An Update on Arches For Science (IPERION HS Workshop October 2021)

So what I wanted to talk about today is to give you an update on a project what we're calling Arches for Science, which is a project of the Getty Conservation Institute that basically tries to answer the question of once you have a handle on the kinds of tools that Joe just talked about, the semantic modeling, how to find and use a IIIF image. How can we start to assemble those types of tools into something that approaches we hope sort of a complete data management system?

I should mention before we dive into deep but, of course, I'm speaking on behalf of our collaborative group of authors to which you see on the screen there. And I also wanted to mention that specifically my colleague Dennis Wuthrich of our developer Farallon Geographics is here on the call with us to answer any technical questions you guys might have about what is about to be a lightning fast preview of what's in Arches for Science.

So those of you that have heard about this project may know it from some work that we did about five-ish years ago now where basically to start thinking about digital data management and how to build a management platform. We started by going and having site visits at a number of organizations and institutions and individual researchers notably on both sides of the Atlantic to try and get a sense of what folks research and data management needs and desires truly were to make sure we understood the scope of the field.

I'm not going to go into this in too much detail, but you can see on the right side of the screen there a bunch of the sort of types of institutions, types of work, types of projects that we tried to have sort of represented in this group while keeping the number of site interviews relatively tidy.

So what did we learn from all of the interviews we did? This year won't be surprised to hear that regardless of sort of the size, scope, scale and resources of researchers in cultural heritage. We all fundamentally kind of want the same things. That is to say, we want to be able to use data. We want to be able to interpret data, whether our own or other people's. We want to be able to reuse it. We want to be able to make comparisons between objects and develop correlations between objects over time between institutions. All of that kind of data sharing and data use kind of material.

But on that note of data sharing, we also heard quite a common refrain that there's a lot of concerns around what access means for the types of objects we all study. So we have a lot of comments that we heard about the need to control access, whether that means understanding who can see what pieces of data, determining whether an entire institution that's sharing all of their scientific or cultural heritage science data or some of it, whether they're sharing sets of data with a specified group of people or the public that kind of thing.

And it was really important just given the cultural sensitivities around some of the data we work with that those processes, those access processes be able to take into account different institutional and cultural practices and policies. The other thing we heard pretty consistently is that the fundamental sort of pain point or pinch point before being able to really fully address data use and data access questions kind of came down to data management.

That as we heard that many institutions of, again, different scales and scopes have-- not all institutions have a uniform way of storing, finding, or sharing their own data even within their own institutions. And so that we can correct that and provide more tools for data organization management. We may be able to do things like enable data sharing, enable data mining, enable more granular control over data access. And importantly, it would also potentially allow us to minimize data loss when something happens like a researcher leads an organization.

So if you'll forgive the metaphor, basically, what we heard is we needed to spend a lot of time thinking about the sort of underlying plumbing of how we store our data to be able to do new and exciting things with. So that's really what the GCI's project was trying to build out was that kind of underlying plumbing.

As you all have intuited from the discussion of modeling data that we just heard one of the first steps to being able to better organize and manage our data, however, is understanding how to break down the idea of a conservation or heritage science project into constituent parts that you can build a data model around. So just as a very quick example, if we're talking about the types of studies where we take scientific data on an object at some point, whether that's a painting we can see here. They're getting a fabulous Jeanne image by Manet or a sample mockup that kind of thing.

The first thing to understand is that you need to be able to break that sort of idea or object down into some concepts that is this painting is a physical thing that exists in the world. It's part of a collection or more than one collection that is it's part of a Getty's actual museum collection, but it may also be, for example, part of a Manet study set that was part of a larger project.

The concept of a person is also kind of deeply embedded in a physical object in the sense that a person painted this-- Manet, right? But it also depicts the person. We know the sitter for this particular portrait was Jeanne Demarsy who is herself a person. So you have to kind of break these parts of a project down into these constituent concepts. And you have to be able to do that for all the parts of a heritage science research project, whether that's the people that do the work, the instruments, or scientific observations that we make, the samples that we may remove from the heritage object. And frankly the metadata around how that sample was removed.

The images, digital files and spectra that those kinds of things that result from it, and all of the amazingly well received digital products that we hope come from our work, such as publications and presentations. So all of that needs to be linked. But as we've said, it's not just those concepts, it's the relationships between them that need to be understood and model in order for us to recapture in our management system the intricacies of our work.

So that's really what Arches for Science is trying to do. It's trying to build a platform that we can use to store, retrieve, visualize, and ultimately share our data through [INAUDIBLE]. And one that does, in fact, capture all of those connections that we just discussed. Importantly, again we have already heard today the whole project of Arches and therefore Arches for Science takes into account the fact that this is all made extraordinarily easier by the use of standards and open source software. Some of what you see here. So that can include things like an underlying semantic CIDOC CRM like Joe mentioned, controlled vocabulary, such as those from the Getty, which we use extensively and things like the IIIF image based handling.

And using all of these means that our data will be protected through time and accessible we hope into the future. The benefits of using a standards and open source approach are many. And everything that you see on the right side of your screen is important, but I'll let you read it while I talk. I just wanted to highlight

a couple of them for our purposes today. One of those is that notion of sustainability and controlling things into the future.

It's critical I think to remember that one reason we're building all of this is to help make sure that the data lasts over time and remove the risks of things like institutional data loss. If we can reduce those risks, we also then reduce the need to reanalyze data, move objects to say a single prompt analysis again those kinds of things who can help preserve our collections in that way.

And the other thing, of course, is this idea of openness and community that we believe strongly that we can get more out of the work that we all do if we can share our resources and data reuse becomes easier. So luckily for us to be getting these ideas and these tools are nothing new. There's a long standing project here are called Arches, which you can access more information about your archesproject.org.

It has tons of features, which I'm not going to get into for time. But what's important at a high level is to know that it contains a series of features that are kind of grouped into three major areas. The first of these is that it has a robust system for data import and data management, which is critical to what we're going to try and do with heritage science data.

It has a rich discovery and searching capability built into it, including because of that semantic standard underneath it all the ability to search for concepts. And importantly, for the practice and conservation CHS. It has built into it what we think of as workflows that make our life on the process of entering data much more straightforward. And I'm going to use the rest of my time today to kind of dive into that a little bit.

So in short, what have we built, what do we do? We created these workflows to help data entry for the most commonly encountered projects and the types of projects where you have a sensible thing that we do science, too. We keep these as simple as possible. I'm going to [INAUDIBLE] Joe's framing and say we do this in a way that hides as much of the complexity as possible because that just makes our lives easier.

So we can do things like start a new project, select this workflow and start entering data in simple things that look like a website that you've seen before where you see a tabbed view at top. You can do things like enter the name of a project or you can follow your institutional naming conventions to name the work that you're doing. You can see a couple of radio button and drop down menu, types of data entry areas where you can start building some of that metadata into the work that you do.

As we slide along this presentation, the next page of the data entry or the workflow allows you to or provides a free text box where you can start to enter again critical metadata for our kinds of work, such as, for example, including the rationale for why a project is undertaken to help contextualize data for future researchers. We can also do things like add metadata about who worked on a particular thing.

I'm going to dwell on this for just a second to point out that we can-- this is where we really start to see the effects of bothering to do semantic data modeling, where you can do things like in the search bar which you may or may not be able to see based on the size of your screen. But basically, to add a person, I started typing the first few letters of my own name, Catherine. And what you may see there is that two individual people named Catherine popped up, and I can select one of them and link that to the project. And you might expect that to work.

But the other thing you can do is search on a concept because of the semantic linkages that have been built into that concept of person. So that is here in this instance, I've started typing into that same search

bar the first letters of the word science. And what's popped up is a list of the scientists that work at the Getty Conservation Institute for me to select them. So you can search in a number of different ways, and that's true across the platform.

You can add objects say that many thinking to the project in a very similar way. So once you've created a project, the next thing you might want to do is tell somebody what you did to that project. So you may need to, for example, define the areas that you analyzed. That's very simple. This is going to all look like something we've all done in many different software packages over time. It's an image of an object with some spots marked on.

But really what I want to mention here is that this leverages everything that you just heard Joe talk about IIIF manifests and everything that comes with that. All the rich metadata behind the image itself is being used here because this is a triple IIIF system. Once you've indicated your analysis spot, we can start to upload actual scientific data.

So what that looks like is again just sort of the simplest data entry form we could come up with. In this case, the first step is to indicate whether this is the type of analysis where all we need to link it to is a discrete area on an object. So something that you do for a non-destructive analysis versus a second radio button that would let you indicate that this is something that you did on a removed sample so that we need to build in a link to a parent object, for example.

Going a little bit further into the non-destructive data upload process. We get to then start to say, well, I created a project. Let me tell you which project this analysis was done for. The system will automatically feed us which objects say paintings are part of that project and what you choose from them which we do. You can do the same thing with the instruments that were used. And then we have the ability to do just a really simple drag and drop of files and spectral data, for example, into the system.

Right now we have the setup to read ASCII type data. So anything we have an ask you form, you can drop here. So you can imagine dropping all 30 XR spectrum, for example, that you took from an object all in one fell swoop. From the next thing the system allows you to do is to select areas that you've identified as areas where analysis is done, which is being done for the screening here and then add group say 30 XR spectrum just uploaded.

Select which ones are associated with that area. And just by clicking these little radio boxes, you're able to link a spectrum, a specific data file to that analysis. And of course, like I said, we heard that everyone wants to be able to dynamically look at their data so we've built in viewers to do that. And I should note here too that you may see here at the bottom right the ability to include things like interpretations. The file parameters, file specific parameters that you use for this specific data piece of data, which becomes searchable and allow people to understand that.

So like us on a lightning fast work sort of what we've been doing, what's coming next? We're in the final stages of this project of testing all of the data models that we have. Testing those workflows, making sure they work the way we expect them to. And that will lead us directly with our first real implementation of the system, which will be here at the GCI. So be on the lookout for one of us coming and tell you whether or not this works the way we intend it to.

We also have a couple workflows that we're still working on including things that would help us track modifications to things like mock up samples. So think artificial aging that kind of thing so you can keep data linked to a specific external event happen to it. And workflows that will help us with reporting out the

results of what we do. And of course, we're always looking for better ways to visualize and search the data once it's in the system.
So I think we're definitely coming up on time here. So I'm just going to stop there. I just want to thank everyone who is a part of what we've been able to accomplish now and open the floor for a few minutes of discussion.
OK, thank you very much, Catherine. I will start asking something to Catherine about, who do you think should use or fill in data in Arches in the system?
The way we've been designing it thus far is for me to do it. That is to say the practicing conservation scientist or heritage scientist that we can do this as we go through the process of collecting data for a project. Legacy data, that's a whole different question and every institution is going to deal with that differently. But the goal is to make it a usable work day tool.
So you envision a situation where you will have somebody operating the XRF and on the same time having next to the desk computer with Arches and typing in stuff, right?
It could happen that way. Yeah. That's the goal. And that way that hopefully what that reduces is the need to sort of try and go backwards and try and get stuff into a system after you've kind of moved on to your next project. We're trying to ease that particular pain point.
And the system would have to be centralized because there might be different many different people working at the same time on the same project, right?
I might toss that question to you, Dennis, if you'd be willing to take that.
Sure. Thanks, Catherine. So the system is designed as a web based system, which means that you interact with it through a browser. And depending on how you choose to deploy it, it could be limited to just internal group of people. You could choose to expose it to external users or partners if you wish to. It's really up to you. But the point is it's a web based system so it's hosted on a central server.
OK, thank you. That's clear. And the last question regarding this system is, how much effort in time you think it's needed to code to-- I mean, the routine work on filling the project?
Million dollar question. I would say a lot of that is going to depend on familiarity. The first time I tried to enter data into our development version of the system, I was a lot slower at it than I am now, you know? Now it's getting pretty easy and so it does not take me any longer than it does to figure out where on my laptop. I want to create a file structure to put the data. It's pretty similar once you're familiar with the system.
And I think the reason for that is this ease of use that we've tried to emphasize throughout. We're trying to make it as easy as is possible with the complex types of data we used to get it into the system.
OK, so I will read-- I'm not sure that you can read the chat, maybe yes because it's the chat and not the Q&A session. So look, I was complimenting you about the tool and he's asking if there are plans in place to avoid that it will be yet another nice tool for heritage content management.
It's a good question. I'm sure the answer is not yet because we haven't actually deployed it. So we're going to have to use it and see what works. The hope is that as I kind of started said at the beginning that having this robust data management system in place provides enough utility in terms of being able to find our data in order to be able to share it to be able to share it with its context. That it will just become so valuable to us as researchers to use that we'll want to use it. Or if we want to add to that, Dennis.
Yeah, I think that is really kind of the key points. I do want to also remind everyone that one of the key things Catherine pointed out in her conversation was this idea of community and we're really, really

committed to building a community around the Arches platform. For a very simple reason and that is the technology on its own is just never sufficient. You need people to use technology properly for things to actually work.
And so while we think Arches and Arches for Science are going to be really interesting, and I think really, really effective technology tools. Ultimately, it's the building the community and connecting members of the community with one another for this to really be more than just another pretty tool. And we're really, really focused on that.
So just though there is also another question, but I wanted to ask you before you mentioned that it's a browser based. But it's a cloud based or it's locally installed in every institution separately.
It can be either. You can choose to deploy it on the cloud or you can choose to deploy it with your own network within your own network. But I do want to make this clear that you as an organization can choose to implement arches. And therefore, you control the way you deploy it and also the data that you put into it. So you really maintain complete control of your own data. It's not what it isn't, it is one-- it's not one system where everybody's data goes into it. It's your own system that you can manage and really deploy as you see most effective.
OK. So I will read also another question here and then I think we should go wrap it up. Is the file parameter box free text or is it possible to use fixed metadata schemas for different techniques?
Can I take that, Catherine?
Oh, sure, go ahead.
So the simple answer is it can be either. What we've chosen to do is deploy the file parameter as a free text box for a number of reasons. But really all kind of all boiling down to that particular implementation best fits the needs of the Getty Conservation Scientists. However, if you want to use a more formal metadata schema for managing file parameters, you're free to do that. Which is does exactly what Joe talked about a few moments ago, which is provide you with semantic data modeling tools.
In our case, visual semantic data modeling tools that allow you to define quite precisely the way you want to structure and model all your heritage data, including metadata around files. So you're free to use whatever you think best meets your own needs.
Yeah, and I think that's just a really nice note to end on is that this whole system really is designed to be open-source and adaptable. So each institution that employs it, if you want more specific data models for every technique, rock on, we can do that. But we're in the way we built it so far, we just haven't made that choice yet. So it is customizable by each institution and your needs.