SPECIAL BIG DATA ISSUE

JANUARY 2022

Trends and Issues in Library Technology

International Federation of Library Associations

IT Section IFLA IT

TRENDS & ISSUES IN LIBRARY TECHNOLOGY

IFLA Information Technology Section

Special Issue on Big Data

Feature Articles

Invited Articles

Departments

Research Data Repositories and Global Scholarly Ecosystem Possibilities

Ray Uzwyshyn, ruzwyshyn@txstate.edu
Director, Collections and Digital Services, Texas State University Libraries, US

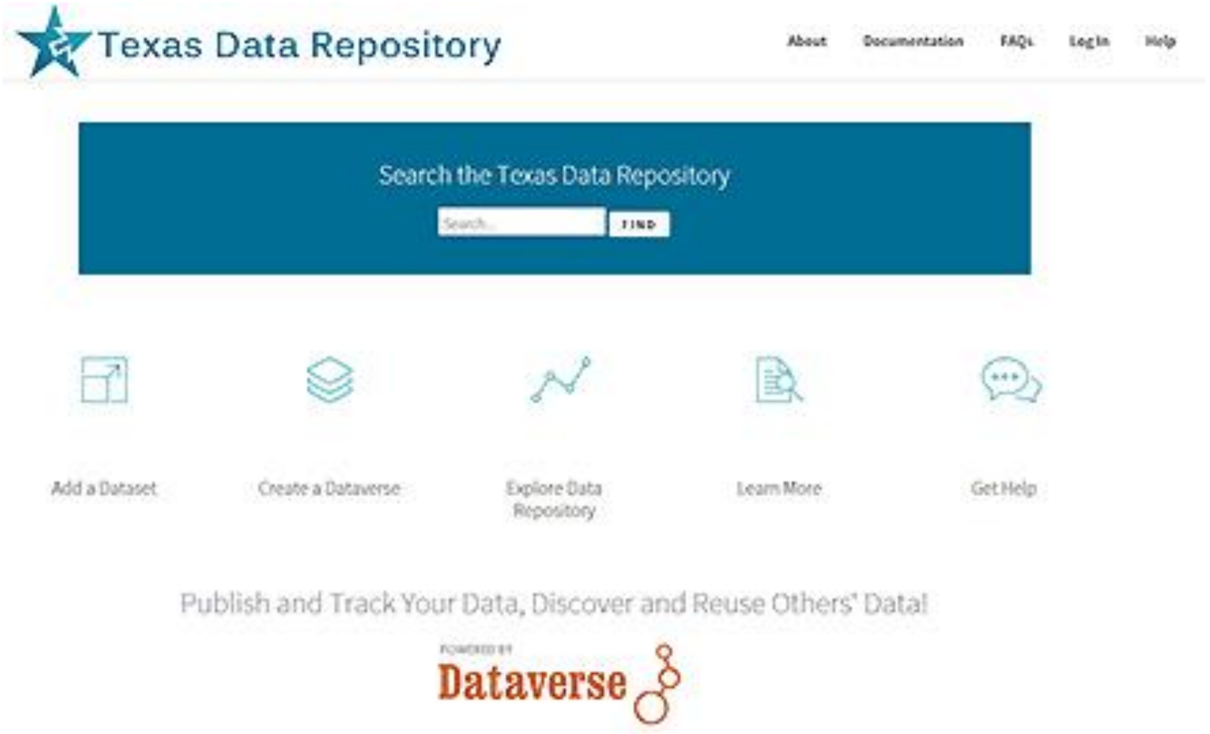


Figure 1. Texas Data Repository, https://data.tdl.org/

Introduction

Online networked data research repositories allow sharing and archiving of research data for experiments and research studies. This opens data to modern interoperability and metadata standards for search and retrieval. Located in open scholarly ecosystems, data research repositories are currently being leveraged to accelerate global research, promote international collaboration, and innovate on levels previously thought impossible.

Research data repositories open possibilities for collaboration with others. They link data to further content from online publications and aggregation tools making associated documents easily accessible. This paper pragmatically overviews such a data-centered ecosystem at Texas State University Libraries, a large state university research library in the United States. The research then goes on to speculate on possibilities for global research data repository networks utilizing the models presented.

Texas State Data Research Repository

A Data Research Repository is now a necessity For STEM disciplines (Science, Technology, Engineering and Math), the Social Sciences, Open Science, or any discipline which utilizes data-driven research methodologies. Data research repositories are suitable for any library, research, or academic institution with large scientific and social science research and datasets.

This new class of open-source software becomes the infrastructure of choice in a digital scholarly ecosystem and academic library infrastructure. Library human resources and information technology models are worked out and promising future possibilities are manifest. Texas State University Libraries uses a customized consortial version of Harvard University’s open-source data repository software - Dataverse. Dataverse is a multi-tiered open-source platform for publishing, archiving, and sharing research data.

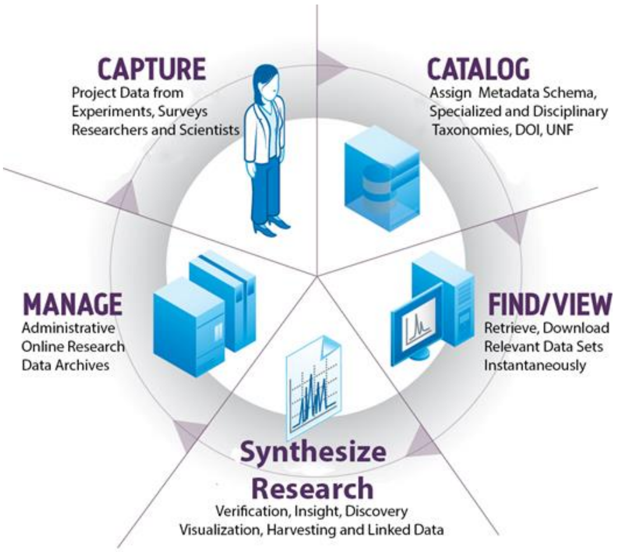

CAPTURE
Project Data from Experiments, Surveys Researchers and Scientists
CATALOG
Assign Metadata Schema, Specialized and Disciplinary Taxonomies, DOI, UNF
MANAGE
Administrative Online Research Data Archives
FIND/VIEW
Retrieve, Download Relevant Data Sets Instantaneously
Synthesize Research
Verification, Insight, Discovery Visualization, Harvesting and Linked Data

Figure 2. The Data Research Cycle

Dataverse and the Texas Data Repository allow researchers to publish, track, discover and reuse other researchers' data and participate in the larger data research cycle. The data repository enables researchers to operate independently, assigning metadata. Specialized disciplinary taxonomies are also possible. A data repository specialist assists in archiving data archives and creating unique metadata schemas. The Dataverse platform allows researchers to search internally within an institution and retrieve institutional data. Simultaneously, it also enables search access across other institutions.

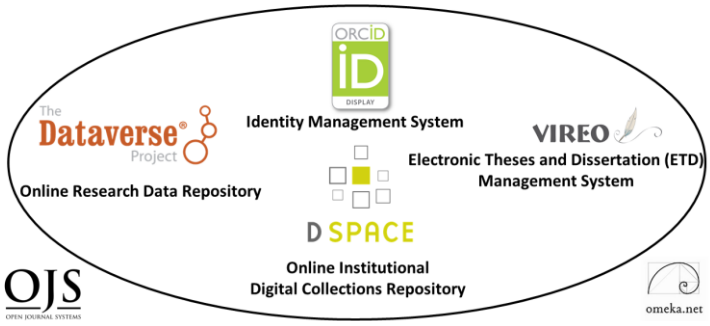

ORCID iD DISPLAY
Identity Management System
The Dataverse Project
Online Research Data Repository
VIREO
Electronic Theses and Dissertation (ETD) Management System
D SPACE
Online Institutional Digital Collections Repository
OJS OPEN JOURNAL SYSTEMS
omeka.net

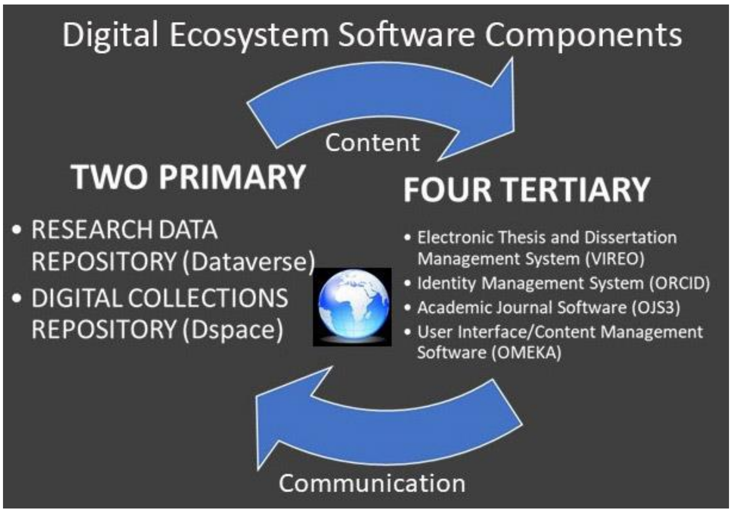
Figure 3. Texas State University Libraries Scholarly Research Ecosystem

Within the Dataverse Structural Model, the researcher may search and retrieve data from other researchers who may be working locally. Alternatively, they may search more inclusively for similar research data at other institutions in the Texas Digital Library Dataverse Consortium.

The Texas State University Libraries Online Data-centric Research Ecosystem consists of six basic software components, two primary components and four tertiary. The primary components allow content to be shared and archived. The tertiary components allow communication among internal and external networks.

Central to this scholarly ecosystem are primary components of an Online Research Data Repository (Texas State University DATAVERSE) and an Online Institutional Collection Repository (Texas State University DSPACE). These primary components serve to house content - research data and research papers.

Four other software communication components comprise tertiary components: an Identity Management System (ORCID), Electronic Theses and Dissertation Management System (VIREO), Academic Journal System (Open Journal Systems 3) and User Interface Software (OMEKA). The tertiary components connect the ecosystem components with each other to various areas internal to the research institution (i.e., Electronic Theses and Dissertation Management System) and externally to larger global networks (ORCID Identity Management Software). Typical characteristics necessary for this type of larger digital research ecosystem are open-source software, active developer communities, stable software releases, easy configurability (i.e., open API's), customizability, and connectivity.



Digital Ecosystem Software Components

Content

TWO PRIMARY

- RESEARCH DATA REPOSITORY (Dataverse)
- DIGITAL COLLECTIONS REPOSITORY (Dspace)

FOUR TERTIARY

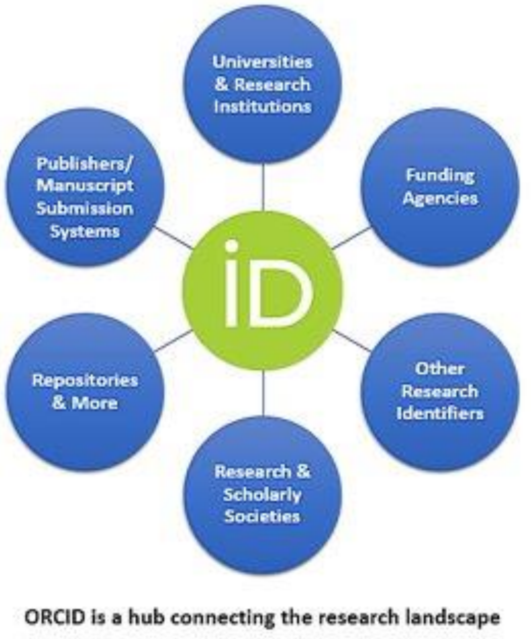
- Electronic Thesis and Dissertation Management System (VIREO)
- Identity Management System (ORCID)
- Academic Journal Software (OJS3)
- User Interface/Content Management Software (OMEKA)

Communication

Figure 4. Digital Research Ecosystem:
Primary and Tertiary Software Components

All ecosystem components may be interlinked with the Data Repository. Research papers residing in the Digital Collections Repository (Dspace) may contain direct links to repository datasets. Datasets within the data repository may reference research papers published through Open Access Academic Journal Software (OJS3). These may also directly reference associated datasets through citation and reference links that the Open Journal Software allows. These allowances enable researchers reviewing a scientific paper to click directly to the data to scrutinize research integrity and possible download.

Further internal linkages are possible through the Electronic Theses and Dissertation Management System (Vireo) and externally, the Researcher Identity Management System (ORCID). ORCID may be used as a network hub for further linkages to funding agencies, other research identifiers and institutions.


Universities & Research Institutions
Funding Agencies
Other Research Identifiers
Research & Scholarly Societies
Repositories & More
Publishers/ Manuscript Submission Systems
iD

ORCID is a hub connecting the research landscape

Figure 5. ORCID https://orcid.org/

Data Research Repository Ecosystems: Global Possibilities

Currently, the Texas State Data Research Repository and ecosystem is part of a statewide 22 university library consortium. Participating members who possess a local instance of the Dataverse repository may search statewide for similar datasets. A researcher working on datacentric subject areas can connect with associated network ecosystems

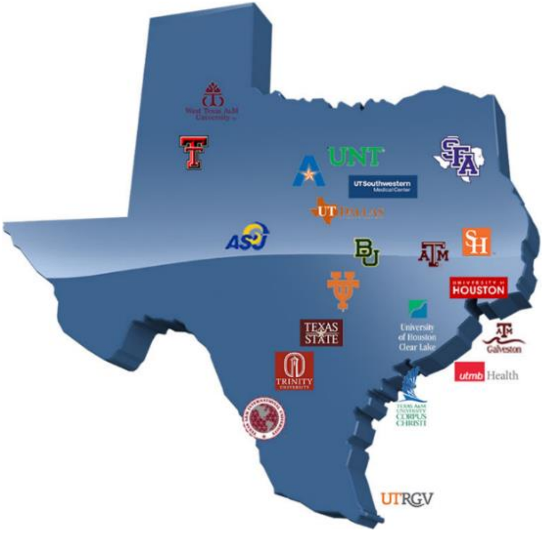


Figure 6. Texas Digital Library University Members https://www.tdl.org/

The possibilities with the Dataverse model for larger geographically disbursed consortia on state, national or international levels, become interesting as data-centered networked research ecosystems scale. Surrounding tertiary software can be connected on consortial, state, national or international levels.

There are currently approximately 266-300 High Research Activity institutions (universities and national research centers) in the United States and Canada. The Times Higher Education Supplement estimates there are 1250 higher level research universities worldwide (2.7 - 4.2% of all universities worldwide). QS World University Rankings estimate that 40% of these higher-level research institutions are in Europe, 26.5% in the Asia/Pacific Corridor, 18% in North America, 9% in Latin America and 6.5% in Africa and the Middle East.

Future Possibilities for Data and Digital Scholarly Ecosystems: Global Data Sharing

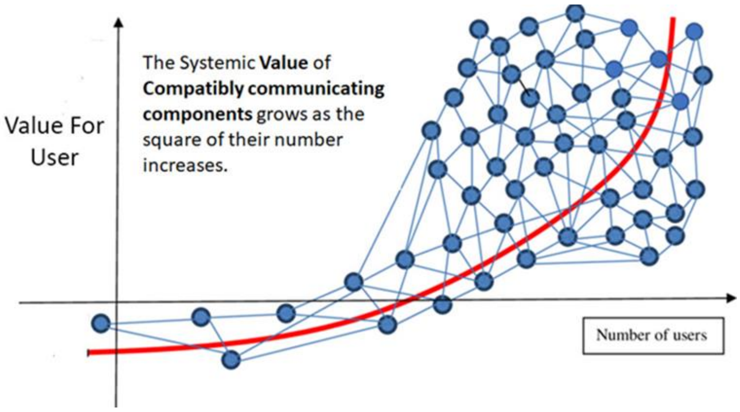


Figure 7. Metcalfe's Law for Networked Components

With the abundance of mature network-enabled open-source online research software possibilities outlined, there is little reason that the other top 2-3% of research libraries globally should not be empowered with these types of digital ecosystems at the least. The larger majority of the top 2% of these research libraries globally (approximately 1000) currently do not yet possess these newer software infrastructures. It would be easy enough with the proper organizational funding to open possibilities for the entire 1250 research level institutions globally.

As a global grand challenge, why not enable these research institutions and their academic libraries with these types of online networked research ecosystems? This would open abilities for dialogue, sharing research data and collaborating with established networks of contact. One need not look far for an immediate precedence for this type of global need beyond the continuing pandemic and needs for more organized future global data-sharing networks.

Future Global Possibilities One Server Per Research University: 2022-2027

With interest and funding it would be beneficial and easily possible to give at least 1000 Research University Libraries globally one configured online research ecosystem server with the six pieces of open-source research software described. Three times a year, weeklong training could be given over five continents and an online help network set up through IFLA as a worthy new millennia initiative. Later, assessment would be done to gauge the network effects of sharing data and research globally.

Physical servers could be given to research institutions with software installed, or fractional server space placed in the cloud over five continents, with mirror sites globally. With current server power, this would not be more than $1000.00 USD/server. One or two large funding agencies, or suitable international organizations, could shepherd the project with IFLA to develop version 1.0 of this Global Digital Scholarly Ecosystem and Research Data-Sharing Network. After one, three- and five-year periods assessment could be done. Subsequent versions could also be produced improving and streamlining the model and utilizing the latest agile information technology project management techniques. The positive effects for research, collaboration, data collection, and science would be enormous. The benefits for the progress of knowledge through research libraries in the 21st century would also be large if even a fraction of the global potential was achieved.

Figure 8. One Server Per Institution
Global Digital Research Ecosystem Initiative

Conclusions

This paper has given a high-level outline of a successful, pragmatically possible, digital ecosystem on local levels (Texas State University Libraries Model) and extended this model towards future global research possibilities. These systems are not overly challenging to set up at local institutions. Example links, papers, presentations, and open-source software for download from the working digital ecosystem links are given in the references below.

Possibilities with current open-source software for data research repositories and digital scholarly research ecosystems are wide open. These systems are pragmatically realizable on local university library levels. They should begin to be prototyped with larger national, international, and global networks. The time has come to begin connecting these types of digital research ecosystems. Present day research networks, libraries and universities are well-suited towards these new millennia digital stewardship infrastructures and roles.

Further References

Papers

Uzwyshyn, R. 2020 Developing an Open Source Digital Scholarship Ecosystem International Conference on Education and Information Technology ICEIT2020. St. Anne's Oxford, United Kingdom. February 2020.

- - -. Open Digital Research Ecosystems: How to Build Them and Why . *Computers in Libraries*, (40) 8. November 2020.

---.Online Research Data Repositories: The What, When, Why and How. Computers in Libraries. 36:3, April 2016. pp. 18-21.

Presentations

Uzwyshyn, R. Digital Research Ecosystems for Open Science (Presentation). AI for Data Discovery, Reuse & Open Science Symposium. Carnegie Mellon University. October 20, 2020.

---.Developing a Digital Scholarship Research Ecosystem. (PDF) Association of Southeastern Research Libraries Members' Meeting Presentation, April 29, 2020.

Open-Source Digital Ecosystem Software (Downloads & Info.)

Data Research Repository: Dataverse https://dataverse.org/
Digital Collections Repository: Dspace https://duraspace.org/dspace/
Content Management System: Omeka https://omeka.org/
Academic Journal Software: OJS3 https://pkp.sfu.ca/ojs/
Identity Management Software: ORCID https://orcid.org/
ETD Management Software: Vireo https://www.tdl.org/etds/

Texas State University Libraries Data Repository & Digital Ecosystem

Texas State Data Research Repository
Texas State University Libraries Digital Scholarship Ecosystem.
Texas State Digital Collections Repository
Texas State Online Research Identity Management System (ORCID)
Open Journal Systems @ Texas State