CHoRUS Bridge2AI Standards Office Hours-20250918_175713UTC-Meeting Recording

September 18, 2025, 5:00PM

57m 2s

Alvarez, Marta 3:56
Hi, everyone. It looks like people are still coming in, but thank you for joining today. Just wanted to give a quick heads up that I did send out the invitation for to extend this standards office hours through to the end of year 4.
And please let me know if for any reason you're on this call and did not receive it, but otherwise I don't have any other.
Things to mention. Jared, I'll let you decide when you want to get started.

Houghtaling, Jared 4:34
OK, I can maybe go and jump in. So I guess I have some good news and some bad news for today. Good news is I have finalized the tool for doing GIS stuff from beginning to end. So that takes location, location, history table.
And gets you all the way through with a set of variables and data sources into an external exposure and does it relatively in a in a lightweight way. And I would like to go through kind of what that means because it was, it's been a, it's been a whole ordeal getting to this point.
So I'm I think I was initially going to spend more time on on geocoding today and I will still spend a few minutes and I can go through some of that today as well. But I think the bulk of what I will do is kind of show you what what the tool is and kind of how it works and why it works in the way that it does.
And then with kind of an eye toward how we can update it in the future and how we can add different data sources. But I think that so I have I'm able to pull in two major data sources, one of which is.
Like social vulnerability index, full data set. So they have essentially that index, but all the kind of component criteria and I think all of which were in Gilles list of 22 terms. So we have those and then I also just out of.
More personal curiosity included a pollution data sets in there as well. Together with the tool, you get 20 addresses, and those addresses are dispersed across four locations in the US. They're all actually like fast food restaurants.
Don't worry about it, but essentially commercial addresses and I've already geocoded them for the purposes of the tool, because the thought is that you'd come to the tool with geocoded addresses and it would run for you. So.
Covered a lot of ground there, so I guess all that is is good news that we have this. The bad news is I really was working on that until like 3 minutes before this call. So I have not really had a lot of chance to prepare. So this could be and likely will be a little bit chaotic.
And it's also going to be highly technical. And so if you are not an engineer or not a clinician or sorry, if you're if you're a clinician that that doesn't wanna like walk through code and understand what's happening on kind of the the back end of this.
Please feel free to tune out or drop off or whatever you would like to do because I am planning for it to be relatively technical and go into like how how we actually do the geocoding and the linkage or not the not the geocoding but the the linkage up with with OMAT terms.
OK, um, is that all sound good to everybody? Before I dive into this, are there any specific questions? Um.
About GIS or STOH? Or are there any other topics of interest that we need to cover before things get, uh, like get kinda carried away here?
Go for an ID.

Schmidt, Heidi E. 7:44
I just want to verify that we're investigating it now, but we're not adding into the central DDL.
The tables that would go along with and we're not actually running this as a process locally for the data submission that we're expecting in November, correct?

Houghtaling, Jared 7:58
Not yet. It's good.
Great. I think today's submission is in October and I, but I don't think that it's it's expected. No, I I believe it is not. The one thing that is expected that we talked about on Tuesday is you all submit kind of this notes dot MD file or kind of.

Schmidt, Heidi E. 8:08
OK.

Houghtaling, Jared 8:23
I don't annotations. I don't remember the name of this file. If somebody's from data acquisition on here and knows the name of the file, please chime in. Maybe Tony, if you're on, I don't see him. There is essentially a way to submit notes about your your upload and in that file one of the things that we will ask in this coming round is.
To include a status of if you can Geocode location data or not. So that's that's essentially the extent to which sites are are expected to to kind of participate in the GIS work in this upcoming round because we're kind of.
Getting towards the deadline here. Um, hopefully that's clear. Um.
OK.
All right, so I'm gonna dive into it. And I think the first thing I'm gonna start with is a repo that I did have. I don't know if anybody from Florida is on the call today. I don't know.
Marty, if you see anybody, if Divya is on or anybody else.

Schmidt, Heidi E. 9:25
I'm here.

Houghtaling, Jared 9:29
Is that Ben? Great. All right. Hey, Ben. OK, sweet. So this is a tool that the the team at the University of Florida is is working on. Um.

Alvarez, Marta 9:30
And I think it's fun.

Rooks,Benjamin J 9:30
This is Ben.

Houghtaling, Jared 9:44
And.
What the tool is is intended to do, it's a couple of different things, actually multiple tools we're talking about here, but it's going to take data out of these public data sets, whether they're from the CDC or or CDAC or which is like NASA or or some other other places.
I need these public data sets, pull them through, standardize them, and then kind of make them available to our kind of OMOP tables. We can join them together and then link all of our OMOP tables with this.
The part of this repo that is currently available to use for you all is the is the geocoding part, and so there's some instructions here that talk about pulling a docker image, although this has actually changed since I looked at it last.
Let me quickly see. Maybe not. Maybe there's no demo video anymore. Oh, it's a new new demo video. Sorry. But essentially you pull a Docker image and that Docker image allows you to convert your location.
Rows into latitude and longitude values, right? So you come at this with a location table, you run this docker command and that is going to produce for you an associated set of lat long values with your location.
This is a Python based code base and it's using this degauss. It's built on this degauss.
You see right this this actually has changed I think. OK anyway so that there there is a a docker image that's been hosted on this Prisma lab or Prismap lab.
Called Xposome Geocoder and as with a lot of these tools, Divya's here. Great. I will maybe Divya if you want to talk toward this a little bit. I was just gonna spend a few minutes on geocoding and then talk about the process that's that's available now for post geocoding. But if you want to go through and.

Vellanki, Divya 11:50
Yeah.

Houghtaling, Jared 11:57
Talk toward any of this. I I think I'm just kind of giving an overview, but if there's anything else you want to highlight.

Vellanki, Divya 12:06
Yeah, so this at this point, I don't wanna highlight anything. I would just request everybody to just go through that manual. We also have a demo video added on there on top of it, like just in the beginning of the manual.
Pretty much once you have the addresses in whatever form you have in the required formats, the code should run smoothly and and just let me know if you encounter any issues and we can hop onto a call and get it resolved.

Houghtaling, Jared 12:41
Awesome. OK and so that and this is this is using there's there's a degauss which is a a geocoder, but there is a degauss Docker image and this is kind of wrapping up some scripts to make calls to that local Docker image. So if you can if you can imagine.

Vellanki, Divya 12:46
OK.
Yes.
Yes, yes.

Houghtaling, Jared 13:00
You're pulling something that's been precompiled and I think with D Gauss you also need to pull a specific regional data set. Is that right? There's like different versions and you have you say like I want Northeast or Northeast US or Southeast US or things like this and depending on which ones you choose, it'll be larger or smaller the size and some of these can be very big like we're talking.

Vellanki, Divya 13:11
Mhm.

Houghtaling, Jared 13:20
10s of gigabytes in size for for these source files. OK, so that's and and we I think we want to maybe another call. Degauss is one option and it's it's a relatively popular option for geocoding, but there's lots of others and there's lots of also other containerized versions, so.

Vellanki, Divya 13:21
Mhm.

Houghtaling, Jared 13:38
What I would maybe encourage you all to do and we can follow up at another time with this, but go through this manual, try it out, see if you can and can run this tool and what would be great is if we can start to have some comparisons with D Gauss and the performance of that to geocode.
Against Nominatim and against Opensnewtmaps or OSI, I don't remember. There's I think I want to say at least five commonly used ones, but I am again not not the chew coding expert here. So but take a look at this. I did want to ask Divi and that you're on. Can we just fork this repository so it's over?
Over on the Chorus AI organization space, just to make it cleaner and clearer for everybody where stuff is at. Are you OK with that or is it OK? OK then maybe what I'll do is I'll move this over.

Vellanki, Divya 14:24
Sure, sure. Yeah, yeah, that's fine.

Houghtaling, Jared 14:33
It could be that I actually do like a sub module that I pull it in into like the container app space alongside the other GIS tool. I'll keep you posted on how I do that.

Vellanki, Divya 14:42
Sure, sure sounds good.

Houghtaling, Jared 14:45
Cool.
OK so that is and and so yeah So what this this will then get you like I said it and if you once you read this this through you'll take your location table locally which if for those of you who wanna refresh on the location table you can pop in here.
Um.
It'll it'll take these addresses in their various forms and however you've kind of pre processed them and and then add it'll essentially fill these two columns, right? So we have like our location ID and then we have our address associated with it. So we got like an address, a city.
State, the zip county and then the the geocoder is going to take the combination of this information and different geocoders require different structured formats to to bring that in and it's going to fill these last two columns for you, the latitude and longitude.
And that's very important because the latitude and longitude are used in by Postgis, which is a GIS Postgres based tool which you can also see here and I believe that it's kind of backing what's happening on Florida's cloud.
Um.
But PostGIS is able to take these latitude and longitude values and place them inside of data sets that you get publicly. So if if let's say the CDC has a data set at the zip code level and it's got the whole, it's mapped out all the zip codes and polygons and you have these like point latitude longitude values. Let's say I live on 50
18th St. It's going to place my particular point, latitude, longitude inside of one of these polygons, and then you can, based on the Polygon you're in, you can look at a bunch of different variables that are associated with that Polygon. So like.
Social vulnerability index or other types of either poverty or kind of health outcome measures associated with that region, so.
All that to say.
Location you can you essentially are augmenting your your location table with this and it is a a critical first step to.
To getting to where we want to be with the CIS and the SDOH data. So let's see. I think this was still active at some point. I don't know if they've shut it down, Marty, if you know.
Nope, it's still here. OK, so this is kind of a hack. You can't get to this anywhere through the website, but if you switch the URL to 60 then you can see what was going to be the next version of the common data model before it was scrapped. And in that version you have this other special table called location history.
And this is the other critical table for doing any SLIH stuff with kind of the GIS work and it's it's a simple table. It's only has what, 6 cols of which one of them is a a foreign key to.
To the location table, then you have something that says OK, how is this history related to location? And then here you have kind of like a two-part, a two-part foreign key, kind of like we have in the measurement observation table where we're saying.
This is the person domain and then this is the primary key of that. So if I wanted to to say a person was at a location for a given time, I would need this domain ID to be the person ID and we can look at that just a second.
Let's keep pressing.
And I swear to. I got to think about this.
Metadata, I think. Yeah, so this is the one. So weirdly enough in the in the OMOP vocabulary, all of the tables are represented by concept of these as well in this metadata domain.
And so for the head location history table, your domain ID is going to then refer in in most cases that we're concerned with, it's going to refer to this 1147314. I wouldn't worry about. It's possible as well to to like use provider and care site for histories and stuff, but for us let's not be.
Even go down that road, let's just say, OK, this is pretty much a constant value filled with this 1147314. I'm gonna put that in the chat just so you all are aware.
Um, OK.
All right. And so.
Once you do that, then you have your entity ID, which is the primary key of that table. In our case, this is going to be the person ID of your person that you're concerned with. And then you simply just have a start date and an end date, and that's when the person was at that location. Maybe they lived there for six months or maybe.
For 60 years. And so this these dates should report on that and there's various ways to get this information in your EHR system. And I I I do have some code at least for how we've done it with our Epic instance that I could potentially share with folks if interested.
But what you need to be able to to look at is kind of the address history more or less of the person in your system. And from that address history you can populate your.
When people have moved or updated their their address information and that gives you an indication of at least where they lived. You can of course go wild with this and try to track like a workplace location and all these other things if you have that info. But for now probably residential addresses is great if you if you have it available.
But you need to have some some metric of of historical location. Go for a city ship.

Dandayudhapani, Satish 20:54
So I'm assuming these dates will also be date shifted.

Houghtaling, Jared 21:01
Oh, good question. I guess yes, that is a that's a great. It's a great question though. In theory they they don't need to be because this is not a delivered table. This is only being used to pull the information. It's only an intermediate table for you. So this is not something you're gonna ever send centrally.
We.

Dandayudhapani, Satish 21:22
OK, that's.

Houghtaling, Jared 21:22
So I I guess preferably they would actually not be date shifted if possible. There's there's a discussion that we're gonna have about like the output that comes out of this and making sure that we're able to to properly de identify that output, but the dates themselves will not appear in.
No, actually I guess the dates will appear in the external exposure table. OK, so I guess I guess my what I would say to do then is this. These dates should not be shifted full stop, but.

Dandayudhapani, Satish 21:56
OK.

Houghtaling, Jared 21:57
They will be used in combination with the public data sets and the resulting table, the external exposure will have a start and a stop date of that exposure and those will be then shifted for you. You should shift those I guess before delivering if that makes sense. So let's say somebody lived.
Somewhere from 2011 to 2013 and your date shift is like a five year shift or something. I don't know what you're doing. We would still want the the proper census data for that time period, that 2011 to 2013 associated with your exposure. So then that's what you would use.
For the process, but then everything would get shifted in the external exposure and in theory would like, yeah, I don't. Anyway, I'm, I'm, I feel like it's just getting more confusing the more I'm talking about it, but I would say.
Short answer is don't shift the location history, but shift the external exposure and then it's up to us. Yeah, sorry.

Dandayudhapani, Satish 22:54
So.
Yeah, so one one question about this location history when we are building locally, are we going to hit any of the OMOP tables like the visit occurrence to go and check some things or any of the other OMOP tables to?
Use these dates.

Houghtaling, Jared 23:14
I probably. I mean, you could do some checks in theory, right? Like you could say, I mean for sure the address is.

Dandayudhapani, Satish 23:15
To check.

Houghtaling, Jared 23:25
I mean, even that I I guess it's likely the data that you're gonna use to populate this are are from different parts of your EHR system than what you're using for. It's gonna probably be from like demographic information, address history and things like this that get checked.

Dandayudhapani, Satish 23:33
OK.

Houghtaling, Jared 23:40
Regularly on encounters, but are not necessarily referenced in our general EHR ETLS.
But if if you do want to do checks to make sure like there's like not wild dates in there, like a person's lived there for 50 years, but they're like 20 years old, like that would be also probably a good good thing to do, but yeah.

Dandayudhapani, Satish 23:48
Oh, yeah. Thanks.
Yeah, I I was referring to is there the the external table you are talking about? Is there any reference to a linkage to the OMAP data we are submitting like the the visit occurrence ID or those kind of stuff?

Houghtaling, Jared 24:16
It'll it'll refer to the person ID and I can show you. I don't is. I don't think external exposure is in here. Marty correct. I don't think so. That's only on the CIS repo. Let me quickly pull that up because that'll also make it clear.

Dandayudhapani, Satish 24:19
Of course an idea.

Houghtaling, Jared 24:33
Um.
I wanna see.
SDI.
OK.
OK.
OK, so the external exposure table.
Uh.
Is like this, right? So this is. I'll drop this link in the chat 'cause this is honestly difficult for me to find as well sometimes. Um, and I'll also put this link in the chat just for.
If you do want to come take a look at these again, OK, but this is the external exposure table and you'll see there are not visits associated with this. It does have essentially like a lot of the measurement related columns like units, values, concept, values, number, operator, quantity.
There's even a dosage in here. Um.
But you don't have any other linkages apart from person ID and location ID. But in this case the location ID is not necessary for your delivery. Like we won't. You don't need to deliver that.
So it is. It does say it's required, but because we're not delivering the location table, that's probably the biggest change that you're gonna make on this is is location ID can be empty.
If that makes sense, like you'll have it in in the output of what you're gonna do, but you can just nullify the column if you want. It doesn't matter, but for the rest it's it looks pretty much like the measurement table, right? And it it just doesn't have any sort of visit associated. So I guess the person is really the only thing that you would need to link it to.
And from then an ETL perspective, that is nice because what I have published here, which is essentially a single shell script that does this and goes through, takes location, location, history, pulls in.
Data source values and variable values and then outputs then this external exposure. So it outputs this exact table for you with the proper structure. Yeah so.
But yeah, good question, Satish. Um, very good question as well about the date shifting. Um.
I'm sorry I got a little muddled in the response, but I do still think if you can do the original dates when you run things through the through the GIS tool and then you shift it after, I think that's probably gonna be the the preferred option so.

Dandayudhapani, Satish 27:23
Yeah, thanks.

Houghtaling, Jared 27:24
OK, and so now now we come to this guy. You know, I think I'm actually gonna switch. I don't know if you all care if if it's dark or light. I do have my PI charm in dark mode, but I prefer to look.

Add code in there, so maybe I will do that if that's OK with you all.
Unless anybody has strong objections, but um.
Not hearing any. I guess I'll move forward. OK, So what? What happens here? And maybe I'll take this real quick. I think I just commented some stuff out.
So what I have what I put together is this is a based on a post GIS docker image and like I said, post GIS is is kind of an extension of Postgres that includes functionality that can handle.
Different projections and geometric representations of points on a map, and so we're starting with that as our base image in this Docker file. We're adding all of these kind of dependent components for what we need to run, specifically GDAO.
So if you look at any sort of GIS stuff, GDAL is is like one of the the main tools that folks are using. In our case we need GDAL in order to load these specific GIS database files into our our Postgres database and so.
Difference data sources. I don't know why my thing is going so slowly. Hope you can you all even still see me and hear me? My computer is lagging very badly, but.

Alvarez, Marta 29:05
could see and hear you well.

Houghtaling, Jared 29:06
We close out of.

Schmidt, Heidi E. 29:07
Your audio is coming through very clear. Sorry about your computer though.

Houghtaling, Jared 29:10
OK, OK. No, so it's all good. I just will pull that up some stuff. I think I just had too many Python windows open potentially.
OK, should be.
Or maybe I'll just share my screen instead.
OK, alright. Can you all see this?

Alvarez, Marta 29:43
Yes.

Houghtaling, Jared 29:44
Cool. OK, and so this is the Docker file, right? And so essentially what we're doing is we're starting with Postgis. We're adding this G Dal, which allows us to load in some special files that these open source GIS data are are.
Pulled in as and then we're getting a couple of different things, so we are.
So these are actually the files that Caia DB or or kind of the old um.
The old workflow had reference, so actually these probably should even be commented on because I've created new versions of those here. So we're essentially getting our our catalog of data sources and variables of interest, and I'll show you what those look like in a second. That's very important.
And then we're also getting a set of vocabulary tables from. In this case, this is the tufts, the CVB, which I think I've showed you all. This is the same tooling that we're using to create a bridge to AI ontology. So we're going and we're fetching all of our delta vocabulary for GIS.
And then we're essentially just copying all of our stuff into our into our app in the right spots. And so we've done this, we're getting all our files, we build this image and this is kind of these are the instructions to build the image that that we would ask you to deploy locally.
That's the first thing.
Probably the the most critical part of this image, apart from kind of the software that's allowing you to do cool stuff, are these two files. So I called them data source simple and variable simple and I'll open these as tables. So data source simple.
Is a subset.
OK, I think it's bigger here. So this is a subset. See if I can zoom in on that for you. Great of what the GIS working group had put together and this includes just three data sources so far and.
Really it's two data sources. There's there's two of them that are dependent. So when it comes to these GIS data sources you can have and this is actually this was shown nicely on the on the Florida's manual, but you can have things that represents.
Pure geographies like different ways to represent physical space, whether it be like a voting district or a zip code or a.
A county or whatever it might be, right? And in our case, this TIGER census is exactly that. So this is a a GIS data source that just lists all of the counties in the United States.
As of 2018 in this case, and it like places bounding boxes, polygons around all the counties based on where they essentially begin and end. That's all it does though. It doesn't have any other fun information for us to use. It just says here's where your counties are like, that's it.
That's the first one. The second one is a combination of both things. So the second one has not only the variables of interest that we are that we want to use, so whether that be in this case it's PM 2.5 levels.
Annual average for a bunch of different cities. And so it got those average values as well as the associated geometry. So it's a it's like kind of a combination of what values we're interested in and where they are actually physically present.
At a given time, and so this is this is the easiest one for us to work with from like a technical perspective. And then this third one is very rich, but it doesn't have much in the way of.
Um.
Kind of this, this geographic detail. So it it has a county level SVI details. So again all these different metrics of of social vulnerability and economics and things like that. It's got a ton of those in this case for the year 2018.
But what it doesn't have are the the geometric pieces. So what we have to do in order to make use of this is we need to combine it with our place related stuff.
And we can do that then also in this in this database fashion, right? So.
To recap on that, this data source simple has three rows of which there are really two data sources. There's a a kind of a a generic county geography, and then we have our pollution data source with geography, and we have our SVI data source without geography that needs to be combined with the counties.
All right. And in total we have, I want to say.
Something like.
75 variables I haven't. No, it's like 130 maybe in variables that we can we can take a look at here. So this is our first file. That's the data sources that we're pulling. What you'll notice in this file is there is a download URL and so part of this this process this.
This.
Bash script here is we go and we fetch the this part. Actually specifically we go and we fetch our file from the Internet publicly available and then we unzip it or we untar it I guess and then.
Uh.
And we move forward there. So like we have this nice call out in our table here. It's not gonna.
Some time.
Um.
So we have like this nice download format, right? So we know what it's coming at us as and then we have the download URL. So we use W get this utility to pull a file from the Internet and then we unzip it based on this guy and that's gonna actually get us the raw files but.
We're not quite yet to the point where we can do cool stuff with that, and so the next thing we have to actually do. So I guess if we're walking through how this this code all works, right?
We have our source files which are the location location history that I talked about and then we have our like data sources that we're pulling publicly and those are represented in this table and this is a table that we can iterate on and we can add new data sources and update versions and as long as we keep the format.
The tool will continue to work and this has been kind of a big focus of the GIS working group on Odyssey was is essentially expanding this table. For our purposes, I actually reduced the number of rows just for testing and to make sure that what is in the table. I know that we can.
Use and use effectively. Um, so yeah, we'll expand on this eventually, but this covers kind of our first pilot set.
All right, so location, location, history over here and then data source from, like I said, CDC or in this case NASA, those are over on the side. And so the first thing we do is we want to pull the data source right with these links, we get our data source and then what we're going to do.
Is in those data sources we.
Do this.
Sweet.
What's this one up now? And so this is now getting closer to what you would expect like in OMOP, right? So you start to see some OMOP stuff come up here. So we have like.
A couple of key key columns. So we have this geometric dependency and we have a data source dependency.
Um, so CS and C2 here. These are I would say the most important columns in this and these are essentially telling us what uh rows from this other table that we just downloaded. We need to be able to reference what which data files we're coming from.
And so these are all of our like SVI ones, right? You see, I don't know, different numbers of people below a poverty estimate, household estimate, housing units, population, education stuff, right? Like these are all STOH variables that we're we're interested in.
And those are dependent on our county geography data set and our SVI CDC data set, or the census data set, but it's I think controlled by the CDC in this case.
And so in order to actually place these all together, right? So we have to say we have to to link output from those two data sets together with this variable in order to get where we want to go. And I'll I'll show you kind of how that works, but.
Um.
The idea there is that this this variable list is essentially, where you would start. If you're let's say you're a researcher coming at this, you want to do some research. You're going to look at this variable list and you're going to come in and say, OK, I'm interested in 145 and I'm interested in 151 and I'm interested in 161.
Those are then going to be the three. And then you're going to launch this process with those 3 variables in mind. And the process is it's going to go and it's going to fetch your data sets, and it's going to then just use those 3 variables in the external exposure calculation. And I'll show how that happens in just a second, OK.
So those are kind of our two key files and those are what I would refer to kind of as the catalog. I think in the GIS or the GIS working group, sorry, the catalog is a is a broadly used and often.
Like diversely used term, I see the catalog as essentially a list of what the data sources are that we're interested in and what variables those data sources contain and where you can go grab that data. And so that's essentially we've made kind of a simple catalog here in the CSV folder.
That's covering our pilot use cases.
And so then what's cool is these two files actually really the data source alone file is is telling us how to get a pretty tremendous amount of information. So we go, we fetch the raw files, we unzip them. They're like, let's say hundreds of megabytes.
Those get loaded into database tables in postgis, which we can then just reference in our queries and so.
At this point, right, we've we've done that. Yeah, sorry, I'm kind of struggling with this. OK, so we go, we fetch our source files and then we ingest them. So we ingest them in two different ways. And if you we're looking closely at those files.
There are different data standards that these GIS data sets are represented in. Most common I would say probably is our shape files. They're called, which is a specific GIS file set. It's actually a combination of files.
That contain information and then there are also these GBD files. That's when we talked about that gdal. This is the this is a gdal based file and so in order to load our shape files in we have to use this post just.
Command line tool which is essentially saying I want to bring shape to PostgreSQL and we send it the proper commands and we're good to go and then we we do some updates on our resulting tables.
And again, not going too deep into the weeds here, but different data sets have different standard projection representations. So you can represent of course you like a like a mercat projection, or there's lots of different types of ways to put a sphere onto a flat plane and map it out.
And anyway, and So what we want to do though is we want to make sure that the projection that our data set is referencing is consistent with the project like our kind of geographical projection of our latitude and longitude points. And so then we essentially after we ingest those into the database, we use this St. transform which is a very.
Very powerful function to reproject our data set into this new standard, this new SRID it's called. So we bring our data set in, we transform it, essentially project it onto this new space and now it's essentially ready for us to compare.
With our location, latitude, longitude values, right? So now we've gone through this process for all of our data sets. In this case, we've transformed our geometry here. We did not, and that's why did we not do that? I have to double check as to why.
Don't totally remember at this moment, but in any event, this is OCR to OCR is another tool that allows you to bring in.
Yeah, that's your to to to bring in these these GDP files, right? And This is why we had to add those GDAL specific requirements.
And at the end of this we end up with all all of our data sets in post GIS. And at that point what we end up doing is we loop across all of our variables that we've selected. And again, if I go here, right, we have all of these nice variables, we we make a selection, maybe we want all of them, maybe just a subset.
And what it's gonna do is it's gonna go row by row and then it's going to pull the associated columns. You'll notice that there's like a variable name and then you have like a specific.
Essentially attribute source value. So these are these are referring specifically to columns in the tables that we've just loaded, and so in this case they're going to be columns in our data source table. You have a value as number because each of these columns containing a numeric response.
That we want to capture in a value source value. One thing I don't know if Paulina is on the call today. I do need to still do some mapping. I know that these are all represented in the GIS vocab, but this at least the file that I was referencing for Caia DB didn't have them. Maybe you know of a new file that does have these maps.
It shouldn't be a lot of work. I think I can even do literal strings on a lot of these, but this is an attribute concept ID and this is essentially what's gonna populate the.
Yeah, the the external exposure concept ID, that's that's kind of key. And so this variable table is is allowing us to convert what's in those data sets and the associated variables in them into OMOP representations of those values. And this is kind of the key point, right?
So we have, we're looping over each of these variables, and as we're doing that, we're referencing our location and location history tables that we've loaded. And in my case, I make what I call like a merge version of those, like a location merge that essentially just has.
The latitude and longitude values for my location table together with my location history information. That's essentially all we need, right? And with that look at these these values essentially I can run.
And this is kind of where the magic happens of of joining this all together. But what we do is we and this is gonna look kind of silly, but.
We really the function that you should take a take a look at is this within function. So Postgres has this St. within and what that's doing is it's saying I want to join my tables based on whether my point is inside.
Of a polygon geometry, and that's a super weird thing to try to like visualize or imagine, but our data source tables have a bunch of polygons that that are representing them, and our location history table has a bunch of points that represent the history.
And so we're essentially joining and placing our points into the the relevant polygons and returning the values of those polygons. And that's what this this guy is doing, and it's doing that variable by variable. And as you see, this is a programmatic SQL script that gets run.
Because it's kind of a pivot effectively where we are taking this wide table that has a lot of different variable columns and then we're pivoting the values of those columns.
Together with the name of the column essentially into into our exposure occurrence or external exposure. Sorry, I can continue to get that wrong. OK, and so all that to say, right in theory, and this is what I'm going to start to write up now, let's.
You will initialize the database. So this is how the tool will work. You'll initialize the database. It's gonna be built and I have already. Yeah, so I can make this bigger.
So if you look on the container apps, I don't yet have it built into an image, so I will try to do that soon, but at least the code is up. It's called Postgis dash STOH and that will create for you a Postgis image that has the G dial requirements and that also has.
The Cove files of our kind of simle catalog as well as the QL files to do these things.
What's then needs to happen is and I I will write this, you will actually run a you'll you'll kind of bring this up. It has to initialize and then it's going to be running maybe locally or wherever it's going to run and and on another call we can show that actual process in action.
And then you'll execute this file against it. And there's a Docker command that you can run to execute this file, and that just goes through the process that I walked you through. So it first fetches your data sources that you're interested in, it ingests them into the database, and then it goes row by row through your variables and it says.
Which data source do I need? And it writes then like a a SQL code to reference that data source, do the spatial join and end up with an external exposure that has the the the variables of of interest or kind of the associated values in the OMOP specific stuff.
This database I mentioned briefly right, but it it has the GIS vocabulary in it and so one of the things I'm also doing is updating the the chorus vocab that's central so that it will also contain this GIS vocabulary terms.
There's some additional updates on that JS stuff as well that I'm working on, so it's gonna be gonna be a push to try to get that out, but I've been able to run this end to end now and my hope is that.
We can essentially test this out at a few sites at list and then after this delivery we can probably spend more time seeing how we can either deploy it at your site or or see what ways that we can essentially help you help you with this process.
Um.
OK, but yeah, like I said, if you want to take a look at the source code, it's here. The test files I mentioned also offhand at the beginning. This is what they look like. Like I said, I I took four different cities from the.
The NASA pollution data sets and I just grabbed 5 addresses in those cities and essentially just chose commercial addresses of like restaurants. So it does serve as a nice case. They're not very ugly though, like your addresses probably will look uglier than this.
These are pretty clean, so you should be able to to practice geocoding with them without too much issue and then you can test my against my geocoding if you would like. I also created a test location history.
Which is more or less like assigning a random person ID in my case to a long 22 year period that surrounds my data sets of interest and.
Covers all the locations essentially, so don't worry too much about this one. This is this do whatever you want with. It's again pretty simplistic. I think the location table is probably going to be more useful to you and as kind of an exercising tool.
Yeah. OK, So there's there's those pieces. I I feel like I'm missing one of the thing.
But yeah, maybe I'll I'll come back to it. Um.
All right, I'm gonna pause there because I'm sorry, I kind of got lost for a second, but I want to see if there's any questions because I that was a pretty intense technical session and I want to make sure that if anybody does have questions or wants to follow up on something that I might see we do have 10 minutes, so.
It'd be just more, yeah.

Schmidt, Heidi E. 51:26
I just wanted. I just wanted to say thanks for taking the time to go deep diving into it. I probably would have to just kind of sit there and noodle with it because that's what I was doing locally and finding that, you know, there were files that.
You know, figuring out the puzzle and seeing where the pieces fit just by doing it. And so I'm I'm grateful that you showed us, but I think I'm going to have to play with it first.

Houghtaling, Jared 51:56
Awesome. Yeah, I think. And I do need to make documentation and do some other stuff as well, but I plan to to do that and have kind of a step by step of how you can how you can run this.
I I will say, I mean this could be incredibly powerful or a powerful way to augment your OMOP instances considering.
Most of this information is public, like a lot of the the data that you're you're pulling that into your OMAP instance is just publicly available and accessible data. All you need to be able to do is provide a location and a time and place.
A location and a time that somebody was at that place and you can then just pull and then kind of join with all these these interesting variables, so.
Yeah, I I appreciate you all staying with me through that. I feel like I've I'm maybe too deep into it, but I will say the more you get into the GIS stuff, the more you realize that it's like a very it has its own.
Kind of ecosystem and and set up and software like the it's it's own kind of realm of of computation and and science honestly. And so again I'm just only scratching the surface on some of these things but.
I would encourage you all if you are interested in this, like there's some pretty fantastic resources and and definitely come check out the GIS work group because some some good stuff going over there as well.

Alvarez, Marta 53:33
There is the course container apps GitHub public. It's probably a dumb question. OK, I can't seem to find it. Do you mind sharing the link with me? I'm sorry I I must not have gone there before and I'm trying to hide it from GitHub.

Houghtaling, Jared 53:37
It is, yeah, yeah.
Definitely.
This is.
This is the this will point you to the actual directory where all this stuff is is sitting.

Alvarez, Marta 53:54
Perfect. Thank you. I'll include this in the
Meeting summary.

Houghtaling, Jared 54:04
Awesome.

Alvarez, Marta 54:05
Uh, there it is. Okay, thank you.

Houghtaling, Jared 54:09
All right, but like I said, so I haven't built this yet into an image that's kind of next on my list. So I guess the if we talk about next steps of what I will do is build it into an image that gets hosted on Chorus GHCR like the all the other packages we're hosting.
And then write up documentation on how you can use it and do some testing in that. It is like I said, it's super lightweight in terms of it's an Alpine image. It's like a small, very lightweight version of Linux.
And the the code to run all this is just like a bash script. It's not like Python or anything. There's there's almost no dependencies. So hopefully, I don't know if that's gonna make it easier or harder, honestly to to update things or understand what's going on. My hope is that it's easier.
But yeah, please take a look for those of you who are interested and I would love any feedback or any any help to to push it forward so.
All right.
Well, that is all I have today. Thanks again for staying with me and I guess I'll see you all next week. I don't know what our topic is, is planned to be, but.

Alvarez, Marta 55:31
We had talked a little bit about potentially following up about concept sets, but I don't know if that's going to change between now and then.

Houghtaling, Jared 55:40
I did have another kind of public service announcement. I guess for those of you who have tried to access the DCS space, you'll have noticed that there is a construction message on it. I have migrated the space up elsewhere, but we have to wait for MCB to.
Kind of grant us the.
A.
The rights to have that public address be accessible through their authentication. So they have to essentially.
I don't know. Check a box that says that we can have that address available and as soon as they do that then I will send you all that new link for the DCS space. It'll look exactly the same.
But the old link will just not work. It'll work, but it's going to show that construction page and it'll have a link to the new page essentially. So just a heads up, I know that UCSF team was trying to access earlier this week and they could not. I'm assuming at least one or two others probably have ran into the same thing.
But um, getting there so.

Alvarez, Marta 56:43
If you if when the new link is ready, if you'd like, you can send it to me and I can just send it to the whole to the to the office hours crew, OK.

Houghtaling, Jared 56:49
Awesome. That'd be great.

Alvarez, Marta 56:54
Thank you, Jared. Thanks, everyone, for coming. Take care.

Houghtaling, Jared 56:57
Thank you all.

Alvarez, Marta stopped transcription