ITS4US Data Management Plan Training Session Transcript (05-11-21)

Speaker: Kate Hartman

Welcome today to the ITS4US Data Management Plan training. I'm Kate Hartman, and I work at the USDOT in the ITS Joint Program Office.

Next slide please.

So here's the agenda for today.

We're going to go over purpose and outcomes, a very brief program overview, just in case people see this out of out of context, and then we're going to go to the meat of the presentation, which is the data management plan, which I will also refer to as a DMP.

We're going to go through a template with a focus on the overview in the sections, and then a few slides on resources that may be useful to you as you work through your data management plan.

Next slide please.

So the purpose in the outcomes is really to just review the DMP template that we will be providing to ensure that project teams can properly address the requirements and instructions in drafting their DMP's.

And one side note that I want to make is this is about documenting the data that is to be collected as part of your project. It is not necessarily to direct your project that focuses on data and that's it.

We may have some questions about that, but we're very aware that some of these projects really are focused on collecting, standardizing, using, figuring out data sources, and this isn't about documenting your project per say.

It's about the data that you collect in your project.

So again, if it isn't clear or I stumble over something that doesn't make sense, please ask in the chat box and we will do our best.

So hopefully the outcomes will be understanding how the data management plan fits into the broader set of project deliverables, and then a clear understanding of the various sections that USDOT is looking for.

Next slide please.

So next slide we're going to do a brief program overview, and I promise it'll be brief because most of the folks listening today have seen these slides before now.

You know, so the complete trip ITS4US deployment program is a multimodal deployment effort led by the ITS JPO and in partnership with the Office of the Secretary, Federal Highway Administration, and the Federal Transit Administration.

It supports multiple large scale replicable deployments addressing challenges of planning and executing all segments of a complete trip and you can see the various pieces and parts that make up a trip and that that the vision of a complete trip is that it's an innovative and integrated deployment to support seamless travel for all users across all modes, regardless of location, income, or disability.

Next slide, please.

And the goals are high impact integrated complete trip deployments, identify needs and challenges, develop and deploy mobility solutions that meet these challenges and needs, measure the impact of these deployments, and then identify replicable solutions and disseminate this information and lessons learned.

Next slide.

So the complete trip.

Phase one awardees. We have five of them. University of Washington, California Association of Coordinated Transportation, Heart of Iowa Regional Transit Agency, ICF working with Buffalo, and then the Atlantic Regional Commission.

Next slide, please.

We also have these deployment phases that folks are aware of.

We're currently in phase one, the concept development, transitioning within 12 months to phase two which is a design test and build phase, which then leads into a phase three of operate and evaluate, and then with the post deployment hope of a five-year operation after that.

Next slide please.

So here we are at the main portion of today's presentation, which is an overview of the data management plan.

Next slide.

So the data management plan is a document that describes the data you expect to acquire or generate during the course of your ITS4US project.

It will also describe how you manage, describe, analyze, protect, and store that data, and what mechanisms you will use during your project to share and preserve your data.

So again, this is about the data that you collect, use, and store for the project. It's not about your data project per site.

And again if this isn't clear or something is confusing please put your questions in the in the chat box and we will address it.

So this data management plan, like most of the other deliverables in this program, has a draft and a final. The draft is due July 26, with the final 508 compliant data management plan due August 23rd.

Next slide.

Uh, so the data management plan.

What is it? It describes needs related to protecting the privacy of users.

It assists future researchers and employers with understanding and using the data.

It is aimed at people with some technical background and data collection and analysis.

And it's a living document that we expect will be updated during the life of the project.

But the data management plan does not supersede other project documentation like the PMP, the ConOps, the human use and approval, or the systems engineering management plan.

And it needs to include all the details that will not be done.

It needs to include as much detail, even want details that you may not know fully until phase two or three.

Next slide please.

Uh, this again is something that may look familiar to most of you in the schedule within the phase one deliverables. You can see the data management plan is task three.

You should be working on it now with the final deliverable due, I thought it was in August? But maybe it's July.

There is the schedule and how it fits in with the other plans that are being required from these sites, the five sites.

Next slide please.

So the major components of the data management plan are listed here.

You can see data summary, PII information system security, context diagram, standards, metadata, data license, and the data summary is a summary of the types and nature, scope and scale of the data you are collecting.

You will then document all the PII data, meta data elements and how they will be handled, because personally identifying information needs to be handled very carefully.

You also need to document the system or systems you will be using for collecting, monitoring, and storing the data. As well as how the system will provide security and privacy controls.

Further, the context diagram you need to add data flows to from the ConOps. You also need to document any standards used for collection, storage or transport of the data that you know of.

You need to provide metadata to address USDOT needs, and we will get into that.

And then the data created, if it is covered under a documented license that needs to be addressed, and again, you may not know everything, but as I stated earlier, this is a living document, and we expect that it will be updated.

You need to, though put what you do know into the document.

Next slide, please.

So you can see that there are some inputs and outputs into the data management plan.

Some of the inputs revolve around the ConOps and the performance measures, as well as the context diagram from the ConOps.

What will come out of the data management plan or data definitions that impact the safety management plan?

Performance measures, systems requirements, the training plan, the ICTD plan, and the deployment briefing.

So it's a fairly important document that informs some significant tasks.

Next slide please.

So the deployment phases and the data management plan in interdependencies are listed below.

So in phase one it's the initial assessment of the internal and external data format and sources.

Potential PII is identified.

These data management processes should be clear and data agreements, though may not be confirmed yet, they should be identified. Same with the IRB requirements.

You may not know them all, but you should at least address the fact that an IRB may or will be necessary.

So then in phase two you have more information, and you start getting sample data that you provide to the USDOT and we can work with the sites in terms of what further information is needed.

You also have your data schema and metadata defined in phase two.

Data agreements should all be confirmed.

IRB requirements should be documented.

The systems are fully defined, and baseline data may start being collected.

So then in phase three you are actually collecting live data, streaming it, cut to public data source that USDOT will provide, and you will have a finalized DMP because at this point you should have all the processes in place and documented.

Next slide, please.

Ok, so I'm going to talk here a little bit about USDOT and its preference for open data.

So next slide.

So what is open data?

Open data is data that is freely available to everyone to use and republish as they wish without restriction from copyright, patents, or other mechanisms of control.

Technically, open means available in the machine readable, non-proprietary standard format.

Legally, open means explicitly licensed in a way that permits commercial and non-commercial use and reuse without restrictions.

So these definitions matter and that's why we wanted to put them in writing in front of you within this training.

Next slide, please.

So why is USDOT interested in open data? Well, there are a number of different reasons.

I'm going to start with the bottom one that is probably the most important. It's the law. Through the foundations for evidence-based policymaking act of 2018, Title 2. (Also known as open Government Data Act) we are required to have open data.

So it also allows others to build upon USDOT funded development work.

It provides transparency into development of resources to support applications and software.

It promotes collaboration on development activities.

And it facilitates sharing of common code across projects and deployments.

Next slide, please.

So USDOT data sharing.

Data, including data provided by partners from the project shall be provided for public access to the data collected by default, unless specific privacy, confidentiality, security, or other valid restrictions are identified and documented to the USDOT.

Some of the data must be made available to the public, at least at an aggregate level or an anonymized format.

Data rights for data generated, created, captured by project partners should be determined and documented early in the process.

Data must also include proper documentation and metadata, and I suspect we may have some questions around proprietary data that is part of a program and data collected under the USDOT government funding.

So if there are issues about that please put them in the chat and we can address, but basically if there is data paid for and if data is collected, paid for under a USDOT funding source, we require that it be open and shared unless there is some reason to not do that.

And if there are some valid reasons they need to be documented and approved by the USDOT.

Next slide please.

So now we are going to get into the details of the actual data management plan template that everyone should have seen by now.

Next slide, please.

So here are the six template sections. Introduction, project overview, data overview, data stewardship, data standards and a glossary of terms.

So now I'm going to go into more details about these 6 sections.

Uh ok section, do we skip over section one?

I think it's just the intro, so section 2 is the project overview. This section will provide a summary about the research project and its goals, as well as how the data helps achieve the USDOT research goals for the ITS4US program.

In this presentation, as I go over the sections, we're going to highlight the some of the subsections, not necessarily all of them, but the ones that are probably more concerning or difficult to understand, or may have some questions and concerns, and hopefully we can put any issues to rest.

So the first subsection in the project overview is section 2.1, called Change Control. It describes plans for modifications and updates to the data management plan and includes plans for how changes in any of the data will be logged.

So you just need to track what you're doing, and this change control section is where you describe how you're doing that track.

Section 2.2, relevant sources, and this is where you will list any reference documents or sources with information relevant to your data management.

And subsection 2.3, the data schedule, provides schedule documenting key milestones pertaining to the data.

Next slide, please.

Section 3, data overview.

This provides a summary of the data flows at a high level and documents all the different datasets planned for the system.

So there are two key subsections to pay attention to in section 3.

3.1, the data needs summary, which is a high-level extension of the ConOps context diagram showing the data flows.

And subsection 3.2, data overview, which provides a description of the nature, scope, and scale of the data that is collected and or produced.

Again in phase one you may not have all of this information. You can do what you can, rough out what you can, put notes that you expect more information, but you need to start thinking about these things.

Next slide, please.

So here is an example of an initial context diagram taken from a ConOps, so don't get worried about the specific words on here, it was just taken as an as an example, but this is something that we would expect to see in a data management plan.

Next slide, please.

So the data needs summary.

Again, here is an example. It extends the context diagram at a high level by providing a summary of the types, nature, scopes, and scale of the data expected to flow among the system entities, and it provides a single location for a high-level view for the data flows.

There should be a way to go quickly to look at to make sure that the data that you've identified fits in there somewhere and is moving and doing what it's supposed to.

There are a couple of examples of data flows. I'm not going to read them, but you can see what we're trying to capture here in kind of an easy to reference graphic.

Next slide, please.

So subsection 3.2 is a data overview and it provides a description of the nature, scope, and scale of the data that's collected and are produced. Each unique data set should be included in this section.

So here are some recommended elements for this section and having a table is a potentially good idea to manage and organize the data that you're listing.

So the elements that are recommended are dataset title.

A description with purpose, externality, value, and relevance to performance measures.

There's the type of data, the collection method, and the data file format.

So you have a table that is a graphic representation of one place to go to look for all the data or the all the types of data that you are collecting in your project.

Next slide.

So here's an example. Again, just an example. You can see how this could lend itself to a table format where you have the data set title, the description, the type and scale, collection method, and the data file format.

So fairly straightforward. And again, if there are any questions, put them in the chat box and we can hopefully address them.

Next slide please.

So changes that may happen that create unique datasets.

You can see here there might be different data types but the same data, so a user profile may have account information and a different user profile may have account information, time information coded as strings and not date formats.

Yes, you may have different aggregation levels, so on a project that is collecting sensors on weather, you may have 5-minute weather or daily weather and so those would be unique datasets.

And then another example is update in format. There may be a warning log and then an updated warning log.

So you would have to have different datasets and again these are from other projects.

Your project will have its own unique set datasets.

Next slide, please.

So moving on to section four, data stewardship.

This provides details concerning data stewardship, maintaining data quality, and safeguarding data.

And the subsections to highlight are data ownership and stewardship, access level, re-use, redistribution, and derivative products policies.

And then data storage and retention. And I'll go into a little bit more detail.

So next slide please.

4.2 access level, which is a brief summary of the different access levels for each of the different datasets, relating back to the context diagram where possible.

So you can see the subcategories here. Private datasets, access request, related tools, software, and/or code, and relevant privacy and or security agreements.

Next slide, please.

So what are private datasets? And you may have to restrict access on these.

So data may contain PII, such as Social Security number, personal location, medical conditions, anything that would personally identify an individual would potentially be restricted data access.

You could also have data that contains confidential business information. Delivery information and location for business and third-party data with licensing that can't be shared outside of the project.

And then another example of a restriction for private data set would be that the data contains any information that may threaten privacy or security of any individual or group.

Location of explosive materials or vocational private religious centers. Again, these are examples we don't necessarily think that any or all of the projects in ITS4US will use, but these are examples of types of private datasets.

Next slide, please.

This slide goes into some more level of detail on PII, personally identifiable information.

So, for example, non PII would be something like traffic count information and general trends on network conditions, date, time, and weather. None of those things would identify somebody in particular.

What could potentially identify somebody is if you tracked internet cookies, IP addresses, and vehicle characteristics.

An actual PII would be names, addresses, telephone numbers, vehicle identification numbers.

You can also have locational PII, such as GPS tracking information, which would be used to pinpoint where somebody is coming and going from and therefore be able to identify an individual.

You could also have roadway video data, video of faces, and in vehicle video which would be considered personally identifiable information.

And then sensitive PII would be medical records, Social Security, bank account, and passport numbers.

So again, just some examples of PII, non PII, and potential PII for you to consider as you create your data management plan.

Next slide, please.

So here are some of the PII challenges that we think may potentially come up.

So one of them is in survey data. You know, the issue is can you include PII data such as name, home address, home addresses, etc.?

Well sure, but what are you going to do about protecting it? It is clearly PII, and a possible strategy to address that is to get IRB approval and keep the data separate from the research data.

Another potential challenge is the GPS trajectories, which can identify where an individual lives, works, goes to the doctors.

So a possible strategy is to De-identify those sensitive locations. And there are methods to do that which, if you're not familiar with them, I'm sure we would be able to assist you by providing some technical assistance there.

Another PII challenge is personally identifiable information. Tracking of an individual or stealing their identity can be accomplished through stolen or hacked PII.

So you need to really make sure that the data you're collecting you need to have and it's not just nice to have. Or yeah, you think you might use it. It needs to be justified, and then you really do need to lock it down and protect it.

The last kind of high-level PII challenge is these agreements covering third party data.

It's often unclear how much or at what level a third party's data will be shared for a project, so you need to discuss this data sharing upfront and make sure to have a written agreement with the third party early in the project and documented in your data management plan.

Next slide please.

So the next subsection that I'm going to hit on is the re-use, redistribution, and derivative products policies that you need to document in your data management plan.

You have to assign open licenses to federally funded data and custom developed source code.

USDOT recommends using the Creative Commons Attribution 4.0 international. There is a link. Slides are available and they will be something that you can get and use the link if you're not familiar with it.

Suggested elements include the dataset title, the license or licenses used, and reasons for non-open license if applicable.

Next slide, please.

Data storage and retention.

So this is where you list all data storage systems that will be used to store the project’s data, with details of those systems, and specifying how long the data will be stored in each system.

Where possible, reference the data needs summary diagram to provide additional context.

And then you can see that there are these five subsections in there, and these are all things that may be easy to just write something, but I would strongly encourage you to think through carefully. And again you may not have all the information in phase one, but data can become like gold and without having clear storage systems, descriptions, cyber policies, backup and recovery, and data retention policies, it can turn into a bit of a mess. So think these things through.

Again, questions, concerns, clarifications, put them in the chat or ask your core AOR as you move along.

Next slide please.

A storage system.

So each unique dataset could be stored in different systems, and/or location, and be updated at varying frequencies.

So again, this is a chart. The suggestion is to use a chart table to track these things where you have the data storage system, the dataset files, the initial storage date, the frequency of update, and then any archiving and preservation period.

All listed in in one easily found chart. Again examples in there are from other projects.

Next slide please.

Security needs.

So confidentiality, availability, integrity, authenticity, are kind of the high-level security needs.

With respect to confidentiality, data is not disclosed to unauthorized users or systems. It's disclosed to people who have a right to it.

And again you need to document this availability. Data is available and functioning when it needs to be, and again it's up to you to think this through and document what that means.

Integrity. The data needs to be accurate and consistent to meet the system needs. How do you assure that?

And authenticity. The data source can be confirmed, and document what has been sent and received.

Next slide, please.

So section 5, so we're getting near the end.

Section 5 is on data standards. This section will discuss the standards that will be used for data, as well as detailing the support documents related to data analysis.

So the three subsections.

They are data standards, versioning, metadata, and data dictionary.

Next slide please.

Data standards introduction. So in this subsection you should provide details on data standards used for each data set as they exist.

So here again suggested elements. Dataset title, data standard, open or proprietary, data standard rationale.

Next slide please.

Ok, I forgot there wasn't a table there, but you can see based on some of the previous slides the format in a table or chart.

So subsection 5.2 is versioning. This is where you outline procedures for version control, document how older data will be updated if required, and document how data changes will be recorded for scheduled and unscheduled events.

Next slide, please.

Metadata types. So there are business metadata for discovery and licensing, and then there is technical metadata which includes schema, processing, impact log, static. And those need to be identified.

Next slide, please.

Ok, so final thoughts.

Data management plan challenges.

Ensuring proper amount of data is collected.

So, how are you going to deal with the data collection that may get disrupted by various items reducing the amount of data collected?

So one thought or strategy is to provide data buffering for both the before and after case data to ensure adequate data is collected. You need to monitor data processes for changes or disruptions. So you may have another strategy, but this may actually be something that with a little bit of planning becomes a non-issue.

So another challenge would be ensuring current data information is shared.

So sometimes data documentation lags behind collection which can cause issues with analysis and research on the data collected by the project.

So one strategy to address this is to have a set plan for updating the DMP and other data related documentations which include notifications to users working with the data. So USDOT will be able to help identify any USDOT users of the data and there should be a data stakeholder list that you notify when there are updates to the documentation.

Next slide, please.

So here we have a number of useful references for you to consider. Everything from the pre bid webinar to some NIST standards, and some other fairly helpful potential references for you to use.

And I believe that is it on this slide deck.

And here is the usual stay connected information.

We will go to questions now.