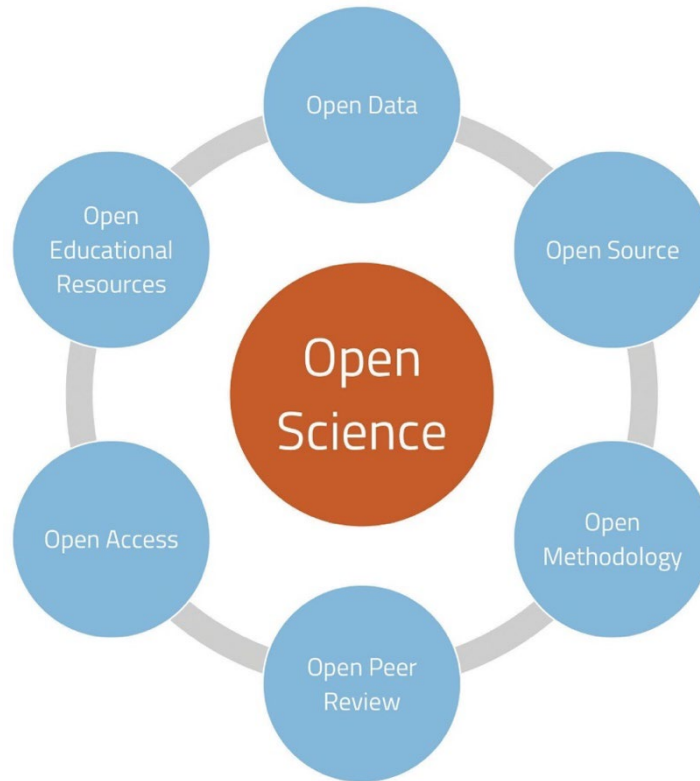


From Open Science and Datasets to AI and Discovery

Dr. Raymond Uzwyshyn, rju13@msstate.edu

Associate Dean, Research Collections

Mississippi State University Libraries, US



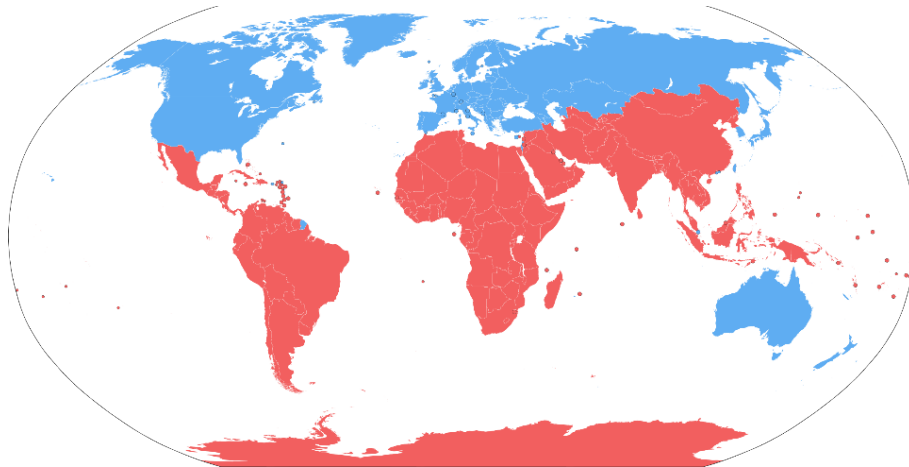
The Many Facets of Open Science Digital Library Ecosystems

Introduction

Exciting new online library infrastructures are now available for open science and experimental research data. These possibilities are being enabled by online data repositories, research ecosystems and artificial intelligence discovery. This research overviews and outlines these recommended new infrastructures focusing upon online research data repositories and open science scholarly ecosystems. Together, these better orient scientific research towards discovery and new possibilities.

This paper utilizes a university library research data ecosystem to overview the utility of placing an online research data repository within an open science research ecosystem framework. Specific examples of online data and open discovery within Artificial Intelligence from Deep Learning are provided. Two online medical related image dataset library examples foreground the importance of online data repository ecosystems towards open science within digital library

ecosystems, and particularly, Artificial Intelligence and new discovery. The first example is from a recent discovery (2017) from scientific deep learning neural net towards cancer detection. The example utilizes object recognition and big data for machine learning and neural net training from a US university (Stanford). The second example (2022) builds on the first model's methodology utilizing an undergraduate student thesis from BRAC University from Dhaka, Bangladesh. Stanford's earlier AI neural net model and enabling affinities with geographically dispersed ecosystem methodologies open AI neural net research with online available datasets. Both examples give compelling evidence illustrating the value of open science. Online open data research repositories within data-centered scholarly ecosystems enable the future progress of science and discovery in our new millennia. They bridge collaboration possibilities among traditional global north and south divides.



World Map Showing Traditional North-South Global Divide

These new potentials for open science are enabled by: recent constellations of global networks, powerful new algorithmic AI Neural Network Deep Learning models and online storage and retrieval capacity of data research repositories. This increasing computing processing power pragmatically enables these paths. Rudiments of an online data research repository and open science research ecosystem are outlined. Easily implementable digital library examples are utilized. These examples show how these new infrastructures may be used to enable future AI methodologies for scientific discovery in the 21st century. While this is not

the only utility of these ecosystems and repositories, it is an important pathway which shows large promise.

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 at once a platform to manage a researcher's and institution's data and metadata. It is also a perma-linking strategy for Data Citation, a way to manage mandated large grant compliance, and an efficacious global data archiving and sharing strategy.



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
Trinity University Dataverse



utmb Health
Working together to save lives™
UT Medical Branch Dataverse



TEXAS
University of Texas Dataverse



TEXAS
STATE
UNIVERSITY
Texas State University Dataverse

>

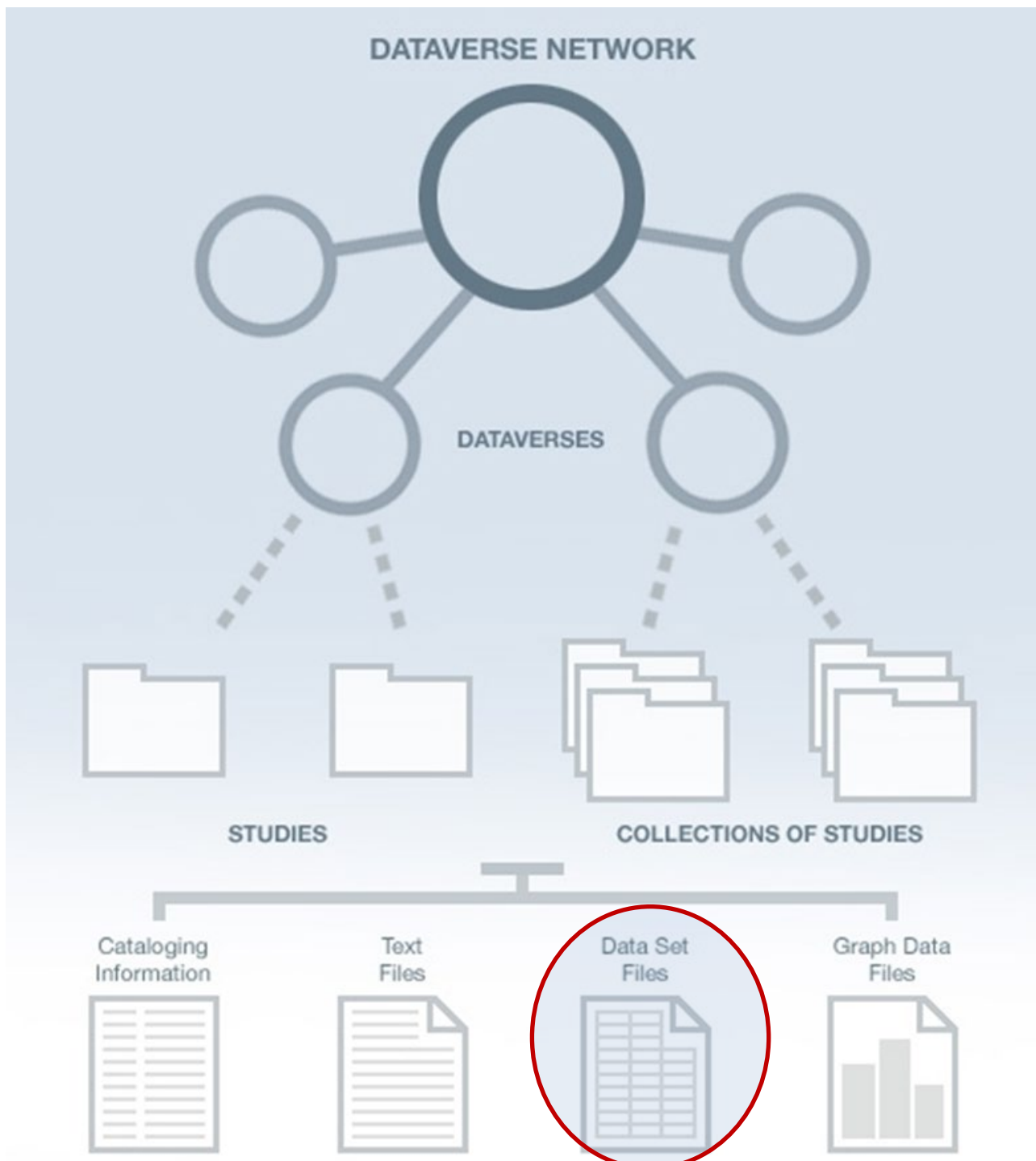
🔍 Find [Advanced Search](#)

+ Add Data

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

The Texas Data Repository is a good example of an online data repository. It utilizes Harvard's open source Dataverse software, customized towards a multi-university strategy.¹ This Data Repository aggregates various 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. The repository can also easily be configured on consortial, state, or international levels.



Texas Data Repository Consortial Architecture

¹ See Uzwyshyn, Online Data Repositories (2016).

Digital Scholarship Ecosystems

A digital repository may also be placed within a larger digital scholarship ecosystem. This enables a horizon of content and further global network communication. The prototypical digital scholarship ecosystem utilizes well-known open-source digital repository software (i.e., Dspace, etc.), for the university's digital collections repository. Four other tertiary components are then utilized by researchers to better enable online 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 unified digital scholarship ecosystem comprising larger thematic elements. Synergistically, these enable technologies of content and communication.²

This ecosystem allows for great facility in enabling data-centered methodologies. It continues to build on strong foundations. It provides foundational training data for later AI pathways that may be needed. The general common 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 for the software enable an exchange of new possibilities with regards to continuing innovation. The open-source code allows bridges among systems. The sum of the system's capabilities exceeds the separate parts. Collocating open-source digital components in a networked research ecosystem enables large connections, network effects and untapped possibility.

Cooperatively, these digital ecosystem components enable a larger open science research cycle. This cycle moves from original search and retrieval of data and content to gathering and analysis of data, to later writing, publishing, sharing and further discovery.



Open Science Research Cycle.

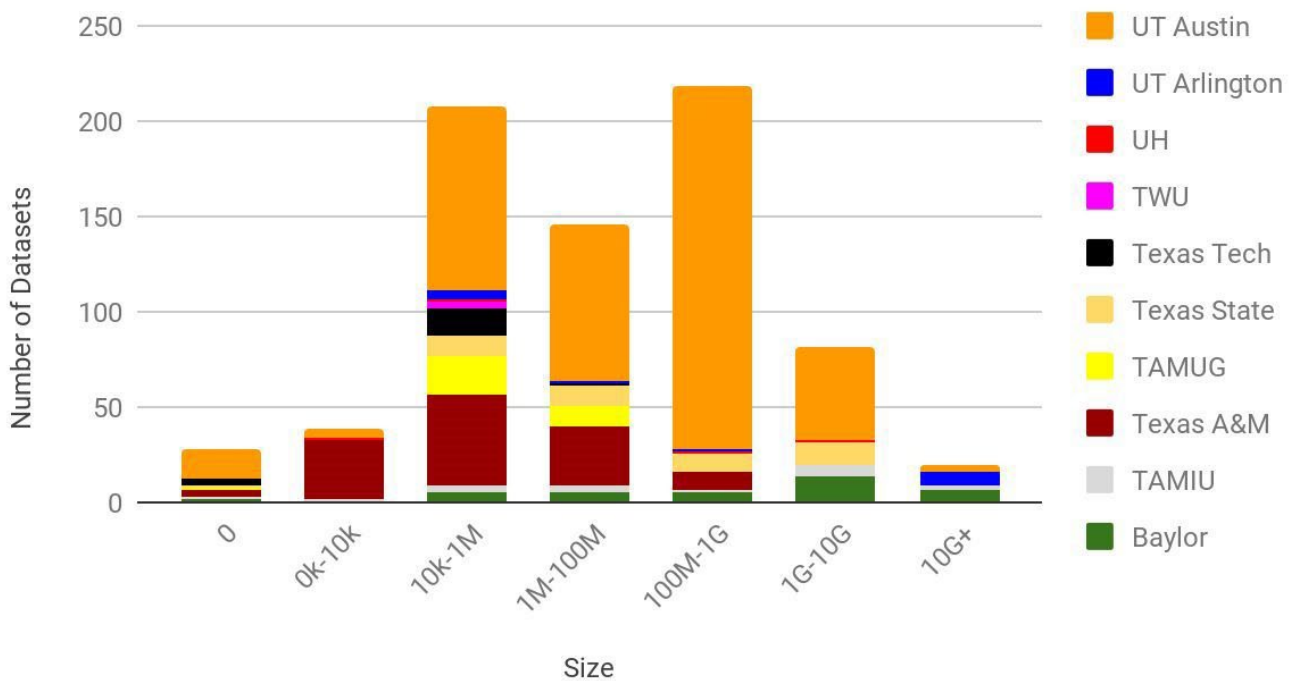
Data, Datasets, Big Data

Data comes in a variety of file types, formats, media, and sizes. For AI and, particularly recent Deep Learning, labelled and unlabeled datasets become important for machine learning and AI model training. Within open science frameworks, metadata (labelling) is key. One size also does not fit all for various open science data research repository project needs. There are many types of sizes for data projects and repositories. Repositories utilizing Dataverse can typically upload currently up to 4GB data for individual files and 10GB Datasets. This may not seem particularly large currently, considering recent examples. There are now mammoth level natural language processing datasets, or image/video modelling datasets, used for training Google's DeepMind or Microsoft's Open AI (see Mitchell, 2022). These models utilize Terabytes and Exabytes of data. Smaller datasets though, serve the needs of still many academic researchers and have served researchers well for the last six years (2017-2023).

² See Uzwyshyn, 2020. Available at:

https://www.researchgate.net/publication/336923249_Developing_an_Open_Source_Digital_Scholarship_Ecosystem

Size of Datasets

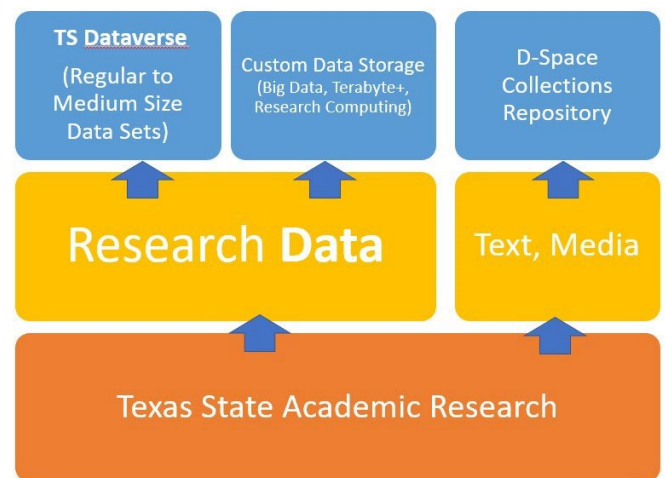


General Size Preferences for Data Research Repository Datasets Example (See Waugh, 2020)

Most experimental researchers' open science datasets for upload have been between the 1 < 1000 MB range. Currently, there is the growing recognition by researchers that 'bigger' data repositories are needed. These begin in the GB and TB ranges, though preparation must be taken now for the next phase. Many researchers are also working extensively with specialized media or GIS datasets. In these cases, for larger and custom data storage, it is still not yet feasible to place these huge datasets online, especially those in the Exabyte range. These are typically 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.

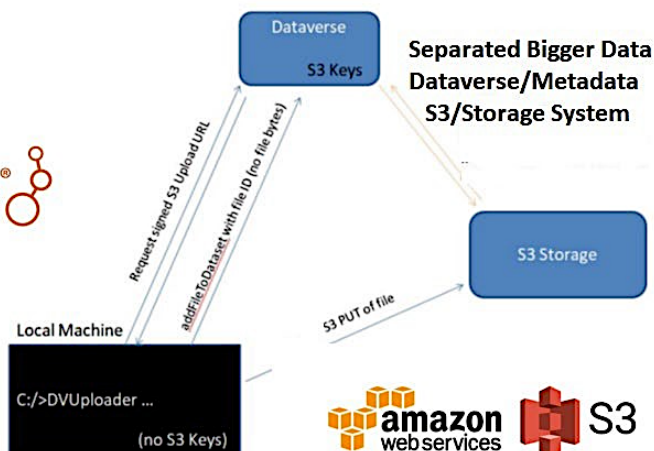
Beyond custom big data storage needs, the requests for very 'big data' (Terabytes, Exabyte storage) are still few, but these requests are increasing.

In this regard, libraries have been exploring various 'bigger data options and beta prototypes (2020-2022) with partnerships.



General University Big Data Storage Model

This ranges from commercial partnerships (Amazon Web Services S3 storage) to state efforts (i.e., Texas Advanced Computing Center, TACC) to hybrid metadata/storage pointer systems and more fee-based institutional models filling middle ground space needs (i.e. 300GB/dataset, Data Dryad).



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



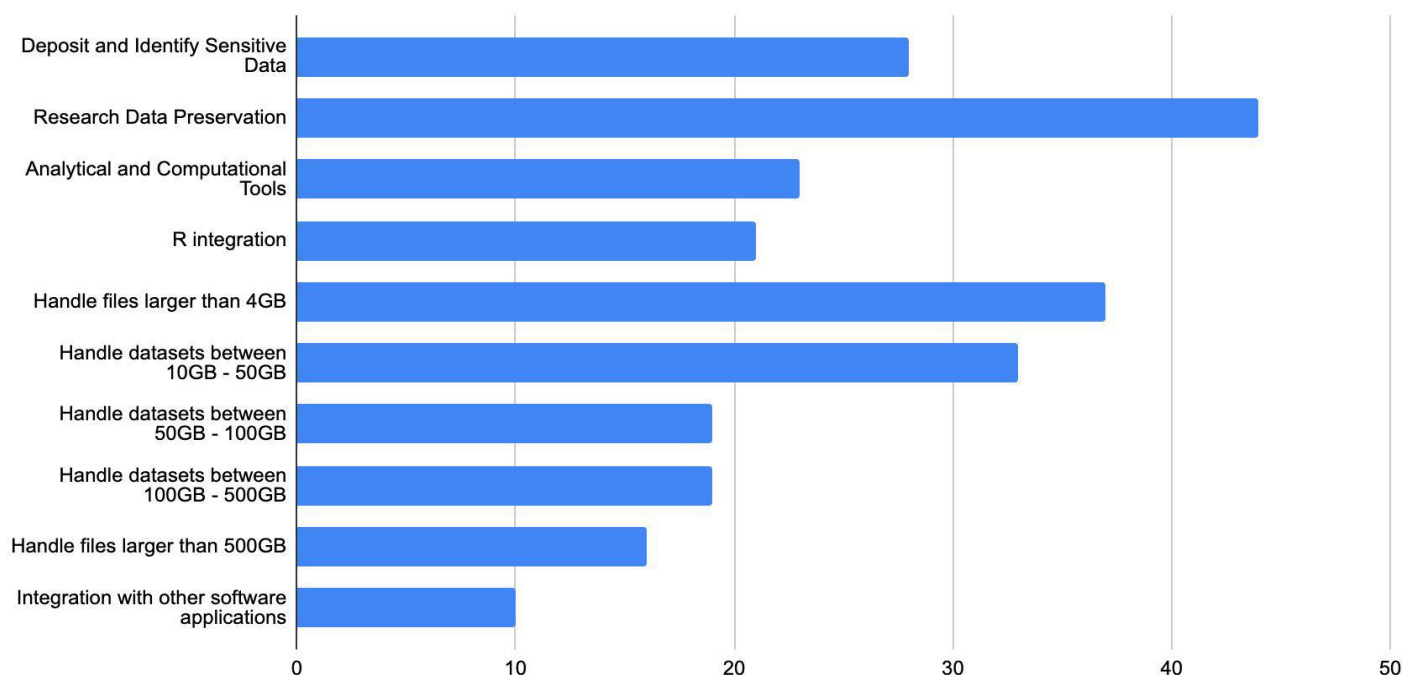
<20 GB Upload
 (Download Challenges)

Beta Prototyping Bigger Data Online Texas Data Repository Architectures. 2020-2022, TACC:

<https://www.tacc.utexas.edu/>, Data Dryad <https://datadryad.org/stash>

Currently, 'Big Data' (Exabyte, Terabyte) is among, but not at the top of, the list of new data research repositories feature-set requests that most researchers would like to see. Higher on this list of new features is long-term research digital data preservation³. Also ranking high, is managing middle ground data files (4-10 GB range) and datasets in the 10-50 GB range as well as being able to safely deposit and clean sensitive data (i.e.,

medical related, etc., see data survey below). Support for analytical and computational tools also comes high on the list. Ranging from data analytics and visualization, these tool and data literacy requests help to enable researchers from non-computer science disciplines towards new AI methodologies such as those being forwarded through neural net and deep learning.⁴

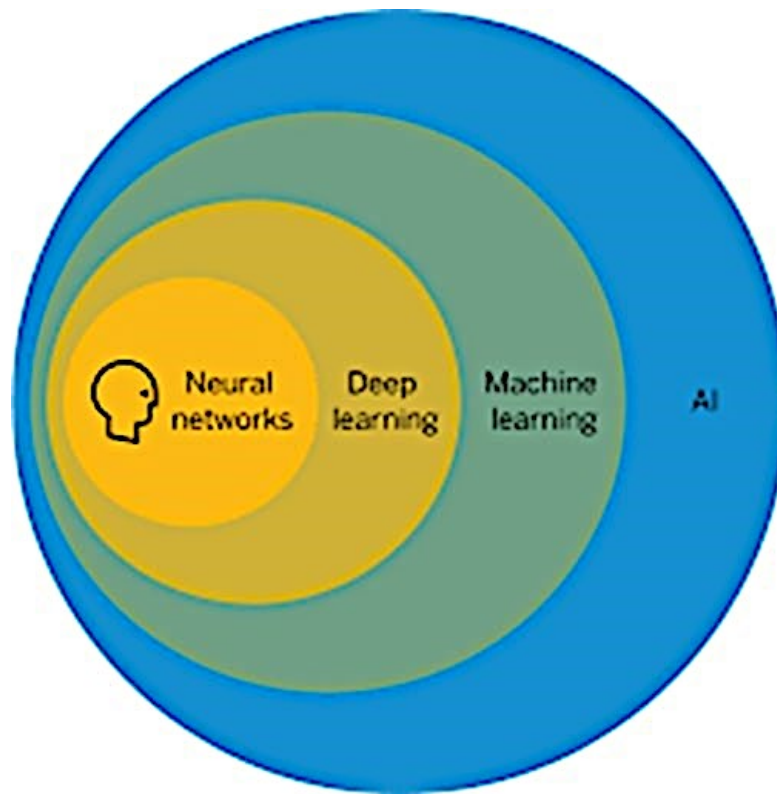


What New Data Research Repository Features Would Users Like to See? (Chan-Park, Sare and Waugh, 2022)

³ See Uzwyshyn, 2021. Frameworks for Long Term Digital Preservation. Computers in Libraries.

⁴ An important learning infrastructure for these new library/university researcher algorithmic literacy needs is filled by international communities arising such as The Carpentries <https://carpentries.org/>

Data Research Repositories, Digital Ecosystems and AI



Venn Diagram of relationships among AI and its subdomains of Machine Learning, Deep Learning and Neural Nets

The last five years (2017-2022) have shown incredible progress and gains in analytic computational tools and discovery. This is particularly true with methodologies associated with new domains of Artificial Intelligence. Machine learning, deep learning and neural net research has shown incredible potential for open science paradigm breakthroughs (Mitchell, 2022). These breakthroughs range from Computer Vision (Facial/Object Recognition) to Natural Language Processing (speech to text recognition and translation), to Cybersecurity (Fraud Detection). These advances also include Conversational Chatbots, Robotic Agents and Strategic Reasoning (AlphaGo, Game Theory).

Breakthroughs have been enabled through a fortuitous combination of better algorithms, greater computing processing power (Compute), more precise metadata schemas, online datasets and, increasingly, open science research data repositories and ecosystems.

The following section utilizes recent discoveries from Neural Net object identification to illustrate how online data research repositories and online data research ecosystems are facilitating the next generation of global collaboration with networked ecosystems research, discovery, and open science.

Cancer Detection, Library Image Data Repositories, AI

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

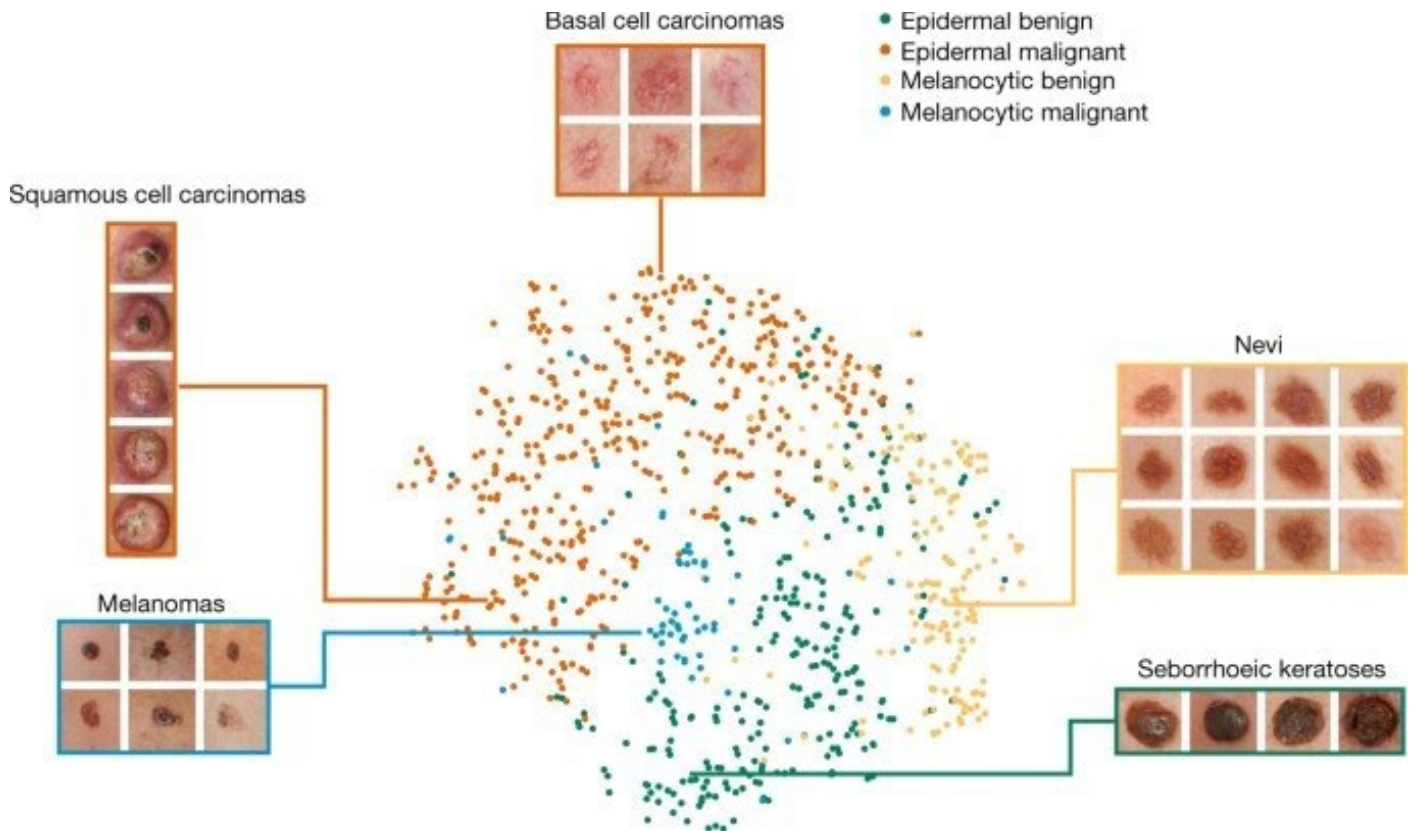


Image from *Dermatologist Level Classification of Skin Cancer with Deep Neural Nets (Esteva et al, 2017)*⁵

The neural net model was able to successfully classify epidermal lesions on mobile phones for early cancer detection into benign and cancerous (malignant) lesions better than credentialed experts. This method involved pixel-level differentiation and training through a multilevel neural net AI model.

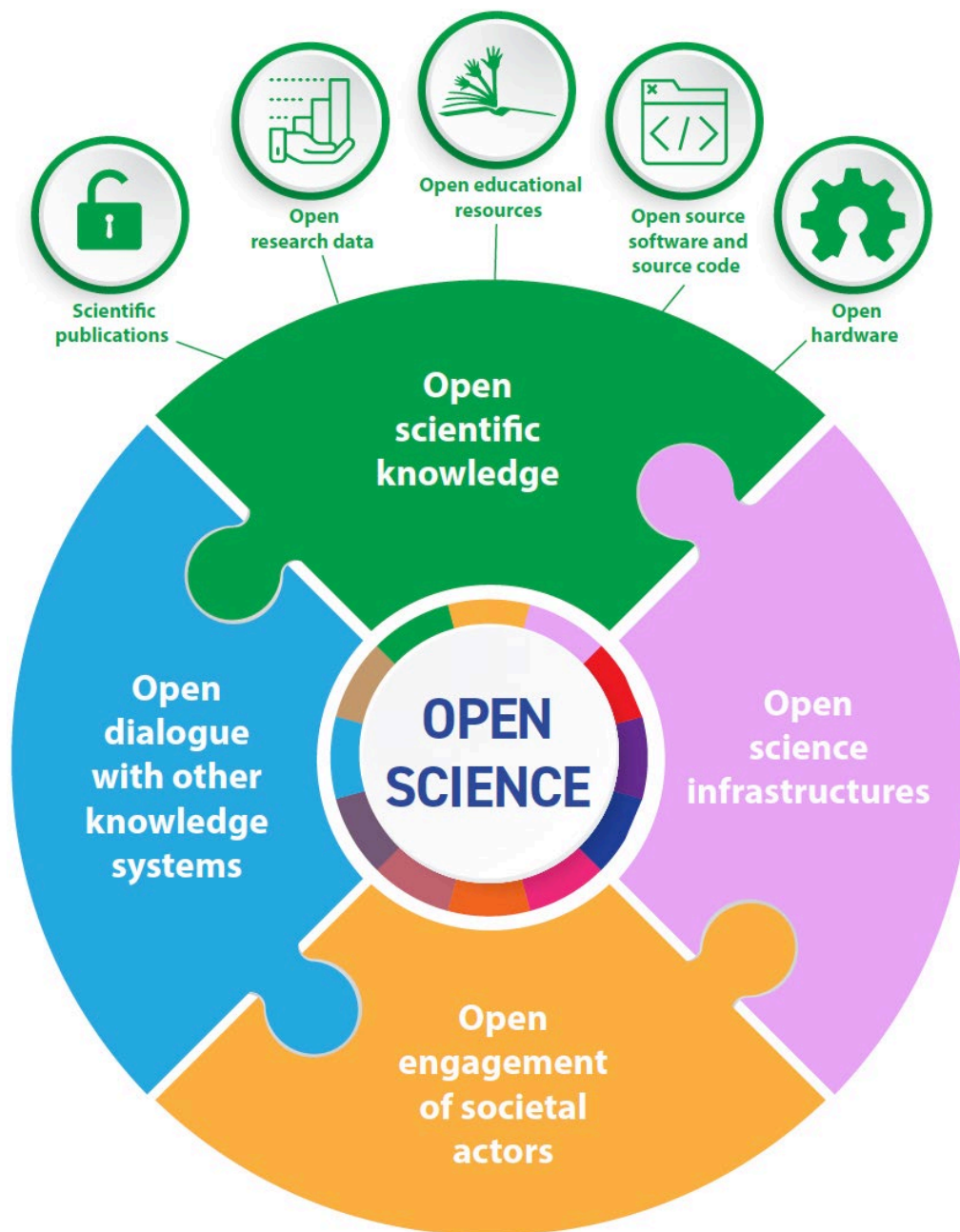
The large relevance of the digital image data repositories for initial training and metadata labelling should not be underestimated for researchers. In a recent article on *Deep Learning in Cancer Pathology Surrounding a New Generation of Clinical Biomarker* (Echle, 2020), the authors emphasize the need for: organized digital data repositories, metadata preprocessing for later accuracy in training and external validation.

Open Science, AI, and Data-Centered Ecosystems

Huge data sets like the Stanford example are not the only and most recent of those able to be utilized through AI and neural net methodologies. Innovative global open science infrastructures are being assertively forwarded. (see UNESCO Figure below). AI machine learning possibilities are also being leveraged efficiently through previous algorithmic training and the application of new regular sized datasets. New affordances are enabled through a confluence of data research repositories and researchers' willingness to share their research and data sets through open science.

Research data libraries open search and retrieval. These allow other researchers globally to apply algorithmic machine learning and building on previous models to available new online research data.

⁵ 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 Echle, 2020. Summary Video: <https://youtu.be/lvmlEq9piJ4>



UNESCO FACETS for Open Science, See: <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>

If a university or research institution does not possess a Data Repository and the researcher is conducting valid academic research, they can utilize the Harvard repository. Harvard's open source Dataverse software may be utilized for the uploading of datasets from other universities globally. Appropriate research datasets may be uploaded for sharing later or use by researchers anywhere. Dataverse is also open-source software. This means any research level library, institution and university should can set up their own instances of data repository and digital ecosystems.

To further trace these innovative discovery example pathways, the HAM10000 image dataset is a diverse 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 Stanford Nature Neural Net algorithmic methodology article appeared.

As can be seen, the images, data and metadata can be easily downloaded, unzipped, and used by researchers for neural net training purposes.

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 Data

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 lentigines / seborrheic keratoses and lichen-planus like keratoses, `bk1`), 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>

Files Metadata Terms Versions

Search this dataset... 🔍

Filter by
File Type: All ▾ Access: All ▾

1 to 6 of 6 Files

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

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

One may use data research repositories to facilitate open science around the globe through online dataset reuse. This occurs through researchers in other areas of the globe training and further developing previous deep learning and neural net models. This includes customization of variables, weights and new gathered datasets. This also becomes remarkably interesting with regards to possibilities for open science, globally dispersed academic researchers, and new possibilities for discovery and innovation.

Below is a cover page from BRAC University from Dhaka Bangladesh. They are using DSpace, a well known open source institutional repository software 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 dermatological cancer training images, metadata, and datasets. They utilized the labelled image data as training material to train a deep learning neural net algorithm with further parameterization to recognize cancer growths with

efficiency and accuracy to build on previous results. The example becomes interesting for global possibilities of telemedicine, mobile possibilities, data and global populations which may not have as quick access to trained specialists as those in the West.

This is also a particularly good example of open science and AI possibilities operating on global levels through the enabling power of digital scholarship ecosystems and data repositories' aggregation abilities. Content and specialized image data sets with specialized labelled metadata can be aggregated online that would otherwise be unavailable. This data can now be easily brought together utilized, reviewed, and improved upon with new machine learning algorithmic techniques. An example of new research, and an exceptionally good thesis has been produced below by undergraduates from the global south. Globally dispersed content and data, from three different continents, has been aggregated to advance the pursuit of knowledge and science. There is a speed, dispersion and utility here that would be unimaginable in previous 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, Ashfaqu
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

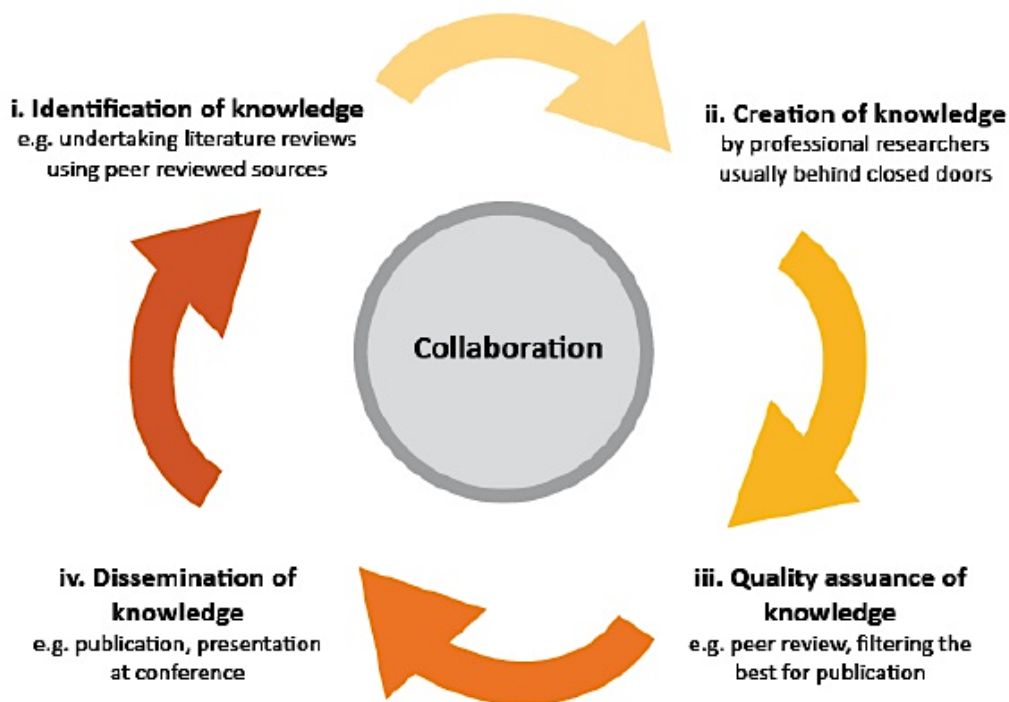
Cancer detection; Convolutional neural networks; Image classification; Deep learning

LC Subject Headings

Machine learning; Cognitive learning theory (Deep learning)

BRAC University Dspace Repository 2021 Deep Learning/AI Thesis
<http://dspace.bracu.ac.bd/xmlui/handle/10361/15932>

Conclusion – AI, Data and the Open Science Research Cycle



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

New data repository and digital scholarly ecosystem possibilities are enabling the academic research cycle, progress of knowledge and discovery in our new millennia in amazing ways. Open science possibilities empower a new global networked generation towards incredible new science and knowledge discovery. This is directly through the enabling power of digital libraries, data research repositories and open science scholarly ecosystems.

The creation of data and knowledge usually occurs hidden away in research labs, file cabinets and computer hard drives. Data sharing has now been enabled through possibilities of networked communication and content technologies. This sharing by researchers on a global stage allows transparency towards the quality assurance of knowledge ranging from online peer review to availability of data and research for citation, discovery, download and pragmatic use.

Paired with other ecosystem possibilities such as open online academic research journals, theses, and dissertation (VIREO) and online identity management

systems (ORCID), and new multimedia user interface possibilities, these tools facilitate global collaboration and intrinsically human creative activities of discovery, invention, innovation, and progress from previous generations of researchers and scholars.

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.

- The Carpentries. Data Science Skills for Libraries and Researchers. 2022. <https://carpentries.org/>
- ColdFusion (2018). *Why Deep Learning Now?* (Documentary Overview). https://www.youtube.com/watch?v=b3lyDNB_cil
- 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/s41416-02001122-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>
- Kleinveldt, 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>
- Marcum C and Donohue, R. *New Guidance to Ensure Federally Funded Research Data Equitably Benefits America*. Office of Science and Technology Policy: Washington DC, The White House. 2022. <https://www.whitehouse.gov/ostp/news-updates/2022/05/26/new-guidance-to-ensure-federally-funded-research-data-equitably-benefits-all-of-america/>
- Mitchell, Tom. 2022 *Where on Earth is AI Headed?* Carnegie Mellon. <https://www.youtube.com/watch?v=ij9vqTb8Rjc>
- NASA. Open Science Overview. 2022. <https://science.nasa.gov/open-science-overview>
- 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/s41591020-0942-0>.
- UNESCO. *Recommendations on Open Science*. 2021. Paris: General Conference. 41/C. Resolution 24. <https://www.unesco.org/en/natural-sciences/open-science>
Final Report: https://unesdoc.unesco.org/ark:/48223/pf0000379949.local_e=en
- Uzwysyhyn, 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.
- Uzwysyhyn, R. 2021. Frameworks for Long Term Digital Preservation Infrastructures. *Computers in Libraries*. September 2021. pp.4-8.
- Uzwysyhyn, 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://rayuzwysyhyn.net/TXU2016/OnlineDataResearchRepositoriesUzwysyhyn.pdf>
- Waugh, L. *Texas State University Annual Usage Report 2020*. TXST Dataverse Repository. Texas Conference on Digital Libraries Presentation. Texas State University.