# CHoRUS Bridge2AI Standards Office Hours-20241114_190349-Meeting Recording

November 14, 2024, 5:58PM

1h 5m 9s

Schmidt, Heidi E. 0:56
Hello.

Alvarez, Marta 0:59
How you doing?

Schmidt, Heidi E. 1:02
Pretty good.
Yeah.
Life is good. Busy with.
Kids and concerts and homework and.
All the normal stuff.
How is life with you?

Alvarez, Marta 1:29
Yeah, it's not good.
It's, you know, it's it's more chill in Salem Post Halloween, so I like that.

Schmidt, Heidi E. 1:42
I think it would be kind of.
It's kind of probably more fun because Halloween gets.
More enhanced I guess I would say.

Alvarez, Marta 1:57
Yeah, yeah.

Schmidt, Heidi E. 1:57
Sort of.
Sort of like my neighbor's street.
And Bainbridge, St. and Malden, if you want.
Every single house to have some sort of inflatable.
Character smoke machines and for them to like go all out and they have the police actually cordon off the streets so everybody can go for about 5 blocks and just trick or treat without worrying about traffic. That is the street to go.

Alvarez, Marta 2:13
Yeah, exactly.
It's awesome, but yeah, that part is fun.
The day I was always fun.
It's the lead up that's a little overwhelming, but the day of is always fun.
That's a great costume and lots of kids enjoy it, you know? So cool.

Schmidt, Heidi E. 2:47
I hope you got a lot of traffic on your street.

Alvarez, Marta 2:51
Not the traffic in my hood.
Yeah, it was good.
It's good.

Schmidt, Heidi E. 3:04
I guess I should come off of.
Being a vague circle with initials.

Alvarez, Marta 3:17
Always appreciated.

Schmidt, Heidi E. 3:19
Yeah, no worries.

Alvarez, Marta 4:05
Alright, I should mention that everybody responded to my e-mail and I have a list of people.
Or yeah, so I'll. I'll send that over to you.

Haughtaling, Jared 4:16
All right, very cool. Thank you.
That sounds great.
And then I can open the tickets to get external access and then in theory I can give you all access to an environment where you can start to upload dqd results.
And we can start to kinda have more interaction between data sites and the central cloud.
So that's that's great. Thank you.

Alvarez, Marta 4:51
Andrew, I just.
I added a like another for the agenda on the notes.
Just, you know in reference to the conversation we we've had several times, but on Tuesday when we talked about, you know just touching base with people to see if there was anything in particular they wanted to.
That would be helpful coming up to this data upload.

Williams, Andrew E 5:15
Thanks.
Since for four minutes and maybe I'll just queue up the conversation, we're gonna have today and touch on that second point that you put in the agenda, Marty, as always, we're really eager to understand if we're using your time spent on these calls and the time spent to.
Prepare for them in an effective way that's meeting your needs and.
We often ask for feedback about stuff we've done and for input on how to be more effective.
So this is just another.
Call for that to do so, either during the call or afterwards in whatever way you see fit. I think there's many ways to let us know.
After last week's office hours, there were, it seemed, as though.
There was going to be some benefit to go into more depth.
And review some of the things related to unit standardization and I think that's going to be the main topic for today.
But we're a little uncomfortable guessing at what it is that's most useful. We we we have a central perspective that gives us some insight.
On some things, but really we need your input to make sure that we're doing the work that is most useful to you.
So please don't be shy.
Let us know and we'll be responsive.
That's it.
And I'll, Jared, pick it away.

Haughtaling, Jared 7:27
All right, awesome. Thank you.
All right. So first, maybe I can preface this to say thank you to all of you who participated last week in kind of the question answer query.
Query session to go through different elements of the measurement table.
I've compiled the results and I think the first in the first section of the call, I'll kind of go through those.
I don't. I'm not.
Not with the intent of singling anybody out, but also to kind of hopefully go through and talk through a little bit like my interpretation of the results and what things and maybe try to highlight some things that either could be fixed or or.
Some, some just general guidelines based on the data that was returned.
So we'll do a little bit of that and then?
Like Andrew said, I'll talk a little bit about unit standardization.
And tie that in with the results that we got from from you folks on your unit specific queries.
But maybe.
I guess that's not.
I don't expect that to take the whole office hours call or even really close to the whole office hours call.
So what I have been working on this week that I that ties in together with some of what's been asked also asked is kind of this idea of a gap analysis and being able to understand.
What you're missing in your source data or things that you might not have been able to find and how you can at least quantify where those are or try to understand how much of them you're capturing. I think this is a very important task in the context of.
Creating ETL.
And So what I've done.
Tufts, at least, is I put together a dashboard that shows per domain data and epic versus data in omop at high level counts.
And you can search that dashboard for key terms like.
Theoplasm, or key drugs or things like this, depending on your domain and you can see what the capture is like for those terms.
And so this is a resource that that I would like to make kind of hopefully generalize and make kind of.
Reusable for you at different sites, especially those sites of Epic.
And so, yeah, at least all I don't know if how much of in a demo, if I'm going to demo the actual tool, but I wanted to kind of just introduce that idea of performing a gap analysis and understanding.
What you're able to capture in Omah versus what you're not, because of course the quality tooling in in Odyssey or from Odyseey on the Omop data looks very nicely at your omop data.
But it doesn't necessarily tell you the whole story of what the source, the underlying source information, was that either you're capturing or capturing inaccurately, or just flat out missing.
So that's the piece that we're trying to understand more of and it's particularly relevant when you know that there's a specific set of data like critical care data that you wanna capture and you have particular data elements of interest and you wanna make sure that you.
Are representing those in your source data appropriately in the output.
So without further ado, I'm gonna dive into the results from last week. I'll get into a little bit of units and then maybe we can continue a little bit on the discussion of gap analysis and I could potentially show something still not not fully decided if I'm going.
To show anything yet or not, but stay tuned for more there.
OK.
Let me share my screen quickly.
Can you all see this?
I'll make it a little bit better aspect ratio hopefully.

Alvarez, Marta 11:06
Yeah.

Haughtaling, Jared 11:08
So this is actually these are the results from the third query.

Williams, Andrew E 11:06
Yes.

Haughtaling, Jared 11:10
So I asked you last week to, for those of you on the call, I ask you to perform in order to execute three different queries, one of which looked at categorical values, one of which looked at units and one of one of which was. I would say a.
General query that combines the mean concept ID fields of the measurement table.
And so this is that third one.
And So what?
The numbers that you're seeing in these columns.
Are distinct concepts in your measurement table.
So these are the sites of course, on the on the in this case kind of Y axis ish of this heat map.
This is per site stratified and then this is the measurement concept.
That's the values concept.
ID this is your type concept, operator concept and unit concept and these are then distinct counts of your concept.
I, OS and so few things jump out at you. I guess in this one is mimic has no measurement mappings.
Actually everything is mapped currently to 0.
It's unmapped.
There's an issue with the handling of the custom vocabulary.
That's another thing. It's being worked on.
There's a couple other things that are interesting that I wanna point out, which are these two boxes in red?
So value is concept ID for both UCSF and University of Florida.
Given the number of distinct measurement concepts.
These are very, very small categorical response values for for that diversity and measurement concept. So I don't know if either of those sites have representatives on the call today, but I would be curious.
My interpretation is is it can be either due to one of two things. Either you've decided to only map measurements that have quantitative or or value based responses, which would be interesting, but I may be unlikely.
And the second is you.
You haven't yet mapped many of the categorical values associated with your measurements. Sorry Azure.
So you've come off mute, so maybe you can.

Williams, Andrew E 13:16
No, I I never put myself on mood. But but Kaspari had her.

Haughtaling, Jared 13:18
OK.

Williams, Andrew E 13:19
There was a hand raised.
So somebody was.

Haughtaling, Jared 13:21
Oh, OK.
Sorry I I great.

Balakrishnan, Kasturi 13:24
Hey, Jared. I just as soon as I noticed that I'd like something.
So our instance of comap is populated from DE identified caboodle.
And so a lot of our string based values are did.
I mean, we have a pattern there and so a lot of times when we map the measurements, we do go ahead and map the measurement values, but the value as concept ID needs to go through a process where we kind of evaluate what the values are.

Haughtaling, Jared 13:39
You.

Balakrishnan, Kasturi 13:54
And desidentify them before we unmask them as a value as concept ID. Because like in your analysis, if there's something like.

Haughtaling, Jared 13:56
Z.

Balakrishnan, Kasturi 14:04
You know, there's a color that could be the name of a person.
It it can be flagged incorrectly as Phi or P I I in our DE identification algorithm and then so our measurements are generally like if somebody asks for certain kind of measurements we evaluate that list and then go ahead and map them as value as concept.
ID like that's the process we followed. It's need based, so I'm not surprised by that, but we definitely need some, some some more analysis there.
But that's my.
Thought about we've we've talked about this at at our site.

Haughtaling, Jared 14:38
Awesome. All right.
Well, thank you that that absolutely clarifies that, because indeed 8000 distinct measurements, I would expect a good chunk of categorical responses for those.
So that's already interesting insight to have.
Maybe a follow up on Florida. So this is an interesting query.
Or is there? Results are interesting in the context of the other. The other queries that you ran and what I know about Florida is that like at least half of your categorical values are actually related to the Glasgow coma scale.
So this 38 is actually even a slightly inflated when you're looking at terms of distinct response values.
Most of these are like half of them are Glasgow coma scale comma 15 or comma 7 or things like this.
So Florida would also be curious.
Maybe it's a similar thing that that it has to do with the DE identification pipeline and you need to manually select or interrogate each string based response for a mapping.
But I don't if Divya is on the call or anybody else from Florida is here today, so.
Maybe not, but we can move on otherwise.
OK, so indeed.
So that I guess that was the 1st.
First big thing that I noticed.
When we talk about type concepts and operators.
Not really a whole lot there. I think having one type concept is also not so unexpected. Having ten, I would actually say is more unexpected.
That's a lot of different measurement sources, so I think a lot of that comes from our legacy data where we had very granular type concept mappings.

Some of which are not to non standard concepts, but in any case.
So I think there's not a lot to pull from there. If we look at the units.
I would say that the sites are relatively consistent on units.
I guess pit doesn't pit has fairly sparse sparse in terms of measurements, at least from this query.
But I think if you look at the ratios of of measurements to units, that's maybe one of the more consistent fields. I think the biggest discrepancies are probably in the value as concept ID field and then just generally the diversity in terms of how many distinct measurements you.
Working with and I guess some of that is coming from.
Some people are running.
I think this against a cohort while others are running it against their overall data sets, and I think we talked a little bit about that last week as well, so.
That's kind of my, let's say general take on this.
I can also.
Post this or we can we can share this.
Go forward Asure. Sorry.

**Williams, Andrew E** 17:36
You say you could post it.

**Haughtaling, Jared** 17:36
I was gonna.
I was gonna shift to the other tabs, but this is a good transition to to step in if you want.

**Williams, Andrew E** 17:42
Just a brief note on the type concept ID.
It's the main way. As you're aware, the, as you're all aware, that a lot of Providence of where the data comes from are preserved because.
Our project here together is really about multimodal data and highlighting the value of getting data from different sources about the same thing.
It's going to be increasingly important that we pay attention to how we are all collectively encoding.
That.
Rovenance in using the Tye concept ID O NLP derived will be one you know.
They'll be instances where the same kind of thing might come either from, you know, some algorithm we apply to a waveform, or an image, or something that comes out of a database that's automatically associated with the generation of those waveforms that might be complimentary.
We will want to be able to distinguish those and so.
As we have in the past said, we're going to need to kind of review the vocabulary for things like ICU type. This is another area where.
We have the coverage we need for type concept IDs for the different sources that are that we need to be able to distinguish, cuz we're gonna wanna enable queries that use that type concept ID to say we're it's not only the thing and the setting and so on.
But the origin of the data as a as a feature of interest in doing work with the data after it's been outputted.
So just a little note.

**Haughtaling, Jared** 19:16
One relevant thing that we've also we talked about a couple weeks ago in the measurement call is what what we should be able to distinguish between ideally is tabular laboratory data like Low Inc.
Data that's sitting in tables and nicely curated versus flow sheets. And so that's where you can have potentially the same measurement being represented both in a low Inc table as well as in a flow sheet.
But it's very important to make be able to make the distinction, especially if you're trying to define cohorts and you want one particular data.
Or another, uh, or data, I guess source.
So and I I believe and I I think where we had landed in terms of the standard concept for representing flow sheets was nursing reports.
But I can double check on that and I think I have those discussion thread about this, but I think that was the the type concept of choice when we had talked about this prior.
I don't know Paulina or anybody else.
You remember.
Confirm that, but otherwise I can just find it offline.
OK. But thank you again for contributing.
I think this is already very interesting to see and what for me I find interesting is this has sites that have not yet or that I at least have it. Not yet that I have not ingested data centrally from your sites, but it also represents something that's maybe.
A different look on your measurements than what I have seen centrally or like what I have computed centrally.
I should say so.
Hopefully some interesting insights there.

**Williams, Andrew E** 20:52
Last comment real quick.

**Haughtaling, Jared** 20:53
Yep.

**Williams, Andrew E** 20:53
I think going back to the type concept ID, if you are aware that you have either state or National death index.
State registry statistics as we do or national death index data as a complimentary source of mortality information that are being that are part of your own map.
It's that's a really crucial one, I would say of of the outcomes that matter in a lot of different studies.
Death is a big one and.
Completeness of data and accuracy of cause of death and other kinds of things vary a fair amount depending on whether you're drawing just on what's in your EHR or you've got a registry like that, a state or a national registry, and so if you're aware of that, just.
Be be sure that it's marked in there. Thanks.

**Haughtaling, Jared** 21:49
All right.
And I'm just following up quickly on.
See on one of Heidi's or Heidi's comment in the chat.
So Heidi mentions.

**Schmidt, Heidi E** 22:05
I.
I I saw the section in red, so I thought I'd give context.

**Haughtaling, Jared** 22:08
Awesome, yeah.
Thank you.
Yeah. So I guess my question then to you is in your measurement table, do you have both flow sheets and lowlink mapping?
So like are you mapping flow sheet data in there?

**Schmidt, Heidi E** 22:20
The majority of them are flowsheets and I'd say 27 millionaire flow sheets.

**Haughtaling, Jared** 22:22
Yeah.

**Schmidt, Heidi E** 22:28
And about four Millionaire Low Inc.
And I put that in the the chat.
After we did our last.
Session just so that people would have kind of like that high level context of.
What was coming from where?
At least from our site.
Help in that gap analysis.
So if we needed more flow sheets to be defined in order to get more information, then that would probably be something that would be handled either in searching with the tool or some sort of Delphi analysis I was thinking.

**Haughtaling, Jared** 23:13
Awesome Yep.
Thanks.
OK, I'm gonna jump over here quickly to the value query.
So this was the initial query that I had sent, which is more or less trying to get a gauge about what are the most common.
Categorical values in your instance and I I included four of the responses here. There is, I think one or two that I don't know if something got switched in the query or what, but I have excluded those.
So this is MCB pit. UF and Tufts.
Couple of quick notes here.
So first couple I guess for pats.
But Pitt has a very large fraction of your values conceptually.
You're actually map to O.
I see Mike is on the call.
I don't know, Mike, if if these are just by default zero is filling in your values concept ID or is this something that hasn't you haven't yet mapped, but you have like a source value that you're planning to map.

**Kelley, Mike** 24:18
Yeah, there are source values that were just working through the mapping for them right now.

**Haughtaling, Jared** 24:20
Yeah.
Awesome. OK, then that's, that's perfect. Then I think that that's that also helps you keep track of the the the missing mapping essentially. So that's that's to be expected.
These ones, though, will need to be updated, so these are.
I did a quick check in Athena.
So these are numeric values represented as categorical.
I think sometimes you could use this like as a temperature response or something. I think at least for the top two I didn't check all four, but they were both non standard so.
Take a quick look if you have a response.
That's like 37.4.
It's likely that that's gonna go into the value as number fields and your values concept ID will be empty because that is a numeric resp.
I'm guessing for temperature I don't know what it's actually for.
Weirdly enough, there are non standard concepts in CNA that are like numbers like this. So I understand why you would map it to that, but.
Yeah, if it's if it's numeric in the sense, especially with the decimal and everything and it's, it's probably going to be going to be a valuous number.
OK.
Quick note then for UF.
Of course, we talked about this in one of the prior calls, but.
There it is possible that you have identical concept names that are both standard concepts.
So what you have at UF is there's actually of all of the negative response measurements like 1/3 of them are mapped to.
I think this is the Low Inc code and 2/3 are mapped to the Snow Ink version of negative.
I think we talked about this a couple weeks ago, but.
If the and I guess if your measurement itself is low Inc, then definitely defaults to the low Inc response.
And I would by default if you if you have one of them in your set, I would try to consolidate and not have two different representations of negative. When you talk about.
Creating.
Pillars down the line and you wanna have a very specific measurement with a specific response value.
It's important that these are consolidated and so part of the process of that, we're gonna try to do where we provide additional guidance with standardizing mappings is is exactly this. If there are semantically very similar or even in this case, identical concepts that we guide you to choose.
The one that is gonna be consistent across the consumption.
So just a brief note there.
Interesting though.
Again, you see the Glasgow coma scale.
So I think that that's some targeted mapping going on in Florida to get that.
Those responses in particular.
Otherwise, always kind of interesting just to see what you're actually getting in this. In this values concept of E column.
So I'm gonna jump to the last tab, which is units, unless there are any questions on either of those first two or any other comments.
Percentages are also just to hide the truecounts, but they also are percentages out of the total of the top 10.
So it's kind of a weird percent, but.
Just for reference there. OK then I will jump to units. And so this last one, it's a bit of an interesting one. Takes a look at.
Your and I can maybe add an additional insert here.
Takes a look at your measurement concept.
IDs.
So kind of your lab test in this case and it isolates the ones that have the most distinct units mapped.
So we're not talking about distinct unit source values now.
That are unmapped like we talked about with pit. But we're talk talking now about.
Measurement concept IDs and then associated distinct units.
And So what we have here is I guess for both UF and MCB, pretty consistent.
There's not a lot of mapping or at least unit diversity amongst your.
Your measurements.
But when we get down into Columbia.
So first you'll see that Columbia has a good chunk. I think we've been over.
Is it 100 million hundred million measurements or so mapped to 0 and you as expected within that kind of?
Group you have a bunch of different unit mappings that make sense.
But then one that I wanted to mention here was referral lab results.
So we also.
See this. At least I don't know if Columbia's on epic.
I can't remember, but we see this referral lab at our in our Tufts data as well.
We have chosen to exclude.
These data points because at least as they are represented in Epic, we don't have any additional information about the lab that was taken.
We just have.
It shows up in in. in our case an epic as like this is something.
Unknown lab that the patient had and it has a row entry, but that's it.
So we we don't have this, just to know there's a lot of units associated with it. If you have somehow additional information that you that they could somehow tie this into a legitimate lab that would have then a a response that you could use.
Or if this ties with some other concept, I don't know the Columbia ETL or how this is working.
Maybe worth taking a look.
It's it's a sizable number of measurements.
But it is something that I've seen as well and maybe worth just a quick investigation on.
Umm.

**Williams, Andrew E** 30:18
Comment real briefly.
I think in general when you have something like this at any site, regardless of your EHR.

**Haughtaling, Jared** 30:19
Yeah.

**Williams, Andrew E** 30:28
There are probably stewards, sometimes more than one at your organization, who are whose job it is to know about this stuff.
People who are in charge of laboratory information systems or in charge of pharmacy data or in charge of pack systems or whatever it is, and I think.
Sometimes there's not a lot you can do just looking at your data and what you need. Your first question might be who would the organization could tell me this? And if you don't already have access to them, who do I need to help me arrange a a meeting?
With them, just to be really practical about how you might investigate something like this, sometimes it's not something you you do with code. Sometime it's something you do with a with a person who has the intimate knowledge of the of the relevant data source.
And sure, you all know that just kind of covering that in general. Like what?
What some of the options are and that's it.

**Haughtaling, Jared** 31:27
Thank you.
All right.
And so then the next set of highlighted cells were a pretty.
Large number of distinct units, although there are a large number of.
Accounts of these, but these are a lot of.
A lot of units associated with these different cell count types.
And so these in their low Inc code. I didn't this was I think a screenshot that I just transcribed, but.
They are a number per volume and so it could be the case that it's a lot of different volume sizes like number per microliter, number per milliliter, things like that.
But to have six with that like prescribed unit, I I would double check and make sure and look at your unit source values if that is really the case. Maybe you already have and this is a moot point, but anything typically above honestly above like 3 or 4.
Distinct units I would I would investigate and take a look and see what where that's coming from.
Because that's that's a. That's a pretty big diversity for a single measurement.
So just something to note there for the Columbia folks.
And in the last bit was kind of what I'll touch on a little bit where.
When you are trying to investigate.
What those might be?
What you can do and what I have done actually adds with the Tufts data is you can use like an aggregating like depending on the the SQL flavor you're using in data bricks you have a function called collect set that's going to aggregate all of the values in.
A row into the the distinct values available.
It's like an array AC on unique values essentially.
And so I've done this for these top 10.
Like measurements at Tufts that have a lot of distinct units, and what you'll see is I don't.
Maybe I can zoom in here a little bit.
I realize it's probably kind of small.
But in our case you see that there's. Oh, man, what's going on with my excel? I don't.
I'll try to get over there for you here.
There's diversity in like the ways of representing, so you'll have like 1000 per micro liter cells per UL K per UL, and then you have some stuff that's clearly not a unit.
So this is stuff that's coming out of flow sheets.
And again, my apologies for the size of this.
But we have like values in our unit field or like stuff that's clearly not ten out of 10.
So you have stuff that ends up in the unit field of flow sheets. That is very much not intended to be there and so.

I'm guessing as you go through and do this exercise of trying to to understand and collect the expecially for these measurements that do you do see a lot of of diversity in the units?
You'll find some of this type of unexpected data as well, so hopefully that was clear what I just described. But what I'm doing is I'm grouping my table in this case.
By this measurement. So I'm saying everything that is this measurement I'm collapsing.
I'm counting how often it occurs.
I'm counting the distinct mapped units and here that I'm aggregating all of my distinct unit source values and so like. Yeah, you have a question mark. You have numbers you have, like, you're all over the place with this, with this particular measurement.
So.
The synopsis there are like just the data is messy, especially the flow sheet data that we have seen or that we work with with regard to the unit entry is messy.
It is not well constrained and you end up with stuff that that you wouldn't expect to be there so.
Yeah, I have the query so I you can make an addendum to the query that I had sent you that does this.
At least the databricks version of that most other flavors will have some sort of uniquely aggregating function.
I don't know, Heidi, if you know offhand, collect set.
Is that also in Snowflake?
I.
I thought it might have been but.

**Schmidt, Heidi E** 35:54
It'll aggregate.
Will do, yeah.

**Houghtaling, Jared** 35:57
Aggregate always already takes the unique.

**Schmidt, Heidi E** 36:00
I think so.
I did some of that when I had to do the lateral and split.

**Houghtaling, Jared** 36:10
OK.

**Schmidt, Heidi E** 36:11
So.
But I didn't have to then take column F and see if I could match it up to.
A specific unit somewhere.
That's I think that would be splitting and doing some sort of pattern matching.

**Houghtaling, Jared** 36:36
OK.
But yeah, so I guess on this on this last example, this is one where you do see.
So there's three units that we mapped these all.
Are I mean?
One weird thing is that milligrams per deciliter is of course not moles per volume.
That's, but we're still mapping it as its own unit.
There's an element here of if you have a measurement that specifies the type of units that you expect.
And the units do not fall under that.
Like what do you do?
And I think there's a kind of a case by case dependency here, but if you see something like that where?
All right, I have 3 distinct units and one missing, two of which follow my moles per volume expectation based on the the low ink measurements and one of which does not.
But it is a valid unit, so to speak.
What do I do about this guy? So I think that's.
Kinda going back to what Andrew was saying, there's kind of follow-ups like this where you can investigate further with.
The like the the subject matter expertise experts at your sites that know the lab systems in and out and can tell you, hey. This is why we have an inconsistent unit with our measurements because this is not the only case where that happens. I can tell you.
And even amongst the Odyssey community, there is, I've seen multiple threads in the the Odyssey forums talking about whether you should accept.
Inconsistent units that are verified for measurements that they conflict with.
So sorry, go for it.

**Schmidt, Heidi E** 38:26
My eyes were opened when?
When someone in our group pointed me to lowinc.org, I'm blanking on names because I'm so focused on one of the entries that you pointed out. And so like, if if even though it's a flow sheet.
It's a free account and I was just about to type that.
I went spelunking, but this is the URL and so.
If it pretty much matches the flow sheet name.
Then when you click on.
Creatine mass volume in urine. It could be low ink. Code 2149-3, for example, and you Scroll down to the example units it will have.
For for that.
It will have.
This unit.
For.
That.
So that I know it's.
I didn't play around with like doing it as an API, but if if that helps kind of with phishing, I thought I'd mention it.

**Houghtaling, Jared** 39:53
Yeah, thanks for for posting that that link for those who haven't browsed in all the link sites you think you have to create an account, you can do so for free and then you can search through the link relationships and kind of underlying.
It's a data structure or data model of of low link ontology.
So thank you for that.
And that was specifically what Heidi was talking about was this one.
This one.
Which is I think this is the one that you were referring to, right?
I don't remember.
That's exactly the one.
Or not in any case.

**Schmidt, Heidi E** 40:28
I guess that it might be similar.

**Houghtaling, Jared** 40:30
Yeah.

**Schmidt, Heidi E** 40:31
There are two entries with two different low Inc codes, but this that one says that it's milligrams per deciliter.

**Houghtaling, Jared** 40:35
'Cause that would not be a mass ratio then.
So I think it's probably the other one that that you're referring to. It's OK.

**Schmidt, Heidi E** 40:43
OK.
Yeah, yeah. Mass volume, it says.
Sorry, but the that's why I trust and verify.

**Houghtaling, Jared** 40:52
Awesome.
Go for nature.

**Williams, Andrew E** 40:58
Wondering if you or Paulina had advice for people who want to do any kind of browsing of linked whether the it's worth the effort to, you know, download realma and use it, or you get most of what you need without it. Just if anybody's kind of follow.
Up on.
Optimal ways to browse for stuff outside of, I guess what's provided via Athena for relevant mappings for loinc.
Do you have any thoughts on that?

**Talapova, Polina** 41:32
Yeah, it's possible to download the link sources directly into the database and work with these sources there.
Another options is allowing search, but you should be registered in the link system.
Also.
So it's the two best ways I recommend.
I would recommend the first option when you just leverage the sources of flowing that they provide.
As on the their website.
There is download link option. You can get this for instance.
I just can put it into the chat.

**Williams, Andrew E** 42:16
Realma you when you say? Is that what you're talking about?

**Talapova, Polina** 42:19
I didn't set realm.
I'm just set the login source files.

**Williams, Andrew E** 42:24
I see.

**Talapova, Polina** 42:25
Where? Yeah.
And it's quite useful for work within the database when you don't need to search on the web.

**Williams, Andrew E** 42:34
I see. Thanks.

**Houghtaling, Jared** 42:41
Right. But yeah. indeed, I I would recommend checking out.
It's I found it to be helpful to go kind of to the source ontology, both in the case of Snow Med and Low Inc.
And and other ontologies to browse and kind of interrogate there as well, especially for very particular use cases like investigating the units of a precise limit measurement so.
Definitely worth knowing that those resources are there and and taking a look at them if you if you need.
OK.
That's so that's the extent. I just also wanted to give a little bit of an example on.
Diversity that you do see in pop up in these unit fields.
I've already mentioned we talked a little bit about inferring units.
You'll see that I do that in a couple of these cases.
So this is a.
This is one where a percent is expected, and so the ones that are missing, I infer that it's percent.
You'll notice this is one that I need to probably follow up more on.
These are actually almost the same concept, one of which is specifically through HPLC.
One of which is not specified.
This one actually has a possible unit of of milligrams per deciliter, and it also has a missing unit.
Which I'm guessing is coming out of our legacy data.
Because otherwise it would be inferred.
But this is one that I will need to take a look at in more detail but.
Yeah, I guess that's all I have to say about.
About units for now, unless there are any questions on the sheet or other otherwise.
And then I can go to the last topic if if not.

**Schmidt, Heidi E** 44:30
Do you have the logic for the collect set? Even if it's not gonna translate into?
All sequel, maybe.

**Houghtaling, Jared** 44:38
Yeah, I'll post it in the chat.
It's the same query actually that I used before.
But actually no, it's slightly different.
I'll. I'll send this to you because I did break it apart.
So just for speed reasons.
So that's that's the query.
Get the collect set piece.
Actually that one.
Sorry that has array AG which you don't wanna run.
Let me modify that for you real quick.
That was an old version.
So.
I think the the most difficult translation will be to like SQL Server probably, but I don't know how they handle.
Their the handling of array aggregation on SQL Server is always slightly different but.
Snowflake should be pretty quick.
OK.
So I'm gonna start my sharing, so maybe I'll take a quick pause and say.
I can go through.
Briefly, this idea of a gap analysis and a dashboard and my thoughts on that use that would show Epic specific data and epic tables and names that we would then have to exclude entirely from this recording.
I think we could chop the recording probably at this point.
I don't know though.
I'm looking a little bit toward them, kind of throwing Andrew and and others at curveball here with this. Would this be valuable or should we?
Should we hang on to this and and try to do with something more full-fledged in the future office hours?

**Williams, Andrew E** 46:53
You are the best judge of that.
I think it's really exciting.
And as long as people don't.
Take it to be anything other than what you show them.
I think it's really OK.
I think where we've hesitated in the past to show things that are in progress is because people might.
Not understand the full implications of it being in progress, and I you're the best person to kinda judge that risk and what's what's there so.
What? What are your thoughts?
I'm. I'm throwing the curveball back at you.

**Houghtaling, Jared** 47:09
Then I think it.
I think it could be useful, so I would I would.
I would error on the side of of showing it's and discussing the gap analysis, so I will. I'll go for it, OK.

**Williams, Andrew E** 47:19
It.

**Houghtaling, Jared** 47:23
Then let me share my screen for those of you who have already seen the.
The data upload dashboard that we have, I should update just a second here that we put together for on the central cloud.
This should look familiar to you.
But I'll kind of give a very brief overview of what what we're doing here.
So one of the questions that that I get a lot is or that's kind of comes up as you're trying to set up a study.
Are you trying to investigate a set of omop data? Is we're looking for?
Let's say ECMO data and we know that we're capturing it in a hospital, but we have nothing in omop.
Like, why is that?
Where is it and how can we make sure that we get it into omop?
And so in that case, right?
Like I can go here into my device domain and I see already that I have a very low device coverage, at least at Tufts, we are not capturing. We're capturing a small percentage of our devices at the moment.
Working on that.
This kind of gives you a feel for if you look at the number of rows of potential devices and you look at the number of actual device exposure data, how much are you missing?
And then that's contextualized a little bit by looking at how many unique concepts you have between omop and epic.
How many unique persons you have and how many visits you're actually capturing so.

In this case, we have a lot of people in and a lot of visits, so it's the issue is not coming from just missing people, it's actually coming likely from missing mappings and what you see if I go to something like ECMO. I know I can search that.
Now in this table and I have a lot of data and again.
This is not.
This is maybe a bit misleading in the sense that I'm looking at flow sheets and these are not direct device entries.
They're like indicative that a device is being used.
It's like a secondary way of understanding a device.
But we have, at least in our Tufts epic instance, we have a lot of flow sheet data related to ECMO.
But you'll see if I Scroll down here, there's. There's nothing with omop.
So if like this data ID we don't have any ECMO data yet mapped and that's just one example like.
If you wanna come in and search for something you can do so here and you can see where that term exists in your omop versus where it exists in your source.
And you can do the same. What I've also done in the second table is I create.
I actually just took all of the unique terms in the vocabulary in this case in the device domain, but per domain, and I've then mapped like how often those terms appear in epic versus in omop.
Uh, so you can actually search this field for like. I don't know if proteins gonna be in here, but like, uh, prosthetic or something like this. And you can look and and compare at a term by term level how often something appears.
And then this is just broken down by the the various domains.
So you can do the same for drugs.
So what is interesting actually in the drug domain is we are actually we have additional granularity beyond what Epic has in this whole load and just actually it's filtered. So that's why it's not loading.
And this is kind of a weird artifact from the way that handles its medication mapping to RX norm.
But we still there's a lot of drug, at least by Rose.
So we we we can investigate and there's potential rationale for that.
All that to say though, you could then search drugs here and you could see like.
If you do something like this, you'll see a lot of different options from EPIC, but then I can.
You'll see the omop entries and you can again filter these tables as you wish.
And again, I did the same thing here where I have different drug terms of interest and you can compare those between Epic and omop.
Same for each domain so.
This dashboard is intended to serve as a little bit of like a first step into understanding.
On either a term level or a condition based level.
The the disparities or the gaps that you have between your underlying source data and your omop instance.
And it does so by essentially pre populating a set of three tables per domain and it uses in our case either staging tables that we create during our ETL, or really the raw epic clarity tables.
To get ads.
What what you're looking for?
Like kind of this this general overview that you can search through and identify where things might be missing.
See you and again, this is the first version, but the intent is to better understand what things were missing and identify why we're missing them, or how we can potentially map them and and what part of the logic that exists. So we have like references to the log.
Right where this actually is drawn from or where that code would need to be changed.
So I see this is like a another tool to help update and upgrade an ETL to capture more elements of interest.

**Williams, Andrew E** 53:03
fantastic, Jared.
It's my first time seeing that.
That's so exciting as usual. you.
Make things happen so well and so quickly.
It's kind of takes my breath away.
Some of the context is that we are a site that is looking to be as comprehensive as possible with respect to the available data, and we also played this role in helping to expand the vocabulary to capture things that currently can't be mapped.
Both of those things are probably considerations for people who are thinking about.
The value of this tool for your work, if you're at a site that's primarily trying to capture what's easy to get in the broad domains and then the specific things that we're requiring for bridge to AI.
There's less probably of an interest in.
Close, comprehensive representation possible, and it's a nontrivial amount of work to go after things.
It's also the case that a lot of things that people are interested in are going to be.
Identifying new gaps in the omop vocabulary that need to be filled in order to represent them.
And so there's a an additional piece of work that we can work we can pursue together on things kitsur.

**Balakrishnan, Kasturi** 54:21
I.
Jared, I wanted to say this.
This is this is really what I'm working towards here at UCSF.
So it's it's it's pretty close to my heart and we've struggled with it because of the the domain transformations that happen as you pull data into omop.
So but my really nice question is you don't see provenance information for this. You see literal concept level like when you populate those percentages.
It's based on the the literal concept, like a medication ID or like an epic ID.
To Omop concept and and not like I've tried to do it.
At at with the prominence information in place, because it it's just very difficult to capture all of the filters that the ETL actually accounts for, like missing dates and things like that. To do it at the concept level and the fact that you move across domains as you.
Transform data from one table to like 5 or 6 tables in omop. But I I would be very excited to talk to you about this.
This is like on top of my list as as we improve our instance. So I really wanna hear your thoughts about it.

**Houghtaling, Jared** 55:28
Yeah. No, absolutely.
I think you you hit the nail on the head with the the the biggest issue, right?
And that's why I need needed to use a lot of our intermediate like our staging tables, because the way that our ETL is structured is the majority of the filters, with the exception of the mappings themselves, happen at the staging table level.
So like we filter to make sure that we have an encounter that that is good. And so also these numbers are in context like.
We're assuming that we've captured the encounters appropriately in OMA.
And we're saying that the encounters that make it encounters that make it into our staging table have already been vetted upstream.
So that's like the first big filter and then we have all these other things that like, was the record disqualified or like was the record?
I don't know invalid for whatever reason.
And those also get chopped out at this in the entry to the staging table.
So there is some.
Upstream logic that populates this so the the numbers that you're seeing are.
Not total counts of like perfect provenance.
Everything in the diagnostic table.
It's like what exists in our staging table that then we're able to capture assuming that the staging table is more or less like our ground truth.
So there's additional kind of context needed for us, but you're actually right. It's very difficult to.
Take a look at just an epic table which has all of its idiosyncrasies and columns that that negate things and missing dates and all this, and then be able to then compare that with something like an omopt table where there's a lot more constraints.
On certain fields cannot be missing and and things like that so.
But absolutely would love to talk to you more about this offline.
I think you're muted, Andrew.

**Williams, Andrew E** 57:15
I'm curious how many sites do something?
Row in a table that's persisted.
So you don't have to re query the original table multiple times depending on which part of your ETL you're working in.
You save, you save yourself a lot of re querying of the same data by creating that first mash up table which I think is the staging table. You're you're referring to in general.
There's other approaches that don't do that and.
More than one, but that is a basic category of like saving a lot of requiring work.
And I was just wondering how many?
Sites do their ETL that way. If you're doing it yourself and you're not getting it from someone else, how many sites do that initial mashup table and then draw from that?
You all do kasturi at UCSF or other Satish.

**Houghtaling, Jared** 58:28
Mgb does as well, at least to some extent. with the staging tables in EPIC.

**Williams, Andrew E** 58:36
It is.
It does save both on processing and potential to do something that's meant to be equivalent in an unintentionally different way.
Would you go back to the same allegedly in the same way at the same source? But there's some slight difference in the code because you're you're not using the same code in the same source.
So there's a couple of different problems that it solves, but I know other people feel like it's optimized in other ways, so.
Anyhow, that's interesting.

**Houghtaling, Jared** 59:06
Yeah. So the I guess I should say the the way that I've structured the dashboard.
Is it's a single Python file. Will run the dashboard for you, the data to to actually support the dashboard. This is all just created with SQL queries and the like.
It just has.
There's 3 tables that have a particular date structure, but you can populate them however you want, and you can. The intent is that it's flexible to accommodate not just specific epic tables or epic staging tables, but you could.
You could place.
Content from other source system tables as you as you would like, so hopefully it's fairly extensible, but I can share the code both for the dashboard and for the.
Dashboard for sure.
For the underlying sequel, that's gonna be epic, but I could maybe put it on epic user web or do something like that to to disseminate that so.
The dashboard shouldn't have any epic specific stuff I think.
So we should be good there.

**Schmidt, Heidi E** 1:00:07
Yeah. I mean if if the queries that are being run underneath are.
Epic user web.
Then I'm assuming that like we have to plug in whatever SQL logic.
Would be anyway.
And then if if someone didn't have an epic site, they'd have to plug in their their source sequel.

**Houghtaling, Jared** 1:00:37
Yeah, you just have to be able to to to create three different CSV files for the domain with a particular structure, and it's more or less like what is the table that you're talking about.
What's the like the the field that you're talking about and the distinct concept in that and then the counts of people and visits and the concept like rows of concept itself?

**Marta's Fathom Notetaker** 1:00:55
Play me a song.

**Houghtaling, Jared** 1:00:58
That's really the main thing.
And then if you can create a sequel query that generates that then.
Then this dashboard will show it for you essentially.

**Williams, Andrew E** 1:01:09
It would be wonderful.
I don't know that we wanna acquire this.
It would be a great.
Piece of metadata about.
The data set that we end up producing from the project, if we could estimate the overall coverage of what got mapped in the relevant domains, in addition to saying you know what we did to validate those mappings and assure their quality and so on so forth just to.
Say.
Because it's it's possible to have a ton of things and really only cover a fairly small amount of what was potentially available in a given area, and I think that's often the case when you're when you're doing a big ETL. And so having some characterization of that would.
Be fantastic from as many sites as we could get it and this would really enable us to understand that.
So it's a really.
Jared, have you ever. you know, you had a whole bunch of notes about Rogers original ETL?
Have you ever added those or communicated those to him?

**Houghtaling, Jared** 1:02:21
So yeah, I need to.
I need to actually send an e-mail with that I was hoping to.
Kind of refine my my notes on that. I did send it to Satish I think.
I think Satish should have them in Seattle and I've maybe distributed them to some other people like individually, but I haven't put them in a format that is I don't know, like a slightly less blunt.
I think I was maybe a bit crude in my in my evaluation of certain things and I want to make it a little more nuanced so.
I think that's once I once I'm able to refine those notes, then I will send them to him and prioritize that probably.
So.

**Williams, Andrew E** 1:02:58
It's great, I think.
Yeah, as other people are using it a fair amount. And so where there's knowledge to be gained about things to be alert to and perhaps correct.
Be good to get that.
Maybe I don't know anybody else. You've shared it with wants to kind of help in wordsmithing some of is it really descriptors associated with the things that you note that you want to or is it?

**Houghtaling, Jared** 1:03:24
I just kind of exported the set of issues that I had created.
In our GitHub repository that were just like, there weren't a lot of context around the issues and it like, I think having additional because now also with the updates that I've made sense those issues were created and resolved.
I think that there's additional information that I can add and kind of contextualize them more.
So I think what I gave was a start, but there's there's more to be said.

**Williams, Andrew E** 1:03:53
Yeah. I guess if there's a way anybody who's using.
Rogers ETL as a stub to annotate it. Whether it's building on what you did or indevendent from it, that's a good thing. I guess for us.
All right to do.
And it's great that you've you've done that, but anybody else who's done it either.
Building on what you did or indevendent of it, I think it would be great for us to encourage that.

**Houghtaling, Jared** 1:04:22
Awesome. Well, thank you all very much.
Hopefully this was a useful useful session.
And look forward to talking to you all.
Actually, I'm gonna be probably out next week, but hopefully in a couple weeks I'll be back so.
Yeah. Thank you all very much.

**Williams, Andrew E** 1:04:40
You. Jared.

**Houghtaling, Jared** 1:04:43
Have a nice rest of your day and.

**Williams, Andrew E** 1:04:43
Hope you have a great great week.

**Houghtaling, Jared** 1:04:46
Thank you. And Kasturi, please reach out.
Let's let's chat more about this. I would love to get your thoughts and also love to collaborate more on this because it's a work in progress and I think there's a lot of potential to do some cool stuff so.

**Balakrishnan, Kasturi** 1:04:59
That's great.
Thank you. Jeff.
This was great.

**Houghtaling, Jared** 1:05:00
All right.
Yep. thanks all.

**Alvarez, Marta** stopped transcription