

C-Level View | Feature

Coming of Age: The Online Research Data Repository

A Q&A with Ray Uzwyshyn

- By Mary Grush
- 12/13/16



"Online research data repositories are now pragmatic realities. Most discovery in the future will be predicated on the sharing and synthesis of data." — Ray Uzwyshyn

Technology advancements surround us, and sometimes the sheer volume of new tools and services is overwhelming. Can we identify which technologies are poised to make significant changes in the way we work? One technology that's been relatively under the radar may be about to make a huge difference in scholarly research practice and has the potential to help move scientific and social scientific discovery ahead as never before.

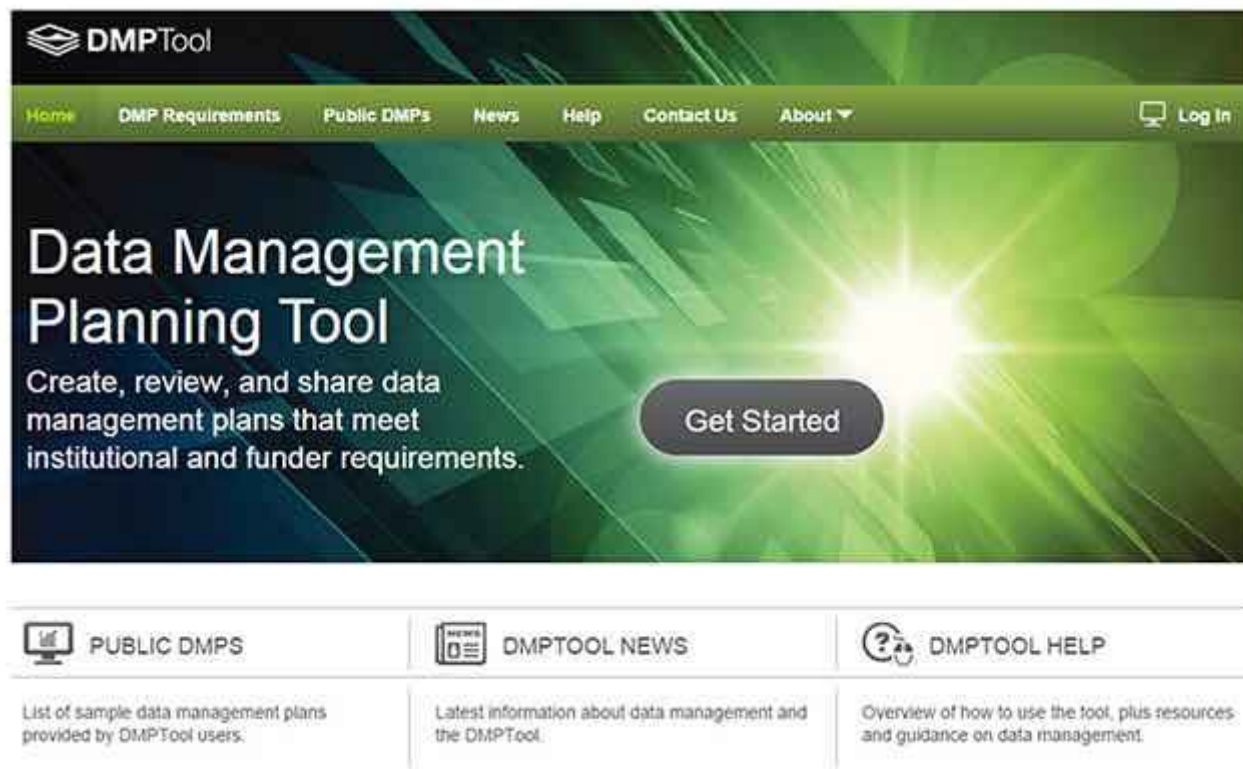
Here, CT asks Ray Uzwyshyn, the director of digital and collection services for the Texas State University library, about research data repositories — a technology that is just coming into its own. Uzwyshyn served on the implementation, planning, and policy committees for the Texas Data Repository, which launched December 2016. He offers both a current view of the technology and insight into its impact.

Mary Grush: What's the main objective of an online academic research data repository?

Ray Uzwyshyn: Research data repositories enable academic researchers to access, cite, and share data for a particular project — not just the final paper or project summary, but the actual data and paratextual material associated with it. They house both datasets and the material surrounding the data: field notes, documents, multimedia, and even specialized software used to process this data.

Most scholars today are working with an online community of colleagues who may be geographically dispersed around the globe. For them, the online data repository is becoming the storage application of choice — access speeds and software no longer present significant barriers to entry.

Grush: Why would an academic researcher choose to share their data via an online research data repository?



Grush: What are the important steps for research faculty — or for that matter, graduate students — as they create datasets and utilize a research data repository?

Uzwyshyn: All research data has an associated data repository life cycle path. Typically, researchers capture project data from experiments, instruments, surveys, and field work. They assign a disciplinary taxonomy, classification, or what we call a metadata schema to the data — essentially more data or description describing the primary data. This schema is key for the repository's search capabilities and later database search.

Ultimately, this classificatory work done properly allows effective searching across repositories so that datasets can be aggregated and harvested for later insight. Most researchers are looking for datasets similar to theirs: Do the datasets I found confirm my data? Can I use other researchers' results to build on my own experiments and data?

By searching research data repositories, researchers can also find examples of "negative data" or experiments that have failed — so they do not have to recreate the wheel and go down paths that previously have been dead ends or found to be unproductive. They can avoid duplicating such work that has already been done.

The final, and very important stage in the research data repository life cycle, is the long-term archiving and storage of datasets. This is both for the historical record and so that experiments won't be needlessly repeated. Basic research done today may not find its use value until twenty years hence. Because of this, it's important that the data be more transparently archived, stored, and kept accessible through file normalization and updates to software formats. This is especially true for time series data that tracks changes over time.

Grush: What types of data and data formats will you find in research data repositories?

Uzwyshyn: Most repositories are format-neutral and accept wide ranges of data formats. Of course, everyone knows Excel, but different disciplines also have their own specific data formats and preferences. For example, biochemistry may have specific data formats and software for data that the repository needs to accept.

There are also many specific types of data repositories: *project-specific*, *discipline-specific*, *institutional*, or *consortial* data repositories. Project-specific repositories are project-oriented and typically contain research data created by a single faculty or a small team. Discipline-specific data repositories are usually subject-focused and

aggregate data from a certain discipline, say experiments surrounding nanotechnology — Purdue's [Nanohub](#) is an example. Institutions may also possess their own research data repository that goes across disciplines — this is an increasing trend. And finally, there are consortial research data repositories. The [Texas Data Repository](#) (TDR), which launched in December 2016, is the first statewide academic consortial repository. My institution, Texas State University, is one of the institutional repositories that make up the larger network within this consortial repository.

Grush: How big are these various types of data repositories, and what size datasets do they accept?

Uzwyshyn: There is wide variation among repositories, depending on storage requirements and the sizes of datasets being gathered. The majority of online research data repositories for academic institutions accept what we might think of as regular or medium-sized datasets. These are typically of a size small enough to allow that the data may be housed right in the repository itself. A researcher or research group can upload their data from their desktop computer or research group server. To help you get your bearings on this question, for the Texas Data Repository each researcher may currently add as many files as they like up to 2GB in total, and research data groups within repositories may possess up to 10GB. These are very loose and flexible size limitations though, and it's safe to say they are constantly expanding and being re-evaluated in light of researchers' needs.

If there are larger datasets, the data repository might be considered a specialized project-specific or discipline-specific variant rather than institutional or consortial. Very large datasets — with voluminous amounts of data, like the Seti project generates — might be more effectively treated with pointers from the repository to the actual storage places where the project data is housed. In such cases, the research data repository becomes a metadata repository — a place for data describing the data. Again, the value here is that the data repository enables researcher discovery, searchability, interoperability, and aggregation of datasets for further research.

The screenshot shows the Texas Data Repository website. At the top left is the logo, a blue star with a white 'E' inside, followed by the text "Texas Data Repository". To the right of the logo are navigation links: "About", "Documentation", "FAQs", "Log In", and "Help". Below the navigation is a large blue search bar with the text "Search the Texas Data Repository" and a search input field with a "FIND" button. Below the search bar are five icons representing different actions: "Add a Dataset" (a folder icon), "Create a Dataverse" (a stack of three books), "Explore Data Repository" (a line graph), "Learn More" (a magnifying glass over a document), and "Get Help" (a speech bubble). Below these icons is the text "Publish and Track Your Data, Discover and Reuse Others' Data!" and the "POWERED BY Dataverse" logo, which includes the word "Dataverse" in orange and a graphic of three connected circles.

Grush: What are the search capabilities in general? And what are a few of the benefits of searchability?

Uzwyshyn: Searchability is a primary value of research data repositories — the scholar is able to search across institutions, a consortium, or an entire discipline's experiments in specialized areas. Researchers can identify someone else's data that may help validate their findings; they can share data in partnerships with other researchers for new levels of discovery in their field; and they may aggregate or mash up data from various fields to create new knowledge and insight.

The concept of the research data mashup is similar to most software mashups, where, for example, one database provides GIS geospatial data, another provides real estate data, and a common field provides a link to combine disparate knowledge sets. By mashing these together, greater relational insight is achieved.

The analogy holds for scientific and social scientific experimentation through this linking field. This becomes especially interesting in linking datasets for disciplines that wouldn't normally "talk" to each other, academically speaking, but have commonality of one or more data fields. Research can then be synthesized, validated, or invalidated through the examination of a global scholarly community. Researchers can also gain insight from access to previously unavailable relevant datasets. It's probably also important to mention that data visualization technology becomes an important tool and infrastructure within the data repository ecology.

Grush: How can institutions approach all this? What kind of infrastructure do you need to provide, and what factors should you consider in choosing to build a repository? Do you have to build this from scratch?

Uzwyshyn: Today you'll find a variety of new solutions for housing and sharing your data, both open source and proprietary. A good data repository should have a permalinking strategy — citation and access capabilities, typically with a Digital Object Identifier (DOI) or a Universal Numerical Fingerprint (UNF) that give the data a permanent location on the Internet. The repository could either be installed on a university server or hosted somewhere else, and a good solution will include administrative and collaborative options. Capacity for ingesting a wide range of data types, from Excel, to SPSS, to various discipline-specific data formats is also an important factor.

The number of good examples and models to investigate is increasing over time. Here are a couple links to what we did in Texas as we created the Texas Data Repository: <http://tinyurl.com/h36w93v> and <http://tinyurl.com/j5pcccz>.

Grush: What is the landscape now, for research data repositories? Is this a good time for institutions to think of getting "in" on this?

Uzwyshyn: The top research institutions in the U.S. have adopted, so the early adopters are all in. The early majority adoption is presently occurring and we're somewhere in the middle of this cycle. It's an excellent time to start thinking about adoption, especially if your institution has research faculty or aspires to be a research institution.

Even before selecting or implementing anything though, the best place to begin is with an environmental scan to examine your institution's needs. This means both polling your researchers and reviewing the state of current focused solutions. Harvard is very much behind a product called [Dataverse](#) — the product is flexible across disciplines and has its roots in the social sciences. Purdue University came out with a different orientation and originally advocated a more discipline-centered approach — their product, [HubZero](#), is used for more specific research interests, often in highly specialized technical scientific research areas.

Grush: Given that researchers in any field may be far-flung geographically, are there any global organizations that are pointing the way to help join data repositories together, promote better access to them, or help them interoperate via universal standards?

Uzwyshyn: In general, this idea of the research data repository is still fairly new and hasn't yet adopted official, universally accepted bodies of standards. But there are plenty of organizations beginning to think very seriously about data repositories, both present realities and emergent possibilities.

There are also emerging standards strongly in place for various specialized disciplinary metadata schema. This standardization of metadata schema ranges from the more general Dublin Core standard, to other, more specialized schema — for example, geophysical, life sciences, or astrophysics data. All of these enable interoperability.

Another organization advocating open data, SPARC, has just published an important [tool that helps researchers navigate federal data requirements](#) for their grants in the U.S.

Grush: When institutions take the plunge, if you will, and select a research data repository solution, is there any way they can plan for agility in the future, or even just have a reasonable exit strategy?

Uzwyshyn: Well, the evolving academic record of how research is being carried out today is rapidly changing, so the hard answer is that you must be thinking with both this moving target and a new generation of researchers in mind. You shouldn't be focusing on exit strategies these days, but rather thinking about evolutionary and developmental scenarios and those variables that will allow migration down the road.

To generalize, academics are not going to cease research efforts, and the possibilities for organizing, sharing, and housing research data have exponentially expanded through technology. You just need to keep your eyes open and more importantly, keep an open mind to the technological possibilities as the research data repository continues to evolve.



Grush: Are there other factors behind understanding and planning for a research data repository? What about planning for staff and the multiple roles that will be needed to build and support a repository?

Uzwyshyn: We've already mentioned several of the characteristics of research data repositories and what unique services and discovery advances they can bring to an institution. Beyond things we've highlighted specific to

this technology, other factors campus leaders should consider are more typical of any major technology initiative.

Behind the research data repository lie technical factors like emerging data and metadata standards, QC standards, and a range of technology issues; administrative, policy, and legal issues such as copyright and intellectual property; and outreach information, user education, and operational and service expectations. Challenges in implementing a research data repository will be similar to those you find with any important technology initiative. An institution should plan to leverage expertise from its previous technology implementation successes and be prepared for the research data repository to dialogue with various levels of the university campus.

Human resource expertise in research data repositories does also especially need to be developed over time. All research institutions will have to do something at some point, to plan for and create their repository infrastructure, and the need for staffing and staff expertise is a reality. Now is a very good time for leadership to begin considering staff roles, along with the related discussion of whether they want their institution at the back of the pack tomorrow or in the middle today.

Grush: What is the outlook for research data repository consortia? Can institutions gain advantage through consortia — maybe in their own region or even globally?

Uzwyshyn: Historically, most research collaborations among academic researchers have been more localized, with nearby universities, states in their region, or collegial institutional networks. Consortial efforts increasingly allow researchers to enhance the possibilities opened by aggregated datasets, leveraging visibility and the expertise of colleagues both locally and globally. Sharing data globally leads to recognition, grant and project collaboration, and traction for new areas of investigation. These new paradigms have the power to move research ahead more quickly in the disciplines.

On technological levels, consortia also often build a community of technological human resource expertise regarding implementing and building technology offerings like research data repositories. Often these are state or interstate technology groups that can be very helpful in navigating the myriad of issues that will arise. Leveraging the cooperation of numerous institutions as a repository is created has many benefits.

Grush: Are researchers ready, in general, to share their research data more openly and work towards shared discovery? Are we looking ahead at good changes in research practice?

Uzwyshyn: There are great possibilities and I believe most researchers see or will see the value here. Currently our largest granting agencies (including NSF, NIH, or USDA) mandate and encourage data sharing processes, so the future is bright on pragmatic levels. Historically, the advancement of scientific discovery has been predicated on the sharing of knowledge and data. This is true from Newton to Einstein. As Newton put this, "If I have been able to see a little further, it is because I have been allowed to stand on the shoulders of Giants". The larger idea is that no researcher is working in a vacuum, but rather within a scholarly community with a past trajectory and a forward telos. Because of this, I'm encouraged by these new technological possibilities for organization and sharing from this great ocean of data opening before us. Our software infrastructures are now able to allow this next renaissance of discovery, enabling new insights and synthesis from current results. Hopefully this will also allow a few of those next intrepid explorers to stand on the shoulders of giants.

About the Author

Mary Grush is Editor and Conference Program Director, Campus Technology.



©2001-2016 1105 Media Inc, Ed-Tech Group.