Cyberseminar Transcript

Date: April 5, 2017

Series: VIReC Corporate Data Warehouse

Session: CDW: Locating Its Documentation 2017

Presenter: Margaret Gonsoulin, PhD

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http: //www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///\\vhaslcnasres.v19.med.va.gov\l%20)

Moderator: Hi, everyone. Welcome to VIReC’s Corporate Data Warehouse Cyberseminar series. Thank you to CIDER for providing the technical and promotional support. Today’s session is titled CDW: locating its documentation 2017. This is the second part of VIReC's CDW Cyberseminar mini-series, A Beginner’s Guide to CDW and Relational Databases. The first session was presented last week and it reviewed the terms and concepts associated with relational databases and the CDW. If you missed it, you can see the archived presentation on HSR&D Cyberseminar archive. Today’s seminar will walk you through the documentation side of things. It provides 2017 overview of the offices, departments, and websites that maintain documentation on the CDW. Here to present this session we have Dr. Margaret Gonsoulin. Dr. Gonsoulin is a sociologist who studied at the University of Virginia and taught for eight and a half years in the California State University system. Since joining the VA Information Resource Center, VIReC, in 2014, she has focused her effort on the Corporate Data Warehouse. If you have any questions for Dr. Gonsoulin during the presentation, please send them in using the question box. I will present them to her at the end of session. And now I am pleased to welcome Dr. Margaret Gonsoulin.

Dr. Margaret Gonsoulin: Thank you, Hira, and thank you everyone for joining us today. Before I begin, I’d like to take a moment to thank Richard, Mark, Heidi, and Hira for all their help in organizing this session and for mentoring over the years.

And I’d like to begin today’s talk by reviewing a few acronyms. The first set of acronyms that I’ll discuss today have to do with the data and its documentation. The first one is VistA or the Veterans Health Information Systems and Technology Architecture, which is an information system that has many applications, and one of its main purposes is to store the electronic health records. The CDW stands for the Corporate Data Warehouse and it is a national level database that houses clinical, administrative, and financial information from the VHA. And then finally the word metadata is a word that you’ll see used when referring to documentation of the data itself. So why I put these three words together on the slide is because you’ll often see in the documentation that the contents of the Corporate Data Warehouse tend to be derived from VistA, and they’ll keep a record of metadata both for the Corporate Data Warehouse and also inside the VistA system. So we’ll be seeing that as we work through today’s discussion.

So a few more acronyms before we get started having to do with the offices and the resource center that are involved in creating the data and the documentation of the data. First is the Business Intelligence Service Line, or BISL, and this is the group of people who, among other things, architect the contents of the CDW. VINCI, or the VA Informatics and Computing Infrastructure, group that in part maintains a set of servers that hold copies of the Corporate Data Warehouse and provide data extracts to researchers in the HSR&D community, among other people. And finally, we here at VIReC, the VA Information Resource Center, are a group that in part investigates and documents the contents of the Corporate Data Warehouse.

So for today’s talk it’ll also be important to remember two concepts from last week’s talk when we talked about the two types of domains that exist in the Corporate Data Warehouse. The first is referred to as production domains, and these are sets of domains and tables that have been structured by BISL, the group that architects the CDW, to support their re-joining. In other words, they’ve done a lot of work on these domains. The second type of domain is called a raw domain, as you’ll recall, and it contains tables that are direct extract from the source system. VistA is the most common source system. But basically they pull in the data from the VistA packages or files, and they house them in the CDW without doing a lot of work on them as far as adding, linking, keeping things like that, that we discussed in last week’s session. So the reason why I’m reminding the audience about these two types of domains is because they are going to really give shape to how we tend to find documentation for those domains.

But before we begin, I’d like to ask the audience to answer a quick poll question about their primary type of work. Would you describe the primary type of work you do as HSR&D, quality improvement, operations work, or research with operations partners, or something else, and I’ll turn it over.

CIDER Staff: Thank you, Margaret. The poll is open, and we’re getting quite a few respondees. I’m going to let it run for a few more moments. While we’re waiting, Margaret, I just want to say that I had one person report in via questions that you’re too quiet. Maybe move the phone a little bit closer or, but it was only one person so I haven’t had a lot of complaints. I’m going to go ahead and close the polls now and share that out. And you’ll see that 29% answered health service research R&D; only 5% answered quality improvement research; 44% answered operations work/data analytics; only 5% research with other operation, with operation partners; and 17% reported other. Back to you.

Dr. Margaret Gonsoulin: Thank you very much. Thanks, everyone, for participating. Nice to know the kind of work you’re doing so that we can hopefully make better documentation. So in today’s talk I’d like to cover several different aspects of documentation. First, I’d like to spend some time showing you the various sites that hold the documentation of the Corporate Data Warehouse. And I hope that the volume has gotten a little better. I’ve turned up the volume on the phone. And second, I’d like to review the content of the materials found on each of these sites, and then sort of show the basics of how you might go about using the information that you are finding in these sites.

So for the first and largest part of this talk, we’ll review potential sites and sets of documentation as they exist for the production domain, which are more numerous than the options for the raw domains. So that’s why the bulk of the talk tends to be around this part, part of the CDW. So here on this page I have the URL to the CDW metadata report on their SharePoint site that shows you really the most current or up to date list of the available production domains that exist. And we’ll come back to this site as we proceed and look at it in greater detail.

But in general for production domains, there are four main sites that may be useful to you when you are trying to find documentation. The first one that I’ll cover is the work that we’ve done here at VIReC, and then I’ll move into the CDW SharePoint site, the Data Architecture Repository, and then the data portal.

So here at VIReC you would maybe be interested most in the two links that I have circled here, the bottom right-hand side of our home page. The URL for this homepage is at the top left if you wanted to come to it. So the first one that I’ll discuss is the link that is labeled Factbooks. When you click on this link, it actually brings you to our CBW documentation page as a whole where you’ll find a set of documentation. The first thing listed there is getting started using CDW, and that’s really a list, a list or link to all the Cyberseminars that, given for introductory purposes.

So if you’re new, getting started, want to be introduced to the CDW, all of those Cyberseminars that are in this series are listed right here on this page. Second, you see the Factbooks which are documentation of the content found in domains. Obviously, we are still working on this set of Factbooks, and it only includes some domains as you’ll see in a second. The third one is a new series called statistical snapshots, and we have the first statistical snapshot in there. We’ll take a quick look at it. And then you see domain layout and domain descriptions, and then following that what we refer to as sets of data summary tables. So we’ll go through that, and we’ll start with Factbooks.

So in order to see the current list of Factbooks, just press the plus sign to the left of the word Factbooks. And you’ll see down below the available Factbook for the selected domain, so the Consult domain, the CPRS Orders domain, and so forth. So I’ll bring you through examples from a variety of these just to illustrate the kind of content you’ll find in these descriptive books.

So in this case I opened up the Factbooks for the inpatient 2.1 domain, and I’m showing an example of the kind of content that you’ll find in all of the domain Factbooks that we have. You’ll always find, Part 1 will be about the kind of content in the Factbook. Part 2 will be an introduction to the domain. Part 3 is something that I’ve been calling the clinical context, and that’s a sort of background and definitions and maybe processes that happen in a clinical context that are relevant to the data found in that domain. So it’s sort of whatever seemed important about the data from the perspective of care, episodes of care, or clinical context, and things like that. Part 4 will take you through each table, give a description of the table as a sort of high-level description, and then go through each column. And then finally you’ll see a list of primary and foreign keys in Part 5 that will tell you about all of the connections that this domain has to other domains. And Part 6 will be for people who are trying to learn SQL and might like some tips and tricks that are using SQL in this domain. Then you have references and appendix.

So an example of what you would find in the column description might look like this. So we’re still in the inpatient Factbook at the moment, and I’ve gone down to the table called Inpat.Inpatient in the book mark on the left, and that table is labeled four for Part 4.20, the 20th table in the list. And then I’ve chosen to look at the column called DischargeFromService, and in this example you can see, you know, it has a basic definition or description of the field, a few notes about entries that weren’t expected, and then the values and their labels, and finally the VistA source information. So that’s a typical column description.

There are times when column descriptions will have more extensive notes that just defining various aspects of that column that were not so clear in its original definition. So in this care it takes time to give the definition of what does a value of ordering mean? What does a value of resulting mean? And it also shows that a majority of the time this field isn’t being used. Then it has, the majority of the values are NULL.

So let’s move on and take a basic look at what you might find in a statistical snapshot. So in this case, we are looking at the statistical snapshot for patient demographic information. It’s the first in our series. And an example of the kind of content that you would find in this statistical snapshot is here, an example from Table 2 where we look at the values based on patient sex. So, as many people are already familiar, many times there is inconsistent information about demographics inside a database. So in this case, we have the instances where there’s a singleton system value of male, then a stable consistent value of female, followed by two, a value we don’t expect, and only one of those, and then the multiple entries of sex are inconsistent for a single patient, missing, etc., and then as you go across the table you can see the sample size and the percentage for all patients and then for Veterans only. And that’s throughout for the selected set of demographics.

Third, let’s take a quick look at what you’ll find in the domain layout and the domain descriptions. So for the domain layout, this is an Excel spreadsheet that will list on the far left each of the domains as of the date seen at the top of the page, in this case November 2016. And then you’ll see to the right of that each of the tables, how many fields they have in it, and then primary VistA source for that table. And the domain descriptions are quite different. They are text based, high-level descriptions, so what you would find in each of the selected domains that are included in the document.

Finally, let’s take a look at a few of the examples for the data summary tables. So generally you would just click the plus sign to the left of your selected product. In this case, I am looking at the frequencies that were made for discrete columns or discrete variables inside the CDW’s various domains. So in this case I am looking at an example for the Appointment domain. And I would open up the Excel file that’s highlighted here in the pink box.

Once I open up that Excel spreadsheet, I arrive at an About page, or About tab inside the workbook and I can select my table of choice. In this case, I’ve selected the primary fact table in here, A-P-P-T, Appointment.

When I click on the name of that table, it’ll bring me to the correct tab, and I can go scroll through the options of discrete fields or columns. And in this case I selected CancelNoShow, and then for each of these discrete values, you’ll see to the right of it the value label that goes with that. And then as you keep scrolling to the right, you will see the frequencies for each fiscal year.

Another example from the data summary tables are the Record and Null Counts. In this case, I’m going to expand the list next to Record and Null Counts and then select my domain of interest. I have selected Inpatient here. Open the Excel file again. Arrive on the About tab, pick a table. In this case, I picked the Inpat, Inpatient. Go to the correct tab in the workbook, select a column of interest. In this case I selected DischargeFromService. And as I continue to scroll to the right, I can see the number of times that a null value is entered and the percent of all values that represents.

So let’s take a look at the other link that I mentioned at the beginning of the section on VIReC's documentation, and that’s the Researcher’s Notebook. So here you can see our current list of Researcher’s Notebooks. These are intended to be methodological or technical recommendations by a particular researcher for their particular purpose. So they’re intended to be used cautiously, if you will, by other users. So in other words, we would expect that each user would probably modify the code or information and methods that they find in these notebooks for their own purposes, but it’s sort of intended as a starting point.

So here’s an example of one of the Researcher’s Notebooks on dealing with multiple inconsistent values of race as they exist currently in the Corporate Data Warehouse. And as you scroll through the document you’ll find that it’s stepping out logic, explaining what’s happening in just sort of regular text, and then it is showing you SQL code with explanation of what that SQL code is doing. So these can also serve as a tutorial for learning more SQL and for people who are newer to SQL and are trying to increase their skills in this area.

So let’s move onto the second source of documentation that we’ll discuss today, four production domains specifically, and that’s the CDW SharePoint site. Here on the CDW SharePoint site, a user can find information for all of the domains and note every single one that they have so far will have a table contents, data origins, linking keys, the format that the data are stored in, and more.

So you would start out on the BISL or CDW home site. The URL is here at the top. And you could summarize either one of the links to go into the metadata for the production domains that you see circled here.

Once you click on that link for the metadata you’ll arrive at the metadata page, and here you have two main options. At the top in the pink circle you see an option for clicking on this link to launch the metadata report. But below, with the pink bracket, you’ll see other information that’s important to remember is there because it can be really useful. These are the release documents, if you will, that’s what I call them, for various domains. Not all of the domains have release documents here, but a lot of them do. And they are sort of high-level information about the work that got done to create that domain and gives you some basic information that’s really helpful.

So pretending now that I clicked to launch the metadata report, it will bring you into the metadata report page, as you see pictured here on this slide. All of the domains are up to date here for the latest list of production domains in the warehouse. And you’ll be able to scroll in alphabetical order to find the one of interest to you. Now when you’re here, you have two different options of different kinds of information that may be useful to you. The first would be to click on the name of the domain itself, and that will open up what is referred to as an Entity Relationship Diagram, or ER diagram, which is a picture of the tables and their connections. And we’ll take a look at that in a second. Or you can press the plus sign just to the left of the domain names to expand the information so that you can see the list of tables inside there, and then I’ll show you, you can step down drilling into the column and its definition as well after you press that plus sign.

So this is an example of me pressing the button, the link to open up the ER diagram for Inpatient 2.1. When I did that, a new tab will open and you’ll see the Entity Relationship Diagram as pictured here on the right hand bottom of the slide. Of course this is quite small, so in order to enlarge it, you would just click on the image itself, and it will bring you in closer so that you can read what’s in each one of these tables.

So here what we’re looking at is that each one of the tables represents a table in the CDW inside a box. Right? And these lines in between the tables are telling us that we can find connection between these two tables, Inpat.PresentOnAdmission connects to Inpat.Inpatient, and similarly it also connects to Inpat.PatientTransfer down on the right. So we would be, at this point, inclined to look for which of the keys listed in these boxes are going to allow us to connect these two tables that we discussed on last Wednesday. So after some searching I found out that the primary key in this fact table, Inpat.Inpatient, connects on the foreign key of the same name in PresentOnAdmission, and it connects to the foreign key in PatientTransfer by the same name.

So what if I had not chosen to look at that ER diagram but instead had pressed the plus sign just to the left of the domain name? Well that, as I have mentioned earlier, will open up a list of all the tables found in this domain, and so the very first column you can see there the name of each of the tables are views that exist in this domain. And then my third column over is telling me which is the primary VistA source associated with this table. Sometimes the table can have information from more than one VistA file, so this would be the primary. That’s why I’m saying that, the primary VistA source. And then as we discussed last time on Wednesday, the very far right shows a connection that we can open up to find out more information about which keys we can use to join the table of interest to us at the moment to other tables in the CDW.

So pretending like we were scrolling through that list of tables that we see in the Inpatient domain, we could arrive at a table that we’re interested in, such as Inpat.Inpatient, and then click on the name of that table to drill down even deeper and start looking at the columns and information about the individual columns in that table. So pretending that I did that, and then I scrolled around looking for a column I wanted to know more about. Say I finally arrived at DischargeFromService, I would click on what looks like this little blank white piece of paper to the far right to open it up and find out more information about that particular column.

When I open up this column level metadata, it will always have this form. At the very top you see the column name. Right underneath that, you see the CDW view that it’s from, Inpat.Inpatient. Beneath that you see the view version right in the middle, and then the VistA file and field after that. So it’s telling me this is the 20th version of this view and it’s coming from VistA file number 501, so note name 501, number 4502, and then the field that corresponds to DischargeFormService is called Losing Service in VistA and it has a number of 21. So it’s telling you where it came from in that source. And then finally at the bottom you’ll see a description of what this field means.

Let’s go to the third source of documentation for Production domain. The Data Architecture Repository, or the DAR as I call it, serves as a repository for metadata, which is just the documentation of data from around the VA. This includes metadata for VistA, which of course, as you know, is one of the primary sources of data found in the Corporate Data Warehouse.

So by going to the Data Architecture Repository, at times I’m able to help clarify the meaning of things that remain unclear to me when I look at the metadata that are available in the Corporate Data Warehouse’s SharePoint site. So for example, if I were in the CDW SharePoint site and I had been looking through the Outpatient domain, I’ve got an interest in finding out the definition of this linking key here called EligibilitySID. As you’ll recall from last week, linking keys in the CDW have SID at the end of their names, so that’s why I know that it’s a linking key. I opened up the documentation available for this column in the CDW SharePoint site using that, what looks like, that blank piece of paper. And I see the following report that you see here on this slide at the bottom. At the top I see the typical name of my field, EligibilitySID. Below that, I see that this field is, as I know, stored in Outpat Visit. And then I see a related file from VistA called Visit, a related file number from VistA, 9000010, a related field number, and a field name. So what I don’t have is a definition down at the bottom. And that leaves me with of no way of being sure about what EligibilitySID might mean unless I use the Data Architectural Repository. So I’m going to take this information about the related file name and number from VistA and go and find that field name and number in that file.

So here’s the URL at the bottom of this slide for the Data Architectural Repository. I’ve come here and go to the left menu. I would hover over VHA so I get the drop down menu that occurs, and then I would select VistA because I’m looking for metadata for VistA. Once I get in here, I’m going to search for the VISIT file and then go and search for my field of interest.

So I search for it by number, and I see the results down below after I press the Search VistA button. Remember not to press enter. You always have to press Search VistA or it doesn’t do anything. And then I select my matching file, which is the VISIT file, and open up the file level metadata report.

When I do that, the file metadata looks like this. At the top I see the number, I see the name, I see the Owning Package. And then as I look down, I start to see information about the fields that are inside this VistA file. So on the very far left you can see the field name; field number comes after that; the description of that field; help text, which is what people in CPRS see when they’re entering information that’ll eventually be stored in this place, in VistA; and then the, comes the file that it points to because sometimes it will point to another file to go get information out of that file. So I need to scroll down in the field name and the field number in order to find the field that I was interested in.

Once I start scrolling, I do eventually find Eligibility number .21 in that VISIT file. And I can see right away that it is a Pointer, which means it’s going out to another file in VistA to get the information to bring it back to the VISIT file. And if I clicked on this word, Pointer, it will bring me to the file that’s it’s pointing to and show me descriptions of what’s in that file. That’s all, it’s called Eligibility Code and it is file number eight. Then to the right, in the purple box, you see that I don’t really have to go click on the Pointer to find out what this particular field means because it’s defined for me here, and so you can see that in this case I’ve found that the help text select eligibility which best describes the patients entitlement to care for this visit be a little more helpful than its description, but they’re both helpful.

Just a tip for the future, sometimes the value labels can be missing from the CDW metadata, and on occasion when they’re missing they can be found by clicking on this magnifying glass in the Data Architecture Repository's metadata. So this may come in handy on a rare occasion, so you might want to remember that it’s here.

Let’s move to the fourth and final source of metadata or documentation for Production domain, and that’s the VHA Data Portal. This site will provide links to all of the places we’ve mentioned in this talk, so it’s really a great resource for that, and it’ll provide information about how to apply for access to CDW data once you know which of these data may be of interest to you and will work.

First, I want to point out where you would come on the Data Portal page to request access. As you can see, it is the second link here under Data Access. And then from the drop down menu you would want to select whether you’re planning to do operations work, preparatory to research work, or you're planning to start your research project. And that’s gonna define the way in which you request access to the Corporate Data Warehouse.

So also there’s lots of other really great information out here on the Data Portal because lots of offices are putting their documentation here. The one I want to take the time to point out to you is located under Resources, and then from the drop down menu I select Data Reports. And that brings me into a variety of reports, but the first set of reports, the one I want to point out, and this is the VHA Data Quality reports that have been so useful to many, many users at the CDW. So you can now find them by expanding the View Reports button, and you’ll see them there sorted by topic, so Death Data was first, Emergency Department second. But there’s the Lab Reports and the Test Patient's Lab Reports, and all these reports are now out here on the Data Portal.

Before we proceed to the section on raw domains, I’d like to ask the audience one more question. Which of the following best describes your role in the VA?

CIDER Staff: The poll is up now. You are receiving some responses. I’ll let that run for a few more moments. Things are starting to slow down, so I’ll go ahead and close the poll and share it out. And the answers are as follows: 14% replied research investigator/PI, 50% data manager/analyst, 7% project coordinator, 15% operations, and 14% other. Thank you.

Dr. Margaret Gonsoulin: Yes, thank you, everyone. So let’s move into a discussion of the documentation that exists for Raw Domains of the CDW. So I put here the page that pertains to documentation of CDW as it exists on the CDW SharePoint site, so the direct URL is here at the bottom of the screen. But I’ll also show you how to work your way there from the home site. So we’re going to explore the content that we see both on the CDW SharePoint site, and then once again we’ll go and take a quick look at how you would use the Data Architecture Repository to help fill in information about bits of source data.

So from the CDW SharePoint site you would first want to go to the top menu and press the down arrow next to the main menu item labeled CDW. And then under Community tab, you would want to select CDW Raw. Then you would scroll down a little bit underneath the National VX130 table list to find a table containing a list of the “regular Raw Domains.” You could scroll for any topic that you’re interested in among the list.

In this case I am choosing to look at a report or a domain that contains oncology data. By clicking on the name of the domain itself, I’ll open up a report. At the top of the report you can see that it’s called CDW-Raw Oncology Extracts - v1. It has the date. And then you can find a variety of information about this Raw Domain inside this report. What I tend to be looking for first when I arrive here is information about which tables are included in that domain.

So I scrolled down until I found Oncology Tables, and then I start to look as I see there are two of them, and I see that they’re oncology\_patient\_160, oncology\_primary\_165\_5. And what this naming convention is telling me is that in VistA I will find a table called oncology patient and its file number will be 160. So I can use this information to go back into the Data Architecture Repository and get all of the metadata for the field inside these tables.

So just as a quick example, I am here at the Data Architecture Repository [xxxx 42:07] Lindsey discussed in the previous slide. In this case I’m entering in the name of the table and pressing Search VistA, and then I am selecting the matching file from the output.

And I arrive at file 160 ONCOLOGY PATIENT. And down below I can see each of the fields that are located inside this file, and I can see descriptions of them, help text. I can see if there are Pointers to some other files and so on and so forth.

I’d just like to take one second before we close out this talk to mention that you will find instances in the CDW that are not forced from VistA, and as you see here among the raw domains, TMS is a non-VistA exact. In this case you would not be able to follow that process of going into the Data Architecture Repository. You would want to read that report carefully, seeing where you might go to find out more information about a non-VistA domain.

So to summarize, there are a variety of ways of finding documentation for the CDW, and it depends on whether you're dealing with raw or production primarily. However, even all of the information and sites that I’ve covered today don’t always do the trick. So sometimes I end up looking at handbook from the VHA or directives or just the software document library or all sorts of sites looking for the answers I’m seeking in terms of what the data in CDW really means. So just to sort, you know, to keep that in mind. And now I’ll take any questions if there are any.

Moderator: Alright, thank you, Margaret. Yes, we do have several questions in the queue. To the audience, if you have any additional questions you would like to ask Margaret, please send those in. We still have 13 minutes left in our slot right here. Alright, Margaret, first question. Is it correct to say that production domains will have no missing data values and raw domains would have missing data values?

Dr. Margaret Gonsoulin: No, they both might have missing data values, so equally likely. You know, the population of a field is really occurring from the CPRS guide or from the administrators who are entering their administrative data. And so some fields may be used more or less also at various stations. So I think there’s all sorts of factors that come into play with the percent of null in the field. But I’m not sure that there would be a pattern between raw and production.

Moderator: Okay. Okay, the next question, I think the person meant to say Factbooks, but they wrote are fact tables available for viewing for non-VA people?

Dr. Margaret Gonsoulin: Well, they are located on the intranet site for VIReC, so without being behind the firewall, you know, you couldn’t, you couldn’t find these Factbooks. So, sorry, not at the moment.

Moderator: Okay. Alright, the next question is about ER Diagrams. Are ER Diagrams for, are there ER Diagrams for links between domains or between production and raw?

Dr. Margaret Gonsoulin: No. There aren’t. Raw data is housed on a different server from production data. Plus, in general, remember as I was saying that the raw domains are not architected in the same way that the production domains are. And what I mean by that is the production domains have all these linking keys that are added to them, whereas the raw don’t. So when you try to join tables in raw domains, you are using what’s called an internal entry number and a station number to do that joining. So when you’re trying to join tables inside the production domains, you’re using all those fields that end with FIV, the surrogate ID, so there would never be an ER Diagram that would show you which combination that the IEN and Sta3N in the raw, you would need to go over to the SIDs and the production on a different server. So there are guides that CDW have written for working with that raw data, and it is treated, you know, quite differently whenever you try to work with it in terms of writing queries and things like that, too, so I think it would be much more of a extensive process of joining.

Moderator: Alright, thank you. Another question about the ER Diagrams. Is there a good way to print those? Some of them are so large that they don’t fit on the page and everything gets cut off.

Dr. Margaret Gonsoulin: Well, short of having a very large printer I don’t know if there is a way. I have never seen anybody doing it either, so I [inaudible]…

Moderator: Okay.

Dr. Margaret Gonsoulin: ...find it I’d like to know as well.

Moderator: Alright. Okay, so this next person I think was looking at the metadata reports. I, this person wrote, I did not find Inpatient 2.1 Diagram 1 of 3 on the list. Has this diagram been renamed? I think there are diagrams in there that are called Inpatient 2.1 Diagram 2 or 3 and 3 of 3, but this person isn’t seeing the 1 of 3 one.

Dr. Margaret Gonsoulin: Yeah, it probably just got taken off. They were updating the metadata report, and they have updated the Inpatient Domain to 3.0. So now you should see Inpatient 3.0 on there instead where they have made upgrades and changes to some of the structures of the table, which means that now, you know, we need to go back and update the corresponding Factbook. But this is one of the things that sort of happens over and over again because really, the Corporate Data Warehouse is still in the process of being built. It’s not a finished product. And sometimes there is a need for changes, and so yeah, you’re absolutely right. The Inpatient 2.1 has now been replaced with Inpatient 3.0. And that’s where you’ll look for the...

Moderator: Alright.

Dr. Margaret Gonsoulin: ...current diagram.

Moderator: Okay, thank you. Is there information on any of these sites you mentioned about how to get [Inaudible 49: 49] SAS to access CDW, either outside or inside VINCI? Are there instructions here on how to request access to VINCI?

Dr. Margaret Gonsoulin: So I think, I believe VINCI has some, some guides around SAS and in the movement of data and then requesting access will be on the date, on the data portal that we covered. So it’s sort of, I think the answers are a little bit all over the place. I don’t, I would definitely go to VINCI for their stuff, guides. I think they have some nice ones. And then, I don’t know if, you know, you can just find general guidance for writing a Proc SQL statement inside SAS if you’re looking for that. I usually just Google things like that. And you know, otherwise you can find lessons on how to write basic SQL and then you can pretty well use those lessons inside your Proc SQL statement, too. So there’s sort of a variety of guidance, but if you’re trying to find information about working within VINCI, VINCI has all kinds of great documents on their site. And if you’re trying to do SAS, there’s tons of expertise out in the field that can be really useful in how to pass through that to get into SQL. A lot of colleagues out there can really help. I’m an SPSS user, historically, that [inaudible 51: 26]...

Moderator: Okay.

Dr. Margaret Gonsoulin: ...so yeah.

Moderator: Alright. Do you feel that there is documentation for every table or column somewhere in the VA?

Dr. Margaret Gonsoulin: If it is a VistA source, I do feel like it must exist somewhere. Sometimes I have a very difficult time tracking things down, and I have had some absolute inability to find a handful of things. But I assume that it is there somewhere. When it is not sourced from VistA, like if it’s coming from some kind of outside source, like you know, through a contractor or something like that, then I’m not sure whether, you know, you could be sure about the existence of documentation. You know, I feel a little bit less clear on how you would sort that out, but for VistA sourcing, the VA origin data, I feel like it must be there somewhere.

Moderator: Okay, thank you. Alright, we still have five minutes for questions and we still have a few more questions left. Are there any good resources on comparisons between different data sources? For instance, the CDW fee data versus a MedSAS v. outpatient tables?

Dr. Margaret Gonsoulin: Well, I know that there’s some really good investigations from HERC, and that’s where I would go looking. I’m not sure if they’ve done that exact comparison, but the Health Economics Resource Center, HERC, does spend a lot of time with things like fee basis data, and they have some really great reports on that. So I would go there first, specifically for that, that comparison. But I, not sure that I can speak definitively about whether there is one that they’ve done.

Moderator: Okay. This might be in your slide deck somewhere, but what is the website for DAR?

Dr. Margaret Gonsoulin: Oh, yes, it is in the slide. So there’s a URL, and I see a slide showing, see if I could quickly find that slide. So here’s the URL.

Moderator: Alright, there it is. Thank you. Okay, let’s move on to the next question. Is the raw data technically more up to date than the production data?

Dr. Margaret Gonsoulin: No. The production data is pulled in every single night. The raw domains of data are pulled in on different schedules. So from the CDW SharePoint site, when you see that list of tables among the raw domains, you can see, I went too far. You can see the update schedule is part of the report on the CDW SharePoint site. Let’s see. Okay, so here on this slide, you can see the schedule of updates is the third column in this report. So it’s telling you that beneficiary travel is updated on a biweekly basis. Then you can see right below that bill claims is updated on a quarterly basis every year. So they have different update schedules of those data in each of the raw domains. But I would say production would be more recent, more up to date than raw.

Moderator: Alright, thank you. I think we have maybe time for one question. This might be a little bit off topic, but can you provide any guidance about finding metadata related to CPRS access and entry use?

Dr. Margaret Gonsoulin: Well, I would go, I have some links that I’ve saved to, you know, the training for CPRS. I could send you that offline. But you know, as far as, you know, how the interface, what, you know, how the interface interacts with VistA, I feel like that is something that CACs tend to know, Clinical Application Coordinators. But you know, so I guess depending on exactly what kind of information you're hoping to find about it. But...

Moderator: Okay.

Dr. Margaret Gonsoulin: ...[inaudible 56:48] for a longer conversation.

Moderator: Okay. I like this next question that just came up. Is there an easy way to determine if a primary key or a foreign key is unique to a patient? Sometimes patients have more than one SID depending on the station they’re at.

Dr. Margaret Gonsoulin: Oh, you’re talking about patient SID. May, I [inaudible 57:10]...

Moderator: Yes.

Dr. Margaret Gonsoulin: ...at this moment. So use patient. There’s a great report actually, the reports that I pointed out on VHA Data Portal from the data quality group, did a really great report on dealing with patient SID and patient ICN. And it brings you through that process of, you know, how to use, sort out all the multiple, like if a patient goes to five stations, and they have five patient records or five patient SIDs. You use Patient ICN to help you bring that together into a single record. But then, of course, you’re going to have to deal with all the potential diversity and entries around different demographics and things like that. So, and that’s where those Researchers Notebooks tend to try to speak, try to speak to, but they’re also really wonderful data quality reports on those as well. I would say it’s a common problem and if, there’re are a number of things we talked about today hopefully can help with that.

Moderator: Alright, that’s all we have time for today. Margaret, thank you so much for taking the time to put together today’s session. For members of the audience, if your question was not answered or you have additional questions, you can contact the VIReC Help Desk at [VIReC@VA.gov](mailto:VIReC@VA.gov).

VIReC’s next session will be for our partnered research Cyberseminar series. Dr. Nasia Safdar will be with us on Tuesday, April 18th, at 12 pm Eastern, and she will present on her experience conducting a needs assessment for VA healthcare-associated infection prevention. We hope you can join us.

[ END OF AUDIO ]