Colorado State University Libraries
Conference Proceedings and Events
CI Days: Cyberinfrastructure 2010 in the Rockies
Transcription of Cyberinfrastructure 2010: data curation and digital repositories panel, 2010

Collection: CI Days: Cyberinfrastructure 2010 in the Rockies
Title: Cyberinfrastructure 2010: data curation and digital repositories panel
Date: 2010
File Name: CI_Days_2010_Panel_Disc.mp4
Date Transcribed: November 2024
Transcription Platform: Konch AI

BEGIN TRANSCRIPTION

[00:02 - 01:37] Dawn Paschal: Well, as Jeff mentioned earlier, I am Dawn Paschal. I'm one of the assistant deans at the libraries here at Colorado State, and I'd like to thank you all for coming out today for CI days event. We have arranged for a great panel of speakers for you on the topic of data curation. As you know, the importance of managing data throughout its lifecycle of interest and value to research, scholarship, and teaching has been recognized by NSF and other research funding bodies, institutions, and information agencies. Data curation is a highly complex, exciting, and emergent field distinguished by technological, sociological, and intellectual property challenges. It is my pleasure to introduce on my immediate left, Mary Merlino. Mary is the Director of science and the library at the National Center for Atmospheric Research. And I want to pause here and mention that NCAR is the first NSF federal lab to mandate open access of its scholarship, and this initiative was led by the NCAR library. Next, we have Greg Newman. Greg is a research associate at the Natural Resource Ecology Laboratory here on the CSU campus. And we have Jessica Branco Colati. Jessica is director of digital Repository Services. And each will present on various aspects of data curation. And after they have all finished, then we will take your questions at the end of the-, this program. Mary.

[01:41 - 03:50] Mary Merlino: Well, first of all, I want to thank Russ and all the people who asked questions for setting out some of the challenges, the issues involved with Cyber Infrastructure. Whereas, Russ, I'd sit through a lot of these presentations and very rarely have I heard Cyber Infrastructure described so clearly and succinctly. So you should go to NSF and perhaps give the same presentation. What I'm going to do today in my short time with you is talk about one particular Cyber Infrastructure Initiative, the Data Net Initiative. This is the first large NSF funded solicitation to

come out of OCI, the Office of Cyber Infrastructure, which was created, I don't know, 4 or 5 years ago. It's a relatively new office. And talk about in general, the solicitation talk about our particular-, let me see if I can do this here. Okay. What I'm going to do is talk about the general NSF program, the data net program, talk about our particular response to that program, the Data Conservancy, talk about implications for libraries and then finally, how you in this room might get involved. So first of all, I think many of you probably have seen this quote before. What is cyber infrastructure? This is one of the-, one of those touchstone visions that people refer to. Everybody writes this in their grant. The vision of science and engineering. Digital data are routinely deposited in well-documented form, regularly and easily consulted and analyzed, etc., etc. So NSF came out with this report in 2007. Interestingly enough, a month before this report came out, another report came out that said, "Oh, by the way, cyber infrastructure is hard. There is no one way to do it. There are multiple avenues, and interestingly enough, the solution space is often much larger than we may originally think. And I think that was actually quite visionary because it is actually turning out to be true.

[03:52 - 05:57] Mary Merlino: In terms of the NSF data net programs and goals. This is directly from the solicitation. These are not my words, but the NSF initiative is designed to provide systemic, long term preservation, access and analysis capabilities. Engage at the frontiers of science. We want to advance science and serve as part of an interoperable network. Again, going back to that idea, there is no one way to create cyber infrastructure. This is a living, growing, very amorphous thing. The requirements for the solicitation required-, nd this, this is very interesting. This is the first time, to my mind that NSF specifically said, "You libraries, you must be involved." And it's very interesting to me that to date, all of the awards have been led by the library community as opposed to the technology community or the scientific research community. So this is an acknowledgement on the part of NSF of the important and very necessary roles that traditional libraries and library skills bring to the problem. So we're looking for people to combine-, or looking for program solutions that combine expertise in library and archival sciences, computer computational information sciences, etc. Bringing together heretofore unrelated disciplines. Develop models for economic and technological sustainability. This is a huge push on the part of NSF in particular in the past five years. Everybody wants to know how you're going to sustain your grant. How are we going to sustain this network once it is developed? And then finally, of course, they want us to work cooperatively. So our particular, our particular grant the Data Conservancy is one of the first two awards it was actually awarded last year, and it is a five year, $20 million award. These are big, big awards. It is led by Saeed Chowdhury at the Sheraton Library at Johns Hopkins. The second award was Data one, led by Bill Michener at New Mexico University.

[05:57 - 07:57] Mary Merlino: And the next round of three data net solicitations will be announced hopefully by the end of this month. Our particular ward is a network of domain scientists-, a domain scientists, information and computer science researchers, enterprise experts, librarians and engineers. Actually, as I look at this slide, I see I should have put librarians first here. These are our partners wonderful partners from Cornell, University of Illinois Marine Biology Lab and other very important and very energetic institutions. We have a group of international partners. And the goal of our particular project is to support new forms of scientific inquiry and learning through the creation, implementation and sustained management. Going back to Suzanne's question about-, I think it was Suzanne who asked about sustainability of an integrated and comprehensive data curation strategy. And this last, this second bullet. You know, I talk about this a lot. You know, this, this this very noble vision of data curation is not an end, but rather a means to collect, organize, validate and preserve data to address the grand scientific challenges of our society. And you-, I say things like this. And every once in a while I say, is this real? Is this pie in the sky? So this morning and some of you may have seen this. Has anybody seen this article from today's New York Times? Okay. It really-, it brings chills to my spine. Rare sharing of data led to results on Alzheimer's and basically what they've announced in the New York Times today is it's a story behind the story that was announced earlier this week. Early, earlier this week. They've identified the, the biomarkers which we think contribute to Alzheimer's. How did they do it? They did it through cyber infrastructure. And I'm just going to read one-, two sentences from here.

[07:57 - 09:53] Mary Merlino: The key to the Alzheimer's project was an agreement as ambitious as its goal. Not just to raise money, not just to do research on a vast scale, but also to share all the data, making every single finding public immediately, making every single funding public immediately available to anyone with a computer anywhere in the world. This is cyber infrastructure. This is the result of cyber infrastructure. One more sentence. It was unbelievable, says Doctor John Cue Trojanski (ph) and Alzheimer's researcher at the University of Pennsylvania. It is not science the way most of us have practiced it in our careers. But when we realized that we would never get biomarkers unless all of us parked, our egos and intellectual property noses out the door and agreed that all of our data would be public immediately. So again, it is very-, it's-, it is it is just a miracle to me that this was announced that cyber infrastructure, although they don't say CIA cyber infrastructure, it works and it is real and it's encouraging me-, to me. And I hope to all of you. Okay. Back to my presentation. We are looking at-, over our first year, we are looking at identifying the technical requirements, the scientific and user requirements. And I'm going to spend a few minutes talking about that defining broader impacts. What do we want this project to do? And looking at that all important issue of sustainability. We are working with 3 or 4 different domains astronomy, life sciences, earth sciences and social sciences. And the way we are defining our user requirements is

really a mixed method. We have wonderful colleagues at UCLA led by Chris Borgman. Chris is doing a deep ethnic-, ethnography, a more traditional social science approach. Looking at the astronomy community. She's looking at how they share their data, what their work practices are.

[09:53 - 12:02] Mary Merlino: We at UCR are letting are leading the user centered design approach across life sciences, earth and social sciences. Specifically, we're looking at several questions. What are the data practices? What are data management, curation and sharing practices? Who uses what networks? Why? Why are they important and what data-, what specific data are important to curate? For whom and why? Astronomy is a great exemplar community. Again, this is a group that the UCLA group is looking at and that the astronomers have really cracked early on the issue of data standards and practices. Our initial goal for the first year was to ingest astronomy data from a variety of sources, including the Sloan Digital Sky survey, into a preservation archive, connect that data and connect it with the existing services by astronomers. The work that I am most closely affiliated and really excited about is this connection of the hard sciences with the social sciences, and Rus early on showed the example of your of the hypothetical researcher. I think her name was Jane, and she uses one of the NCAR models, the CSM model, and what we are trying at NCAR. We are really on the forefront of trying to look at the impacts of on society, of climate change and the impact of bringing all this together. So we have one of our researchers, Patty Romero, who's interested in looking at megacities and the vulnerability and adaptive capacity to climate change. Now, this is really hard. This is really hard because first of all, vulnerability is one of those squishy kinds of kinds of notions. And Patty is doing a meta analysis of 100 studies and nobody actually agrees what vulnerability is. But even if you define vulnerability, think about all the kinds of data that Patty has to deal with on top of everything that Russell already showed you.

[12:02 - 14:53] Mary Merlino: I mean, again, this was a great setup. So we know how to run those models. We're doing those models. But on top of that Patty has to look at an array of hazards. So it's it's hurricane data tornado data snow and ice data temperature data. She's looking at different units of analysis and looking at, urban data. And this is, this is really the at risk data. There's demographic data, census data. But local, local rainfall data. And a lot of this, again, is the kind of data that has been sitting on individual researchers computers. So this is one of those grand challenges that we are trying to to address. We're also looking at broader impacts. Looking at broader impacts, educational outreach, and again, ensuring the widest community use and penetration of the tools and services and the infrastructure that's being developed .W orking closely with the wonderful people at University of Illinois School who, you know, University of Illinois Library School, they have a masters in digital data curation, which I'm sure some of you are aware where of we're setting up mentoring experiences and summer boot camps, field work, including some field work with, graduate

students working with our data scientists at NCAR. So what does all this mean for libraries? I encourage you, as I know many of you are already thinking about libraries as a distributed network. We are a network, many other networks. Data, as I think about data as collections. I love the the quote that Winston Taub from, JHU Data Centers are the new library stacks. data is services librarians making, making, handshakes with data managers, libraries, librarians themselves becoming data managers. And as we've talked about before, the new requirements for the NSF data management plan lots and lots of opportunities for libraries to reengage in the conversation in terms of how you might get involved. Again, I think awareness beyond denial is always the first step. Investigate curriculum and education programs. To my knowledge, there are four doctoral programs in the US that offer a degree in scientific data curation Illinois, North Carolina, Tennessee and Syracuse. Attend workshops. Lots and lots of opportunities out there, in particular the upcoming DCC meeting in Chicago in December. And finally, stay informed of the Data Conservancy and other data net project developments. I think my time is up. I want to thank my colleagues and you especially. Thank you. [applause].

[15:04 - 16:56] Greg Newman: Thanks Mary for that awesome presentation. And also thanks for us for the great setup. I think his setup was fantastic for all of these talks today let's see. So my name is Greg Newman. I work at the Natural Resource Ecology Lab at Colorado State University here. And it's a pleasure to learn from you guys more than even present. So I'm just super excited to be here. We received the National Science Foundation PSI team grant to develop a small one of many, as Mary alluded to, cyber infrastructure to contribute to these efforts. And this cyber infrastructure was built for citizen science, which is an open kind of perspective. It's the public participating in data collection and contribution and sharing so that scientists can then subsequently analyze their data and use their data for valuable scientific analysis. So today I'm going to talk about the art and science of multiscale citizen science support. I'll try to give you some guidelines for what we learned through the process. So to give you context, there are many Jains out there in the world. There's Jane, Sam and John as Ross alluded to. And Jane is one of many collecting data. In our case, we had a bunch of citizen science organizations collecting a variety of information about birds, bats, plants, and even invasive worms in the Great Lakes. So the public was involved in collecting these data. And so there was a wealth of community based monitoring programs and citizen science programs generating volumes of ecological data that we have to curate and manage and store. Each of these projects, each of these genes have different goals and objectives. Some might be collecting data just because they like to get outside and record where a bird was found.

[16:56 - 19:03] Greg Newman: Others might be looking at more global grand challenges, as Mary alluded to. So each of these people are in different domains operating across different spatial and

temporal scales. So Jane might be part of a BioBlitz here locally that only occurs one day. And they record all of the taxon-, taxonomic species that they find at a natural area. So that's just a very short time frame. And it's occurring in a very, very small spatial scale. But maybe Jane or Bob or Bill might be-, maybe Bill's involved in the Christmas bird count a Cornell project, that is measuring birds every day annually over the last, you know, 200 years or so. So the temporal scale changes dramatically through, through that project. So it's a large temporal scale over great spatial extent. So we have a lot of citizen science projects and community based monitoring projects operating in different domains, collecting volumes of ecological data across many time and space scales. Each of these have different data curation needs. So our goal was to develop a cyber infrastructure to support the variety of citizen science projects operating across the US today and actually globally. So-, and then finally, I once we did this, I wanted to share some of the lessons learned we had in that process. And it's a learning for all. So what we did is we built a website that sits on top of our cyber infrastructure in support of citizen science projects at multiple spatial and temporal scales. This website can be found at sit side. org, short for Citizen Science. And as Russ and Mary both alluded to, there's many cyber infrastructures and the goal and the key is to make sure that they're interoperable with each other and that the data collected by, say, Jane, for that BioBlitz is interoperable and accessible to a researcher who might be integrating data from the Christmas bird count and Jane's BioBlitz data.

[19:03 - 20:59] Greg Newman: So what we try to do is achieve that goal through sit side.org. We followed a user centered design approach as Jane-, as Mary alluded to at nikal. And we found that useful. And we used an iterative software development lifecycle. So those software developers in the, in the room might understand these different workflows. But basically we found we needed to be very responsive and iterative. We needed to do constant requirements specification. What are the requirements of the cyber infrastructure to support the citizen science organizations. And then we also had to iteratively do design, investigation, development, testing, maintenance support, robustness testing, performance testing so that the website would respond to the user's needs at a, at a reasonable rate. So they're not doing say, a semantic search across the web that just doesn't give the, the information to the user and it quickly-, in a quick response time. So what we did through this iterative software development lifecycle was build site side.org and we developed the website for the NSF grant, which was a three year grant. And we were told at the outset at the time that NSF had no intentions of long term support and could not provide that funding. So I think that's been talked about already today, and that's the challenge, is to make sure that these cyber infrastructures can find that long term sustainability and that support. But nonetheless, we pursued on our track and we wanted to build a cyber infrastructure to support Jane, Bill and Bob. So we built sit side.org using this, this software design lifecycle. And we created a cyber infrastructure that was flexible enough to

create web skins, because we realized the multitude of different citizen science projects required some flexibility and our ability to create user interface designs targeted at each of those needs.

[20:59 - 22:56] Greg Newman: And then we hoped to learn from our mistakes, and I hope to share some of those mistakes with you today. So what we did is we built a schema that was very, very flexible and realizing that we had to deal with birds, bats, frogs and bugs, we needed to make sure that we accommodated the data in these disparate citizen science projects. So our schema was very general in the sense that we, we realized that all of these organizations had an object found at a location, an area at some point in time, a visit to that area. And that core database schema has been useful to us in supporting the needs of a variety of citizen science projects. So-, and then for each of those organisms, attributes were to be recorded. So what we did then is augmented this core database schema with a whole bunch of database tables for metadata management, spatial data management, environmental data management, media management. A lot of people like to take photo verification. So the data that we think of data isn't always numbers. It might be a video, it might be a textual description of a site. It might be a picture taken of that organism that the citizen scientists found. So we wanted to enlarge our view of data to encompass a variety of different data types and formats. So anyway, we built this, this relational database management system as the backbone of our cyber infrastructure. And so visually, this amounts to-, what we did is we organized each of the projects on sit side.org into different projects. So, for example, the Great Lakes Worm Watch is a program that's a nonprofit organization run through the University of Minnesota Duluth, and they're having citizens record using mustard extraction. Actually, what it does is it shows the worms crawl to the surface because they don't like the mustard.

[22:56 - 24:49] Greg Newman: And so it shows the, the diversity and species of invasive worms in the forest floor. And this is an important invasive species in the Midwest. So this organization is recording these data. And so they built a project on sit side.org where they can contribute those data. And then other researchers, through meta analyses can pull those data down and run, you know, maybe a regional analysis throughout the whole upper Great Lakes ecosystem. So that was one of many projects using sit side.org. We've since had others looking at pika populations in response to climate change in the high country here in Colorado and the Intermountain West. And then we have people looking at zebra mussels and Eurasian water mill, Foyle and the Great Lakes and even invasive crabs in Maine. So the idea was to organize these data into projects on sit side.org so that people could be flexible enough to support different data collection protocols, but still standardized within those projects, so that meta analyses and data sharing can occur. So this is an example of a customized data entry form. These are created by the projects themselves not us. And so they log in and they-, in on-, in an online fashion, actually create through a wizard kind of the box

model that you can drag and drop boxes that I alluded to. They can drag and, drag and drop form elements to their data entry form, so that the user can then enter the data they need their volunteers to collect. So this is just an example of one of those forms. Then the data needs to be visualized. We talk about this kind of data lifecycle. There's the curation, the visualization, the analysis, the metadata. Lots of pieces go into this. And so what we did is we created a custom geospatial map application for this cyber infrastructure, which uses some open source capabilities.

[24:49 - 26:48] Greg Newman: You'll recognize Google Maps here. But it's actually custom on top of the Google Maps API, some, some custom development on outside so that you can augment the features of the open source API with your additional features. So what we did is we allow people to search by project by species. And this cuts across projects. So you can look at all like invasive tamarisk collected by all organizations, not just one organization. So it cuts across the data collected by the community as a whole. And then we wanted to generate automatic reports for these citizen science organizations and analyze those data so that the data can come back to them in meaningful, useful ways. Data that is stored and archived and curated but not used is in data that's useful. So our goal was to develop ways that they could actually answer questions. A good example of that would be the City of Fort Collins adopted our system for an amphibian monitoring project. And so they had volunteers collect and record amphibians, and they gave them CD-ROMs to, to hear the, the croaks of the frogs and identify them by species. And then they submitted that data. Well, the Division of Wildlife didn't know the, the magnitude of the problem of invasive bullfrogs in relation to the Woodhouse, Toad Hill and Fort Collins in the ponds in this region. And so the Division of Wildlife used the data that the citizens collected to learn and investigate the magnitude of the species diversity problem here in Fort Collins. So however, we built sit side.org thinking that we could organize data into these projects and meet the multiscale nature of the citizen science community's needs. But that was great. But it didn't quite meet all of our needs. And so some of our projects came to us and said, "Well, we actually need some additional features beyond what sit side.org could offer."

[26:48 - 28:42] Greg Newman: So what we did is we built the cyber infrastructure to adapt to that. So it also has in its cyber infrastructure system and its database, the ability to store information about specific websites that are tailored to specific problems. And so an example of this would be like the Africanized honey bees project that came to us from NASA. And so they were wanting to have beekeepers report hive weight data and pollination data to us through the cyber infrastructure. But this, these the nature of what they were asking for was very, very sophisticated. And so what we did is we built a targeted web skin, which basically changed the appearance of the web pages to make it look like an Africanized honey bee website that's focused and dedicated on that particular domain at

that spatial extent and time scale. And then we could then use the cyber infrastructure to very quickly create a website for that need. These are appropriate when the required features go beyond the basic functionality of your existing cyber infrastructure. And they were appropriate when a user base dictates that that need and when those funding available to support it. So an example of that would be, as I mentioned, the Africanized honeybee website. This was actually created online by myself in a couple of days, actually even shorter than that, simply by using the cyber infrastructure to dictate the menu items and the features that we needed for this specific domain. Another example that we recently got asked to do is have community based monitoring, forest health monitoring data be entered. So the Colorado Forest Restoration Institute here at CSU came to us to develop a skin specific to the needs of community based monitoring programs that are concerned with forest health, especially in light of our bark beetle problem here in the West. So we built this skin for that need. And we're actually we're just building it now.

[28:42 - 30:44] Greg Newman: However, so these skins and these projects sound great, but a paper I read that really rung true for me was a paper in the Journal of Association for Information Systems, written by Rives and Finn Holt in 2009. And they say, "The hubris surrounding new technological solutions for effective data standards, data sharing and solutions and cyber infrastructure development may mask complications experienced by developers, and novel platforms often lack the human resources required to maintain and upgrade this technology." So I think this speaks to the challenges that this whole conference is about and why CSI days are important is that the the, the, the, the need is there for CI and the need is there for cyber infrastructure support systems. And in my case, the need is clearly there for citizen science organizations needing these capabilities. But the, the challenge is supporting these systems and building them and not relying on out of the box solutions. But, but instead flipping the table and going from the need and building to that need, as opposed to saying what's capable and trying to fit the user's needs into what's, what's already capable. So I think today we can brainstorm collectively out of the box to say, "What are the needs? What do we need out of these tools rather than what, what existing tools work for what they work for today?" So that led us to our, our experiences led us to these guidelines. Follow a UCD approach. User centered design. Use iterative development. Stay flexible. Light on your feet. Ready to adapt. Create short and simple documentation. We could have a volume of metadata, standards and specs and schema designs, and that's great. But we got so much to do that we just want to make sure it's targeted and clear and simple documentation. Focus on metadata, avoid feature creep because everybody wants a new feature.

[30:44 - 32:57] Greg Newman: But that can bog your progress down and build capacity. The education of teaching students. I think that Mary talked about that, the institutions that actually have

programs that train students on how to develop effective cyber infrastructure systems are important. So we're building a class here that says, "Get control of your data." I think it's a catchy little name because we all need to get control of our data. And keep it simple. So with, with that, I'm a little bit running along here, but I'll finish with some recommendations that we learned. Start with a requirement specification. What must your cyber infrastructure system do? Organize data into projects. Use skins where appropriate. Allow custom data attributes by projects. Add volunteer management information so that user management, the security that Ross talked about, make sure that data come back to the volunteers and be used, and make sure the system is as participatory web 2.0 as possible. And standardize within projects. Use existing national data standards, link with the Global Information-, the global Invasive Species Information Network for example. We deal with a lot of invasive species data. Those are those international networks that were talked about, such as the Global Biodiversity Information Facility, for example. Keep humans involved in shared data or interoperable. Use social media where appropriate, focus on data sharing and data use and data reuse. Use human, human computer interface testing and make it fun and easy. So finally, to conclude, we found that systems were flexible enough to meet the needs of citizen science organizations, the ones we built, and that the standards we use using standards, controlled vocabularies and mutually exclusive attributes allowed for efficient data exchange, and that integrating program evaluation into the system improved our ability to track effectiveness. So with that, I hope I didn't go too long. I want to thank my principal investigator, Doctor Jim Graham, here at CSU and of course, NSF, who funded the research, and a million other people that I can't even list. So thanks, thanks for listening. [applause].

[33:10 - 35:12] Jessica Branco Colati: Well, good morning. I have-, well, yesterday I thought I had the very tall task of following Russ, Mary and Greg on this panel, knowing I have the very, very tall task of following them. And I am only five two. But I've been [laughs] I've been told that I play big at times, and I like to think that digital repositories can also play big in regards to data curation. Like Russ said about cyber infrastructure, we're not fully there yet. Repositories are relatively young, and they're still developing an identity in the CI space. And when I look at some of Greg's recommendations, which I'd like a copy of, I think we have so much more to do. But we've started and we're in a good place and we're moving forward. I'm going to speak very generally about digital repositories this morning. In specifically repository services, repository services provide architecture and platforms for object creation, management and use. They manage how data is stored, provide publishing workflows and persistent identification of resources. They support preservation planning, enable and enforce access policies, and protect intellectual property all, all while providing highly visible and reliable spaces for discovery, access and use. Repository services are being integrated into larger cyber infrastructure initiatives as they support sharing of scientific, institutional and

cultural data within and among institutions and across disciplines. As Dawn mentioned earlier, I coordinate digital repository services. For those of you who I don't know, I work at the Colorado Alliance of Research Libraries, and there are many members of our alliance and many members of our repository steering committees, working groups, study groups here. If you could just raise your hand if you work with the Alliance Digital Repository Services in some way to give an idea of who's in the room.

[35:14 - 37:24] Jessica Branco Colati: Hi, guys. At the Alliance, we work with our member institutions, with you all to develop and provide the technical infrastructure, hardware and software components of a shared repository service. These are some of the portals to the repository that our members are currently making publicly available. And all of these. There's a link at the end of my presentation where you can then connect to any of our publicly available member repository portals. And we offer assistance to the alliance with data migration and preparation following best practices and standards, as well as act as a support resource for members when they're developing their local collections, their local policies, their local workflows, in their practices. Our members are our users. So when we think about user centered design, we're working with the librarians and trying to develop the services they need to work with their communities. We work closely with these librarians to develop practices, functions and services that support their local digital repository and data curation goals and needs. A collaborative repository service like the Alliance has many benefits, including savings through cost and resource sharing, and the development of a base of community knowledge and expertise to draw on when any institution needs additional support and help. But in many ways, the Alliance Consortium initiative is no different in its needs and activities than any single institutions or inter into citizens repository service. I wanted to take just a moment to talk about repository functions without going into great detail. This is a diagram from the Open Archives Information Systems, the reference model that's usually referenced when talking about how we architect a repository service. And I just wanted to touch on some of these functions. Within each or within OAS. Each entity is broken out into much more specific functions. The functions are in blue preservation planning, data management, ingest, access, archival, storage, and administration.

[37:24 - 39:39] Jessica Branco Colati: But at a high level, this is what we're focused on when we're talking about repositories. And this aligns well with data curation activities. Russ mentioned earlier, and there was some question about how the cloud integrates in some of these functions. And we at the Alliance at least are looking at how to and I think many people are-, I know many people are. How to leverage the cloud and the viability of it and the legal considerations and using the cloud. And we're also looking at the capacity of data centers, as well as how to establish private clouds where we leverage our existing trust relationships. So I think that we'll see variations on the cloud as

we move forward and try to meet some of our archival storage access and administrative components. When considering the roles that repositories play in data curation, it's important to recognize the user communities and that make use of repository functions and benefit from them. For the sake of this presentation, I'll loosely categorize these users into three groups. Creators, curators, and consumers. Each has its own set of goals, needs, benefits, and expectations. In reality, there's often overlap among these communities. We've heard a good bit about user communities the citizen scientists, the researchers already. So I'm just going to briefly profile them. Curators are charged with establishing and maintaining repository services, frequently focus on the content as objects to potentially keep and focus resources on ensuring access and using data that's sustainable. As the vision of the self archiving creator dims, curators pursue and carry out the lion's share of deposits into institutional repositories, often through batching guest actions. This isn't what we originally envisioned, and this is one of the things we've learned in our repository activities. Curious creators of data are typically faculty researchers, citizen scientists, authors, departmental staff. Pretty much anyone who can create a digital object is a creator of content.

[39:39 - 42:03] Jessica Branco Colati: They generate data from their research, their publications, business functions, or as part of digital conversion activities. In our special collections archives and unique holdings. Creators can find deposit requirements either due to the software or the policy driven requirements challenging, but after deposit, they can benefit from the visibility, the reliability and the flexibility, as well as the reusability of repository offers to its data. In some cases, the opportunity to leverage new publishing opportunities while retaining copyright of intellectual work, or the ability to control access to data temporarily, permanently, or to a specific group of consumers is attractive to the creator. The ability to disseminate scholarly research to numerous discipline based repositories and aggregators is also attractive. The ability to assign a persistent identifier to a resource, as well as preservation activities, has long been identified as key benefits of participating in a repository. The added benefit may be that once you've coordinated with the curator, you as a creator are no longer required to maintain, migrate, or manage your deposits. Consumers. They're the ones with their heads on the laptop, maybe considered the traditional end user repository services. When we think about who's going to use this. They have a focus on discoverability and accessibility of repository data. Consumers of data found or managed in repositories look for digital content that's reusable and reliable, with clear rights and usage information available. They want to know what they can do, how, what forms it's in, how they can work with it to ensure that the needs of the creator, the consumer and the curator are all being met by repository services. We need to align those functional requirements with the data curation lifecycles. To understand. The curation life cycle of a digital content helps to ensure its continuity of the data, its integrity and its availability. This model of a curation lifecycle is from the UK's digital curation centre, the DCC, and it graphically

identifies the actions of curation from the moment of creation and appraisal through to migration, potential reappraisal and possible disposition.

[42:04 - 44:24] Jessica Branco Colati: Within this model, activities related to identifying, maintaining integrity, storage and use and reuse are noted and can follow-, can be followed to establish a sustainable repository service, and this aligns with the OIS model in the functionalities. So it's very graphical. And they-, there are several resources that DCC recommends that are used to, to be available to repository administrators such as Trac, Jamboree, Ada, Platt or Plato Nestor, to measure, record, assess and evaluate trustworthiness, risk, functionality, or organizational soundness of a repository service when it's performing these lifecycle activities, supporting the lifecycle activities, or providing the functionalities. These can be usually used in whole or in part as repository services. Look to improve their functionality, expand their services, just merely be established or reposition themselves within an organization. This is a lot for an organization to do when it's dealing with repositories, and I found a model that I think is really interesting to illustrate some of the challenges that I've seen working with our members, although this isn't one of our member institutions. This is the curation continuum as proposed by Monash University in Australia. Monash has spent a significant amount of time evaluating its repositories usefulness. It proposed the concept of the curation continuum into response to its findings, and understanding that repository relationships have impacts on the broader academic community's needs. Working from the notion that as information moves through its lifecycle, it also moves through different community's uses and expectations of persistence and quality. Monash identified continua between two end points of different aspects of repository service and the content stored there in. So as you look at, at the object level, we have less metadata to more metadata. Objects continually updated to objects that are static and moving all the way down through access, whether it's completely closed or completely open.

[44:25 - 46:42] Jessica Branco Colati: Along each of these paths, there's some point where local repository needs are determined and defined. This continuum concept is interesting to consider and can help other repository services, like the Alliance, make choices about where and how to store data, how to manage it, and what risks to tolerate. In considering how the different continuum can interact, some of our traditional approaches to managing analog information objects are challenged. For instance, more items in a repository, every bit of data, every record component of a data set that needs to be curated, will likely lead to less specific metadata. So it's moving on that continuum. The more objects, the less metadata, or the larger the objects, the less metadata that we may be dealing with. This continues to impact also not only how we design our repositories, but also our data store needs. I think we all would like to pile into Pat's car, get those slim drives, because that would be

much easier than dealing with the large objects and how we move them around. In their conceptualization of the repository service at Monash. They wanted to meet their emerging curation needs and some of these issues of large data sets and temporary data or not long term static-, long term attainable static data. But they didn't want to lose that data. And that was mentioned earlier today. So they went along these continuums and created dividing lines and use those lines to create a multi repository service to meet the creator, the consumer and the curator needs at their institution. This is their private, collaborative and public repository model. By considering that based on where and how the creator, the consumer and the curator communities interacted with each continua, Manish identified three types of repositories which overlay the progressions of the continua from large, numerous controlled access with less described data to more extensively described, more selective, more static in smaller objects in public repositories, and typically the public repository would be an open access space.

[46:43 - 49:00] Jessica Branco Colati: So I'm not sure how clearly you can see in the back. But the private research domains are where the researchers that Mary was discussing, as well as what Greg was talking about, citizen scientists, these large data sets are being stored in a research data store, and there's some question of whether or not you have to have three distinct repository data stores, or if we can, again, leverage an existing shared data store and use access controls and other functions of the repository to ensure that the large data sets in the static published open access deposits of articles and findings aren't-, oops. Aren't inadvertently co-mingled or subject to the wrong policies of management. So they've gone with three. But I think you could argue, as we understand more about our large computing capacities, you could use one and layer on top of it. Um, in considering sorry. In there. Sorry. At each repository intersection, there's a curation boundary where the curator mediates the movement of data from one community to domain to the other another, and it moves through its life cycle, its purposes and its uses are redirected. And that's a very human interaction point here. We have to have the curators working with both the creators and the consumers and the creators as consumers, especially in the shared research domain, where we have to add the authentication and authorizations, the administrative controls, so that these groups of researchers can work together. Um, Saeed Chowdhury, who as mentioned earlier is the Pi of the Data Conservancy project, has in the past talked about challenges of human interoperability, not just the data interoperability. And here it Manish, they've tried to insert where the data is moving across different domains or disciplines or spaces from the creator researcher to the scholar student casual learner. So if I take this Monash model and I bring it back to the data-, DCS data curation lifecycle,

[49:02 - 50:38] Jessica Branco Colati: I see that if we have something well defined and well developed and well supported, such as the modest repository service, it would address many, if not

all, of the aspects of this creation life cycle by acquiring, managing, versioning, identifying, migrating needs and, when necessary, disposing of digital information objects. Repository services could fulfill the curation commitment. We're working towards this. I'm keeping the curator and consumer's needs at the forefront is one of the most important aspects of the work that we're doing, and I think Greg talked about that. Ideally, repository services that provide mechanisms to create, modify, reuse and exchange descriptive and representation information about data, i.e. the various streams and types of metadata, as well as support preservation planning, can and need to anticipate interoperability with the future. What we're doing today is to the standards we have today, but we have to be flexible and consider where we'll be tomorrow. We also need to actively monitor and respond to the developments in changing expectations about long term use and preservation, as well as underlying software, hardware and infrastructure. Repository services can support generically depending on their design and architecture, the curation activities of data, large and small, scientific and cultural, and those activities in turn respond to the needs of the communities of users, the creators, the consumers, and the curators themselves. Thank you. This is-, [applause].

[50:50 - 50:00] Moderator: Questions, observations from the group. Anybody?

[51:07 - 51:46] Speaker 1: So this is to our guest, to the whole panel, the, you know, the great barrier to the sort of data creation you're talking about, I think, as Microsoft Access and Microsoft Excel, because that makes it so easy for anybody just to open up anything and start doing anything and not care about anybody else. And I'm sort of wondering sort of, how do you wean people off sort of those self-serving singular tools and start to adopt the kind of more global perspectives. What have you done to sort of make it worth their while to not use Access and Excel?

[51:50 - 51:53] Mary Merlino: That's a toughie. In my-, this is on?

[51:55 - 51:55] Moderator: Yeah.

[51:56 - 52:26] Mary Merlino: Okay. In my community, the atmospheric science community, we actually have a long history of data sharing. And you know, you can't study a slice of the atmosphere, and you aren't-, it just doesn't work like that. So I don't think I can really speak to that. That's not our dominant, our dominant paradigm. However. Going back to today's revelation behind the breakthrough, I mean, I want to think that these kinds of stories might be inspiring.

[52:31 - 54:03] Greg Newman: [unintelligible] there we go. [unintelligible]. Why not? Here we go. Sorry about that. In our community, we don't have a history of sharing with citizen science. It's a new field. And how-, well, I shouldn't say that. It's not new. It's actually been going on for a long time. But

from a digital data curation standpoint, it's new. And in a sense, our volunteers are welcoming for our-, for the most part, we found that they're welcoming, welcoming to the idea of going online instead of using a local Excel file. And the reason is that the usually retired folks who are somewhat new to computers and therefore, when trained, actually embraced the idea because they don't have necessarily the sophistication of the user who wants to kind of customize his or her thing on his or her access or Excel file. So we've actually had pretty good luck in having them participate. That said, there's a technology value with our users that we found through research that is preeminent, preeminent and problematic, and that is the use of a GPS unit. We need to know where our object was found because it was a spatial, temporal kind of perspective. And so we need to know where, what and when and by whom. And so the weather has been problematic because although they're, they're welcoming to a simple user interface online, they're less so adept at using a GPS unit. So those are my thoughts.

[54:07 - 55:41] Jessica Branco Colati: For us we actually have a lot of legacy data and metadata that we work to convert, and it exists in those Excel spreadsheets and access tables and depending because we're consortia and when we have the local challenges that we try to assist our members with, we encourage them to use the tools that they have available, even if it is those tools. And sitting at the table behind or to the side of you is our member Support Services coordinator, Robin Dean, recently hired. We're very happy to have her. And Robin and I will take those tables and those spreadsheets and we'll work with the owning institution to say, "Okay, how do we get from this metadata record that you have in your local schema or in your local tools, or in a standard schema, but organized in this delimited file to a viable digital repository object." And we'll use very basic tools mappings, transformations, exports to build those objects from that data. We also have web forms, and I'm really interested in Greg's web forms that will help them capture once we've worked with the legacy data. And just to add some fun, we have thousands of mark records where data was captured by the librarians for some early digital objects. So we've got lots of conversion going on, but it's a great opportunity to bring that data up to a common standard practice schema. And we have a normalized object model of what metadata schemas we're using currently in the repository.

[55:44 - 55:46] Moderator: Okay, well, one more question.

[55:49 - 57:29] Speaker 2: [unintelligible] guy at the Colorado School of Mines. Before I moved to Colorado, I was a social sciences librarian at the University of Wisconsin, where we had a member-, a campus wide membership to an organization which I think Jessica may know about. All of you may know about is that the University of Michigan called the Inter University Consortium for Political and Social Research. They are-, it's been I've lived here for 15 years, so I've sort of lost track of what

ICPSR is doing. But I believe ICPSR is still very much in the business of archiving and making available massive data sets in numerous forms. I know when I was involved with them, first, we started all of our data on IBM cart, and that's how long ago that was. But one of the initiatives that they have that we're working on, and I believe they probably have been around for 30 years now, maybe longer is standardization of formats and making data available in a usable format, whether that would be raw data or, or in SAS or SPSS format or, or whatever format researchers need. But I just wanted to bring that up. What I noticed, and maybe to Colorado, is that I see ICPSR memberships at Lake FCU are often more based in departments than that are campus wide, which was the model in Wisconsin. So I don't know if ICPSR maybe is less well known in Colorado or not, but they do present a very viable model.

[57:30 - 57:47] Moderator: I will point out that through the Colorado Alliance, we have a consortium membership to ICPSR. So most of the libraries are members. And here at CSU we are and and yes, ICPSR is a great model, roughly 40 some years old for data archiving and distribution and social sciences. So-,

[57:47 - 57:47] Speaker 3: [unintelligible]

[57:48 - 57:58] Moderator: There is an ability to buy in to the membership, whether or not each institution has decided to. That's a different model. So-, okay. We need to get back on track-

END TRANSCRIPTION